Automatic Identification of Topic Tags from Texts Based on Expansion-Extraction Approach

Seungwon Yang

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computer Science and Applications

Edward A. Fox, *Chair*
Weiguo Fan
John F. Moore
Naren Ramakrishnan
Barbara M. Wildemuth

December 4, 2013
Blacksburg, Virginia

Keywords: topic identification, tagging, cognitive informatics, vector space model, knowledge sources, natural language processing, digital libraries, usability study

Abstract

Title: Automatic Identification of Topic Tags from Texts Based on Expansion-Extraction Approach

Seungwon Yang

Identifying topics of a textual document is useful for many purposes. We can organize the documents by topics in digital libraries. Then, we could browse and search for the documents with specific topics. By examining the topics of a document, we can quickly understand what the document is about. To augment the traditional manual way of topic tagging tasks, which is labor-intensive, solutions using computers have been developed.

This dissertation describes the design and development of a topic identification approach, in this case applied to disaster events. In a sense, this study represents the marriage of research analysis with an engineering effort in that it combines inspiration from Cognitive Informatics with a practical model from Information Retrieval. One of the design constraints, however, is that the Web was used as a universal knowledge source, which was essential in accessing the required information for inferring topics from texts.

Retrieving specific information of interest from such a vast information source was achieved by querying a search engine's application programming interface. Specifically, the information gathered was processed mainly by incorporating the Vector Space Model from the Information Retrieval field. As a proof of concept, we subsequently developed and evaluated a prototype tool, Xpantrac, which is able to run in a batch mode to automatically process text documents. A user interface of Xpantrac also was constructed to support an interactive semi-automatic topic tagging application, which was subsequently assessed via a usability study.

Throughout the design, development, and evaluation of these various study components, we detail how the hypotheses and research questions of this dissertation have been supported and answered. We also present that our overarching goal, which was the identification of topics in a human-comparable way without depending on a large training set or a corpus, has been achieved.

Dedication

To my parents, Jeong Hyun Yang and Keum Soon Lim

Acknowledgments

It has been more than 10 years since I transferred to Virginia Tech as an undergraduate student in 2002. Since then, I received a B.S., a Master's, and, this time, a Ph.D. degree, all in Computer Science. I believe that this ten-year journey of mine could not have been possible without persistent support and care from many people, who directly and indirectly interacted with me. Among them, my advisor, Edward A. Fox, is the one to whom I am most indebted.

I met Dr. Fox in the last semester of my undergraduate program at Virginia Tech when I took his course, CS4624: Multimedia, Hypertext, and Information Access. I liked the way he managed the class with various projects for student groups. He frequently interacted with students with a warm manner and insights for scaffolding the learning of them. Since I became a graduate student in his Digital Library Research Laboratory (DLRL) in 2005, Dr. Fox has been my role model in academia, who patiently encouraged and supported his graduate students to go through the difficulties and distractions in pursuing their degrees.

I also want to thank my doctoral committee members. Barbara M. Wildemuth has provided me with valuable feedback and encouragement for my doctoral study with her expertise in Library and Information Science (LIS). Thanks also go to Weiguo (Patrick) Fan, Naren Ramakrishnan, and John F. Moore, who were supportive and prompted me to consider different aspects of my research.

For the last two and a half years, I was very lucky to work with the members of the Crisis, Tragedy, and Recovery Network (CTRnet) project, of which my doctoral work is a part. I felt that the project co-PIs, Adrea L. Kavanaugh, Steven D. Sheetz, and Donald J. Shoemaker, were like my friends, and at the same time, I was impressed by their professionalism in research.

Colleagues in the past and present in DLRL deserve my great praise. They are Uma Murthy, Jonathan Leidig, Seonho Kim, Mohamed Magdy, Sunshin Lee, Paul Mather, Sherief H. Elmeligy, Tarek Kan'an, Monika Akbar, Yinlin Chen, Mohammed Saquib Akmal Khan, Eric Fouh, Sung Hee Park, Noha ElSherbiny, Xiaomo Liu, Jian Jiao, Spencer J. Lee, Venkat Srinivasan, Kiran Chitturi, Anita Walz, and S. M. Shamimul Hasan. They were my friends and critics, who attended my presentations, reviewed draft papers, played Ping-Pong, and collaborated on various tasks in the lab. Also I appreciate Susie Marion for being my friend and helping with scheduling the meetings, preparing documents for conference travel, and having a chat about adventures and travels.

I want to say thanks to many others: Haeyong Chung and Yoonsuk Lee for sharing and discussing ideas, as well as exploring the restaurants in Blacksburg together; Sanghee Oh, Barbara M. Wildemuth, and Jane Greenberg for helping to recruit study participants from their classes and schools; Taeho Kim, who is my long-time friend and has been my housemate for the last three months, for discussing the statistical methods in my dissertation, as well as recommending me nice songs to relieve stress; and Laurie Good for carefully proofreading my dissertation draft.

Finally, huge thanks go to my Ba Gua Zhang shifu, Bok-Nam Park, as well as my Zen master, Kyung-Hoon Kim, who gave me life lessons not only with their words, but also by showing their incessant effort in pursuing their goals.

*This dissertation work received support from the National Science Foundation (Grant: IIS-0916733, IIS-1319578) for the Crisis, Tragedy, and Recovery Network (CTRnet) project, and the Integrated Digital Event Archiving and Library (IDEAL) project, respectively. The Center for Human Computer Interaction (CHCI) as well as the Digital Library Research Laboratory (DLRL) at Virginia Tech also provided fund for experiments and the usability study in this dissertation.

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

This chapter begins with the main motivation of the study. There is the need for developing a flexible topic tagging tool, which works well with minimal human intervention, and also is capable of accessing the Web as a universal knowledge source to enhance the tagging. Existing popular topic identification approaches have limitations when directly applied to our context. Thus, we explored how we can effectively design, develop, and evaluate a topic tagging approach that could satisfy our needs. From such research questions, we derived several hypotheses that can(not) be supported by our experiments and evaluations. This chapter ends with our contributions, and with an overview of the organization of the dissertation.

1.1 Motivation

Two primary research motivations drove the design and development of this study. The first has to do with the need for a flexible topic-tagging technique that can be used for archiving and organizing resources related to a specific topic, e.g., disaster events. For organizing topic-related information, identifying topical tags describing a resource might be very useful to supplement other metadata such as title, authors, publication dates, locations, and so on. The second motivation for carrying out this study stems from a general curiosity about how to best exploit information on the Web, which is considered to be a universal knowledge source, in a more sophisticated way.

Disaster events have been increasing in number both nationally and internationally. Understandably, they are of significant interest to a wide range of individuals (e.g., scientists, actuaries, meteorologists, and emergency workers). Depending on the size and scale of the disaster event, available resources on the Web, and the popularity of a given event on social networking sites, disaster archives can be complex. Thus, it can be very demanding to archive and organize various online resources related to disaster events, particularly if the goal is to develop a digital library for providing services related to this body of knowledge. Information resources for large-scale disasters tend to be collected continuously over a long period of time; this means that the stories and themes associated with disaster events change dynamically as a disaster unfolds and related events occur.

The 2011 Japan Tohoku earthquake and subsequent tsunami is an example of an event requiring long-term archiving. Immediately following the earthquake, most online resources were focused on the wide scale damage caused by the ensuing tsunami. In time, however, the world's attention turned to recovery efforts on both a macro and micro level, but more importantly to the still unfolding story of the Fukushima nuclear power plant and the damage resulting from reactor explosions and meltdown. Indeed, as of this writing nuclear pollutants are still escaping into the seawater. Clearly, this is an information-rich event of significant interest to a worldwide audience.

To assign topic tags to archived news webpages and reports, we need an approach that is sufficiently flexible to assign quality tags regardless of the changing nature of the data. Moreover, the human intervention in this process should be as minimal as possible considering that the archiving and tagging process is likely to be performed on a vast scale, and over a long period of time.

When a human being summarizes texts and infers topics by reading and thinking, related information in the brain is recalled and used in the process. Also important to this activity is identifying the significant and frequently used words in a given text. However, using the Web as a source of information is inherently complex due to the massive amount of information that can be found online, not to mention the fact that new information is continuously being added to it. Typically, retrieving information from the Web in any methodical way requires a search engine, which in some way resembles the way we, as humans, retrieve information from the brain. As search engine technologies evolve over time, more sophisticated access and use of the seemingly limitless information on the Web will become available.

1.2 Problem Statement

Our challenge is devising a system requiring little setup, for finding suitably descriptive topics, for an individual document in a large archive about a disaster event.

Important to this discussion is the fact there are various topic identification approaches currently available. One of the most popular ones is Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), which uses statistical analysis of a large text corpus to extract groups of terms. Using these terms, a person should in theory be able to infer overarching topics for the text.

However, this approach may be less efficient when applied to disaster archives, where events are dynamic and incoming resources are unpredictable in their frequency and content. In addition, inferring a (specific) topic from a group of generated terms that have the potential to be quite varied is not always easy. Further, if the size of the document corpus is small (e.g., 5,000 documents might be considered to be a "small" corpus), care should be taken to adjust the convergence parameters and number of topics until useful results are generated. The characteristics of the documents also can impact the results. In short, when applying LDA to disaster archives, much human intervention is needed to achieve optimal results.

Approaches that are based on supervised learning often attempt to identify categories of documents within a document corpus (Dumais & Chen, 2000; Sebastiani, 2002). The category labels are determined by the researchers based on the researchers' interests before the classification begins. These labels become the topics of the documents that are classified into those categories. In contrast, an unsupervised learning approach groups documents in a corpus into several clusters based on content similarity (He, Ding, Zha, & Simon, 2001; Stein & Meyer zu Eissen, 2011); the assumption here is that documents in the same clusters share topics. However, these learning-based algorithms contain some inherent difficulties. A training set should be developed for supervised learning, which is time consuming and tedious. In addition, re-training or re-clustering might be needed in instances when new input texts are added to the document corpus, for example, when a disaster event progresses and news webpages are updated. In addition, after unsupervised learning (i.e., clustering) occurs, "labels" to the document clusters should be assigned since those labels are the topics that represent the documents in the cluster. Given the fact that many labeling techniques are available, this means that additional effort is required in identifying topics using unsupervised learning (Stein & Meyer zu Eissen, 2011).

1.3 Research Questions

Following the literature review, four research questions, along with their sub-questions, were developed to guide this study. The first question has to do with designing and developing a topic tagging approach that could overcome current constraints and satisfy the need to produce quality topic tags. The second question involves determining effective evaluation metrics and examining the performance of the developed topic tagging approach, principally by comparing it to other approaches. The third question examines whether this study's approach can still produce quality topics using a larger corpus with the parameters used in prior studies. The last question examines the usability of the prototype user interface for the tool that supports interactive semi-automatic tagging tasks. These four questions are listed as follows:

Research Question 1
Can one design a semi-automatic approach (that requires only minimal human intervention) for generating human-comparable topic tags without relying on a prepared large text corpus?

> *Research Sub-question 1-1*
> If such a tagging system were built, what would be its optimal parameters?

Research Question 2
What is an effective evaluation approach for comparing the quality of topics generated by the prototype system with other methods?

> *Research Sub-Question 2-1*
> What is the measured quality of topics generated by our system or other methods?

Research Question 3
Can one apply the optimal parameters, developed in response to *Research Question 1-1*, to our tagging system in order to produce quality topics using a larger data set?

> *Research Sub-Question 3-1*
> Does this approach outperform other approaches?

Research Question 4
If we build a front-end user interface for our system, what is an effective approach and how useful will it be in supporting interactive semi-automatic topic tagging tasks?

1.4 Hypotheses

A principal hypothesis was developed for this study. In addition, four sub-hypotheses that correspond to the research questions are proposed and are listed below.

Principal Hypothesis

The proposed approach, which uses a Web search engine, can produce tags of similar quality to human tags without depending on a large text corpus or a large training set. The tags will be comparable in quality to those obtained from a state-of-the-art topic identification application programming interface (API). It is also hypothesized that the front-end user interface for the system will be usable and satisfactory for suggesting topic tags during interactive semi-automatic tagging tasks.

Hypothesis 1

Optimal parameters can be identified by examining the cosine similarity between the topics generated by our system utilizing different combinations of parameters and the topics assigned by multiple human indexers. The average Inter-Indexer Consistency (IIC) values between our system with optimal parameters and human indexers will be greater or equal to the average of those between human indexers only.

Hypothesis 2

The average relevance rating for the topics generated by our system (using a five-point Likert scale) will be significantly higher than the average relevance rating for those extracted using the baseline, TF * IDF, for both the CTR_30 (i.e., homogeneous documents) and VARIOUS_30 (i.e., heterogeneous documents) data sets.

Hypothesis 3

The average F_1 score for the topics generated by our system using 1000 randomly-selected New York Times test collection articles will be significantly higher than or equal to the average F_1 score for those extracted using either the baseline or OpenCalais natural language processing API.

Hypothesis 4

The usability and user satisfaction for our prototype tool will exceed a “1: Useful and Satisfactory” rating according to a 5-point Likert scale from -2 to 2. In addition, the topic tags generated by human indexers, consulting the suggested tags from our prototype system, will have a higher average F_1 score when compared with the F_1 scores of those generated by either the baseline or OpenCalais.

1.5 Contributions

Findings from this dissertation are expected to contribute to several aspects of topic identification research, as detailed below. All the code, data sets, and outputs of this study are downloadable:

- *A Survey of Topic Identification Studies*

Studies related to topic identification have been assessed, the results of which are provided along four lines of inquiry. The first has to do with the field of Cognitive Informatics, with a particular focus on studies pertaining to topic tagging. The second area of inquiry concerns the widely-used vector-based information retrieval model while the third group includes corpus-based statistical approaches. The fourth group involves studies pertaining to evaluation aspects of the topic identification approaches.

- *A Novel Approach for Topic Identification*

This study also introduces a new approach for topic identification, which combines a methodology based on Cognitive Informatics with an application of the popular Vector Space Model.

- *A Prototype System*

A prototype system, which implements our topic identification approach, was developed as a two-part system. The first corresponds to a back-end system that was developed in Python, which runs in batch mode, processing documents automatically. The second part of the prototype system corresponds to the front-end user interface, which was developed using JavaScript.

- *The Data Sets Indexed with Human-Assigned Topics*

Two data sets, CTR_30 and VARIOUS_30, were provided along with their topic tags assigned by multiple human indexers. The CTR_30 data set included homogeneous documents, whereas the VARIOUS_30 data set contained heterogeneous documents.

- *Scripts for Text Pre-Processing and Analyses of the Results*

A number of scripts written in Python for pre-processing the data, and for analyzing various outputs were developed and made downloadable.

- *Resources for the Usability Study*

The following documents pertaining to the usability study of our prototype tool were developed and are included in this dissertation: a set of questionnaires, a recruitment letter for study participants, a tutorial for using the prototype system, the study's abstract, detailed study procedures, and an informed consent form.

- *A Set of Findings from the Evaluation of Our System*

All the data and related findings generated during the evaluation process are reported in this dissertation and may be useful for informing other current and future studies.

- *A Set of Findings from the Usability Study of Our System*

A detailed analysis of the results obtained from the usability study of our system is presented in this dissertation.

1.6 Organization of the Dissertation

Following the Introduction to this study, a detailed review of the literature is presented in Chapter 2. It includes an assessment of studies pertaining to Cognitive Informatics (see Section 2.1) and the Vector Space Model (VSM, see Section 2.2). In Section 2.3, three widely-used external knowledge sources, namely, Wikipedia, WordNet, and the Web, are presented with specific items and approaches for using them. The corpus-based statistical approaches such as Latent Dirichlet Allocation and Latent Semantic Analysis are explained in Section 2.4. The literature regarding the evaluation of the identified topics is also presented (Section 2.5).

Chapter 3 describes the development of the prototype system. The two principal components of the system, expansion and extraction, are detailed in Section 3.1, while Section 3.2 provides algorithm details. The metrics, processes, and data sets for identifying the optimal parameters of the prototype system are illustrated in Section 3.3. Chapter 3 concludes with a discussion of the benefits and limitations of our use of the VSM, followed by the chapter summary.

A detailed evaluation of the prototype system is provided in Chapter 4, which includes discussions of the experimental settings, the baseline approach, and the popular natural language processing API, OpenCalais. The quality of topics generated using the system is examined in three ways. The first involves collecting and analyzing relevance ratings for the topics that are assigned by human indexers. Secondly, the Inter-Rater Consistency is computed to assess the consistency of the relevance ratings assigned by the indexers. Thirdly, the prototype system is evaluated using a larger data set, namely a subset of the *New York Times* corpus.

The tagging system for this study is designed as a tool to run in batch mode, automatically processing documents with very little human intervention. However, a Web-based user interface also was implemented so that it could provide topic suggestions for interactive semi-automatic tagging tasks. The user interface design and development details, as well as the usability study, are described in Chapter 5.

Chapter 6 summarizes the findings of the study with respect to the research questions and associated hypotheses. Limitations and suggestions for future research also are addressed. There are five appendices in this dissertation, giving additional details. Appendix A includes resources for the usability study of the Xpantrac UI system, including the study abstract, a participant recruitment letter, an informed consent form, an IRB approval letter, and all the questionnaires. Appendix B and C contain resources for the manual topic tagging study and topic relevance rating study, respectively. The types of the included documents are the same as the ones in Appendix A. A tutorial for the Xpantrac User Interface is attached in Appendix D. Finally, Appendix E includes download links and documentation for the software (i.e., Xpantra command line mode, and Xpantrac UI), scripts (i.e., for text preprocessing, and for computing evaluation metrics), and data sets (i.e., CTR_30, VARIOUS_30, document IDs for NYT_1000) used in this study.

Chapter 2. Literature Review

This chapter begins with an examination of Cognitive Informatics and the Vector Space Model (VSM), particularly as they apply to this dissertation study. Cognitive Informatics (CI), as detailed in Section 2.1, is a multi-disciplinary field involving areas such as Cognitive Psychology, Computer Science, and Linguistics (Becker & Kuropka, 2003; Salton, Wong, & Yang, 1975; Shuda, Jiangping, & Riu, 2009; Y. Wang, 2002; Y. Wang et al., 2011; Yingxu Wang & Ying Wang, 2006). The idea of expanding texts using the Web as a knowledge source is based on studies from CI.

Section 2.2 details the Vector Space Model (VSM) (Becker & Kuropka, 2003; Salton et al., 1975; Shuda et al., 2009; Y. Wang, 2002; Y. Wang et al., 2011; Yingxu Wang & Ying Wang, 2006), which is one of the most important and widely used models in the Information Retrieval (IR) field for efficient processing of text. VSM involves a three-step process. The first is to convert texts into document vectors after obtaining an index list from the text corpus. The document vectors and the term index lead to a term-document matrix, where term-weighting and normalization can be applied. To compute the similarity between any two documents, the cosine similarity is computed between the two column vectors that represent the two documents in the term-document matrix.

The topic identification approach proposed in this dissertation uses information on the Web, which is considered to be a universal knowledge source. Studies incorporating external knowledge sources such as the Web, Wikipedia, and WordNet (G. A. Miller, 1995) are discussed in Section 2.3. Corpus-based statistical topic models—in particular Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) (Deerwester, Dumais, Landauer, & Furnas, 1990)—are explained in Section 2.4, followed in Section 2.5 by the evaluation metrics utilized to measure the quality of topics generated. The chapter summary is provided in Section 2.6.

2.1 Cognitive Informatics

2.1.1 Overview

Cognitive Informatics (CI) is a multi-disciplinary field that combines computer science, cognitive science, and intelligence science. Researchers in CI investigate how the human brain processes information. They also address processes of natural intelligence (Y. Wang, 2002; 2007; Y. Wang et al., 2011; Yingxu Wang & Ying Wang, 2006). CI encompasses a variety of topics including thinking, reasoning, remembering, language understanding, perception, learning, consciousness, and emotions. It is typically explored from a computational point of view, based on a solid foundation established by scholars investigating the brain and neuroscience (Z. Shi & Shi, 2003).

One significant CI study conducted by Wang et al. (2006) explored models of the brain. According to their findings, understanding human memory and its functions are key, since as they asserted, memory represents the foundation for natural and artificial intelligence. It is well known that there are two memory categories: short-term memory (STM) and long-term memory (LTM), both of which play a role in the Functional Models of Memories. Basically, STM retains information for a short time—generally from a few minutes to a few hours. After this time span, the information in STM will either be removed from STM or moved to LTM. In contrast, LTM has been considered as a static storage space for information with a large capacity, since the majority of neurons in an adult's brain will not change significantly over the lifespan.

2.1.2 Inspiration

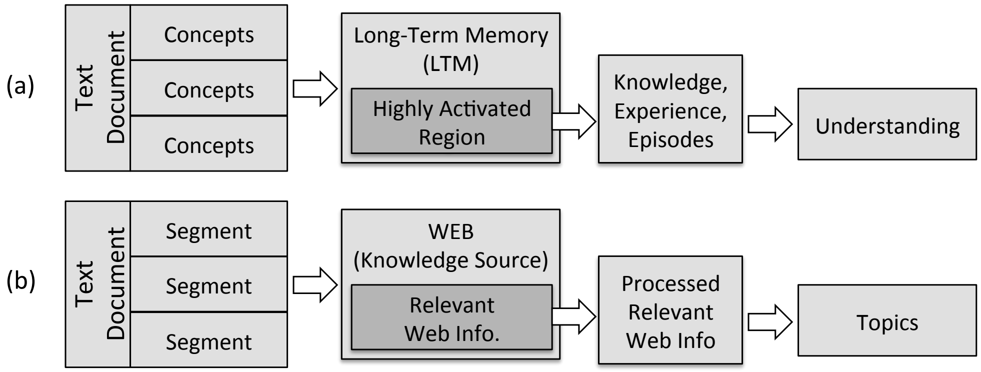


Figure 1. Understanding and finding topics from text documents: (a) Human memory retrieval and understanding; (b) Algorithmic process that is analogous to (a).

When we read, certain stimulating words and concepts that appear in sentences remind us of relevant knowledge and experience that we possess in our memory, whether we consciously realize it or not (Lange & Wharton, 1994; Wharton, Holyoak, Downing, & Lange, 1994) (Figure

1(a)). Many psychology studies have indicated that superficial similarity or thematic similarity (or both) between the cue thoughts in a given text and content in memory influence this memory retrieval process (Gentner, 1989; Gentner & Markman, 1997; Gentner, Rattermann, & Forbus, 1993; Markman & Gentner, 2001; Medin, Goldstone, & Gentner, 1993). Briefly, the higher the similarities between, on the one hand, the elements of cue thoughts, actors, actions, and places, and, on the other hand, memory episodes, the more likely we are to retrieve episodes accurately.

Based on this model of human memory retrieval and understanding, efforts have been made to model an analogous algorithmic process using computers (Figure 1 (b)). For example, Massey presented a framework for natural language understanding called *ReAD*, which stands for "Retrieval of conceptual knowledge from long-term memory and Activation and Decay of this knowledge" (Massey, 2011), based on the memory retrieval, activation, and decay inspired by Cognitive Informatics (L. Ogiela, Tadeusiewicz, & Ogiela, 2007; Y. Wang et al., 2011; Yingxu Wang & Ying Wang, 2006). In his framework, words from a text that a person reads stay in his/her STM temporarily. Then, each word sequentially triggers the retrieval and activation of ideas or concepts stored in LTM. This triggering action by each word is emulated by retrieving definitions of that word from a knowledge source, such as a dictionary or WordNet (Fellbaum, 2010; G. A. Miller, 1995). In this framework, ideas or concepts in LTM are considered as being expressed as word definitions. Retrieved ideas often trigger related ideas via cascading cognition. Thus, these actions are modeled by adding the words in the definitions back to STM.

As we process words in STM, recently accessed ideas in LTM become more activated, which could cause ideas that had been previously accessed to decay. This activation and decay is modeled as incrementing or decrementing the weight associated with each word that a person has seen either in STM or in the word definition. After entire words are processed in this fashion, the words whose weights are higher than a threshold value would remain as topics of a read text. Massey's algorithm, which is based on the activation and decaying of the weight of words, can be used to determine the main topics of a text document. However, from a discourse analysis perspective, for the task of finding the topics that best summarize a full document, decay might be inappropriate. Other benefits of Massey's algorithm are that (a) it is not dependent on a large text collection to extract the statistics of words, and (b) it does not require training a machine learning algorithm.

The design of our approach is similar to Massey's framework in that the information from input texts is expanded by accessing an external knowledge source. However, our approach differs in two significant ways:

1. Input texts are grouped into multi-word units, and each such unit, instead of each single word, accesses a knowledge source (e.g., Web) to retrieve relevant and rich information.
2. Once the retrieval of relevant information for all of the multi-word units is completed, the Vector Space Model (see Section 2.2 for details) is applied to the retrieved information in order to efficiently identify topics. In addition, memory decay is not modeled in our approach, except due to the fact that the Web's older content disappears over time.

2.2 Vector Space Model

The Vector Space Model (VSM) (V. V. Raghavan & Wong, 1986; Salton et al., 1975; S. K. M. Wong & Raghavan, 1984) is a widely-used algebraic model that is valuable for ranking relevant documents, uncovering significant features for classification, filtering information, and so on. The idea of converting texts into their vector representation in a multi-dimensional space also has become the basis for various other techniques. To name a few examples, Latent Semantic Indexing (LSI) (Deerwester et al., 1990; Foltz & Foltz, 1990; Ozsoy, Alpaslan, & Cicekli, 2011) and Support Vector Machines (Hearst, Dumais, Osman, Platt, & Scholkopf, 1998) use VSM as the first step in converting documents in their text corpus into vector representations, and then to apply a term-weighting scheme in order to identify effective features.

2.2.1 Term-Document Matrix

The first step of the VSM is to convert text documents into their vector representations. For example, Figure 2 shows two document vectors, D_1, and D_2, whose values are represented in a three dimensional vector space. For this, an index list, *Index,* is constructed with n unique terms that are present in the corpus:

$$Index = \{t_1, t_2, \ldots, t_n\}$$

Based on *Index* of size n, and all of m documents in the corpus (Figure 3 (a)), we can construct a term-document matrix (*TDM)* of size $n * m$ as shown in Figure 3 (b). The cell value of row i and column j is the frequency of Term i appearing in Doc j. For example, Term 1 is found three times in Doc 1. The next step is to optionally apply a term-weighting scheme to the *TDM*.

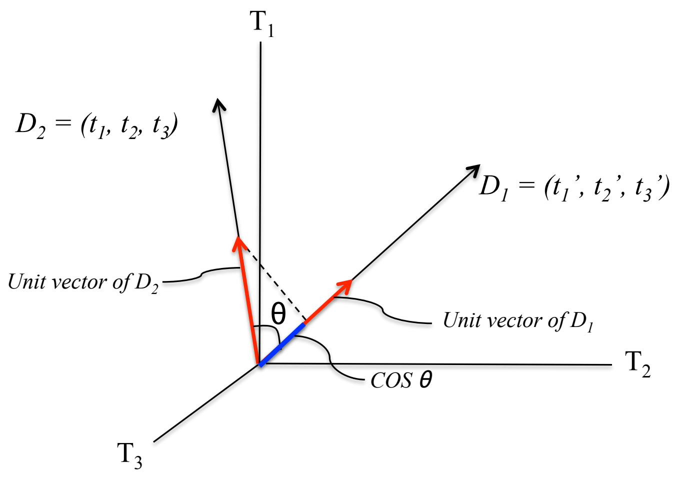
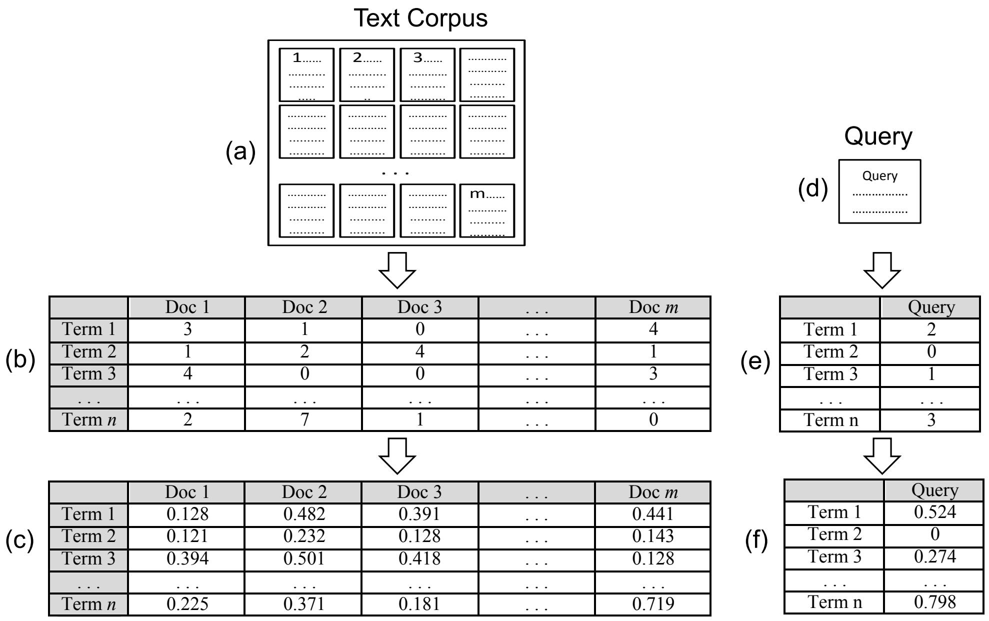


Figure 2. A visualization of document vectors in a three-dimensional vector space. The red arrows denote unit vectors of D_1 and D_2. The blue line is the magnitude of the unit vector of D_2 projected onto the unit vector of D_1. (The figure is adopted from (Salton et al., 1975))

2.2.2 Term Weighting



	Doc 1	Doc 2	Doc 3	. . .	Doc *m*
Term 1	3	1	0	. . .	4
Term 2	1	2	4	. . .	1
Term 3	4	0	0	. . .	3
. . .	. . .	. . .	. . .	. . .	. . .
Term *n*	2	7	1	. . .	0

	Query
Term 1	2
Term 2	0
Term 3	1
. . .	. . .
Term n	3

	Doc 1	Doc 2	Doc 3	. . .	Doc *m*
Term 1	0.128	0.482	0.391	. . .	0.441
Term 2	0.121	0.232	0.128	. . .	0.143
Term 3	0.394	0.501	0.418	. . .	0.128
. . .	. . .	. . .	. . .	. . .	. . .
Term *n*	0.225	0.371	0.181	. . .	0.719

	Query
Term 1	0.524
Term 2	0
Term 3	0.274
. . .	. . .
Term n	0.798

Figure 3. A diagram showing the steps of a text corpus being converted into a TF*IDF weighted term-document matrix. (a) An input text corpus; (b) a term-document matrix developed from the corpus with term frequencies (TF) in the cells; (c) a term-document matrix with TF*IDF and normalization; (d) a user query; (e) a query term vector; and (f) a query term vector with TF*IDF and normalization.

A term weighting scheme, *Term Frequency * Inverse Document Frequency* (TF*IDF), computes a term's significance based both on the document level (TF part) and the corpus level (IDF part) (Papineni, 2001; Salton & Buckley, 1988; Zhanguo, Jing, Xiangyi, Yanqin, & Liang, 2011).

As explained above, a *TDM* of size *n*m* is shown in Figure 3 (b). At this point, each matrix cell contains the frequency of the term in each document. We then multiplied the cell values with each term's IDF (Sparck-Jones, 1972). IDF is effective in identifying a term's corpus-wide significance by considering how many documents contain the term. The more documents in which a term appears in, the less significant the term is in the corpus. Therefore, the IDF value of a term is defined as:

$$IDF = \log \frac{N}{m}$$

where N is the total number of documents in a document set (i.e., corpus), and m is the number of documents that contain the term of interest.

A term frequency *tf* considers frequent terms to be more important in a document. Thus, by combining a term frequency *tf* with IDF, TF*IDF is defined as:

$$TF * IDF = tf * \log \frac{N}{m}$$

The TF*IDF values for the terms in a longer document would generally be larger than those in a shorter document—simply because longer documents have more terms with large TF values than shorter documents. To reduce this effect, TF*IDF is often "normalized" by dividing it by the Euclidian length of the document vector (Salton & Buckley, 1988):

$$Normalized\ TF * IDF = \frac{tf * \log \frac{N}{m}}{\sqrt{\sum_{vector} \left(tf_i * \log \frac{N}{m_i} \right)^2}}$$

The normalized TF*IDF values in each column in Figure 3 (c) showed which terms were significant in the document represented by that column. For example, Term 3 had the highest cell value of 0.418 in Doc 3, followed by Term 1 (cell value of 0.391) and Term *n* (cell value of 0.181). Therefore, we selected the topics in the order of Term 3, Term 1, and Term *n*, as shown in Figure 3.

2.2.3 Measuring Similarity

Once a weighted and normalized *TDM* is obtained, as in Figure 3 (c), one can assess the similarity between two documents by computing the cosine value between the two document vectors. For example, for the two document vectors, *A* and *B*, their cosine similarity is computed as:

$$\cos\theta = \frac{A \cdot B}{|A||B|}$$

where $A \cdot B$ is the dot product, and $|A||B|$ is the product of the magnitudes of A and B.

Since the document vector in each column of *TDM* is normalized, their magnitude is 1 (i.e., $|A||B| = 1$). Therefore, the cosine similarity computation is simplified as the dot product of the two document vectors. To find the most relevant document for a user search query (Figure 3 (d)), one simply converts the query into a term vector using the same index list, *Index*. After applying TF*IDF and normalization to the term vector in the same manner, one can obtain the weighted and normalized term vectors in Figure 3 (f). Finding the most relevant document for a user query is a matter of finding the highest cosine similarity between Figure 3 (f) and the document vectors in Figure 3 (c).

The VSM was incorporated as part of our topic identification approach. As explained briefly in the last paragraph of Section 2.1.2, information retrieved from the Web using the CI-inspired method can be further processed with natural language processing techniques, as well as with VSM. Essentially, the retrieved information forms a webpage description corpus that is derived from a single text document. The VSM converts this corpus into a TDM, and a term-weighting scheme is applied and the cell values are normalized as well. Since the goal is to identify significant terms in the derived corpus, the cell values in each row of the TDM are summed to produce a significance score for the term associated with each row. For a detailed description of the process for identifying topics, refer to Section 3.3.2.3 on "Term Selector."

2.3 Use of External Knowledge Sources

External knowledge sources such as Wikipedia, WordNet, and the Web are used in many text analysis and ontology studies. They are often combined with statistical, machine learning, or natural language processing (NLP) methods in order to produce topics and key phrases. These topics and key phrases are used to summarize texts, index resources in digital libraries, and construct domain ontologies (semi-)automatically. Selected resources for research and several methods that utilize such resources are shown in Table 1. Details of these knowledge sources and relevant examples are explained in the following paragraphs.

Table 1. External knowledge sources and their uses for research.

	Wikipedia	WordNet	Web
Resource	• Titles • Internal link texts • Categories • Redirects • Infobox	• Synonymy (i.e., synonyms) • Hyponymy (i.e., sub-concepts) • Hypernymy (i.e., broader/upper concept)	• Webpages • Descriptions of webpages
Method	• Key phrase extraction • Training set building for key phrase extraction • Training set building for classification • Redirects as synonyms • Semantic information extraction	• Semantic conflation (e.g., ontology concept extension) • Hierarchical classification of concepts	• Semi-automatic domain ontology construction • Query expansion

Wikipedia

Wikipedia is a widely-utilized source for information on a vast array of topics. In his survey study, Owens (2011) presented comprehensive details regarding the components of Wikipedia articles, as well as how the articles are produced and their use in research. Selected components of a Wikipedia article, which are often used for topic identification studies and semantic Web studies, are shown in the “Resource” row of the “Wikipedia” column in Table 1.

Schönhofen (2008) elucidated a reliable technique for labeling words and phrases appearing in texts with corresponding titles from Wikipedia articles. The basic idea is that the title is considered as a concept, and the content of an article is regarded as the description of the concept. Since multiple titles might be associated with the same words and phrases, a heuristic approach is used to select the most relevant title of a given Wikipedia article. For example, if the title of an article is an “official” one (i.e., not a redirected title), it earns a higher score compared to an informal (i.e., redirected) title. Additionally, a longer Wikipedia title also earns a higher score since it is more specific and refers to only a handful of concepts. If the title is multiply linked to the target (i.e., input) text document, the title (i.e., concept) is considered valid as well. Among several titles, the one with the highest score of the combination of such scores is finally selected, after which it is attached to the word or phrase in the target document. The results show that annotation accuracy is affected by the length of the article's title. For example, the accuracy is 77% when a title is a single-word title. In contrast, the accuracy increases to 86% with two-word titles. In summary, this technique assigns Wikipedia titles to every matching word and phrase in a given text document; thus, it is a type of tagging task. It should be noted, however, that this process differs from topic tagging, wherein the primary importance lies in the identification of significant and representative topics in the target document.

Another interesting use of Wikipedia is associated with a key phrase extraction study by Medelyan (2009, p. 83). In this investigation, significant key phrases were selected through filtering as topic indexes of the target document. Then, the redirect titles and anchor texts of internal links in several Wikipedia articles were used as a controlled vocabulary. This vocabulary can be considered as domain-independent due to its large volume of over several million titles and links. Instead of labeling every word and phrases matching the Wikipedia articles (as in Schönhofen's study), Medelyan's approach identified candidate key phrases that described the content of the target texts, after which the most relevant ones were selected as topics of the target texts. Basically, the process of selecting candidates starts from extracting all the n-grams (i.e., phrases consisting of *n* number of terms) up to a pre-defined size from a target text. Both the n-grams and vocabulary terms go through the normalization process involving stopword removal, conversion to lowercase, stemming (Kantrowitz, Mohit, & Mittal, 2000), and re-ordering of the words. The n-grams that match vocabulary terms are then selected, and a semantic conflation is applied. The term "semantic conflation" refers to a process of associating a term with synonyms (e.g., redirect titles) of a concept or a topic in order to improve the consistency and reduce issues involved with direct matching. Once the candidate key phrases are prepared, a filtering process is applied based on machine learning. A score is assigned to each candidate based on the combination of the candidate properties (features) such as TF*IDF, and the (first) position of the key phrase in a document. The candidates that have the higher scores are selected as the final key phrases.

In a study by Coursey et al. (2009), an encyclopedic graph derived from Wikipedia was used for topic identification. This investigation of topic identification had a broader scope in comparison to the key phrase extraction study by Medelyan (described above) in that the topics did not have to be present in the target document; instead, they could be obtained from any external sources. In Coursey et al.'s study, encyclopedic concepts were automatically identified by their system, *Wikify!*. Similar to the Medelyan approach, *Wikify!* selects candidate key phrases in the target text. Using a training corpus, it then computes the probability of an n-gram to be selected as a key phrase in a document, namely the "Key phraseness" metric:

$$P(keyphrase \mid n-gram) = \frac{Count(D_{keyphrase})}{Count(D_{n-gram})}$$

where $D_{key\ phrase}$ is the document that has "key phrase" as its one of key phrases, $D_{n\text{-}gram}$ is the document where the n-gram appears in its content.

The metric simply illustrates that if Candidate A appears more often as a key phrase (assigned by the author) in Wikipedia articles than Candidate B, then A has a higher probability of being selected as a key phrase for an unseen document. Once a list of candidates is identified, an encyclopedic graph based on the entities and categories of the Wikipedia articles is constructed. A biased (toward those candidates) graph centrality algorithm is applied to the entire graph. The graph nodes that match the candidates are ranked based on relevance to the target texts, after which the ones with higher rankings are selected as the topic key phrases of the target document. The results of the Coursey investigation showed that the average consistency of the topics with respect to 15 human index teams was 34.5%, outperforming the consistency of Medelyan's (30.5%).

WordNet

WordNet is a manually-constructed large lexical English database (Fellbaum, 2010; G. A. Miller, 1995). Selected semantic relations provided by WordNet are presented in the second column of Table 1. Miller (G. A. Miller, 1995) described the semantic relationships in WordNet, as follows. *Synonymy* refers to basic relations that use one or more sets of synonyms (i.e., synsets). *Hyponymy* and *hypernymy* have inverse relations, which refers to sub-concepts and super-concepts, respectively. Typically, there is only one hypernym for multiple hyponyms; thus, the meanings of nouns are organized into a hierarchical structure. These three relationships are mostly used for text analysis and natural language processing studies. Other relevant relationships in WordNet include *antonymy* (i.e., opposing terms), *meronymy* (i.e., part of something) and its inverse *holonymy* (i.e., whole of something), *troponymy* (i.e., hyponymy equivalent for verbs), and *entailment*. Similar lexical databases have been built in other languages. For example, EuroWordNet[1] contains wordnets in Dutch, Spanish, Italian, French, German, Czech, Estonian, and Turkish (Fellbaum, 2010, p. 238). WordNet also is considered as a lexical ontology; it has been incorporated in studies such as a domain-ontology development, and also

[1] http://www.illc.uva.nl/EuroWordNet/

has been linked with formal ontologies (e.g., Suggested Upper Merged Ontology (SUMO)) (Pease & Fellbaum, 2010). Such ontologies then are used in topic identification and information extraction studies.

In an automatic topic identification study, Tiun et al. (2001) employed an ontology hierarchy, and then used WordNet as a means to extend the concepts in this ontology to enhance the match between candidate topic terms in the target text and ontology concepts. Specifically, they used Yahoo! Web Directory[2] as their ontology. Their rationale for choosing this directory has to do with the fact that it is the largest subject directory for Web documents, and was developed manually based on human knowledge of the Web. The researchers mapped the concepts in their ontology to three WordNet semantic relationships (synonyms, hyponyms/hypernyms, and meronyms/holonyms), which are considered as the "extension" of their ontology. Using both the Yahoo! Web ontology and its extended ontology, corresponding ontological concepts of the terms from an input Web document were identified, and the weights of those terms were assigned. The term weight favors frequent terms in an input document, and penalizes terms 50% if the term matches the hyponyms/hypernyms or meronyms/holonyms existing in the extended ontology. However, no penalty is assigned for synonym relations in the extended ontology. Since their ontology and extended ontology have a hierarchical structure, the term weights of the children nodes propagate upward, and are accumulated. Among the hypernyms, the ones with higher accumulated weights are ultimately selected as topics of the input Web document.

In another example of using WordNet for ontology expansion and knowledge extraction, Alani et al. (2003) also employed WordNet as an information extraction (IE) tool to develop a domain ontology for artists and paintings. In their study, key information items such as entities and sentences were identified from Web documents using IE tools (i.e., GATE[3]). Since these Web documents could be selected both automatically and manually, we can assume that a search engine's application programming interface (API) was utilized for the automatic selection of relevant webpages (see "Web" column in Table 1). Extracted entities were then further enhanced with lexicons from WordNet. For example, for the sentence "Rembrandt Harmenszoon van Rijn was born on 15 July 1606 in Leiden, The Netherlands," WordNet associated "Leiden" as a city, and "The Netherlands" as a country. WordNet also was used to reduce linguistic variations of relations, expressed in verbs, in the ontology. For example, a verb "depict" is matched with

[2] http://dir.yahoo.com/

[3] General Architecture for Text Engineering (GATE): http://gate.ac.uk/ie/

"portray (synonym)" as well as "represent (hypernym)." The extracted information and enhanced lexicons then can be inserted into biography templates and the constructed biographies of the queried artists are delivered to users of their system.

WordNet was employed as an underlying reference ontology to retrieve suspicious emails in Du et al.'s digital forensics study (Du, Jin, de Vel, & Liu, 2008). The researchers conducted query expansion and query reduction applying WordNet and Latent Semantic Analysis (LSA) (Deerwester et al., 1990), respectively, in order to improve the retrieval of emails that forensics experts might search for using Boolean queries. Realizing the limitations of exact matching between the user queries and emails of interest, each query term was expanded with relevant hypernyms, hyponyms, and synonyms using WordNet. Only the first sense (i.e., synonym) of each query term was used to reduce the inclusion of noise senses because WordNet provides senses in the order of their estimated use frequency. To further refine the expanded queries, the similarity between each expanded term and the original term was computed using LSA. Then, the expanded term, for which the similarity value was lower than a defined threshold, was discarded from the final query. Additional details of LSA use in this study are presented in Section 2.4.2.

The World Wide Web

The Web is considered to be a large, rich, unstructured, and heterogeneous resource of (textual) information (Banko & Etzioni, 2007). Massey et al. (2011) developed a prototype system for topic identification using a search engine API to access information on the Web. This approach is very similar to their *ReAD* method explained in Section 2.1.2, except for the fact that their external knowledge source was changed from a dictionary for retrieving word definitions to the Yahoo! Search API for accessing relevant information from the Web. Other uses of information from the Web include information extraction (IE) as well as ontology development and enrichment.

Agirre et al. (2000) explored an approach for enriching WordNet using the Web in order to improve three known shortcomings of WordNet. For example, semantically-varied concepts in WordNet lack explicit links between them (e.g., *song-to* and *sing* are not related). In addition, topically-related concepts may not have explicit links (e.g., *bat* and *baseball*, *fork* and *dinner*). Also, the diversity of word definitions (i.e., "senses" in WordNet term) attached to each WordNet concept might be problematic at times in identifying the appropriate sense (e.g., word *line* has 32 senses). Therefore, the researchers attached a list of closely-related terms, namely the *topic* *signature*, to each WordNet concept. For this investigation, queries were constructed from the WordNet concepts and their synsets, hypernyms/hyponyms, and meronyms/holonyms. Then, 100 relevant Web documents were retrieved from the Web, after which they were processed to extract the *topic signature* for each query. This process continued for all the concepts in WordNet.

Use of external knowledge sources is becoming an essential part of many text analysis, natural language processing, and ontology-related studies. Considering the continued advancement of Wikipedia in its volume and quality—not to mention the proliferation of highly-sophisticated APIs of various kinds to access information on the Web—it is obvious that users will have access to increasingly advanced and creative information systems in the near future.

2.4 OpenCalais Natural Language Processing API

Origin

Thomson Reuters acquired a text analytics company called ClearForest, in April 2007. Then, it began to provide a service called “Calais[4]”, which is capable of automatically extracting semantic metadata from text documents, in January 2008. There are two versions of this. The free version is called “OpenCalais”, and its counterpart commercial version is called “ProfessionalCalais.” Basically, the differences between the two are in their text processing performance. OpenCalais can process 50,000 submissions of texts per day; however, ProfessionalCalais with a basic contract could process 100,000 submissions. OpenCalais allows submission of texts maximum four times per second. In contrast, ProfessionalCalais accepts submissions five times faster, i.e., 20 submissions per second.

What it does

Calais reads unstructured texts and extracts a set of metadata such as entities (and their relevance scores in the range 0-1 with 3-digit decimal points), facts, events, relations, and social tags, to name a few, in Resource Description Format (RDF) based on natural language processing and machine learning technologies. Currently supported types of entities, events, and facts can be found at http://www.opencalais.com/documentation/calais-web-service-api/api-metadata/entity-index-and-definitions.

[4] http://www.opencalais.com/

Its place in the larger scheme of things

The Calais Web service is used by individual developers, researchers, and software companies for various information extraction purposes. It also is incorporated in content management systems such as Drupal (plug-in at http://drupal.org/project/opencalais), WordPress (plug-in called Tagaroo at http://tagaroo.opencalais.com/), and Web browsers such as Firefox and Internet Explorer for improving the browsing experience.

Its role in relation to research questions

In this dissertation, OpenCalais was used to partly address Research Questions 1 and 2. Although it was not clear whether OpenCalais could produce human-comparable topics, it did not require much human intervention to extract information from texts considering that it returned a list of extracted information given a textual document. Therefore, its performance was compared to that of the proposed approach, Xpantrac. Also, the Inter-Indexer Consistency metric was used in this comparison, addressing Research Question 2.

2.5 Corpus-Based Statistical Methods

It is important to discuss two widely-used approaches for extracting topics from a large text corpus—namely, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Latent Semantic Analysis (LSA) (Deerwester et al., 1990). LDA is a probabilistic language model, and LSA is based on the VSM and the singular value decomposition (SVD) method.

Latent Dirichlet Allocation (LDA)

LDA (Blei et al., 2003) is a probabilistic language model that can be applied to a large text corpus in order to output latent (i.e., hidden) topics from text documents. This model assumes that any document is a mixture of latent topics, and that each topic is expressed as a list of words. LDA measures the probability of generating a document d with S_d words by a mixture of latent topics $\{z_1,\dots,z_T\}$ (X. Liu, Wang, Johri, Zhou, & Fan, 2012):

$$p(w_1,\dots,w_{S_d}) = \prod_{i=1}^{S_d} \sum_{j=1}^{T} p(w_i \mid z_j) p(z_j \mid d)$$

where $p(w_i \mid z_j) = \phi^{(j)}$ is a multinomial distribution of a word w_i over a topic z_j with a Dirichlet prior α, and $p(z_j \mid d) = \vartheta^{(d)}$ also is a multinomial distribution of a topic z_j over a document d with a Dirichlet prior β.

Regarding the values for α and β, Griffiths et al. (2008) suggest $50/T$, and a small value of 0.1, respectively, for a find-grained decomposition of the text corpus into topics.

In contrast to the batch LDA algorithm noted above, online LDA (Matthew D Hoffman, 2010) can update the model incrementally as more documents are added to the corpus. In fact, new documents do not have to be stored locally. Online LDA examines new documents, updates the model, and discards the documents. Its efficacy was confirmed in that it outperformed the batch LDA using 98,000 unique Wikipedia articles.

An example output from LDA is presented in Table 2. For example, a single topic (e.g., business) has to be inferred manually from eight words in the third column: service, systems, companies, business, company, billion, health, and industry. As more documents are added and processed, topics have to be updated, requiring more human involvement if one wishes to generate meaningful topics that correspond to "business." Thus, LDA may be less efficient when directly applied to disaster archives where newly created documents are continuously added and have to be topic-indexed with minimal human intervention.

Table 2. Three business-related topics and their top eight words by online LDA (table adopted from (Matthew D. Hoffman, 2010)).

No. Documents	2,048	12,288	65,536
Top Eight Words	systems	service	business
	road	systems	industry
	made	companies	service
	service	business	companies
	announced	company	services
	national	billion	company
	west	health	management
	language	industry	public

Latent Semantic Analysis (LSA)

Latent Semantic Analysis (or Indexing) (Deerwester et al., 1990) is a widely-used technique—especially in the fields of Information Retrieval (IR), text clustering and classification, and topic identification. Although it is based on VSM, LSA is more sophisticated because it is able to apply the singular value decomposition (SVD) to achieve a large-scale dimensionality reduction; additionally, it allows semantic matching between text documents. LSA features the following steps (Bradford, 2008):

- Documents in a text corpus are transformed into a term-document matrix (TDM) in the VSM, where each row corresponds to a term in the corpus and each column corresponds to a document. Each cell value is the term frequency, *tf*.
- Term-weighting might be applied (e.g., inverse document frequency, *idf*)
- SVD is applied to reduce the dimension of TDM. Three matrices are produced, one of which has non-zero (singular) values only on the diagonal cells. Singular values are organized in the order of their significance from the top-left to the bottom-right.
- Dimensionality is reduced by selecting only the *k* largest values on the diagonal, along with the corresponding columns in the other two matrices. This reduced matrix represents the new vector space, where the similarity between any two objects (i.e., term-term, document-document, and term-document) could be computed with the cosine similarity measure.

One of the challenges of LSA is the difficulty in determining the optimal number of dimensions. In general, for a higher number of dimensions, more specific comparison of concepts can be performed. Bradford (2008) reported optimum values for *k* (i.e., the number of dimensions) appearing in 49 other studies, and suggested that a user employ 300-500 dimensions for the best results. Landauer et al. (Landauer & Dumais, 2008) recommended using a text corpus that included more than 20,000 word types and more than 20,000 passages in order to avoid faulty results. They also suggested using 300 dimensions as an optimal *k*, but mentioned that 200-2,000 dimensions could be considered to be within a useful range. Considering that the optimal *k* should be empirically determined as the number of documents in a corpus changes, LSA also was viewed as a less efficient approach for satisfying the specific conditions of this dissertation study, which required minimal human intervention.

2.6 Metrics for Evaluating Topics

2.6.1 Inter-Indexer Consistency

The purpose of assigning topic tags to webpages and online articles is to improve the "findability" of those resources for human readers. The proposed topic tagging system utilized in this study produces a set of tags for any given text document. However, it is complex and costly to measure the effectiveness and quality of those automatically-indexed topics. One way of evaluating the quality of topic tags generated in this way is to consider the prototype system as a "machine tagger (indexer)," and then compare the tags produced by the machine tagger with those assigned by multiple human indexers. The goal of this exercise is, of course, to compute the consistency between them; this is known as the Inter-Indexer Consistency (IIC).

The Inter-Indexer Consistency is defined as:

> "The degree of agreement in the representation of the essential information content of a document by certain sets of indexing terms selected individually and independently by each of the indexers (Zunde & Dexter, 1969)"

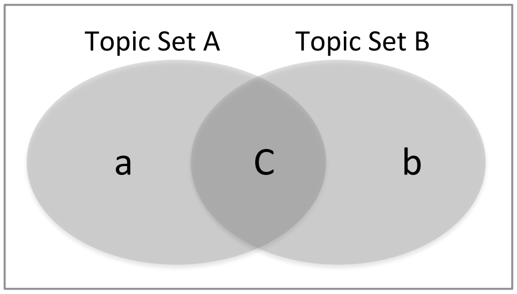


Figure 4. Two topic sets A and B with their intersection topic set C. (Additionally, $a = A - C$, and $b = B - C$).

More specifically, when we have two sets of indexed topics for a same document, A and B, as shown in Figure 4, Rolling's IIC (Rolling, 1981) is computed as:

$$Rolling's\ IIC = \frac{2 \cdot |A \cap B|}{|A| + |B|}$$

Medelyan compares Rolling's IIC with Hooper's IIC, which is computed as:

$$Hooper's\ IIC = \frac{|A \cap B|}{|A \cup B|}$$

The relationships of the two IIC formula can be expressed as:

$$Hooper's\ IIC = \frac{Rolling's\ IIC}{2 - Rolling's\ IIC}$$

Interesting connections can be found between these two IIC metrics and the Jaccard and Dice coefficients in Medelyan's study (2009, p. 24). The Jaccard similarity coefficient is used to compute the similarity between two sample sets. It is computed by dividing the intersection set between two sets with the union set, making this coefficient the same as Hooper's IIC. It should be noted that the Jaccard similarity coefficient corresponds to the Tanimoto Similarity (Rogers & Tanimoto, 1960). For Rolling's IIC, it is exactly the same metric as the Dice's coefficient.

Throughout this dissertation, Rolling's IIC is used extensively to evaluate the quality of topics from the proposed system. For this study, text document collections that could be indexed with topics assigned by multiple humans were developed.

2.6.2 Precision and Recall

Precision and recall represent metrics from the information retrieval field (Baaeza-Yates & Ribeiro-Neto, 1999; Manning, Raghavan, & Schütze, 2008); they are frequently used to evaluate the performance of various algorithms such as document matching, machine learning classification, and topic tagging. The outputs of these algorithms are compared with the "gold standard" outputs by another system.

For example, precision can be computed as the proportion of the matching topic tags (i.e., C) from all the retrieved topics (A) (see Figure 4) by the proposed topic-tagging algorithm:

$$precision = \frac{|C|}{|A|} = P(relevant \mid retrieved)$$

Recall is the proportion of the matching topic tags from all the retrieved topics (B), which are assigned by human topic indexers or exist as the gold standard:

$$recall = \frac{|C|}{|B|} = P(retrieved \mid relevant)$$

In an ideal case, the precision value should be 1, and the recall value also should be 1, which means that all three sets, A, B, and C are exactly the same. However, in most cases, precision and recall are inversely proportional to each other.

2.6.3 F-Measure

F-measure is used when taking both precision and recall into consideration as a single harmonic mean value. By adjusting the parameter value of *β*, either precision or recall could be weighted more compared to the other:

$$F_{\beta} = (1+\beta^2) \cdot \frac{precision \cdot recall}{\beta^2 \cdot precision + recall}$$

The most frequently used value of *β* is 1, which considers precision and recall values equally:

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

In this dissertation, precision, recall, and F1 were utilized for topic evaluation whenever a gold standard data set was available—or when one was developed from topics assigned by multiple human indexers.

2.7 Summary

First, this literature review provides an overview of Cognitive Informatics (CI) and its inspiration for this dissertation study. Then is a discussion of the Vector Space Model (VSM), which is a widely used Information Retrieval model. CI and VSM are especially important since they serve as a foundation for the two major components of the approach utilized in this dissertation. Three types of external knowledge sources (Wikipedia, WordNet, and the Web), their resources, and selected methods to utilize them also are described in this chapter. In particular, Wikipedia and WordNet are becoming an essential component of various text processing and natural language processing studies to improve performance. Studies such as query expansion and automatic ontology construction involve the use of information on the Web. This dissertation investigation also uses the information on the Web through access with a search engine API.

The main concepts associated with LDA and LSA also are explained in this review of the literature—especially the point that they require a large text corpus, as well as human intervention, in order to produce meaningful topics. Additionally, the output of LDA is generated as groups of words, from each of which a topic can be inferred. For these reasons, this strategy might be less efficient when applied to a dynamically changing corpus such as a disaster archive, where additional documents are likely to be added randomly and continuously as a disaster progresses. This chapter concludes with several evaluation metrics used extensively in this dissertation.

Chapter 3. Design and Development

In this chapter, we start by presenting the construction of our tool, and its component details, in Section 3.1. The workflow and algorithm of Xpantrac are given in Section 3.2, followed in Section 3.3 by description of an experiment for finding optimal parameters and evaluating generated topics using the Inter-Indexer Consistency metric. Section 3.4 discusses the benefits and limitations of the VSM, as well as efforts to overcome its limitations. We provide a summary in Section 3.5.

3.1 Components

The architecture of Xpantrac shown in Figure 5 has two parts, Expansion and Extraction. Given an input text, the Expansion part creates a tailored focused text corpus. This corpus is then analyzed by the Extraction part to identify significant topic words that represent the input text.

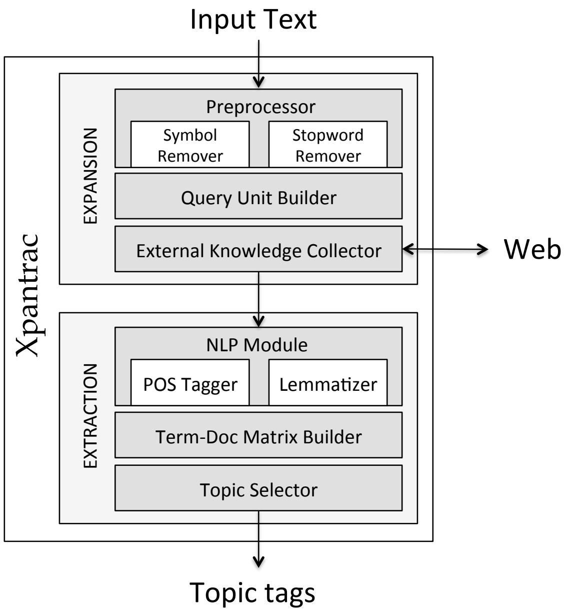


Figure 5. Components of Xpantrac grouped into two parts: Expansion and Extraction.

Due to the modular design of Xpantrac, any component can be flexibly replaced with an upgraded component. In fact, the entire set of components that is responsible for either Expansion or Extraction can be replaced in order to satisfy different needs. For example, a component for extracting Resource Description Framework (RDF) (Lassila, Swick, W3C, 2004; E. Miller, 2005) triples, which generally consist of the subject, verb, and the object, can be used in place of the current topic extraction component in Extraction.

3.1.1 Expansion Part

The Expansion part is responsible for accepting input texts, and expanding them to create a "derived corpus" of relevant information by accessing an external knowledge source, such as the Web. The Expansion part consists of three components; the text data is processed and expanded by going through a Preprocessor, a Query Unit Builder, and an External Knowledge Collector, as shown in the upper box of Figure 5. Once the Expansion part creates a corpus, that corpus becomes an input to the Extraction part displayed in the lower box of Figure 5.

3.1.1.1 Preprocessor

The Preprocessor removes symbol characters (e.g., #, $, %, !, ?) and stopwords (e.g., 'a', 'the', 'and', 'that', 'is', 'he', 'she') to reduce unnecessary processing of less informative character strings. The preprocessed input texts are sent to the Query Unit Builder.

3.1.1.2 Query Unit Builder

The Query Unit Builder segments the preprocessed input texts into uniform sized groups of words, which are eventually sent to a search engine's Application Programming Interface (API). The purpose of grouping words is to keep the contextual information of words by considering their neighboring words together. However, there needs to be a balance in the number of words in a single group because too many words in a single search API query often results in the retrieval of much too specific content together with a small number of results. On the other hand, too few words in a single query, for example, a one-word query, would result in the retrieval of content that is too general and too broad. Therefore, finding an optimal size text of the query unit for the Query Unit Builder is important. The rationale for using multi-word query units and the details of finding optimal parameters are presented in Section 3.2.2 and Section 3.3, respectively.

3.1.1.3 External Knowledge Collector

As its name suggests, the External Knowledge Collector accesses a knowledge source, which is located outside of the Xpantrac system, to search and retrieve relevant information for each query unit provided by the Query Unit Builder. Examples of external knowledge sources include the Web (currently used by Xpantrac), the WordNet lexical database (Fellbaum, 2010; G. A. Miller, 1995), Wikipedia (Coursey et al., 2009; Martin, 2011; Milne, Medelyan, & Witten, 2006; Schonhofen, 2008; Yun, Jing, Yu, Huang, & Zhang, 2011), and any other collection of words and definitions. To access the Web programmatically, Xpantrac uses an API of a search service such as Google Custom Search API (2013), Yahoo! BOSS Search API (2013), or Bing Search API (2013). There are other search APIs provided by different vendors with different API call rate limits and pricing options. For experiments in this dissertation, the Bing Search API, the Yahoo! BOSS Web API, and the Yahoo! BOSS News API were used, and their performance was compared.

3.1.2 Extraction Part

The Extraction part is where a list of significant words is extracted from a derived corpus as representative topics. Part-Of-Speech (POS) tagging (Perkins, 2013a; Sjöbergh, 2003) from the Natural Language Processing (NLP) field is applied during the topic extraction process. Based on the POS tags assigned to the texts in the derived corpus, Xpantrac can extract topics only from nouns, only from verbs, or from both. The assumption is that the topics from nouns may represent concepts and objects very well, whereas topics from verbs may give us better insights about the activities appearing in the texts. The Term-Document Matrix Builder and the Topic Selector also are applied to the POS tagged corpus to finally produce a list of topics.

3.1.2.1 NLP Module

When a derived corpus is loaded into the Extraction part, the NLP Module performs two tasks. The first task is to apply a POS tagger to the corpus in order to select only nouns, only verbs, or a combination of nouns and verbs. The second task is to find "lemmas" of the selected nouns or verbs to resolve singular and plural forms. There exist various designs of POS taggers, including n-gram taggers (e.g., unigram, bigram, or trigram taggers), Affix tagger, Regular Expression tagger, Brill tagger, etc. (see Table 3).

Table 3. The description of the POS taggers used in this dissertation.

Name	Description
Trigram tagger	It uses a word string and the preceding two words' tags. It is based on a second order Markov model.
Bigram tagger	It uses a word string and the preceding one word's tag. It is based on a first order Markov model.
Unigram tagger	It assigns tags based on a word string.

Affix tagger	It considers a leading or trailing substring of a word string. A substring of a token is searched in a table and a matching tag is returned if one exists.
Regular Expression tagger	It uses regular expression patterns of word strings.

One way to build an effective and robust POS tagger is to chain multiple taggers in sequence. For example, one POS tagger becomes a back-off tagger of another. Therefore, if a tagger fails to add tags for certain words, its back-off tagger will attempt to tag those words instead. In this way, we can have high quality POS tagging results (Perkins, 2013a; 2013b). Table 3 presents a list of descriptions and names of the POS taggers chained and used in this dissertation.

Chaining Taggers as a Back-Off of Another Tagger

As mentioned in the previous paragraph, if a tagger could not determine a tag for a given token (e.g., individual word, punctuation marks), its back-off tagger, which is shown on the right side of each tagger in Figure 6, will take over the task and attempt to assign a tag. If this back-off tagger fails again, a back-off of the failed back-off tagger will attempt to tag the token in a cascading manner. For example, if Trigram Tagger fails to assign a tag, Bigram Tagger, which is a back-off of Trigram Tagger, takes over the task; if that fails too, then use is made, as required, of Unigram Tagger, else Affix Tagger, etc.

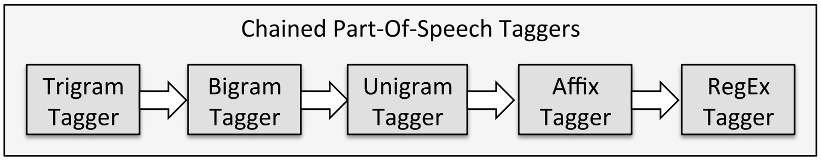
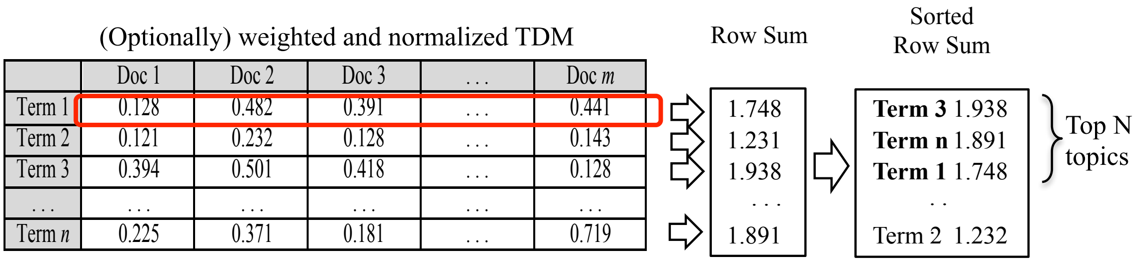


Figure 6. Chained POS taggers for improved accuracy and speed of tagging. If one tagger fails, its back-off tagger (on its right) takes over the task.

3.1.2.2 Term-Document Matrix Builder

The Term-Document Matrix Builder first develops a term index using the unique words from the derived corpus. Using this index, the Matrix Builder constructs a term-document matrix as in the Vector Space Model (Salton, 1971; Salton et al., 1975) by representing each term index with the corresponding row of the matrix, and each document ID with the corresponding column of the matrix, as shown on the left in Figure 7. The matrix cell values show the frequency of a term in a document, and suggest the potential significance of the term in the document. Please see Section 2.2 for details.

3.1.2.3 Topic Selector



(Optionally) weighted and normalized TDM

	Doc 1	Doc 2	Doc 3	. . .	Doc *m*
Term 1	0.128	0.482	0.391	. . .	0.441
Term 2	0.121	0.232	0.128	. . .	0.143
Term 3	0.394	0.501	0.418	. . .	0.128
. . .	. . .	. . .	. . .	. . .	. . .
Term *n*	0.225	0.371	0.181	. . .	0.719

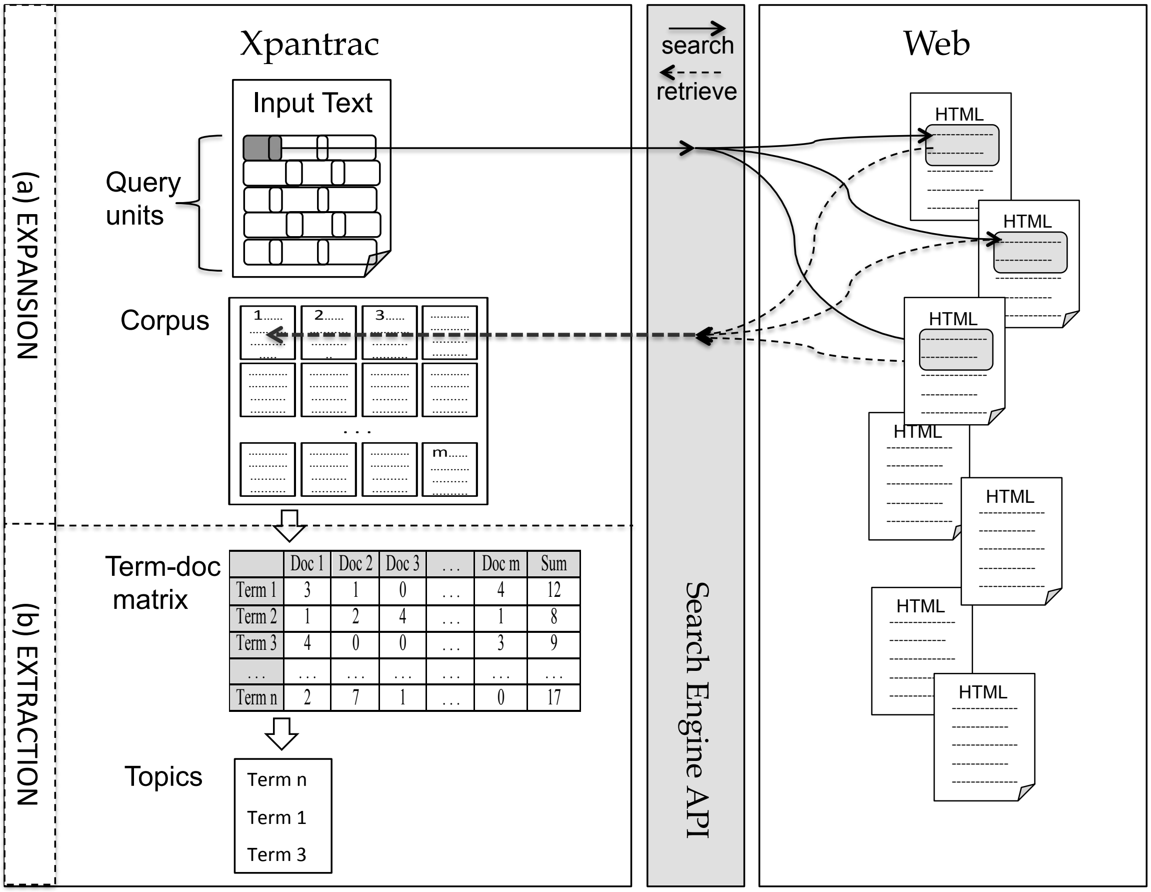
Figure 7. The process of extracting top N topics.

The role of the Topic Selector is to identify significant words, which are representative of an input text, from the term-document matrix developed in the previous step. The Topic Selector optionally applies a term weighting scheme, such as the *document frequency (DF)* or the *term frequency * inverse document frequency (TF*IDF)* (Salton et al., 1975; Salton & Buckley, 1988; W. Zhang, Yoshida, & Tang, 2008; 2011). Term frequencies might be high not because the term is considered significant in a document, but merely due to a longer length of the document. To reduce such effect, normalization is applied. Please see Section 2.2 for details of TF*IDF and normalization.

An example term-document matrix (TDM) with application of term-weighting and normalization is shown on the left of Figure 7. To identify topics, the Topic Selector computes the row sums of the matrix. Each row sum shows the significance of the term across all the documents. The Topic Selector then sorts the terms in descending order based on their row sum values. The terms with the higher row sum values will be located on top of the ordered list, and are more likely to be selected as topics.

3.2 Workflow

Figure 8 graphically explains a workflow of Xpantrac. During (a) Expansion, Xpantrac collects relevant information of its input document from the Web by querying groups of words in the input text to a search engine in sequential order. The volume of the retrieved information is much larger than that of an input document. As the result of this process, a derived document corpus of retrieved Web information is created. Search and retrieve arrows in Figure 8 show the process of expanding a single group of words (i.e., a query unit) from the input text to a derived document in the corpus. This data expansion is mediated by a search engine API. During (b) Extraction, the corpus is transformed into a term-document matrix, where significant words can be extracted as topics.



	Doc 1	Doc 2	Doc 3	…	Doc m	Sum
Term 1	3	1	0	…	4	12
Term 2	1	2	4	…	1	8
Term 3	4	0	0	…	3	9
…	…	…	…	…	…	…
Term n	2	7	1	…	0	17

Figure 8. The workflow of Xpantrac: (a) Expansion, and (b) Extraction processes. Search and retrieve arrows show the direction of the request and retrieval operations.

3.2.1 Algorithm

The workflow described above can be further elaborated in finer granularity as follows:

(a) Expansion Process

1. An input document D_{in} is preprocessed to exclude stopwords and symbol characters. The preprocessed document D_{clean} has a total of N words:

$$D_{clean} = \{w_1, w_2, w_3, \dots, w_N\}$$

2. The words in D_{clean} are grouped into M query units of q_j ($j = 1, 2, ..., M$) with a uniform size i except for the last query, whose size is $1 \leq |q_M| \leq i$:

   $$Q = \{q_1, q_2, q_3, ..., q_M\}$$

   $$q_j = \{w_{j1}, w_{j2}, \ldots, w_{ji}\}, \text{ where } i > 0 \text{ and } 1 \leq j \leq M$$

   Sliding Windows is applied to the word grouping by overlapping a single word between adjacent query units. Therefore, the last word in a certain query unit is the same as the first word in the next query unit:

   $$q_j = \{w_{j1}, w_{j2}, \ldots, w_{ji}\}, q_{j+1} = \{w_{(j+i)1}, w_{(j+i)2}, \ldots, w_{(j+i)i}\}, \text{ where } w_{ji} = w_{(j+i)1}$$

3. Query units q_j $(j = 1, 2, ..., M)$ are sequentially queried to a search engine through the search engine's API. As a result, a maximum of k relevant webpage descriptions v_{jh} $(h = 1, 2, ..., k)$ for each q_j are retrieved from the Web. The maximum and default value of k provided by the search engine APIs used in this dissertation is 50. Therefore, any integer value of k between 1 and 50 can be selected. The retrieved webpage descriptions are concatenated to form a single derived document E_j:

   $$E_j = \{v_{j1,} v_{j2,} v_{j3}, \ldots, v_{jk}\}, \text{ where } 1 \leq k \leq 50$$

4. Corpus C is developed as a result of Step 3, and then becomes an input to the Extraction process:

   $$C = \{E_1, E_2, \ldots, E_M\}$$

(b) Extraction Process

1. For each derived document E_j $(j = 1, 2, ..., M)$, part-of-speech (POS) tags are assigned to optionally select only nouns (or only verbs).
2. Each selected noun (or verb) is lemmatized to resolve the singular and plural forms. Thus, the POS-tagged and lemmatized E_j is:

   $$Epl_j \ (j = 1, 2, ..., M)$$

3. From an updated corpus $Cpl = \{Epl_1, Epl_2, \ldots, Epl_M\}$, an index *Index* is constructed with the z unique terms that are present in *Cpl:*

   $$Index = \{t_1, t_2, \ldots, t_z\}$$

4. Based on *Index* and *Cpl*, a term-document matrix *TDM* with size $z * M$ is constructed.
5. (Optional) A term weighting scheme such as TF-IDF (*term frequency * inverse document frequency*) can be applied to *TDM* and the matrix becomes $TDM_{weighted}$.
6. The matrix cell values are normalized to alleviate the effect from different lengths of Epl_j $(j = 1, 2, ..., M)$. Thus, we have $TDM_{norm.}$ Please see Section 2.2 for details.

7. All *M* cell values in each row of TDM_{norm} are summed as a single significance score Sig_y ($y = 1, 2, 3, \ldots, z$) for each term t_y in *Index*, and a tuple (t_y, Sig_y) is generated.
8. All the generated tuples {(t_1, Sig_1), (t_2, Sig_2), …, (t_z, Sig_z)} are sorted in descending order of Sig_y ($y = 1, 2, 3, \ldots, z$). The top *r* tuples with the highest Sig_y values are selected. Then, terms from those tuples are extracted as topics.

3.2.2 Query Unit

Xpantrac segments an input text into several query units in the expansion process. It is based on an assumption that when a word is grouped together with its neighboring words, the context of the word would be better preserved and its ambiguity caused by the polysemy of the word might be reduced. Several studies of using "multi-word units" instead of single words for Information Retrieval tasks support this assumption (Danielsson, 2003; W. Zhang et al., 2008). Reducing ambiguity in query units is very important in Xpantrac in order to expand an input text with high quality relevant Web information.

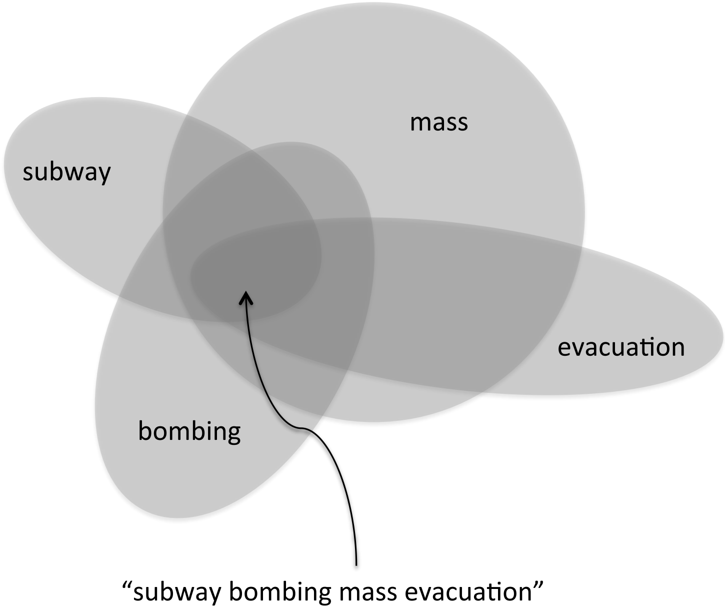


Figure 9. Grouping words as a query unit for improved word meaning disambiguation and for preserving the context.

Figure 9 explains the preservation of the context when using multi-word units with an example. When we query "subway" to a search engine, it returns several sandwich restaurant webpages on top of the search result page. However, the "subway" in "subway bombing mass evacuation" would refer to the underground electric railroad system. In this example, the context of the word "subway" could be preserved when it is used with its neighboring words.

As a means to show the potential effectiveness of using multi-word query units, the text in a news webpage (http://spare05.dlib.vt.edu/~seungwon/various_30/3.pdf) was segmented with varying sizes of query units from 1 to 10, and the first, median, and the second from the last query units were sampled (see Table 4). The reason for sampling the second query unit from the last was that often times the size of the last sample was smaller than the uniform query unit size. Each sample was queried using the Google search engine, and the existence and rank of the input document (i.e., an entire news webpage) was examined in the first three result pages.

The existence of the input document itself, of which the query unit was a part, might indicate that the query unit was effective in retrieving very relevant Web information. Table 4 shows that the search result did not include the input document when the size of the query unit was 1 or 2. This was probably due to the missing contextual information in a short query, which resulted in the retrieval of broad Web information. However, when the size became 3, the input document was retrieved from the first query unit. Considering that the first line of the input webpage was the title, querying using three words from the title was effective to extract the exact news article. For the size of 4, the first and the second from last query units seemed effective. The sample query units in Table 4 with size over 5 were all effective in retrieving the input document.

However, we do not want to use a very large query unit size, either, because the goal is to find a robust set of topics by considering many relevant webpage descriptions but not identical ones. Using large query units may result in retrieving only a small number of Web documents, whose content is almost identical and much too specific to the input.

Table 4. Ranking of an input document in Google search results with varying sizes of query units, which are part of a segmented input document.

Query Unit Size	Sampling position	Query Unit	Input Text Rank
1	First	dozens	n/a
	Median	helping	n/a
	Last -1	heard	n/a
2	First	dozens killed	n/a
	Median	damaged citys	n/a
	Last -1	heard background	n/a
3	First	dozens killed guatemala	1
	Median	damaged citys hospital	n/a

	Last -1	dog heard background	n/a
4	First	dozens killed guatemala earthquake	1
	Median	center rooms damaged citys	n/a
	Last -1	panicked dog heard background	1
5	First	dozens killed guatemala earthquake mexico	1
	Median	damaged citys hospital extensive damage	7
	Last -1	video lamps swinging panicked dog	2
6	First	dozens killed guatemala earthquake mexico city	1
	Median	helping operate relief center rooms damaged	1
	Last -1	dizzying video lamps swinging panicked dog	1
7	First	dozens killed guatemala earthquake mexico city magnitude	1
	Median	center rooms damaged citys hospital extensive damage	1
	Last -1	video lamps swinging panicked dog heard background	1
8	First	dozens killed guatemala earthquake mexico city magnitude earthquake	3
	Median	relief center rooms damaged citys hospital extensive damage	1
	Last -1	landslides dizzying video lamps swinging panicked dog heard	1
9	First	dozens killed guatemala earthquake mexico city magnitude earthquake struck	3
	Median	broke ms miranda helping operate relief center rooms damaged	1
	Last -1	entirely blocked landslides dizzying video lamps swinging panicked dog	1
10	First	dozens killed guatemala earthquake mexico city magnitude earthquake struck guatemala	1
	Median	broke ms miranda helping operate relief center rooms damaged citys	1
	Last -1	blocked landslides dizzying video lamps swinging panicked dog heard background	1

Another characteristic in the input document segmentation was the use of the *Sliding Windows* approach, which is often used in text segmentation studies (Tiedemann & Mur, 2008). When grouping the words, two-word phrases may be separated in half, and each of them could be included in a different (adjacent) query unit, possibly losing information as a single two-word phrase. For this reason, *Sliding Windows* with the step size of (window size - 1), i.e., one-word overlap between adjacent windows, was used to include such phrases. Since two-word phrases were most common, we used one-word overlap in this study.

Figure 10 shows the application of Sliding Windows on a snippet of text. The input text shows “san marcos”, which will lose its meaning when it is split up. Query Unit 1 contains “san” as its last word, which is only a part of “san marcos”. However, due to *Sliding Windows* and one-word overlap, “san marcos” is preserved in Query Unit 2 without being separated. Please also note that a noun phrase “hardest hit” was included in Query Unit 3 without being separated.

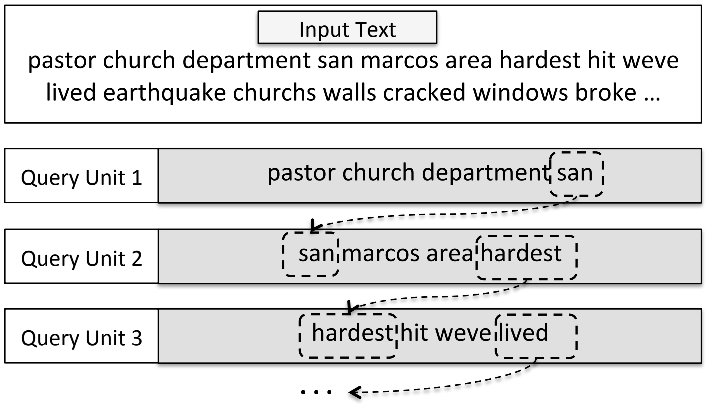
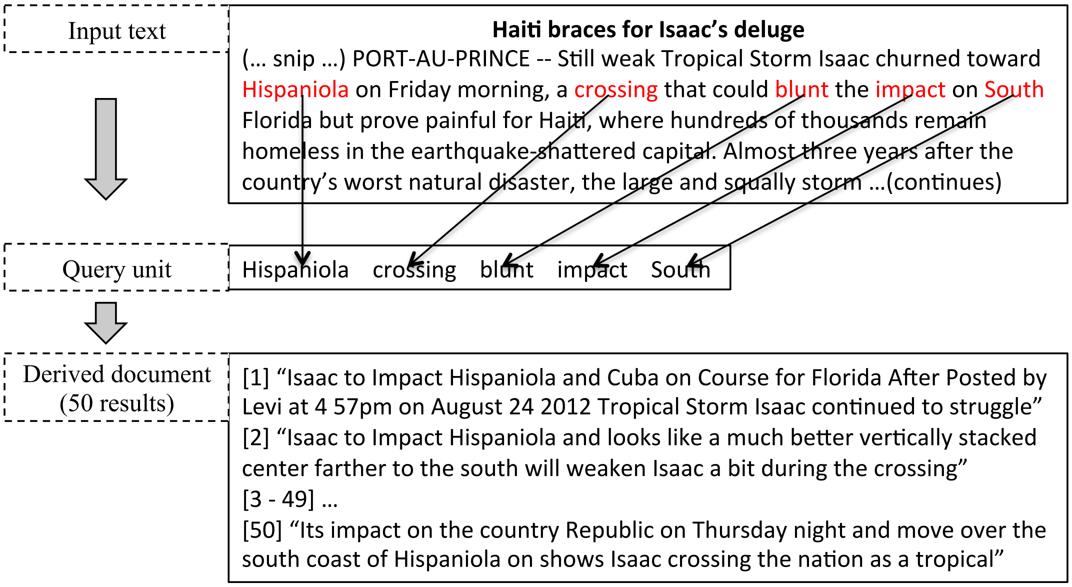

Input Text
pastor church department san marcos area hardest hit weve lived earthquake churchs walls cracked windows broke …
Query Unit 1
pastor church department san
Query Unit 2
san marcos area hardest
Query Unit 3
hardest hit weve lived
. . .

Figure 10. *Sliding Windows* with one word overlap in segmenting input texts. The last word in a query unit becomes the first word in the next query unit.

3.2.3 Derived Corpus


Input text
Query unit
Derived document (50 results)
Hispaniola crossing blunt impact South

Haiti braces for Isaac's deluge

(… snip …) PORT-AU-PRINCE -- Still weak Tropical Storm Isaac churned toward Hispaniola on Friday morning, a crossing that could blunt the impact on South Florida but prove painful for Haiti, where hundreds of thousands remain homeless in the earthquake-shattered capital. Almost three years after the country's worst natural disaster, the large and squally storm …(continues)

[1] "Isaac to Impact Hispaniola and Cuba on Course for Florida After Posted by Levi at 4 57pm on August 24 2012 Tropical Storm Isaac continued to struggle"
[2] "Isaac to Impact Hispaniola and looks like a much better vertically stacked center farther to the south will weaken Isaac a bit during the crossing"
[3 - 49] …
[50] "Its impact on the country Republic on Thursday night and move over the south coast of Hispaniola on shows Isaac crossing the nation as a tropical"

Figure 11. An example of a query unit and its derived document. The default (and maximum) number of results from a search API for a query is 50.

As mentioned in (a) Expansion step 3 in 3.4.1, returned webpage descriptions for each query unit are concatenated into a single derived document. For example, a query unit, "Hispaniola crossing blunt impact South," which was extracted from an input text, is expanded into its derived document (see Figure 11). The derived document consists of 50 webpage descriptions numbered from [1] to [50] in the figure. The collection of such documents becomes a Corpus. The number of webpage descriptions retrieved from a search API is one of the parameters of the Xpantrac system.

Figure 12 shows a line graph of cosine similarity values between the derived documents and the original input document. The input document was segmented into a total of 21 query units. Thus, there exist 21 derived documents numbered from 0 (i.e., the derived document from the first query) to 20 (i.e., the document from the last query) in the graph. Since each query unit contains words from the original input document, the derived document from each query shows some level of similarity to the original input. All such derived documents are used to extract topics in this study; however, excluding the ones whose cosine similarity values are lower than the average might be another approach. Whether this would affect the final output or not will need further investigation, as would ignoring the last derived document, especially when its query unit is smaller and hence likely to have less useful recall.

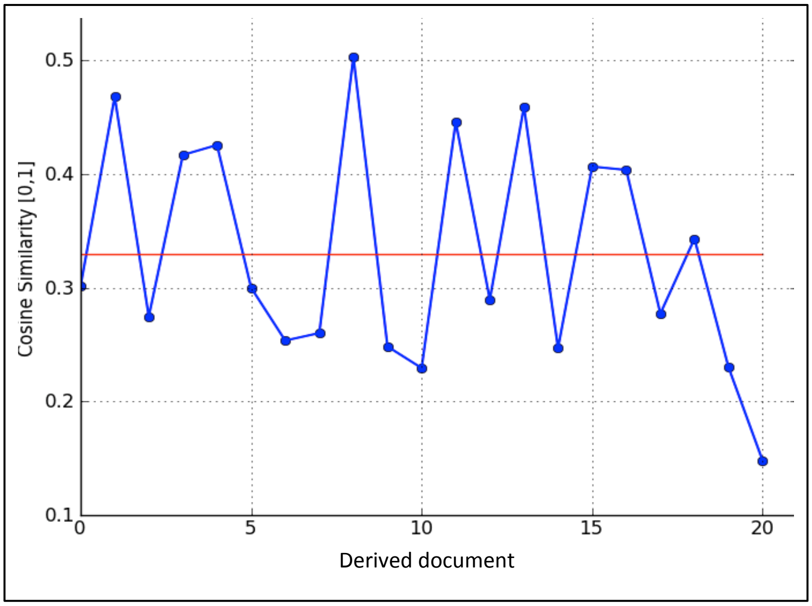


Figure 12. Cosine similarity values between the derived documents and the input document. The red line with cosine similarity value 0.34 denotes the average of the cosine similarity values.

3.3 Optimal Parameters

3.3.1 Data Sets

Table 5 shows two data sets. CTR_30 contains homogeneous documents and VARIOUS_30 contains heterogeneous documents. As the Note column in the table mentions, three human indexers assigned topic tags for the documents in CTR_30, and five indexers worked on VARIOUS_30. All of the documents in CTR_30 are centered on the Hurricane Isaac disaster. In contrast, VARIOUS_30 includes documents of varying scopes, e.g., natural disasters (e.g., earthquake, hurricane), political turmoil (e.g., Hamas-Israel conflict, Syrian conflict), and diseases that would cause personal/family-level emergencies (e.g., diabetes, heart attack, ACS, headaches, and AIDS).

These data sets were used to find optimal parameter configurations for Xpantrac as well as to compute Inter-Indexer Consistency between human indexers and Xpantrac. The reason for using two data sets with different characteristics was to investigate whether there exist optimal parameters that are robust and general enough to be used for different types of content. In the following sections, topic tagging experiments by human indexers, as well as by Xpantrac, are described in detail.

Table 5. Data sets used to find optimal parameters and Inter-Indexer Consistency.

Dataset	Description	Note
CTR_30	It consists of 30 documents, which are webpages selected randomly from the Hurricane Isaac archive. The data set download link is at http://spare05.dlib.vt.edu/~seungwon/various_30.zip	Indexed by 3 humans, homogeneous content, single disaster event, 214 words on average
VARIOUS_30	It contains 30 documents, which also are randomly selected, from several disaster webpage archives: • Guatemala earthquake: 6 docs • Hurricane Sandy: 4 docs • Israel-Hamas conflict: 5 docs • Syrian conflict: 5 docs • Diabetes: 5 docs • Heart attack: 2 docs • Acute Coronary Syndrome (ACS): 1 doc • Headache: 1 doc • AIDS: 1 doc The data set download link is at http://spare05.dlib.vt.edu/~seungwon/various_30.zip	Indexed by 5 humans, heterogeneous content, disaster events with varying scopes (e.g., natural disasters, political turmoil, personal/family emergencies), 418 words on average

* Note: The topic tags assigned by human indexers can be shared upon request.

3.3.2 Tagging

3.3.2.1 Xpantrac Parameter Values

Table 6 shows parameters of Xpantrac, their descriptions, and values. To find optimal parameter values for Xpantrac, we selected several discrete values for the *unit_size* and *num_api_return* parameters due to constraints in using a search engine API service (e.g., rate limits for the number of API calls allowed). For example, the search engine APIs used in this dissertation (i.e., Bing Azure API, Yahoo! Web Search API, and Yahoo! News Search API) could retrieve 50 webpage descriptions as its maximum and default value. Thus, the maximum value of the *num_api_return* parameter was set to 50, and its minimum was set to 1, which was the smallest possible value. Although 25 was the halfway-point between 1 and 50, we chose 10 in order to start examining the performance from the lower value and then to gradually increase it to higher values (in the future study).

Regarding the values of *unit_size*, we first selected 5 and 10 considering that the two numbers were effective in a small pilot study. To examine the performance of the larger *unit_size* values, we then selected 15 and 20. A smaller unit size means segmenting an input text into more and smaller pieces, which then require more API calls. For parameters, *tf_idf* and *only_nouns*, the values were either 1 (apply) or 0 (not apply). When *tf_idf* was set to 1, TF*IDF term weighting was applied. If the *only_nouns* parameter was set to 1, topics were selected only from nouns and proper nouns (i.e., excluding verbs).

Table 6. Parameters of Xpantrac.

Name	Description	Values
unit_size	Number of words to be grouped as a query unit	20, 15, 10, 5
num_api_return	Number of webpage descriptions to retrieve from the Web	50, 10, 1
tf_idf	TF*IDF term weighting	1: apply, 0: not apply
only_ nouns	Only (proper) nouns when extracting topics	1: apply, 0: not apply

In total, as can be seen in Table 7, there were 48 combinations of the parameter values (i.e., 4 x 3 x 2 x 2 = 48). To make it easier to refer to an Xpantrac run with a specific parameter combination, we assigned a machine ID, M1-M48, to each parameter combination. For example, M1 represented an Xpantrac run with *unit_size* = 20, *num_api_return* = 50, *tf_idf* = 1, and *only_nouns* = 1. Each of M1-M48 was considered as an individual "machine" topic indexer.

Table 7. The parameter combinations and their associated machine IDs.

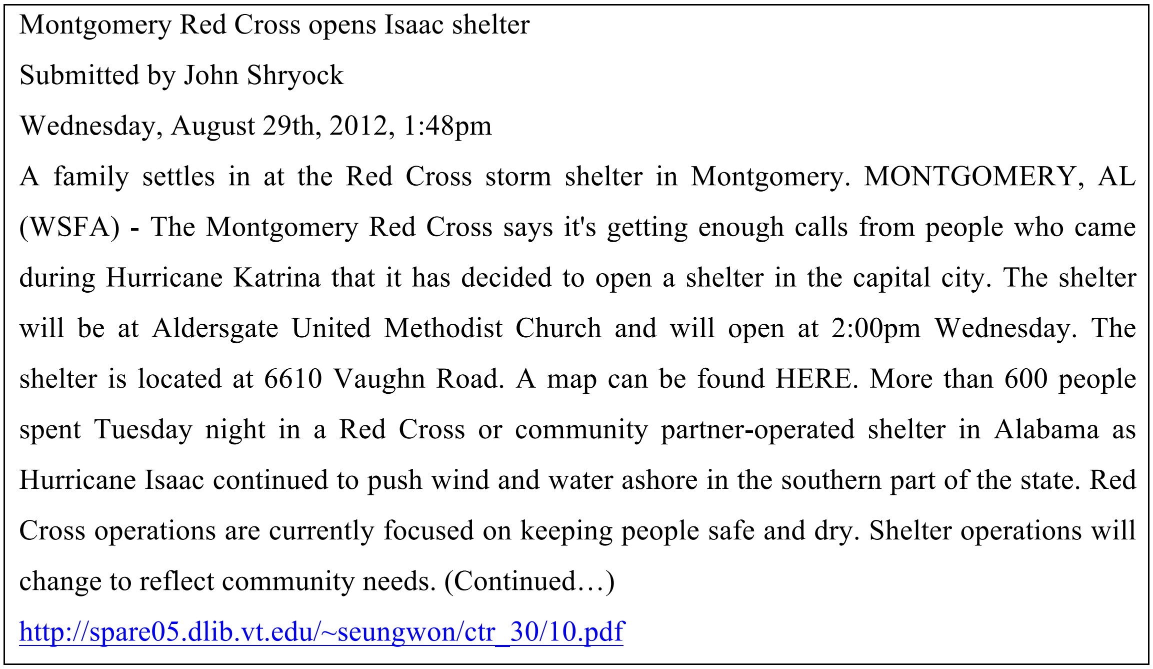
Machine ID	*unit_size*	*num_api_return*	*tf_idf*	*only_nouns*
M1	20	50	1	1
M 2	20	50	1	0
M 3	20	50	0	1
M 4	20	50	0	0
M 5	20	10	1	1
M 6	20	10	1	0
M 7	20	10	0	1
M 8	20	10	0	0
M 9	20	1	1	1
M 10	20	1	1	0
M 11	20	1	0	1
M 12	20	1	0	0
M 13	15	50	1	1
M 14	15	50	1	0
M 15	15	50	0	1
M 16	15	50	0	0
M 17	15	10	1	1
M 18	15	10	1	0
M 19	15	10	0	1
M 20	15	10	0	0
M 21	15	1	1	1
M 22	15	1	1	0
M 23	15	1	0	1
M 24	15	1	0	0
M 25	10	50	1	1
M 26	10	50	1	0
M 27	10	50	0	1
M 28	10	50	0	0
M 29	10	10	1	1
M 30	10	10	1	0
M 31	10	10	0	1
M 32	10	10	0	0
M 33	10	1	1	1
M 34	10	1	1	0
M 35	10	1	0	1
M 36	10	1	0	0
M 37	5	50	1	1
M 38	5	50	1	0
M 39	5	50	0	1
M 40	5	50	0	0
M 41	5	10	1	1
M 42	5	10	1	0
M 43	5	10	0	1
M 44	5	10	0	0
M 45	5	1	1	1
M 46	5	1	1	0
M 47	5	1	0	1
M 48	5	1	0	0

3.3.2.2 CTR_30 Data Set

Topic Tagging by Human Indexers

Three human indexers H1-H3, with backgrounds in the Library and Information Science (LIS) field, were recruited through the JESSE listserv[5] for manual topic labeling tasks. They were all female, and their ages ranged between 26 and 40. They had spent from 1 to 9 years in LIS, specializing in "academic libraries", "archival management", and "information technology education". Regarding their vocabulary size, two of them answered "5: very large", and one person "4: large" on the Likert scale of 1-5. Only one participant, H1, had previous tagging experience, and rated herself as an "intermediate" level tagger, while the other two rated themselves as "amateur".

Regarding the tasks, the human indexers each read 30 documents in CTR_30, deliberated on topics, and then entered several topic tags using an online survey system. On average, they assigned 6.1 topic words and phrases for each document (see Table 8). Interesting to note was that H1 assigned 8.23 topic words per document on average, whereas H2 and H3 assigned 4.52 and 5.16 topic words and phrases, respectively.



Montgomery Red Cross opens Isaac shelter

Submitted by John Shryock

Wednesday, August 29th, 2012, 1:48pm

A family settles in at the Red Cross storm shelter in Montgomery. MONTGOMERY, AL (WSFA) - The Montgomery Red Cross says it's getting enough calls from people who came during Hurricane Katrina that it has decided to open a shelter in the capital city. The shelter will be at Aldersgate United Methodist Church and will open at 2:00pm Wednesday. The shelter is located at 6610 Vaughn Road. A map can be found HERE. More than 600 people spent Tuesday night in a Red Cross or community partner-operated shelter in Alabama as Hurricane Isaac continued to push wind and water ashore in the southern part of the state. Red Cross operations are currently focused on keeping people safe and dry. Shelter operations will change to reflect community needs. (Continued…)

http://spare05.dlib.vt.edu/~seungwon/ctr_30/10.pdf

Figure 13. A document about Hurricane Isaac from the CTR_30 data set.

[5] http://web.utk.edu/~gwhitney/jesse.html

To elaborate on the details of produced topics, we used a text file from the CTR_30 data set, of which the snippet is shown in Figure 13. Its content was about the activities of the Red Cross and the affected community to open a shelter at a church in Montgomery, Alabama. The topics are produced by the human indexers, Xpantrac, and the OpenCalais API, and presented in Table 8, Table 9, and Figure 14, correspondingly.

As shown in Table 8, H1 assigned only topic words, whereas H2 and H3 often assigned topic phrases. All of the topic tags were put into a database after they were cleaned (by removing stopwords) and after tokenizing phrases.

Table 8. Topic tags assigned by H1-H3 for a document in Figure 13.

ID	Topic Tags
H1	Alabama, shelter, Isaac, emergency, preparation, plan, hurricane
H2	Tropical Storm Isaac, Montgomery Red Cross, Aldersgate United Methodist Church, shelter, American Red Cross Safe and Well, Hurricane App
H3	shelters, Hurricane Katrina survivors, Alabama, American Red Cross, app, Red Cross Hurricane app, I'm Safe app

Topic Tagging by Xpantrac

Table 9 shows topic tags, which were extracted by Xpantrac with different (selected) machine IDs and corresponding parameter settings, for a document in Figure 13. Their different parameter settings led to varying topical words. The number of topics was set to 10. As we did for the topics by human indexers, all of the topics by M1-M48 were added to a database for easier evaluation.

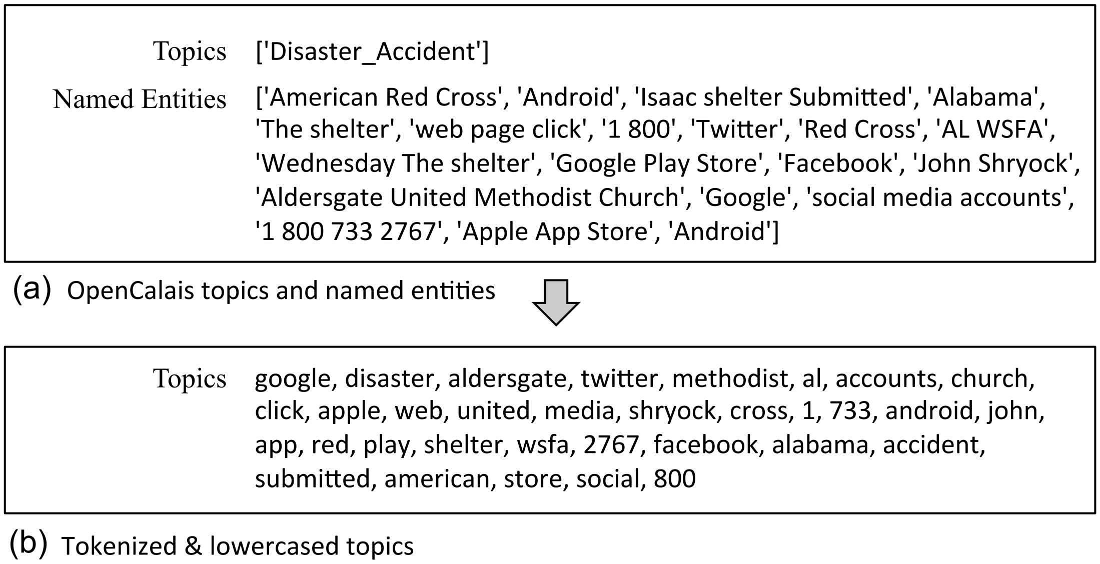
Table 9. Topic tags by (selected) machine IDs for a document in Figure 13.

Machine ID	Topic Tags
M1	account, medium, app, safe, community, response, call, facebook, message, button
M39	shelter, methodist, hurricane, cross, community, american, storm, church, isaac, red
M43	emergency, shelter, hurricane, cross, community, montgomery, american, storm, isaac, red
M48	help, emergency, shelter, safe, cross, montgomery, blood, storm, red, disaster

Topic Tagging by OpenCalais NLP API

The OpenCalais API (*OpenCalais*, 2013) returns topics and named entities of various types such as the location, company, facility, organization, persons, and phone numbers. Often, such returned information includes *n*-grams (e.g., phrases having *n* number of words) and duplicate words, thus we merged the topics and named entities, separated the n-grams into single words, and made all words lowercased (Figure 14).

When we compare those processed OpenCalais topics with the gold standard or the topics from human indexers, we apply lemmatization. Since the lemmatization converted all the words into their root forms, we did not have to stem the words, which was in contrast to what was done in other studies of topic identification (Medelyan, Frank, & Witten, 2009). All the scripts used in this dissertation, including the ones for OpenCalais topic extraction and processing, are downloadable, and the links are provided in the Appendix.



Topics ['Disaster_Accident']

Named Entities ['American Red Cross', 'Android', 'Isaac shelter Submitted', 'Alabama', 'The shelter', 'web page click', '1 800', 'Twitter', 'Red Cross', 'AL WSFA', 'Wednesday The shelter', 'Google Play Store', 'Facebook', 'John Shryock', 'Aldersgate United Methodist Church', 'Google', 'social media accounts', '1 800 733 2767', 'Apple App Store', 'Android']

(a) OpenCalais topics and named entities

Topics google, disaster, aldersgate, twitter, methodist, al, accounts, church, click, apple, web, united, media, shryock, cross, 1, 733, android, john, app, red, play, shelter, wsfa, 2767, facebook, alabama, accident, submitted, american, store, social, 800

(b) Tokenized & lowercased topics

Figure 14. Converting the output from OpenCalais for the document in Figure 13 shown in (a) into the tokenized & lowercased topic words in (b). The topics in (b) are lemmatized before they are compared with the gold standard.

Topic Network Graph Based on the Topics from Human Indexers, Xpantrac (M43), and OpenCalais (CTR_30 data set, document ID is 10)

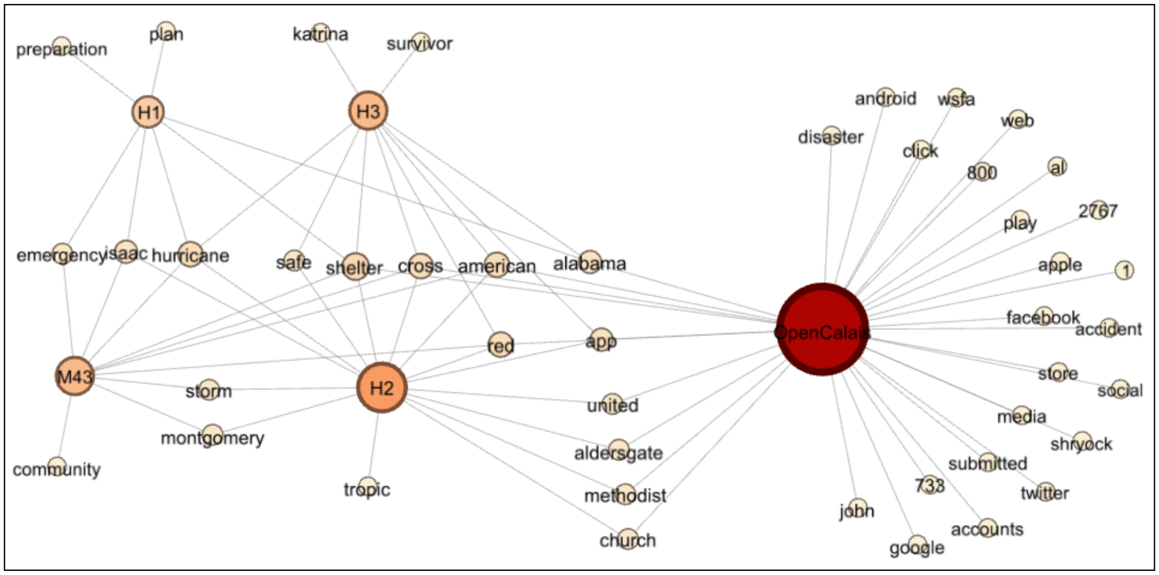


Figure 15. Topic network graph based on Table 8, Table 9 (M43 row), and Figure 14. Although OpenCalais has multiple topics, many of them are not shared.

For this example, Figure 15 presents the topics and their relationships, which are generated by human indexers, Xpantrac, and OpenCalais. It provides an overview of which topics are shared, and which topics are not. In this example, topics from all human indexers (i.e., H1, H2, and H3) and Xpantrac (M43) are mostly shared with each other with the exception of one or two topics. An important observation is that Xpantrac does look very similar to H3 in this network, sharing multiple topics with other indexers, and generates only a single topic, “community,” that is not shared with others. Based on this, we may presume that Xpantrac (M43) generates human indexer comparable topics.

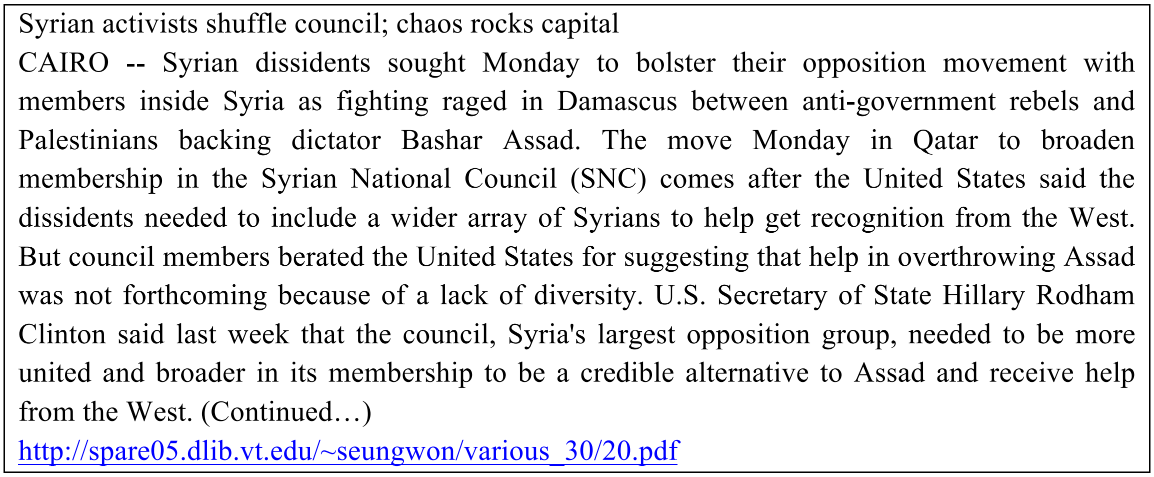
In contrast, note that almost two-thirds of the topics from OpenCalais are not shared with others. This suggests a cause of the low IIC values of OpenCalais when compared with human indexers.

3.3.2.3 VARIOUS_30 Data Set

Topic Tagging by Human Indexers

For this next study, five human indexers H11-H15 were recruited through the JESSE listserv. Again, all participants had a background in LIS. Three of them were male, and two were female. Their ages ranged from 22 to 40. Four participants had been in LIS less than 3 years, and only one had spent from 4 to 6 years in the field. Four participants answered "5: very large," and one participant answered "4: large," when asked about their vocabulary size. One participant had prior experience tagging in a medical library database. Another had experience tagging museum objects and blog posts. Both of them rated themselves as "intermediate" level indexers.

The tagging procedures were the same as that of CTR_30. Human indexers assigned 9.03 topic words and phrases on average. H13 added the most topics (average 15.67), followed by H12 (average 9.17). Both of them had previous experience tagging. H15 added the least number of topics (average 5.93). It seemed that indexers with prior experience tagging tended to assign more topic tags than did the ones without experience.



Syrian activists shuffle council; chaos rocks capital

CAIRO -- Syrian dissidents sought Monday to bolster their opposition movement with members inside Syria as fighting raged in Damascus between anti-government rebels and Palestinians backing dictator Bashar Assad. The move Monday in Qatar to broaden membership in the Syrian National Council (SNC) comes after the United States said the dissidents needed to include a wider array of Syrians to help get recognition from the West. But council members berated the United States for suggesting that help in overthrowing Assad was not forthcoming because of a lack of diversity. U.S. Secretary of State Hillary Rodham Clinton said last week that the council, Syria's largest opposition group, needed to be more united and broader in its membership to be a credible alternative to Assad and receive help from the West. (Continued…)

http://spare05.dlib.vt.edu/~seungwon/various_30/20.pdf

Figure 16. A document about the Syrian conflict from the VARIOUS_30 data set.

As an example text in the VARIOUS_30 data set, we selected the one shown in Figure 16. The content was about the conflict in Syria. The topics are produced by the human indexers, Xpantrac, and the OpenCalais API, and presented in Table 10, Table 11, and Table 12, respectively.

Table 10 shows example topics assigned by H11-H15. All the topics for VARIOUS_30 documents were added to a database. Again, topic phrases were cleaned by removing the stopwords and by tokenizing them into topic words.

Table 10. Topic tags assigned by H11-H15 for a document in Figure 16.

ID	Topic Tags
H11	Syria, activists, United States response, Hillary Clinton, opposition, military aid, revolution
H12	Syria, Qatar, Egypt, activists, Palestine, military, bombings, force
H13	Syria, Damascus, anti-government, government, Palestinians, Qatar, Syrian National Council, bombs, middle east, Bashar Assad
H14	Syria, Syrian National Council, United States, international relations, Western support of Syrian opposition, political change
H15	Syrian National Council, Bashar Assad, Hillary Clinton, Bombings, Border Control

Topic Tagging by Xpantrac

Table 11 shows topics that were automatically extracted by Xpantrac with the settings associated with their machine IDs. The number of topics was set to 10. All the topics by M1-M48 were added to a database for easier evaluation.

Table 11. Topic tags assigned for (selected) machine IDs for the document shown in Figure 16.

Machine ID	Topic Tags
M1	center, eastern, university, joshua, landis, director, united, suicide, rebel, bomber
M39	syrian, syria, council, opposition, assad, united, rebels, members, national, president
M43	syrian, syria, council, opposition, assad, rebels, members, united, national, government
M48	syrian, syria, assad, council, members, opposition, rebels, government, inside, states

Topic Tagging by OpenCalais NLP API

Table 12 shows topics extracted using the OpenCalais API. Again, OpenCalais produced multiple n-grams, but they were tokenized and then lemmatized so that they could be evaluated with the topics from Xpantrac as well as from the gold standard.

Table 12. Tokenized and lowercased topic tags by OpenCalais for the document in Figure 16. These tags are lemmatized before they are compared to the gold standard.

Topic Tags
beata, disaster, deputy, ashworth, washington, states, paul,

fannie, economist, insurance, city, united, flood, electricity, travel, jim, struggles, cent, chief, capital, td, east, strategist, stock, caranci, oil, finance, business, exchange, wells, coast, freddie, mac, train, mae, york, bank, accident, car, economics, paulsen, policies, management, refineries

Topic Network Graph Based on the Topics from Human Indexers, Xpantrac (M43), and OpenCalais (VARIOUS_30 data set, document ID is 20)

The topic network graph in Figure 17 shows a slightly different property than Figure 15, in that there exist more topics from human indexers that are not shared with others. For example, H13 has 5 topics that are not shared, although it shares 9 other topics. H14 has 6 topics that are not shared, and it has only 7 shared topics. This might have been caused by the heterogeneity of the VARIOUS_30 data set, as well as the different set of human indexers who participated in the tagging.

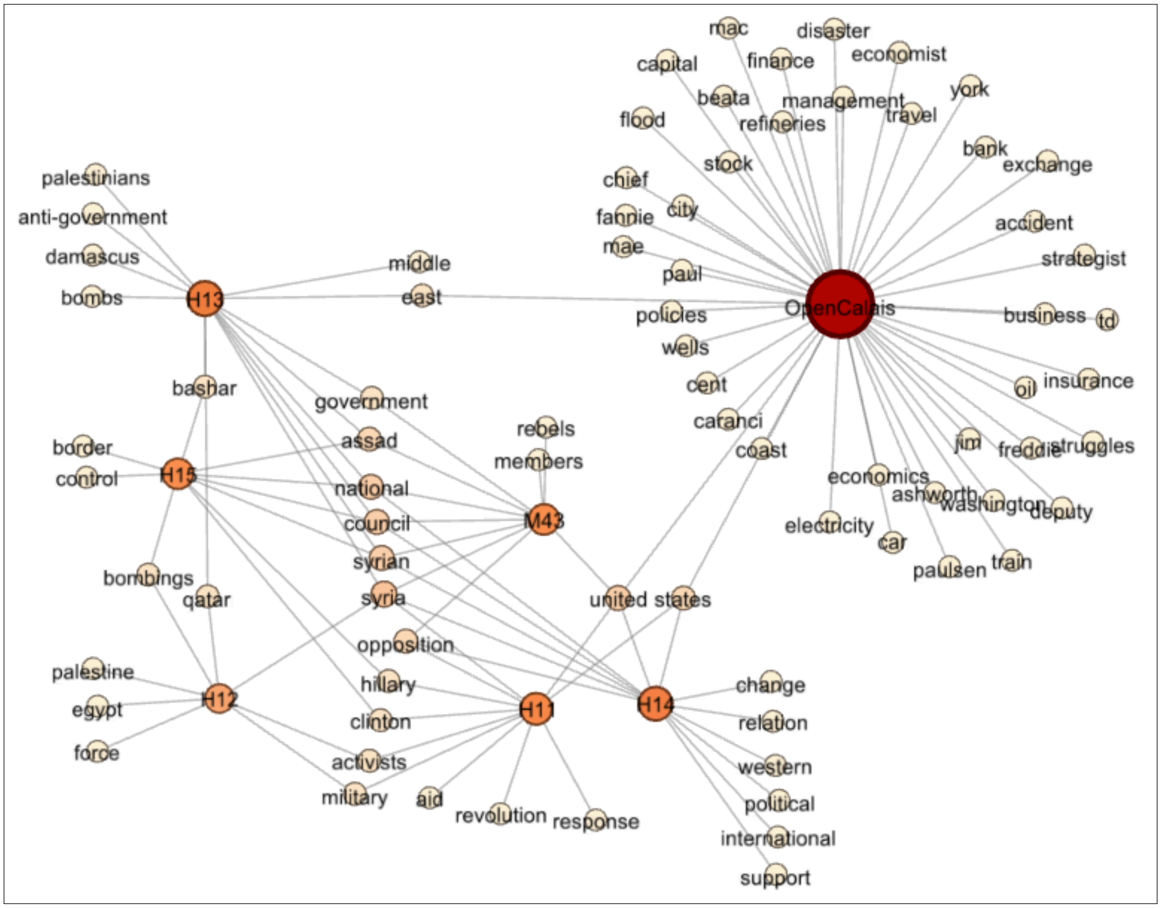


Figure 17. Topic network graph based on Table 10, Table 11 (M43 row), and Table 12. Note that OpenCalais only shares about three topics with others (i.e., east, united, states).

If we see Table 13, Xpantrac (M43) acts similarly to H15 in the network in terms of its shared and non-shared topics with other human indexers. For example, M43 shares 2 topics with H11, 1 topic with H12, 6 topics with H13, and 6 topics with H14. M43 has two non-shared topics as well. These numbers, with the exception of the case for H14, are exactly the same as those of H15.

Table 13. The number of shared and non-shared topic nodes among the human indexers, Xpantrac (M43), and OpenCalais.

	H11	H12	H13	H14	H15	M43	OpenCalais	Non-shared
H15	2	1	6	3	.	4	0	2
M43	2	1	6	6	4	.	1	2
OpenCalais	1	0	1	2	0	1	.	37

In contrast, OpenCalais performs somewhat poorly in this example. A large number (i.e., 42) of topics by OpenCalais, compared to a small number (i.e., 10) of topics by Xpantrac, might have led to this result, since extracting more topics might increase the probability of having more non-shared topics in general. OpenCalais only has 4 shared topics with human indexers, and 37 non-shared topics. Because of this, OpenCalais' IIC values will be low, and we may not be able to say that the topics from OpenCalais are comparable to human-generated topics at least in this document case.

3.3.3 Finding Optimal Parameters

Once both data sets were tagged by human indexers as well as by Xpantrac (i.e., M1-M48), we identified which of M1-M48 were optimal configurations based on the topic cosine similarity to the human topics. The higher the cosine similarity was, the more optimal the configuration must be (i.e., presumably acting like human assigned topics). We illustrate the steps to compute cosine similarities of M1-M48 (see Figure 18):

(1) Repeat the step (2) for all the Xpantrac configurations M_i, where $1 \leq i \leq 48$.

  (2) Repeat the steps from (3) to (6) for all of the documents denoted by D_m, where $1 \leq m \leq 30$. Then, compute an average cosine similarity for M_i over all *m* documents.

    (3) Develop an index of terms from a union set of topics, which were assigned to a document D_m, by multiple human indexers.

(4) Convert topics by each human indexer and by M_i to topic vectors using the index developed in step (3). Results are shown in (d) and (a) respectively.

(5) Develop a human centroid vector (c) by averaging human topic vectors in (d).

(6) Compute a cosine similarity value (b) using the vectors (a) and (c).

Applying the steps above resulted in 48 cosine similarity values for machine IDs, M1-M48. We then selected the top three such configurations in descending order of the highest cosine similarity values. We did this for both CTR_30 and VARIOUS_30.

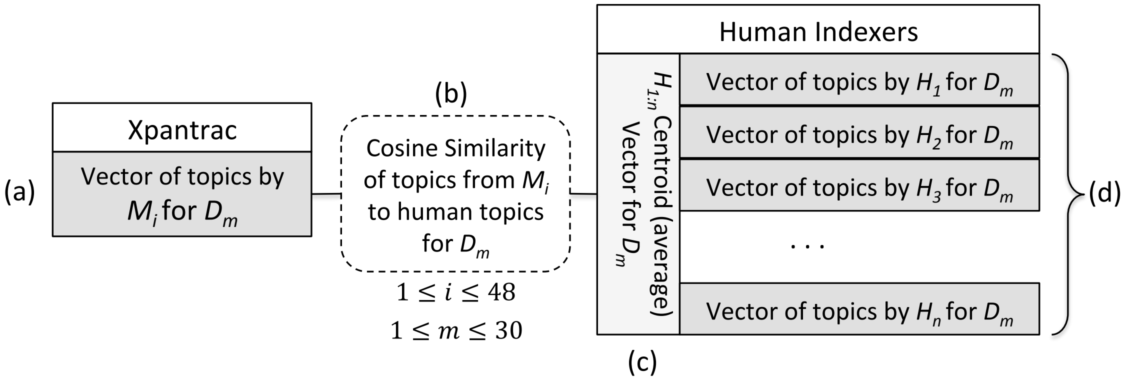

(a)
Xpantrac
Vector of topics by M_i for D_m
(b)
Cosine Similarity of topics from M_i to human topics for D_m
$1 \leq i \leq 48$
$1 \leq m \leq 30$
Human Indexers
$H_{1:n}$ Centroid (average) Vector for D_m
(c)
Vector of topics by H_1 for D_m
Vector of topics by H_2 for D_m
Vector of topics by H_3 for D_m
. . .
Vector of topics by H_n for D_m
(d)

Figure 18. Computing the cosine similarity of topics by Xpantrac (i.e., M_i) to human topics given a document D_m. (a) is a topic vector by M_i, (b) is the Cosine Similarity, (c) is a human topic centroid vector, and (d) is the human topic vectors.

The top three highest average cosine similarity values (in bold) and corresponding machine IDs are shown in Table 14. Interestingly, two machine IDs, M39 and M43, were found in both data sets. The parameter values of M39 and M43 were different only in *num_api_return*, which decides how many webpage descriptions to retrieve for each query unit. M39 retrieves 50 descriptions, whereas M43 retrieves 10 descriptions. If their results were not statistically different, using M43 would be more cost effective considering that M43 retrieves one fifth of what M39 retrieves, resulting in processing only 20% of what M39 would process. Other common parameter values between M39 and M43 were:

- *unit_size* = 5 (group 5 words as a single query to a search API)
- *tf_idf* = 0 (TF*IDF not applied)
- *only_nouns* = 1 (only noun topics extracted)

The *unit_size* for both M39 and M43 was 5, and the *unit_size* of M31 and M27 was 10. Thus, using 5 or 10 as the *unit_size* value works well in CTR_30 and VARIOUS_30. Although more accurate values might be identified, we recommend selecting any value between 5 and 10 as the *unit_size* based on the findings from both Table 4 and Table 14 for the best Xpantrac performance.

We did not include M31 in the study because its similarity value was the third in CTR_30 and the fourth in VARIOUS_30. However, M31 might be another cost effective choice in terms of the network traffic and API service fee. M31 was different from M43 only in its *unit_size*. M31's *unit_size* was 10, which was twice as large as that of M43. This meant that M31 segmented input texts into query units twice as large as that of M43, requiring only half the number of API calls compared to M43. Therefore, if we used M31, the API service fee would have cost only half as much as that of M43.

Table 14. Cosine similarity values between the average of human topic vectors and Xpantrac topic vectors from various configurations (M1-M48). The numbers in bold denote the three highest similarity values from each data set.

MID	CTR_30	VARIOUS_30	MID	CTR_30	VARIOUS_30
1	0.18	0.18	25	0.179	0.135
2	0.317	0.312	26	0.472	0.327
3	0.641	0.549	27	0.717	**0.579**
4	0.577	0.556	28	0.691	0.573
5	0.201	0.219	29	0.224	0.19
6	0.42	0.458	30	0.559	0.472
7	0.695	0.536	31	**0.723**	0.575
8	0.654	0.55	32	0.694	0.576
9	0.334	0.273	33	0.369	0.272
10	0.587	0.451	34	0.517	0.496
11	0.655	0.522	35	0.682	0.534
12	0.638	0.492	36	0.666	0.548
13	0.188	0.198	37	0.134	0.099
14	0.458	0.339	38	0.581	0.473
15	0.691	0.569	39	**0.738**	**0.584**
16	0.647	0.553	40	0.709	0.571
17	0.278	0.199	41	0.185	0.161
18	0.559	0.48	42	0.635	0.551
19	0.708	0.548	43	**0.73**	**0.582**
20	0.679	0.55	44	0.708	0.571
21	0.337	0.256	45	0.319	0.26
22	0.552	0.467	46	0.581	0.52
23	0.675	0.512	47	0.707	0.569
24	0.656	0.517	48	0.696	0.545

3.3.4 Indexing Quality

3.3.4.1 Inter-Indexer Consistency (IIC)

We used IIC by Rolling (1981) to evaluate the indexing quality of Xpantrac with the two optimal configurations, M39 and M43. Recalling from Section 2.6.1, when we have two sets of indexed topics for a same document, whose sizes are denoted as *A* and *B*, and the size of their intersection set is *C*, Rolling's IIC is computed as:

$$Rolling's\ IIC = \frac{2 \times C}{A + B}$$

Using IIC, the tagging consistency values between human indexers and Xpantrac, as well as between human indexers and OpenCalais, were computed using both CTR_30 and VARIOUS_30.

IIC by M39 on CTR_30

Since CTR_30 includes 30 documents, an IIC value for each document was computed, and then the 30 IIC values were averaged. Table 15 shows the averaged IIC values between human indexers and M39. Human indexers' IIC values (e.g., 0.306 and 0.376 in the row of H1 in Table 15) were averaged further as H_mean (e.g., 0.341) so that this value could be compared to that of M39. Overall, M39 performed well considering that its IIC values were higher than H_mean in two-thirds of the cases (see numbers in bold).

The ranks of IIC values by M39 among those of human indexers were shown inside the parentheses in the last column in Table 15. For example, in the second row of the table, the IIC value between H2 and H1 was 0.306, and between H2 and H3 was 0.427. However, the IIC value between H2 and M39 was 0.453. From this, H2's first choice to produce the most consistent topic tags together should be M39. Thus, the IIC value of M39 in the second row has the first rank on that row. Likewise, the IIC values of M39 in the first and third rows have both second ranks in their corresponding rows.

From these ranks, we may imagine that all human indexers would want to work with M39 as their first or second (and never the last) choice collaborator to achieve the most consistency in their topic tags. And the rankings of M39 were never the last among human indexers.

Table 15. Inter-Indexer Consistency values between humans and Xpantrac with M39 configuration for the CTR_30 data set. The numbers in parentheses of the last column denotes the rank of IIC values of M39 among the IIC values of human indexers in each row.

	H1	H2	H3	H_mean	M39

H1		0.306	0.376	0.341	0.333 (2)
H2	0.306		0.427	0.367	**0.453 (1)**
H3	0.376	0.427		0.402	**0.423 (2)**
Mean	0.341	0.366	0.402	0.370	0.403

IIC by M39 on VARIOUS_30

Again, an IIC value for each document in VARIOUS_30 was computed, and then the 30 IIC values were averaged.

Table 16. Inter-Indexer Consistency values between humans and Xpantrac with M39 configuration for the VARIOUS_30 data set. The values inside the parentheses of the last column show the rank of M39 when compared to that of human indexers in the same row.

	H11	H12	H13	H14	H15	H_mean	M39
H11		0.418	0.309	0.387	0.264	0.345	**0.350 (3)**
H12	0.418		0.374	0.409	0.310	0.378	0.367 (4)
H13	0.309	0.374		0.350	0.273	0.327	**0.342 (3)**
H14	0.387	0.409	0.350		0.275	0.355	**0.395 (2)**
H15	0.264	0.310	0.273	0.275		0.281	**0.284 (2)**
Mean	0.345	0.378	0.327	0.355	0.280	0.337	0.348

Table 16 shows the averaged IIC values between human indexers and M39. Five human indexers tagged VARIOUS_30, so four IIC values (e.g., 0.418, 0.309, 0.387, and 0.264 in the row of H11) were averaged as H_mean (e.g., 0.345). Then this value could be compared to that of M39. Overall, M39 performed well considering that its IIC values were higher than H_mean in four-fifths of the cases (see numbers in bold).

If we consider the rank of the IIC values in M39 column, H14 and H15 would want to work with M39 as their second choice collaborator out of five collaborators to achieve the most consistency. For H11-H13, M39 was never the last choice.

Since M43 was another optimal configuration in both data sets, we evaluated the topic consistency of M43 as well.

IIC by M43 on CTR_30

In this case, we computed the IIC values for the documents in CTR_30 using the topics by human indexers and Xpantrac with M43 configuration. Then, those values were averaged. Table 17 shows averaged IIC values between human indexers and M43 across the documents in CTR_30. From the table, the IIC values of M43 were higher than H_mean only in one case out of three. However, if we consider the rank of cosine similarity values in the M43 column, the performance of M43 was not bad even when the IIC values of M43 were lower than H_mean; their ranks were either the first place or the second out of three, but never the last (see the number in bold).

Table 17. Inter-Indexer Consistency between humans and Xpantrac with M43 configuration for the CTR_30 data set. The values in the parentheses of the last column show the rank of M43 when compared to that of human indexers in the same row.

	H1	H2	H3	H_mean	M43
H1		0.306	0.376	0.341	0.338 (2)
H2	0.306		0.427	0.367	**0.452 (1)**
H3	0.376	0.427		0.402	0.397 (2)
Mean	0.341	0.367	0.402	0.370	0.396

IIC by M43 on VARIOUS_30

Table 18 shows IIC values between human indexers and M43, averaged over 30 documents in VARIOUS_30. Overall, M43 performed well considering that its IIC values were higher than H_mean in three-fifths of the cases (see the numbers in bold). In terms of the IIC rank of M43, M43 was never the last choice as a collaborator among human indexers.

Table 18. Inter-Indexer Consistency between humans and Xpantrac with M43 configuration for the VARIOUS_30 data set.

	H11	H12	H13	H14	H15	H_mean	M43
H11		0.418	0.309	0.387	0.264	0.345	0.341 (3)
H12	0.418		0.374	0.409	0.310	0.378	0.372 (4)
H13	0.309	0.374		0.350	0.273	0.327	**0.344 (3)**
H14	0.387	0.409	0.350		0.275	0.355	**0.381 (3)**
H15	0.264	0.310	0.273	0.275		0.281	**0.286 (2)**
Mean	0.345	0.378	0.327	0.355	0.281	0.337	0.345

IIC by OpenCalais on CTR_30

To examine the performance of OpenCalais in terms of the IIC values of its topics, we applied Rolling's IIC again. Table 19 shows averaged IIC values between human indexers and OpenCalais across the documents in CTR_30. The IIC values of OpenCalais were lower than that of H_mean in *all three cases*. Please note that the rank of OpenCalais was the third in every case (rows). In other words, human indexers had better consistency when they worked with the other human indexers, and not with OpenCalais. In addition, there were large gaps between the IIC values of H_mean and OpenCalais.

Table 19. Inter-Indexer Consistency between humans and OpenCalais for CTR_30.

	H1	H2	H3	H_mean	OpenCalais
H1		0.306	0.376	0.341	0.139 (3)
H2	0.306		0.427	0.367	0.214 (3)
H3	0.376	0.427		0.402	0.168 (3)
Mean	0.341	0.367	0.402	0.370	0.174

IIC by OpenCalais on VARIOUS_30

Table 20 shows the IIC values between human indexers and OpenCalais, averaged over 30 documents in VARIOUS_30. Overall, OpenCalais performed not so well considering that its IIC values were lower than any of the human indexers' IIC values in all five cases. Also noticed is that the gaps between the IIC values of H_mean and OpenCalais were large.

Table 20. Inter-Indexer Consistency between humans and OpenCalais for VARIOUS_30.

	H11	H12	H13	H14	H15	H_mean	OpenCalais
H11		0.418	0.309	0.387	0.264	0.345	0.153 (5)
H12	0.418		0.374	0.409	0.310	0.378	0.195 (5)
H13	0.309	0.374		0.350	0.273	0.327	0.234 (5)
H14	0.387	0.409	0.350		0.275	0.355	0.180 (5)
H15	0.264	0.310	0.273	0.275		0.281	0.142 (5)
Mean	0.345	0.378	0.327	0.355	0.281	0.337	0.181

3.3.4.2 Summary of the IIC Analysis

Throughout Table 15, Table 16, Table 17, and Table 18, Xpantrac with M39 and M43 configurations showed a robust performance when their topics were compared with those by the multiple human indexers. In other words, M39 and M43 could produce human-comparable topic tags from the two data sets, CTR_30 and VARIOUS_30. However, further examination is needed to understand whether, and with which settings, Xpantrac would perform the best in other types of text collections.

Regarding the performance of OpenCalais in terms of its IIC values, the results from Table 19 and Table 20 tell us that the topics generated by OpenCalais may not be comparable to those generated by the human indexers at least for this data set. The lower performance of OpenCalais topics in terms of the IIC values was somewhat expected based on the two topic network graphs, Figure 15 and Figure 17, in the previous sections. The main cause of this low performance might originate from the large number of topics extracted, many of which were not relevant to the ones assigned by the human indexers.

3.3.4.3 Tag Set Precision, Recall, and F_1

Several metrics, such as tag set precision, recall, and F_1 from the Information Retrieval (IR) field, were used to determine an optimal number of topics to extract from both CTR_30 and VARIOUS_30. Regarding the settings for this study, we used the M43 configuration for Xpantrac. The gold standard topics were prepared by merging topics from all three human indexers. For each number of topics that were extracted by Xpantrac, we computed the tag set precision, recall, and F_1 scores for each document in the data set. Then, those metric values were averaged over all 30 documents. These values for the number of topics from 1 to 10 are shown in Figure 19.

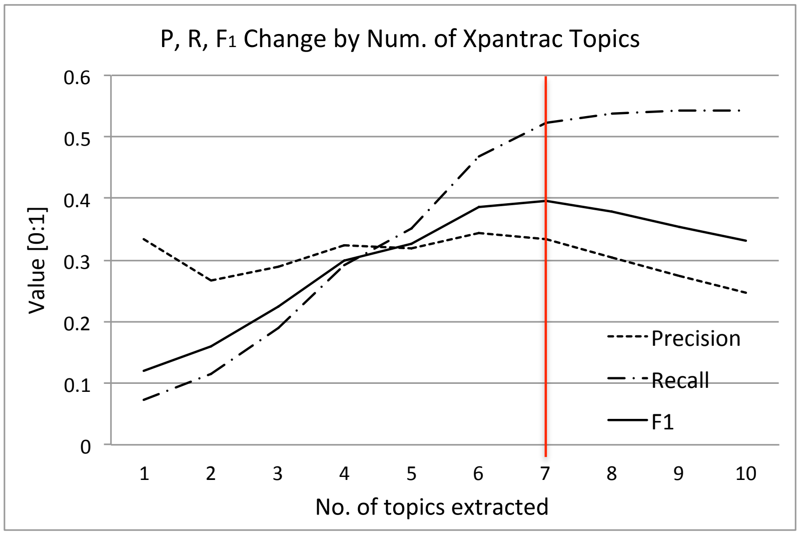


Figure 19. Tag set precision, recall, and F_1 scores change as the number of topics by Xpantrac changes in the CTR_30 data set. The agreement among the human topics was 32.9%.

The optimal number of Xpantrac topics for both data sets turned out to be 7, where the F_1 score was the highest (0.4) in CTR_30 as well as in VARIOUS_30, shown in Figure 19 and Figure 20, respectively. In fact, the highest F_1 score in Figure 20 was 0.4 when the number of topics was 5, 6, or 7. However, selecting 7 would make the most sense considering that it is where both the precision graph and recall graph meet. In other words, it is the point where the two values are balanced. The optimal number of topics, 7, works for both data sets.

Another interesting point is that the agreement among the topics from human indexers was 22.8% in VARIOUS_30, whereas it was 32.9% in CTR_30. The lower agreement value in VARIOUS_30 might originate from two factors. One is the heterogeneous content of the documents in VARIOUS_30, which would require topic tagging using a broader vocabulary set. The other factor might be more human indexers for VARIOUS_30. Unless they have similar backgrounds, more indexers would mean less agreement in their topic tag assignment.

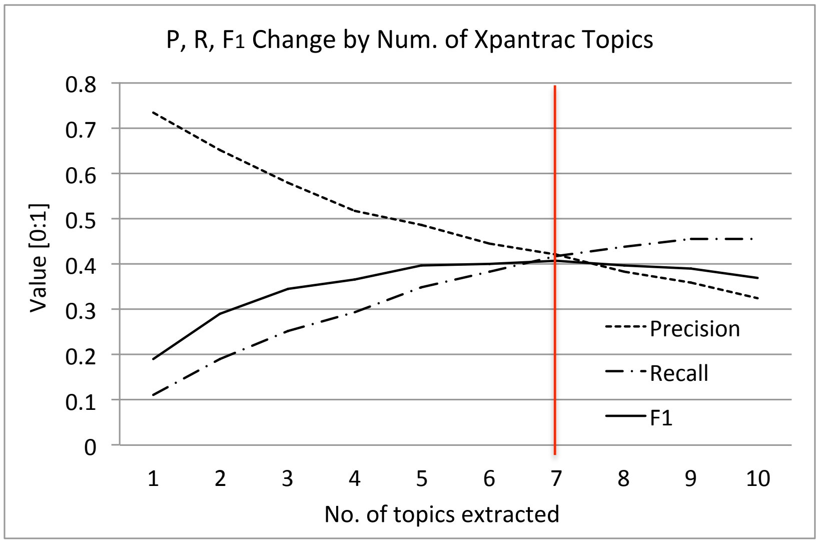


Figure 20. Tag set precision, recall, and F_1 scores change as the number of Xpantrac topics change in VARIOUS_30. The topic agreement among the human topics was 22.8%.

3.4 Discussion

The Vector Space Model is used in the extraction process of Xpantrac in order to uncover the significant terms from the derived corpus. In spite of its usefulness in various text analysis and Information Retrieval tasks, the VSM is said to have the following limitations (Becker & Kuropka, 2003; G. Z. Liu, 1997; V. V. Raghavan & Wong, 1986):

1. Long documents are not represented well in the vector space, possibly resulting in an inaccurate computation of the document similarities, if the normalization is not done carefully.
2. It is based on an exact match of the terms, thus synonyms are not considered to be similar with each other.
3. The bag-of-words model in the VSM is based on an assumption that the words are independent of each other. But, words have dependencies.

In a sense, the limitations recited above may not be applicable for this dissertation. Further, a couple of them are in fact providing benefits. We designed Xpantrac as a topic tagging tool for disaster-related news reports and articles on the Web, which are not very long and can be well represented by the VSM. Therefore, limitation #1 above is not applicable in the context of this study.

In order to overcome limitation #2, existing studies made attempts to incorporate the WordNet lexical database to resolve the synonym issues since it provides a group of synonyms as "synsets" (Fellbaum, 2010). However, this approach also has limitations due to the ambiguities (e.g., missing context information) present in the language. In any case, the Expansion phase of Xpantrac partially addresses synonyms.

Although it is true that we lose the dependency information between words when we tokenized them as in limitation #3, the VSM with the bag-of-words model has simple linear properties such that any two vectors can be added or a scalar value can be multiplied to obtain a new vector. Thus the computations are easy among the elements. It is intuitive and performs well in various tasks such as document ranking and text classification. Therefore, we take advantage of the VSM in this study.

3.5 Summary

In this chapter, we described our design inspiration, which was based on Cognitive Informatics as well as the VSM. The details of the VSM were explained as a three-step process. The first was to convert texts into document vectors after obtaining an index list from the text corpus. The document vectors and the term index lead to a term-document matrix, where term-weighting and normalization can be applied. We then compute cosine similarity between the vectors to find the most relevant documents.

The two parts of Xpantrac were explained. The role of the first part was to expand the input text into relevant information using a search engine API. The actual identification of topics was performed in the second “extraction” part. The structural and functional details were explained using a component diagram (Figure 5), workflow diagram (Figure 8), and the algorithm (Section 3.2.1).

To find the optimal parameters of Xpantrac, we developed two human-indexed data sets, CTR_30 and VARIOUS_30. CTR_30 was homogeneous in that all of the 30 texts in it were randomly selected from a webpage archive of a single disaster event. VARIOUS_30 was heterogeneous because the documents were selected from various health and disaster events. We used the Inter-Indexer Consistency (IIC) as a measure of performance to compare the topics by Xpantrac with those assigned by human indexers. When we compared the IIC scores of Xpantrac with those of the OpenCalais API, Xpantrac outperformed OpenCalais for both data sets. Graphical examples were presented to illustrate the topical relationships and the causes of high/low IIC values among Xpantrac, OpenCalais, and the human-indexed gold standard topics.

Chapter 4. Evaluation

In the previous chapter, we presented the design and development details of the Xpantrac topic tagging system. We also described the process for identifying the optimal parameters of Xpantrac using two human-tagged data sets, one with homogeneous content and the other with heterogeneous content. We expected such optimal parameters to help Xpantrac to produce robust and human-comparable topics. To understand the performance of Xpantrac with those parameter settings, we used metrics such as the Inter-Indexer Consistency (IIC) (Rolling, 1981), precision, recall, and F_1 scores (van Rijsbergen, 1979). Each metric was intended to address a different aspect of Xpantrac's performance.

In this chapter, we present additional evaluations of Xpantrac in terms of its topic tag quality and its performance on a larger text corpus. For topic tag quality, we recruited three human raters to assign a relevance rating for each topic; we then compared the rating scores between Xpantrac and the baseline TF*IDF. We also examined consistency between the three raters. To understand Xpantrac's performance on a larger text set, we used a data set called NYT_1000, which consists of 1,000 *New York Times* articles.

Details concerning the evaluation setting and the baseline system are explained in Section 4.1. Section 4.2 presents the topic quality rating study comparing the relevance rating of topic tags from Xpantrac and the baseline. The rating consistency between the indexers (i.e., raters) also was assessed and is described in Section 4.3. Section 4.4 presents an evaluation of Xpantrac using the NYT_1000 collection, which compares various Xpantrac settings with three different search application programming interfaces (APIs) against the baseline and a popular natural language processing (NLP) API. Further discussions of the results in Sections 4.2 and 4.4 are presented in Section 4.5, followed by the chapter summary in Section 4.6.

4.1 Evaluation Strategy

We present our overall experimental settings in this section. First, we describe which data sets and approaches were used for evaluation. The Xpantrac topics were compared to those of the baseline approach, TF*IDF. In a subsequent section, one of the widely used Natural Language Processing APIs, OpenCalais API, also is compared to Xpantrac. Thus, we present the process of identifying topics using the baseline, as well as how OpenCalais works to process unstructured text to extract topics.

4.1.1 Experimental Settings

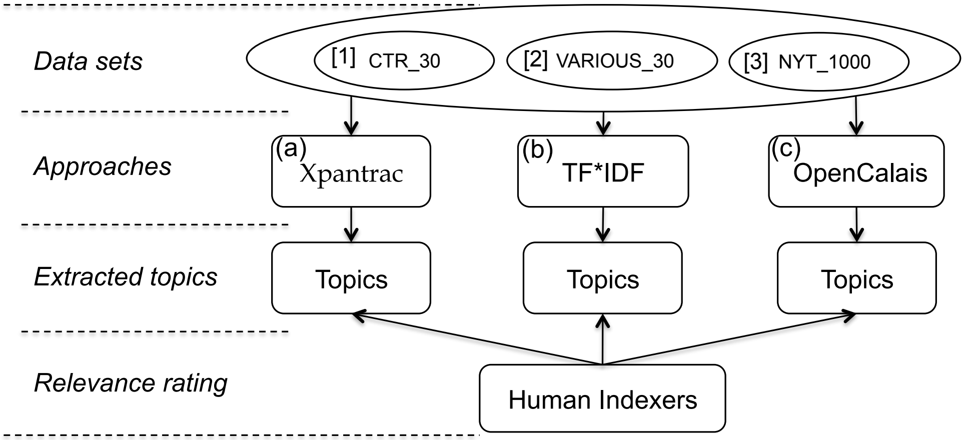
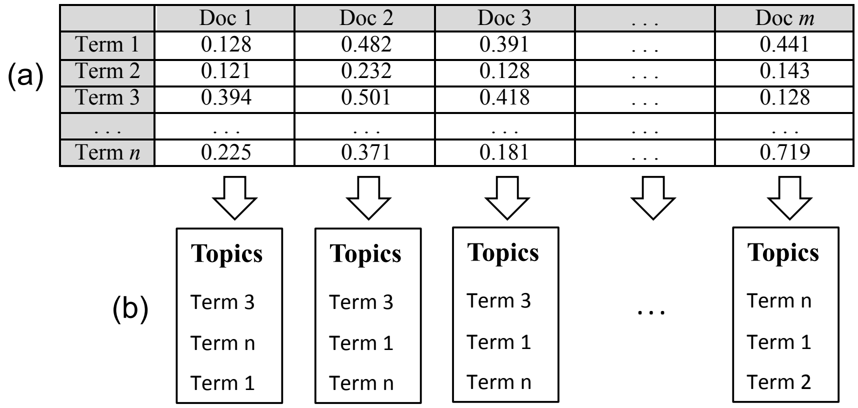


Figure 21. Experimental settings showing three different approaches (a), (b), and (c), and three text data sets [1], [2], and [3].

Figure 21 depicts our overall experimental settings. For the topic quality rating study, which is presented in Section 4.2, we used the two data sets [1] and [2], as well as the two approaches (a) and (b) in Figure 21. For the study of the Inter-Rater Consistency (IRC) presented in Section 4.3, the same data sets and the same approaches were used. For a larger scale evaluation, we used [3] and all three approaches of (a), (b), and (c). Since each NYT article had embedded topics, we used them as the gold standard.

4.1.2 Baseline

A term weighting scheme, *Term Frequency * Inverse Document Frequency* (TF*IDF) (Salton & Buckley, 1988; Sparck-Jones, 1972), was used as a baseline approach to extract topics from the data sets. As explained in detail in Section 3.2 of Chapter 3, TF*IDF computes a term's significance based both on document level (TF part) and corpus level (IDF part) considerations.



	Doc 1	Doc 2	Doc 3	. . .	Doc *m*
Term 1	0.128	0.482	0.391	. . .	0.441
Term 2	0.121	0.232	0.128	. . .	0.143
Term 3	0.394	0.501	0.418	. . .	0.128
. . .	. . .	. . .	. . .	. . .	. . .
Term *n*	0.225	0.371	0.181	. . .	0.719

Figure 22. Extracting topics from each document in a term-document matrix (TDM): (a) weighted and normalized TDM, (b) extracted topics.

An example TDM is shown in Figure 22 (a). The cell values are often decimal numbers after a term weighting (e.g., TF*IDF) and normalization have been applied. The higher cell values in a column represent which terms are more significant in the document. For example, if we look at the cell values in the column for Doc 3, Term 3 had the highest cell value of 0.418, followed by Term 1 (cell value of 0.391) and Term *n* (cell value of 0.181). Therefore, we select the topics in the order of Term 3, Term 1, and Term *n,* as shown in Figure 22 (b) (if we set the number of topics to be three).

4.1.3 OpenCalais

The OpenCalais Web Service (2013) is a popular API for identifying the named entities, facts, and events from unstructured documents in formats such as text, HTML, or XML, as shown in Figure 23. Although not employed in this dissertation study, it also can extract information as a Resource Description Framework (RDF) collection (W3C, 2004), which can be used by Semantic Web technologies (Antoniou & Van Harmelen, 2004; Baker & Cheung, 2007; Breitman, Casanova, & Truszkowski, 2007). Thomson Reuters (2013), the company that provides the OpenCalais service, says that their technology to derive meaning from texts is based on Natural Language Processing (NLP), text analysis, and data mining.

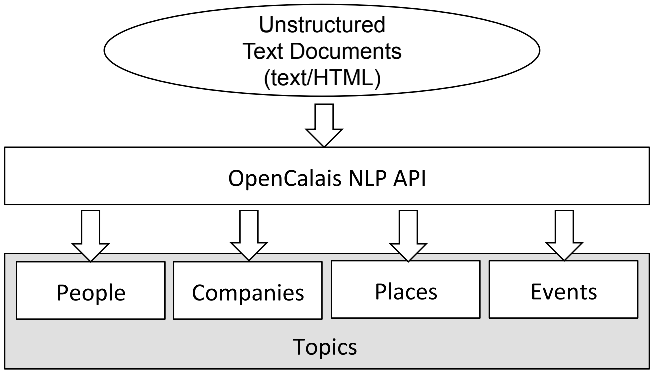


Figure 23. A workflow of the OpenCalais API, which extracts information from unstructured texts to develop topics.

The named entities, such as names of people, events, places, or companies, extracted by OpenCalais are merged as "topics" for the given texts, as shown in Figure 23. These topics were then compared to the gold standard, namely, the NYT_1000 data set. Other examples of similar NLP APIs are AlchemyAPI[6], Evri[7], OpenAmplify[8], and Zemanta[9] to name a few, and different opinions exist for the performance of those APIs (DiCiuccio, 2010; Fagan, 2010). Considering that OpenCalais provides a robust topic analysis and is widely adopted, OpenCalais was used throughout this study.

4.2 Topic Quality Rating

The degree to which extracted topics are "relevant" to their associated document is an important measure for assessing topical quality, as well as the system's performance. To measure topic relevance, we recruited three people as raters. Each was asked to assign a rating score from -2 to 2 in the Likert scale to each topic (Likert, 1932) (see Figure 24), based on how much s/he agreed or disagreed that a topic was relevant to the document. Two out of the three raters were male. One rater had a background in Computer Science, another in Library and Information Science, and the third rater came from the Bioinformatics field. They were all graduate students. Except for one person, English was their first language.

[6] http://www.alchemyapi.com/api/entity-extraction/
[7] http://www.evri.com/
[8] http://www.openamplify.com/
[9] http://developer.zemanta.com/

We generated topics for rating using Xpantrac in four ways. After we had topic sets from both M39 and M43, we developed two more topic sets, one by taking the intersection between the two (i.e., M39_AND_M43), and the other by taking the union of the two (i.e., M39_OR_M43). The topics derived using TF*IDF (see Section 4.1.2) also were rated. Both of CTR_30 and VARIOUS_30 were used in the rating study.

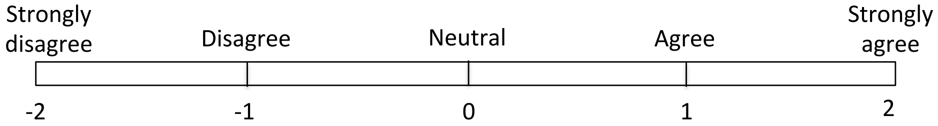


Figure 24. The five-point Likert scale for the topic quality rating.

4.2.1 Topic Relevance of CTR_30

We examined the topic relevance ratings in two ways. First, we computed an average topical relevance rating score for each document in the data set, which enabled us to compare the ratings by different configurations at the document level. Second, we applied Analysis Of Variance (ANOVA) (Gelman, 2005) in order to examine whether there were significant differences between those configurations.

4.2.1.1 Average Relevance Rating

To compute the average relevance rating for the topics from each document, we averaged the ratings assigned by the three human raters. We then further averaged the average from the human raters over all topics from each document in order to produce 30 data points (see Figure 25).

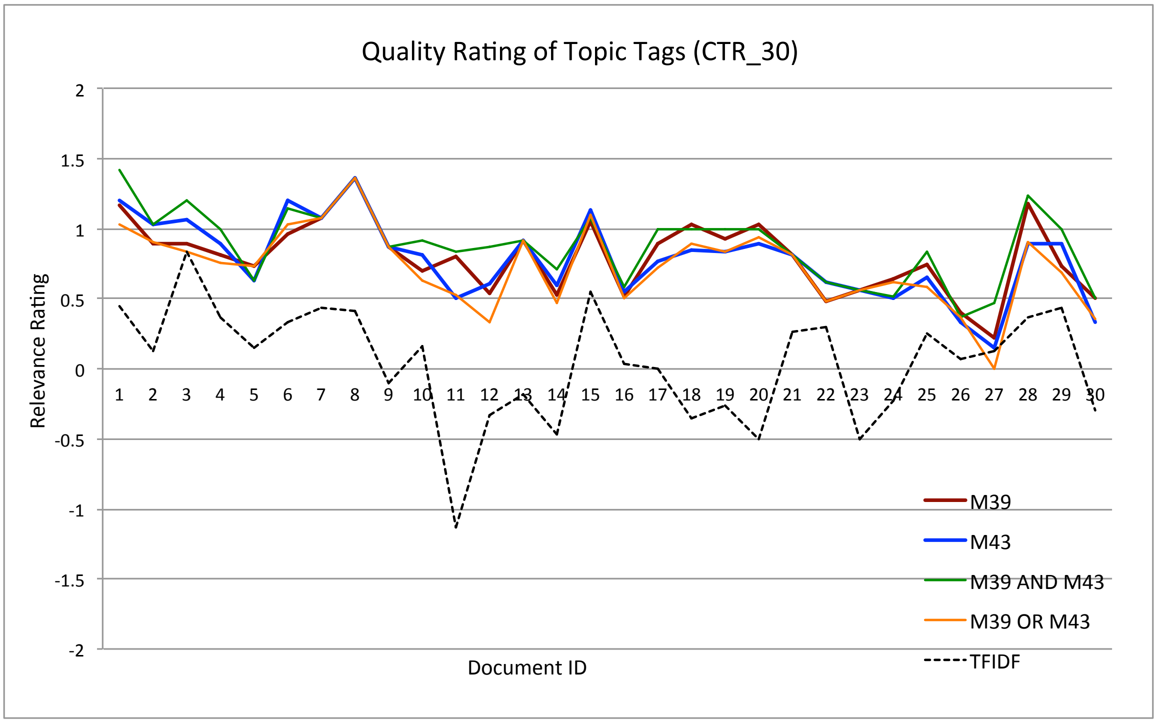


Figure 25. Average topic relevance ratings for the documents in CTR_30. The numbers on the x-axis represent the document IDs.

These data points from the various Xpantrac settings, M39, M43, M39_AND_M43, and M39_OR_M43, tended to fall between 0.5 and 1.5 on the Likert scale, with the exception of the document with ID 27, with rating lower than 0.5. This outcome is associated with the fact that Rater 2 consistently assigned very low ratings to the topics when compared to the other raters, as shown in Table 21.

Table 21. The ratings averaged over all topics of a document (ID 27) by each rater.

Setting	Rater 1	Rater 2	Rater 3	Average
M39	0.889	-1.000	0.778	0.222
M43	0.667	-0.667	0.444	0.148
M39_AND_M43	0.714	-0.286	1.000	0.476
M39_OR_M43	0.818	-1.182	0.364	0.000
Average	0.772	-0.784	0.646	0.211

In the case of TF*IDF (shown as the black dotted line in Figure 25), most of the documents had relevance ratings that were lower than 0.5. In fact, more than one third of the documents had ratings that were below zero, making the TF*IDF graph visibly lower than those from the Xpantrac settings.

Table 22 shows the relevance rating values averaged over all documents. The topics from M39_AND_M43 had the highest rating of 0.887. Considering that the Likert scale rating of 0 corresponded to "neutral" and 1 corresponded to "agree" that a topic was relevant to the document, most human raters in general agreed that the topics generated by Xpantrac were relevant to the document. In contrast, the average rating value for TF*IDF was only 0.044, which was close to 0 or "neutral" on the Likert scale.

Table 22. Topic relevance ratings averaged over 30 documents in CTR_30 and three human raters.

	M39	M43	M39 AND M43	M39 OR M43	TF*IDF
Ave. Rating	0.802	0.786	0.887	0.729	0.044

4.2.1.2 ANOVA on Average Relevance Rating

Although there was a visible difference between the TF*IDF graph and the graphs corresponding to the Xpantrac settings (Figure 25), we applied ANOVA (Gelman, 2005; Vasishth & Broe, 2010) to determine whether there were any significant differences. The JMP tutorial[10] explains the terms in the ANOVA table (e.g., Table 23) and the Connecting Letters report (e.g., Table 24). Please see the (paraphrased) excerpts from the tutorial:

- Source: it lists the three sources of variation, namely the model source (in this case, "Label"), Error, and C. Total.
- DF (degree of freedom): it shows an associated degree of freedom for each source of variation.
- Sum of Squares (SS): it shows a sum of squared distances from each data point to its respective group mean.
- F Ratio: it is a ratio between the model mean square and the error mean square.
- Prob > F: it is the probability of obtaining an F-value by chance, which is greater than the F Ratio computed. If this probability is 0.05 or less, it is an evidence that an analysis of variance model fits the data, in other words, there exists a significant difference among the mean values.
- Connecting Letters Report: each mean value has associated letters. If means does not share a letter, there exists a significant difference between those means.

[10] http://www.jmp.com/support/downloads/pdf/jmp_stat_graph_guide.pdf

As shown in Table 23, there were significant differences ($p < 0.0001$) among the different settings, with TF*IDF (i.e., associated letter is "B") significantly lower than the rest (i.e., their associated letter is "A") (see Table 24).

Table 23. ANOVA showing the significant difference of average topic ratings in Figure 25.

.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Label	4	14.127344	3.53184	37.8613	<.0001*
Error	145	13.526104	0.09328		
C. Total	149	27.653447			

Table 24. The Connecting Letters report (alpha = 0.05) showing the significant difference of TF*IDF topic ratings from others.

Level			Mean
M39_AND_M43	A		0.88683333
M39	A		0.80170000
M43	A		0.78560000
M39_OR_M43	A		0.72916667
TF*IDF		B	0.04406667

4.2.2 Topic Relevance of VARIOUS_30

Similar to Section 4.2.1, this section details the average topic relevance rating scores for the documents in VARIOUS_30. ANOVA also was applied to identify and examine any significant differences.

4.2.2.1 Average Relevance Rating

Figure 26 shows that the data points resulting from using the Xpantrac settings ranged approximately between 0.3 and 1.3. In contrast, the data points associated with the TF*IDF graph ranged between -0.1 and 1.167, and were generally higher than the analogous graph in Figure 25. In other words, the gap between the TF*IDF graph and those corresponding to the Xpantrac settings became smaller than the gap depicted in Figure 25.

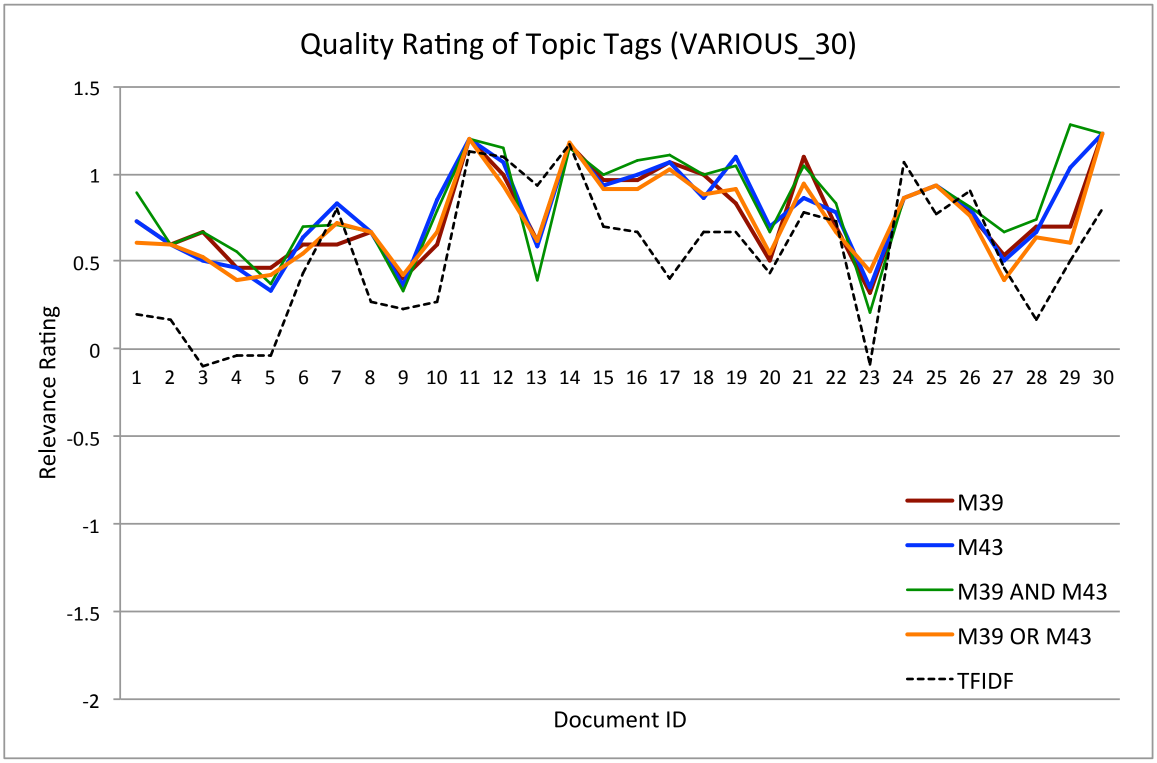


Figure 26. Average topic relevance ratings for the documents in VARIOUS_30.

Table 25 shows the relevance rating values averaged over all documents. The topics from M39_AND_M43 had a rating score of 0.823, which was very close to "agree" with the topics' relevance to the document.

Unlike the prior case described in Section 4.2.1.1 where the average rating was close to "neutral," the average rating of TF*IDF was 0.539, which is approximately a midpoint between "neutral" and "agree" on the Likert scale. A further discussion of this finding is provided in Section 4.5.1.

Table 25. Topic relevance ratings averaged over 30 documents in VARIOUS_30 and three human raters.

	M39	M43	M39 AND M43	M39 OR M43	TF*IDF
Ave. Rating	0.766	0.790	0.823	0.740	0.539

4.2.2.2 ANOVA on Average Relevance Rating

Although there appears to be a smaller visible difference between the TF*IDF graphs and the Xpantrac settings, we still identified a significant difference (*p* = 0.0011) between them (see Table 26). Table 27 further details the significant differences between TF*IDF (denoted as "B") and the three Xpantrac settings, M39_AND_M43, M43, and M39, which were denoted as "A." However, there was no significant difference between TF*IDF and M39_OR_M43 (denoted as "AB" and sharing a letter "B" with TF*IDF).

Table 26. ANOVA showing the significant difference of average topic ratings in Figure 26.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Label	4	1.677634	0.419409	4.8300	0.0011*
Error	145	12.590882	0.086834		
C. Total	149	14.268516			

Table 27. The Connecting Letters report (alpha = 0.05) showing the significant difference of TF*IDF from others except for M39_OR_M43.

Level			Mean
M39_AND_M43	A		0.85126667
M43	A		0.79003333
M39	A		0.76576667
M39_OR_M43	A	B	0.73966667
TF*IDF		B	0.53896667

4.3 Inter-Rater Consistency (IRC)

Since three human raters assigned topic ratings, it was imperative to examine rating consistency. Resulting computations were exactly the same as the Rolling's Inter-Indexer Consistency (explained in Sections 2.5.1 and 3.3.4.1), except that the raters used scores corresponding to the Likert scale (Figure 24).

4.3.1 IRC on CTR_30

Figure 27 shows the IRC values by each pair of raters averaged over all 30 documents in CTR_30. The H2 and H3 pair had the lowest average IRC value compared to other rater pairs in (a)-(d). However, they had the highest IRC value in (e), whereas H1 and H3 had the lowest in this instance.

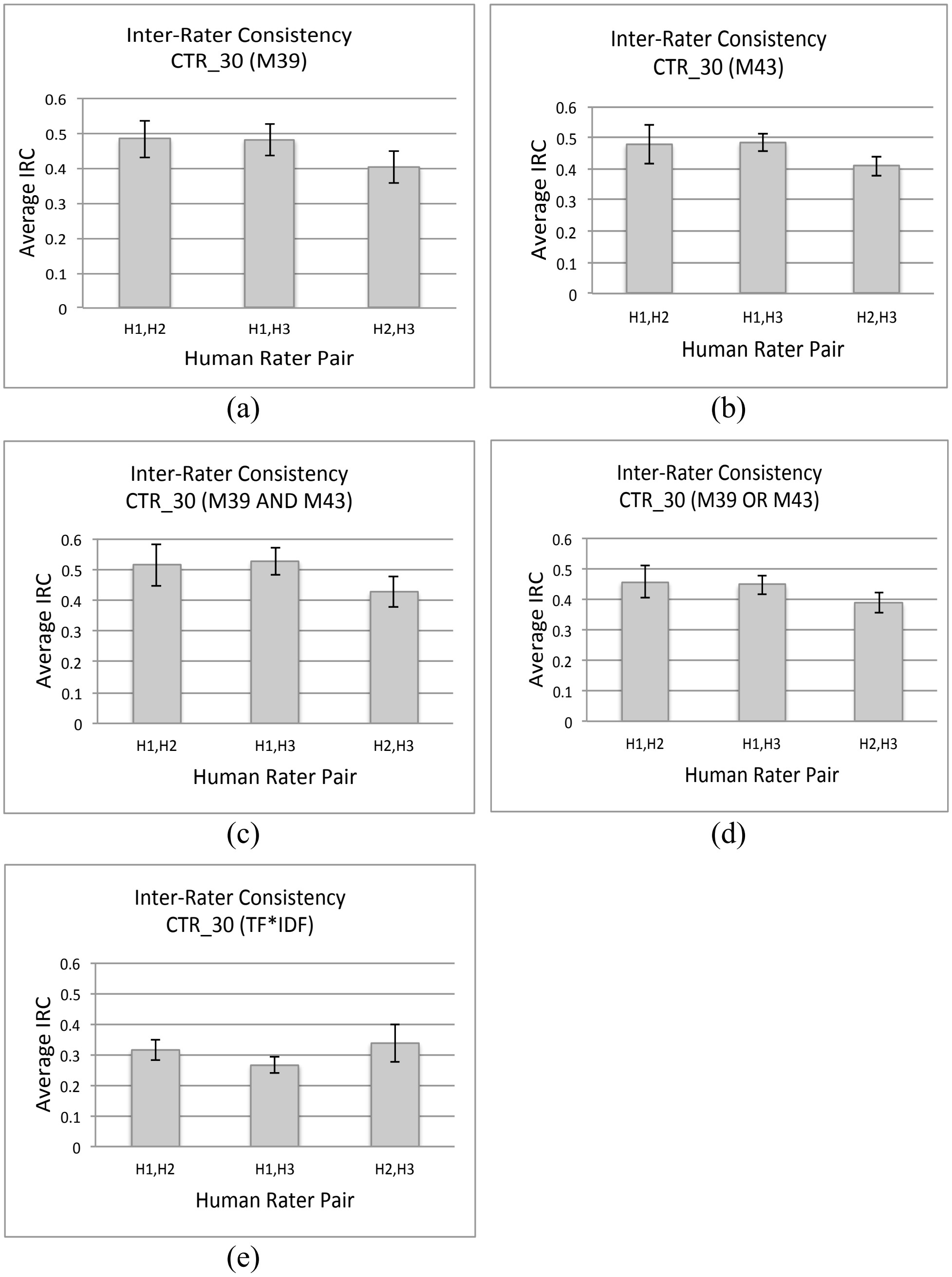


Figure 27. IRC between human topic raters. Topics were generated from CTR_30 by: (a) M39; (b) M43; (c) intersection of topics from M39 and M43; (d) union of topics from M39 and M43; (e) TF*IDF.

As confirmed in Table 7, the IRC values in (e) were lower than those depicted in (a)-(d). To examine this outcome significantly, we applied ANOVA after averaging the IRC values over all three pairs of raters. Significant differences (p < 0.0001) were found (see Table 28), and details are presented in Table 29. We expected that there would be a significant difference, i.e., significantly lower IRC, for TF*IDF based on Figure 27. This was found to be correct considering that only TF*IDF had a letter "C" in the Connecting Letters report in Table 29. Considering that M39_AND_M43 had a letter "A", which was not shared by any other settings in Table 29, and also it had the highest mean value, we were able to determine that M39_AND_M43 displayed a significantly higher IRC than any other Xpantrac setting (e.g., M39, M43, or M39_OR_M43).

Table 28. ANOVA showing the significant difference of IRC values in Figure 27 (a)-(e) averaged over all rater pairs

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Column 2	4	0.59524694	0.148812	106.9657	<.0001*
Error	145	0.20172540	0.001391		
C. Total	149	0.79697233			

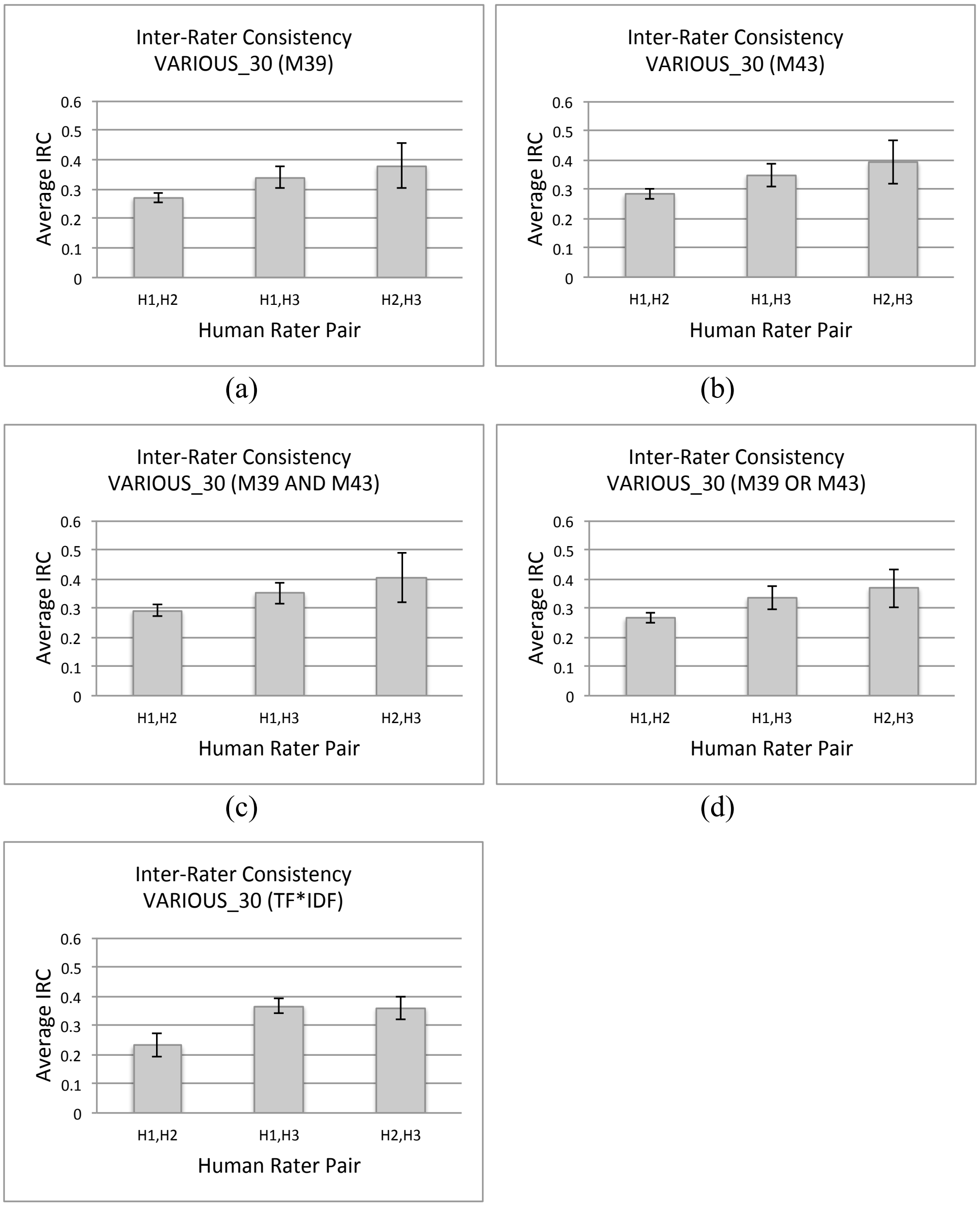
An interesting finding to note is that the descending order of the averaged IRC values by Xpantrac and TF*IDF settings (see Table 29) generally followed the order of the average topic ratings shown in Table 24, except that the order of M39 and M43 was different. In other words, the higher the topic relevance ratings were, the more consistent the human raters were in assigning the ratings. Does this mean that our human raters tended to agree with each other more for the relevant topics, but had difficulty assigning a low rating to less relevant topics? While beyond the scope of this study, this question does merit further investigation.

Table 29. The Connecting Letters report (alpha = 0.05) showing the difference details.

Level				Mean
M39_AND_M43	A			0.48562222
M43		B		0.45748889
M39		B		0.45541111
M39_OR_M43		B		0.43106667
TF*IDF			C	0.30595556

4.3.2 IRC on VARIOUS_30

Figure 28 shows the IRC values by rater pairs, averaged over all 30 documents in VARIOUS_30. The H1 and H2 pair showed the lowest IRC values for (a)-(e), which represents a different pattern from that of CTR_30, where the same pair assigned one of the highest IRC values in (a)-(d).



(a) (b) (c) (d) (e)

Figure 28. IRC between human topic relevance raters. Topics were generated from VARIOUS_30 by: (a) M39; (b) M43; (c) intersection of topics from M39 and M43; (d) union of topics from M39 and M43; (e) TF*IDF.

To determine if Figure 28 (e) was significantly different from the results depicted in (a)-(d), we applied ANOVA after averaging the IRC values over all of the pairs of raters. Indeed, (e) was not significantly different from (a) or (d); however, it was different from M43 and M39_AND_M43, as shown in Table 31.

As indicated in Section 4.3.1, the descending order of the averaged IRC values by Xpantrac and the TF*IDF settings "approximately" followed the order of the average topic ratings. For VARIOUS_30, however, this relationship changed from "approximately" to "exactly." Thus, the order of the content in the "Level" columns in Table 31 and Table 29 was exactly the same.

Table 30. ANOVA showing the significant differences of IRC values in Figure 28 (a)-(e) averaged over all rater pairs.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Column 2	4	0.01929405	0.004824	11.5137	<.0001*
Error	145	0.06074580	0.000419		
C. Total	149	0.08003984			

Table 31. The Connecting Letters report (alpha = 0.05). The IRC values of TF*IDF was not significantly different from that of M39_OR_M43.

Level				Mean
M39_AND_M43	A			0.34982222
M43	A	B		0.34075556
M39		B	C	0.33045556
M39_OR_M43			C	0.32308889
TF*IDF			C	0.31900000

4.4 Evaluation Using a Larger Data Set

In this section, we present the evaluation details of applying Xpantrac to a larger data set, in fact, about 33 times larger by number than either CTR_30 or VARIOUS_30. We used three different search APIs in the expansion process of Xpantrac for comparison purposes. Due to the large volume of the data set, we relied on the evaluation metrics from the Information Retrieval field: *precision, recall*, and F_1 measure, which were applied to tag sets.

4.4.1 New York Times Corpus

The *New York Times Annotated Corpus* data set was released in October 2008 by the Linguistic Data Consortium (LDC) (Sandhaus, 2008). The Corpus consists of over 1.8 million articles,1.5 million of which were manually tagged with a normalized indexing vocabulary for categories such as people, organizations, locations, and topic descriptors. Approximately 275,000 articles were algorithmically-tagged and verified by human staff.

To develop our data set, the NYT_1000, we queried the data set using keywords of interest, principally related to disasters and international events. This search generated about 83,000 articles; from that total we randomly selected 1,000 articles. Since each article had assigned tags in various categories (e.g., organizations, people, locations, descriptors, and names), we merged and used them as the gold standard topics for our evaluation.

4.4.2 Evaluation

To extract topics from the NYT_1000, three different search APIs, Bing, Yahoo! Web, and Yahoo! News (described in Table 32), were employed in the expansion process involving Xpantrac. In the core of the Yahoo! Web API, Microsoft's search platform is running. Therefore, Xpantrac with the Yahoo! Web API showed almost identical performance as Xpantrac with the Bing API.

The four Xpantrac settings (namely, M39, M43, M39_AND_M43, and M39_OR_M43) were used in the evaluation. In addition, we compared the topics from TF*IDF and OpenCalais. For each article in the NYT_1000, we extracted 20 topics for each setting. All of these topics and the gold standard topics (see Section 4.4.1) were stored into a database for easier evaluation. ANOVA and the Tukey and Kramer's Honestly Significant Difference (HSD) (Abdi & Williams, 2010) were computed for examining any significant differences.

Table 32. The descriptions of the search APIs used. The maximum (and default) number of records that can be retrieved with a single API call was 50 for all three APIs.

Search API	Description
Bing	Provided by Microsoft (https://datamarket.azure.com/dataset/bing/searchweb). The "Web results only" option was set in order to retrieve only the relevant webpage descriptions, excluding images, videos, news, and other data types.
Yahoo! Web	Provided by Yahoo! (http://developer.yahoo.com/boss/search/). The Web results are powered by the Microsoft search platform. It retrieves webpage descriptions only.
Yahoo! News	Provided by Yahoo! (http://developer.yahoo.com/boss/search/). It retrieves results from Yahoo!'s own indexed news database.

4.4.2.1 Tag Set Precision

Figure 29 shows the tag set precision scores of the various topic generation settings. Important to note is that all the topics generated by Xpantrac using the Yahoo! News API had the lowest tag set precision. Section 4.5.2 features a more detailed discussion of this outcome. In contrast, the Bing and Yahoo! Web APIs produced the highest tag set precision, with outputs from these two APIs almost identical.

The topics from OpenCalais had a slightly lower tag set precision when compared to the four Xpantrac settings with either Bing or Yahoo! Web APIs. The precision of the TF*IDF topics was even lower than that of OpenCalais.

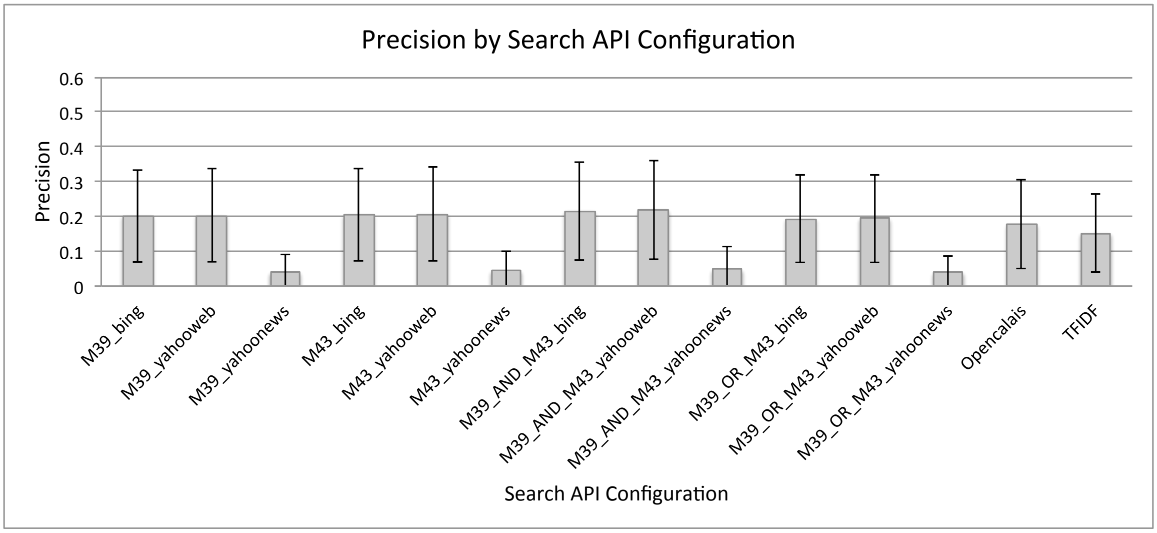


Figure 29. Tag set precision of topics generated from various settings using the NYT_1000. The precision values were averaged over 1000 articles.

4.4.2.2 ANOVA on Tag Set Precision Values

A significant difference was identified among the tag set precision values (Table 33). Table 34 clearly shows that the use of the Yahoo! News API produced topics with the lowest tag set precision in group "E." The topics from an intersection set of M39 and M43 using either the Bing or Yahoo! Web API, had the highest tag set precision in group "A," which was significantly different from the union set of M39 and M43 (Bing or Yahoo! Web APIs), OpenCalais, or TF*IDF. The tag set precision for the TF*IDF topics was significantly lower than any of the Xpantrac settings (with Bing or Yahoo! Web APIs) or OpenCalais.

These results confirmed that a total of six Xpantrac settings, M39, M43, and M39_AND_M43, with either Bing or Yahoo! Web APIs in groups "A" and "A B", outperformed the OpenCalais API for tag set precision scores.

Table 33. An ANOVA table presenting the significant difference among tag set precision scores for various configuration settings with search APIs.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Label	13	70.54064	5.42620	413.3335	0.0000*
Error	13986	183.60691	0.01313		
C. Total	13999	254.14755			

Table 34. The Connecting Letters report (alpha = 0.05) for the details of the significant differences in tag set precision.

Level						Mean
M39_AND_M43_yahooweb	A					0.21861400
M39_AND_M43_bing	A					0.21538800
M43_yahooweb	A	B				0.20650000
M43_bing	A	B				0.20370000
M39_yahooweb	A	B				0.20265000
M39_bing	A	B				0.20165000
M39_OR_M43_yahooweb		B	C			0.19395600
M39_OR_M43_bing		B	C			0.19284900
Opencalais			C			0.17715400
TFIDF				D		0.15140000
M39_AND_M43_yahoonews					E	0.04771800
M43_yahoonews					E	0.04465000
M39_yahoonews					E	0.04040000
M39_OR_M43_yahoonews					E	0.04018900

4.4.2.3 Tag Set Recall

Producing more topics and increasing the number of matching topics would, of course, be helpful in achieving a higher tag set recall score. Indeed, the union topic sets of M39 and M43 using Bing or Yahoo! Web APIs showed one of the highest tag set recall scores (see Figure 30). Again, the Xpantrac settings with the Yahoo! News API scored the lowest in tag set recall (see Section 4.5.2 for further discussion). The topics by OpenCalais showed the highest tag set recall score, followed closely by other Xpantrac settings with the Bing or Yahoo! Web APIs. The TF*IDF topics showed a much smaller tag set recall score compared to the others.

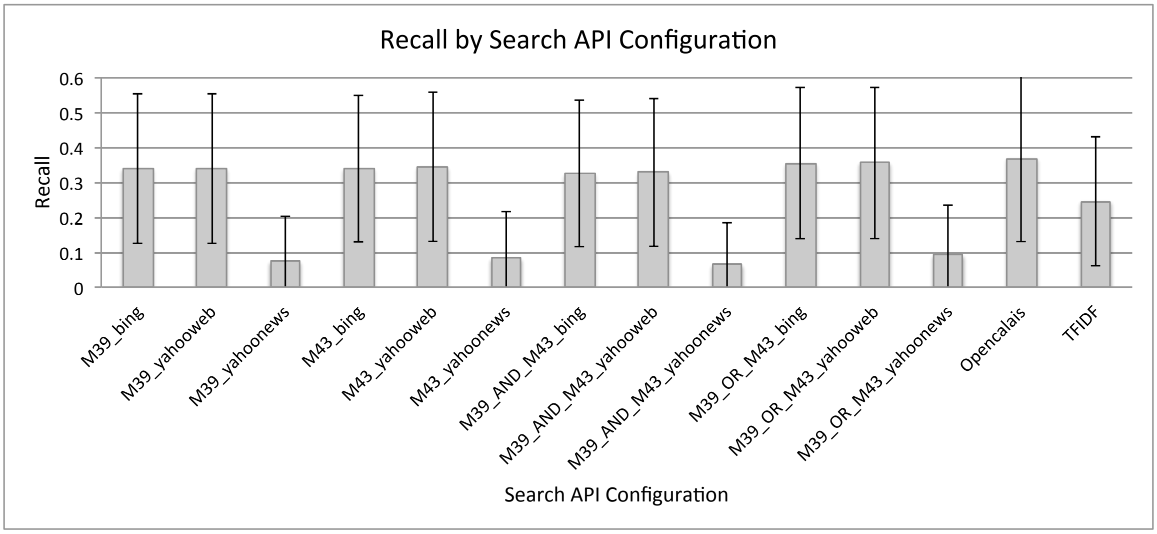


Figure 30. Tag set recall of topics generated from various settings using NYTimes articles. The recall values were averaged over 1000 articles.

4.4.2.4 ANOVA on Tag Set Recall Values

As shown in Table 36, OpenCalais produced topics with the highest tag set recall (0.368). However, four Xpantrac settings (namely, M39_OR_M43_yahooweb, M39_OR_M43_bing, M43_yahooweb, and M39_yahooweb) also performed as well as OpenCalais, as indicated by the fact that they showed no significant difference. The tag set recall score from TF*IDF, however, was significantly lower than any associated with Xpantrac (with Bing or Yahoo! Web APIs) or OpenCalais.

Note that in Table 34, the intersection sets of Xpantrac were among the top tag set precision scores, while the union sets were among the lower ones. However, this ordering was reversed for tag set recall in that the union sets were among the top scores, while the intersection sets were positioned lower on the list among the Xpantrac settings.

Table 35. An ANOVA table presenting the significant difference of tag set recall scores for various configuration settings with search APIs.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Label	13	195.81661	15.0628	403.1710	0.0000*
Error	13986	522.52900	0.0374		
C. Total	13999	718.34561			

Table 36. The Connecting Letters report (alpha = 0.05) for the details of significant difference in tag set recall.

Level						Mean
Opencalais	A					0.36829100
M39_OR_M43_yahooweb	A	B				0.35555000
M39_OR_M43_bing	A	B	C			0.35330100
M43_yahooweb	A	B	C			0.34390500
M39_yahooweb	A	B	C			0.33989300
M39_bing		B	C			0.33914500
M43_bing		B	C			0.33872300
M39_AND_M43_yahooweb		B	C			0.32824900
M39_AND_M43_bing			C			0.32457200
TFIDF				D		0.24563400
M39_OR_M43_yahoonews					E	0.09312900
M43_yahoonews					E	0.08339900
M39_yahoonews					E	0.07477500
M39_AND_M43_yahoonews					E	0.06504400

4.4.2.5 Tag Set F1 Score

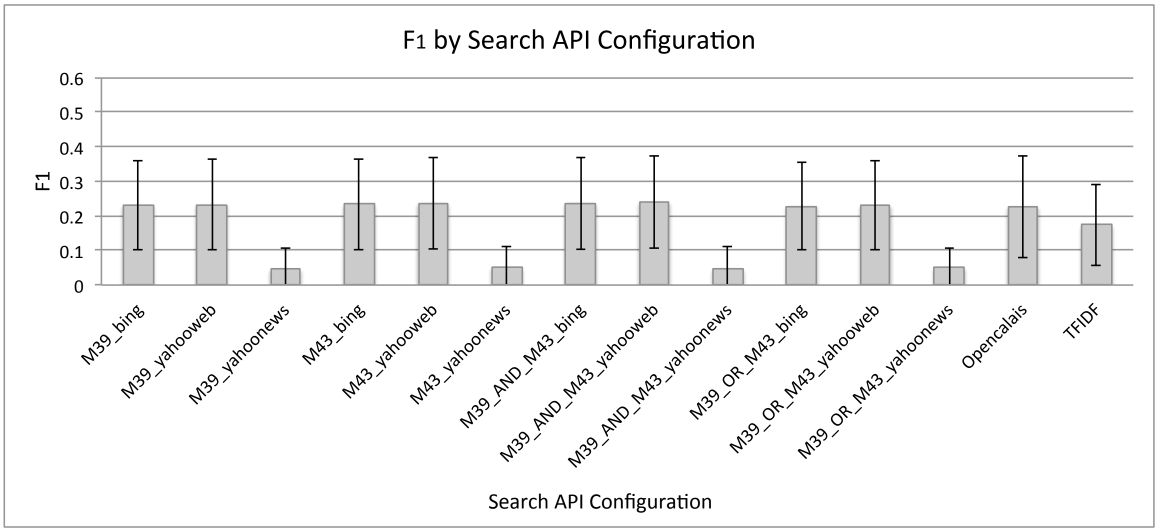


Figure 31. Tag set F1 score of topics generated from various settings using NYTimes articles. The F1 values were averaged over 1000 articles.

The tag set F1 scores for various topic generation settings are shown in Figure 31. The tag set F1 scores using the Xpantrac settings (excluding those with Yahoo! News API) and OpenCalais were among the highest scores. The tag set F1 score of TF*IDF was visibly lower than either of those. Similar to prior results, the tag set F1 scores for Xpantrac with the Yahoo! News API were the lowest. The reason behind this consistently poor performance originated from the fact that the

Yahoo! News API could not always retrieve the webpage descriptions as it was supposed to in the expansion step of the Xpantrac workflow (see Section 3.2). The details of this issue are explained in Section 4.5.2 with an example.

4.4.2.6 ANOVA on Tag Set F1 Score

When compared to the Xpantrac settings with the Yahoo! Web or Bing APIs, OpenCalais performed slightly lower in terms of the tag set F1 score. However, we could not identify a significant difference between them, which is reflected in Table 38. They were all in Group "A." Again, TF*IDF had the significantly lower tag set F1 score in comparison to Xpantrac and OpenCalais; the Xpantrac scores using the Yahoo! News API were the lowest.

Table 37. An ANOVA table showing the significant difference of tag set F1 scores for various configuration settings with search APIs.

Source	DF	Sum of Squares	Mean Square	F Ratio	Prob > F
Label	13	93.03355	7.15643	534.5316	0.0000*
Error	13986	187.24765	0.01339		
C. Total	13999	280.28120			

Table 38. The Connecting Letters report (alpha = 0.05) for the details of the significant differences in tag set F1.

Level				Mean
M39_AND_M43_yahooweb	A			0.23960000
M39_AND_M43_bing	A			0.23673800
M43_yahooweb	A			0.23624000
M43_bing	A			0.23310500
M39_yahooweb	A			0.23224200
M39_bing	A			0.23104200
M39_OR_M43_yahooweb	A			0.22999000
M39_OR_M43_bing	A			0.22832800
Opencalais	A			0.22661300
TFIDF		B		0.17342100
M43_yahoonews			C	0.05217900
M39_OR_M43_yahoonews			C	0.05097600
M39_AND_M43_yahoonews			C	0.04795000
M39_yahoonews			C	0.04701900

4.5 Discussion

4.5.1 Effects of the Data Set Properties on Topic Relevance Ratings

Regarding the topic quality rating study described earlier in Section 4.2, although there were significant differences in the relevance rating of topics between the TF*IDF and Xpantrac settings for both CTR_30 and VARIOUS_30, the "difference gap" was smaller for VARIOUS_30 in comparison to CTR_30. In other words, utilizing TF*IDF to extract topics worked better with VARIOUS_30. As mentioned in Chapter 3, the documents in CTR_30 consisted of homogeneous content (i.e., webpages about a single event, namely, the Hurricane Isaac disaster); in contrast, the content of documents in VARIOUS_30 was rather heterogeneous, involving other natural disasters, as well as epidemics and health issues.

In Section 2.2.2, we explained that TF*IDF favors TF (*Term Frequency)*, meaning that it favors words that appear frequently in the same document. At the same time, TF*IDF penalizes words that appear frequently in other documents due to IDF (*Inverse Document Frequency*). Nonetheless, the TF aspect would not be affected by the homogeneity/heterogeneity of the document content because we compute TF at the document level. However, IDF could be affected by such corpus-level properties. For example, if the data corpus consisted of homogeneous documents, many high frequency words in those documents would be shared by many other documents. This would be likely to occur throughout a homogeneous data set such as CTR_30, reducing the IDF of many potential topics and further decreasing the distinguishing power of them in general. Thus, the words with low relevance might be selected as topics, resulting in low topic quality.

In contrast, if the data set consisted of heterogeneous documents like VARIOUS_30, an overall decrease in the IDF values in the corpus probably would not occur because most terms would not be shared by many other documents. Therefore, the TF*IDF of each term would still maintain its distinguishing power.

What we learned from this issue is that TF*IDF term weighting might be dependent on the corpus properties, i.e., the heterogeneity or homogeneity of the documents. Therefore, care should be taken when applying TF*IDF to uncover significant terms from texts.

4.5.2 The Effectiveness of Search Engine APIs

Two APIs, Bing and Yahoo! Web, showed almost identical performance when they were used in the expansion process of Xpantrac. Using either of these two APIs, Xpantrac generated topics, which resulted in F1 scores that were slightly better than the analogous results using OpenCalais (see Section 4.4.2). However, the topics from Xpantrac with the Yahoo! News API showed the lowest F1 score.

The main reason was most likely the fact that the Yahoo! News API could not always retrieve enough relevant webpage descriptions when requested. Table 39 shows the number of retrieved webpage descriptions for each API query, separated by commas. One can see that both the Bing and Yahoo! Web APIs could retrieve 50 webpage descriptions for each of 31 queries, whereas the Yahoo! News API was unsuccessful in retrieving descriptions, except for the following four queries (7th, 15th, 28th, and 31st). As shown below, the total number of successfully retrieved descriptions using the Yahoo! News API was only 53 when the corresponding values by the other two APIs were 1,550 each.

In other words, Xpantrac was not able to fully expand the input documents. Thus, Xpantrac performed the extraction process based on poorly expanded data, resulting in the production of low quality topics.

Table 39. The number of webpage descriptions retrieved for a sample document (ID: 9224). Default setting of the APIs was to retrieve 50 descriptions per query (for a total of 31 queries).

API Name	Number of retrieved*	Sum
Bing	50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,50, 50,50,50,50,50,50,50,50,50,50,50,50,50,50,50	1,550
Yahoo! Web	50,50,50,50,50,50,50,50,50,50,50,50,50,50,50,50, 50,50,50,50,50,50,50,50,50,50,50,50,50,50,50	1,550
Yahoo! News	0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0 ,1,0,0,50	53

* Separated by a comma for each query return.

4.5.3 Graphical View of Topic Matching

To further assess the topic matching capability of Xpantrac or OpenCalais, which was the baseline for evaluation, we present three examples visualized as a topic network graph using the Gephi visualization platform (Bastian, Heymann, & Jacomy, 2009). The examples feature a best case, a moderate performance case, and a worst case. They showed the relationships between topics that were generated by two Xpantrac settings (i.e., M39_AND_M43, and M39_OR_M43) with Bing Search API, OpenCalais, and the gold standard.

It should be noted that we omitted M39 and M43 from the graph because their F1 scores were usually between the F1 scores of their intersection and union sets, thus they were not interesting to visualize. TF*IDF also was omitted since its F1 scores had been consistently the lowest compared to others, also making it not very interesting to visualize.

An Example of the *Best* Xpantrac Performance

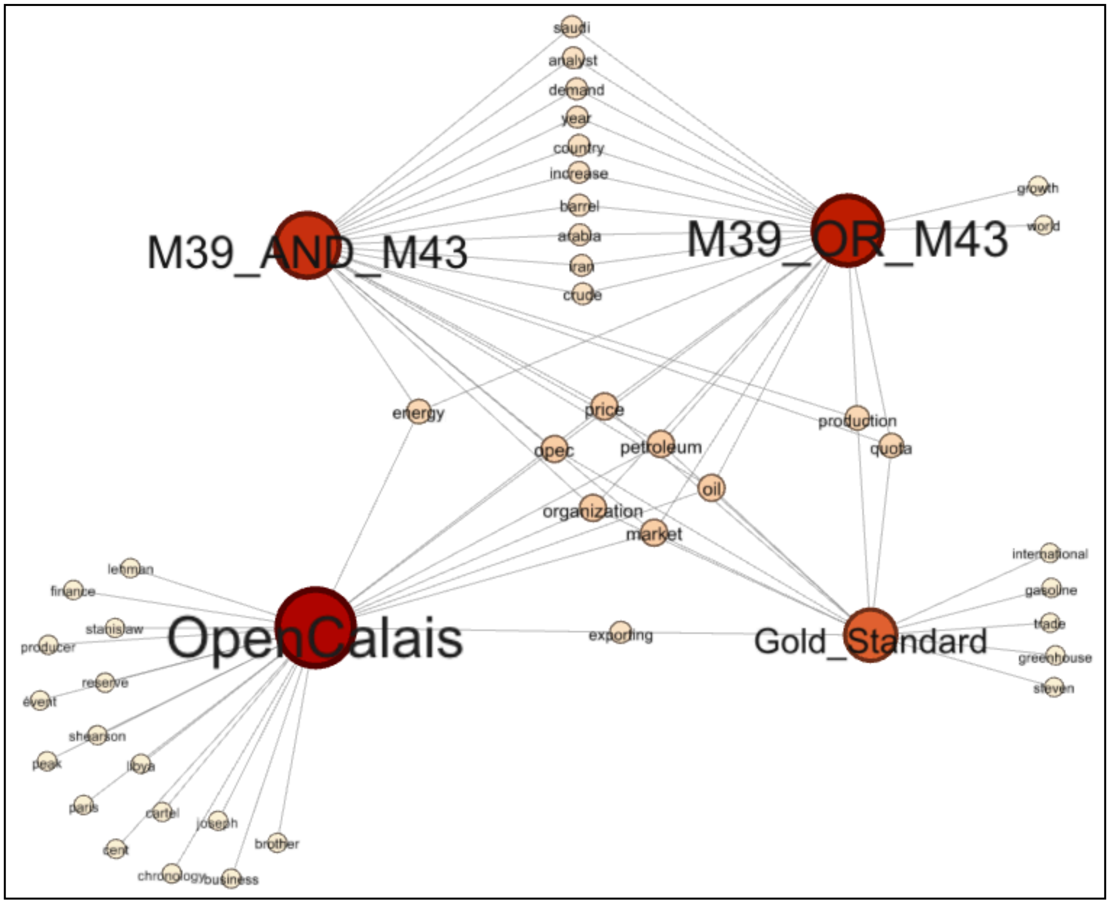


Figure 32. A topic network graph on a document (ID: 51133) in NYT_1000. The graph shows one of the best performing Xpantrac topics.

Figure 32 shows the topical relationships between the four settings. In this example, M39_AND_M43 had the highest F1 score (0.6) among M39_OR_M43 (0.529) or OpenCalais (0.368). Both M39_AND_M43 and M39_OR_M43 had 8 topics that match the Gold_Standard topics. However, the two additional topics (i.e., growth and world) of M39_OR_M43, which did not match any of the Gold_Standard topics, would have contributed to lowering the precision score, thereby resulting in a slightly lower F1 score than that of M39_AND_M43.

OpenCalais matched 7 topics of the Gold_Standard, which is one less than the other two settings. In addition, OpenCalais had 19 topics that did not match the Gold_Standard; in contrast, M39_AND_M43 had 11, and M39_OR_M43 had 13 non-matching topics. Therefore, the lower precision score for OpenCalais that originated from the 19 non-matching topics contributed to the lower F1 score for OpenCalais.

An Example of the *Moderate* Xpantrac Performance

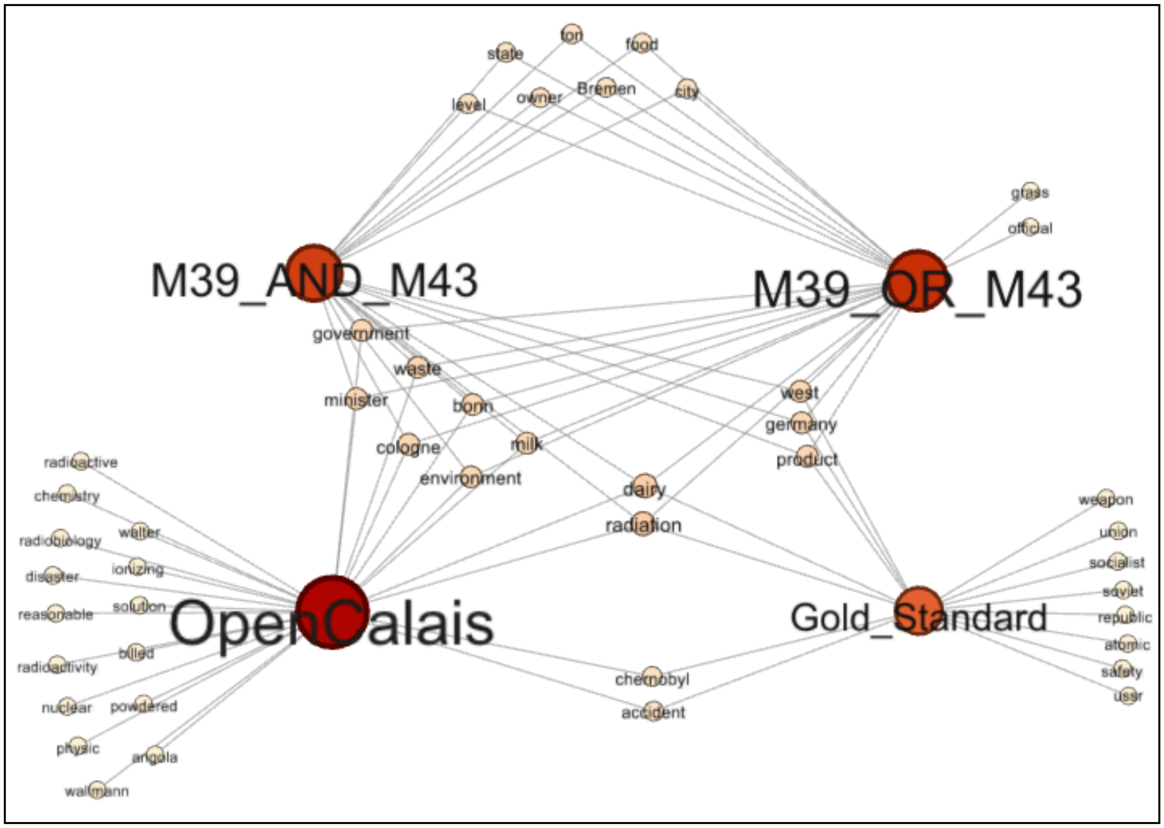


Figure 33. A topic network graph for a document (ID: 9227) in NYT_1000. This is an example of a moderately performing Xpantrac topic group.

The topics generated by the three settings in Figure 33 show moderate performance in terms of the F1 score: M39_AND_M43 (0.242), M39_OR_M43 (0.216), and OpenCalais (0.195). Both M39_AND_M43 and M39_OR_M43 matched 5 of the 13 Gold_Standard topics. However, these two settings also generated 14 and 16 non-matching topics respectively, resulting in a moderate performance. Again, M39_OR_M43 had two more non-matching topics (i.e., "grass", and "official") in comparison to M39_AND_M43, thus M39_OR_M43 had a slightly lower F1.

OpenCalais matched only 4 topics, and had an additional 22 non-matching topics. Therefore, its F1 was even lower than the other two Xpantrac settings.

An Example of the *Worst* Xpantrac Performance

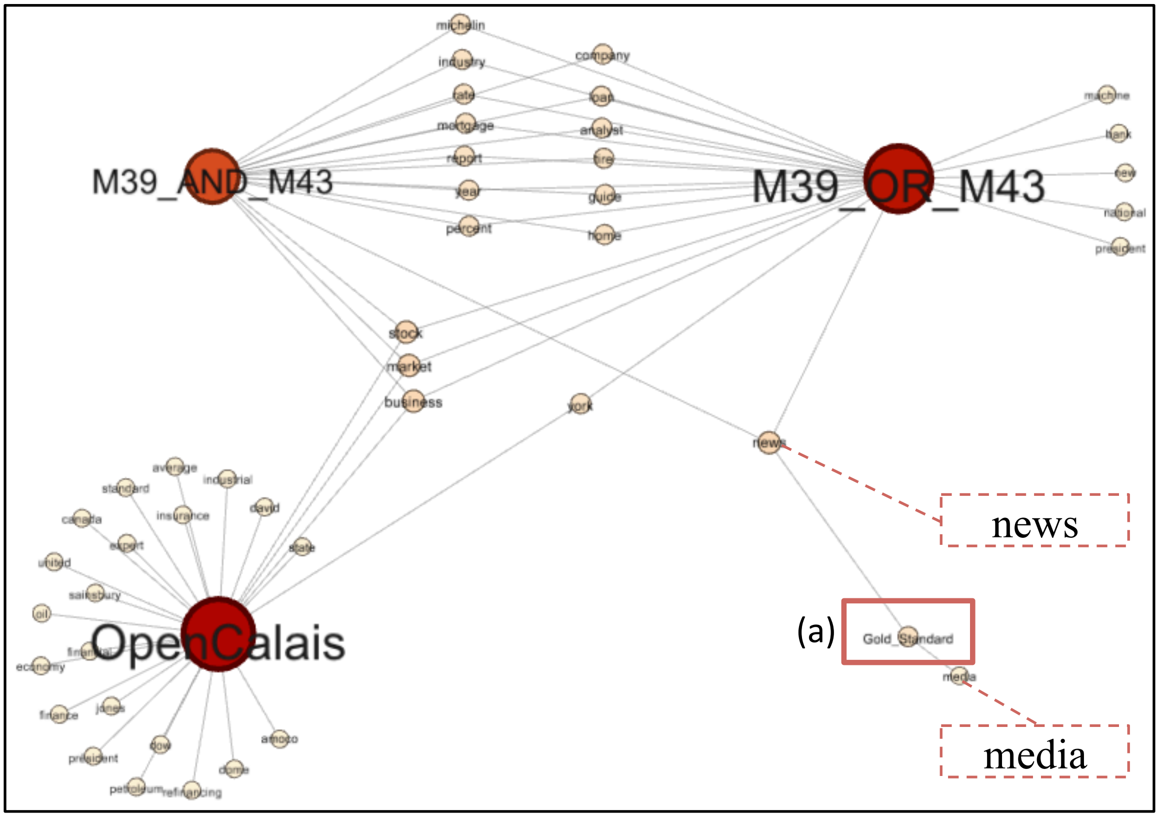
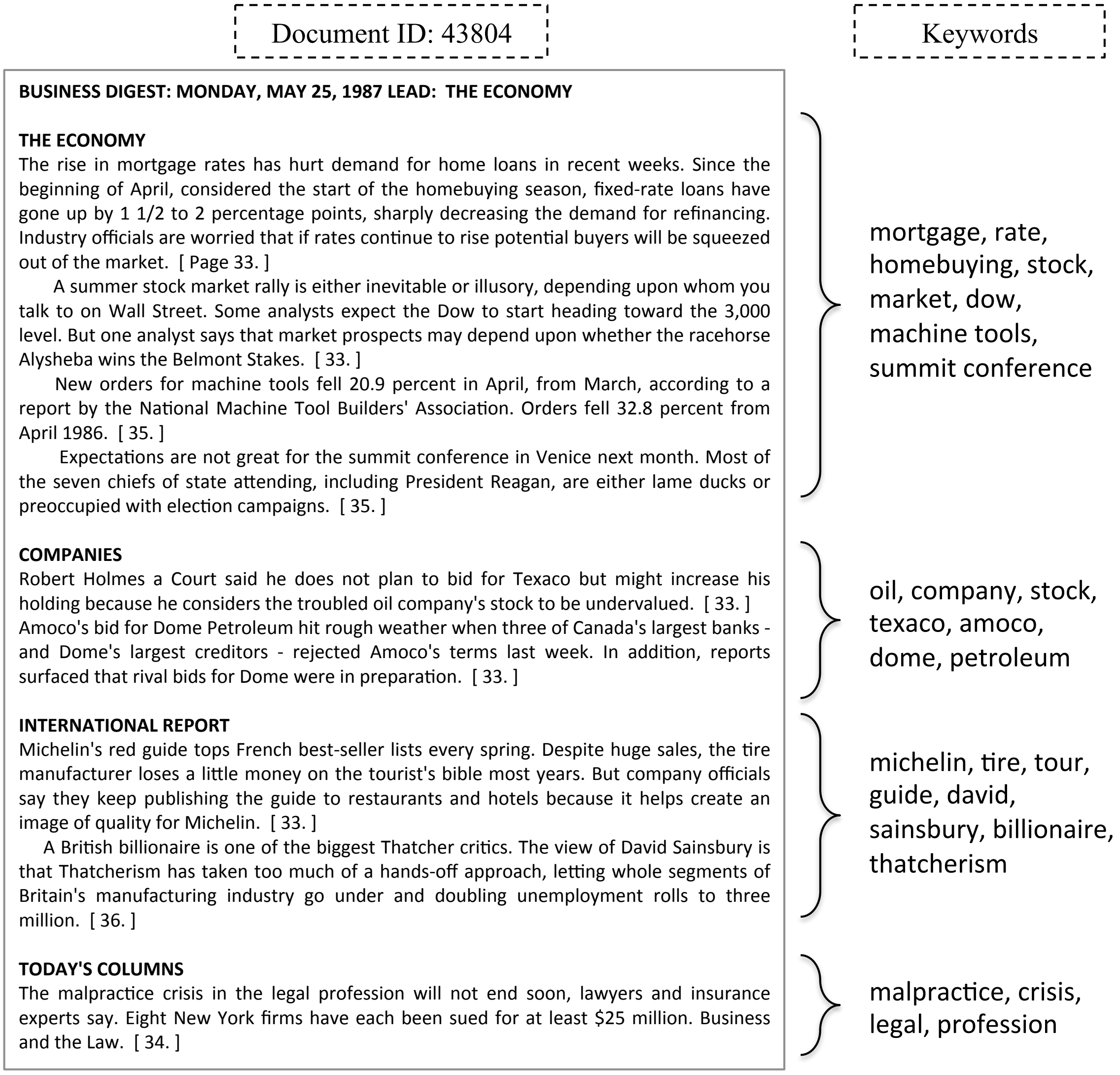


Figure 34. A topic network graph of a document (ID: 43804) in NYT_1000. The gold standard (a) had only two topics "news" and "media".

Figure 34 shows that (a) the Gold_Standard had only two topics "news" and "media," one of which ("news") matched both M39_AND_M43 and M39_OR_M43. These two Xpantrac settings also featured many non-matching topics. Thus, their F1 scores were close to zero. OpenCalais did not share any topic with the Gold_Standard—resulting in an F1 score of zero.

Let's examine this further by looking at the content of one document (ID: 43804) and considering a potential set of topics that might be used as replacements for the Gold Standard.



Document ID: 43804
Keywords

BUSINESS DIGEST: MONDAY, MAY 25, 1987 LEAD: THE ECONOMY

THE ECONOMY

The rise in mortgage rates has hurt demand for home loans in recent weeks. Since the beginning of April, considered the start of the homebuying season, fixed-rate loans have gone up by 1 1/2 to 2 percentage points, sharply decreasing the demand for refinancing. Industry officials are worried that if rates continue to rise potential buyers will be squeezed out of the market. [Page 33.]

A summer stock market rally is either inevitable or illusory, depending upon whom you talk to on Wall Street. Some analysts expect the Dow to start heading toward the 3,000 level. But one analyst says that market prospects may depend upon whether the racehorse Alysheba wins the Belmont Stakes. [33.]

New orders for machine tools fell 20.9 percent in April, from March, according to a report by the National Machine Tool Builders' Association. Orders fell 32.8 percent from April 1986. [35.]

Expectations are not great for the summit conference in Venice next month. Most of the seven chiefs of state attending, including President Reagan, are either lame ducks or preoccupied with election campaigns. [35.]

mortgage, rate, homebuying, stock, market, dow, machine tools, summit conference

COMPANIES

Robert Holmes a Court said he does not plan to bid for Texaco but might increase his holding because he considers the troubled oil company's stock to be undervalued. [33.]

Amoco's bid for Dome Petroleum hit rough weather when three of Canada's largest banks - and Dome's largest creditors - rejected Amoco's terms last week. In addition, reports surfaced that rival bids for Dome were in preparation. [33.]

oil, company, stock, texaco, amoco, dome, petroleum

INTERNATIONAL REPORT

Michelin's red guide tops French best-seller lists every spring. Despite huge sales, the tire manufacturer loses a little money on the tourist's bible most years. But company officials say they keep publishing the guide to restaurants and hotels because it helps create an image of quality for Michelin. [33.]

A British billionaire is one of the biggest Thatcher critics. The view of David Sainsbury is that Thatcherism has taken too much of a hands-off approach, letting whole segments of Britain's manufacturing industry go under and doubling unemployment rolls to three million. [36.]

michelin, tire, tour, guide, david, sainsbury, billionaire, thatcherism

TODAY'S COLUMNS

The malpractice crisis in the legal profession will not end soon, lawyers and insurance experts say. Eight New York firms have each been sued for at least $25 million. Business and the Law. [34.]

malpractice, crisis, legal, profession

Figure 35. The content of a document (ID: 43804). A set of potential keywords is shown on the right. However, the actual gold standard topics are "news" and "media".

Note that the input document (ID: 43804) in Figure 35 has four different sections in bold fonts, "The Economy," "Companies," "International Report," and "Today's Columns." Each section contains a distinct story. The potential keywords that we may come up with for those sections also are shown in Figure 35. When we used those keywords instead of the Gold_Standard topics, the F1 scores for M39_AND_M43, M39_OR_M43, and OpenCalais increased to 0.311, 0.353, and 0.340, respectively. However, as we examined in Figure 34, using the actual topics, "news" and "media", from the Gold_Standard led to very low F1 scores, which were close to zero.

What we learn from this is that while one might assume that the topics in the Gold_Standard are "*the answers*" and compute evaluation metric scores based on them, this might not be an entirely safe hypothesis in some cases. In fact, as our example in Figure 35 indicates, they might be deficient in conveying document content in a comprehensive way.

4.6 Summary

This chapter presents a detailed evaluation of Xpantrac. After explaining the experimental settings, details are provided on how we applied TF*IDF, which was used as a baseline, to extract topics from the documents. We also extracted topics from the data sets using the OpenCalais NLP API, since it is widely employed for various information extraction tasks. In so doing, we were able to compare the performance of Xpantrac and OpenCalais.

To evaluate the quality of the topics, which were extracted using a series of Xpantrac settings, we recruited three topic quality raters. The two men and one woman were asked to assign a number rating from the five-point Likert scale with respect to how much they agreed or disagreed that a given topic was relevant to the document. The results showed that the average topic relevance of Xpantrac outperformed the topic relevance of TF*IDF for both the CTR_30 and VARIOUS_30 data sets. The Inter-Rater Consistency (IRC) values between the raters also were computed.

In the CTR_30 data set, there was a significant difference between the average IRC value for the topics extracted using TF*IDF and the IRC values for the topics extracted using Xpantrac with M39, M43, M39_AND_M43, and M39_OR_M43. However, in the case of VARIOUS_30, the average IRC value of topics extracted using TF*IDF did not show a significant difference when compared with the IRC values of topics extracted using Xpantrac with M43 and M39_OR_M43 settings. In other words, the use of TF*IDF produced topics, of which the quality ratings had higher consistency in VARIOUS_30 than in CTR_30.

To assess the performance of Xpantrac using a larger data set, we utilized 1,000 documents from the *New York Times* corpus. In this experiment, we used three different search APIs (Bing API, Yahoo! Web, and Yahoo! News); these then were combined with the four Xpantrac settings and their intersection topics and union topics. In addition, we also included TF*IDF and OpenCalais. Since the documents in the NYT corpus were manually indexed for several categories, for example, places, people, and descriptions, we combined them and used them as the gold standard for evaluation. We employed evaluation metrics such as the tag set precision, recall, and F1 scores applied to tag sets to compare the topics from Xpantrac, TF*IDF, and OpenCalais. For tag set precision scores, Xpantrac performed significantly better than either TF*IDF or OpenCalais. In contrast, we could not identify any significant differences for tag set recall scores between Xpantrac and OpenCalais.

When the Yahoo! News API was used with Xpantrac, its topics performed the poorest. It was mainly because the API calls using the Yahoo! News did not return enough data consistently. In most cases, no webpage descriptions were returned. Such lack of data in the expansion step of Xpantrac might have contributed to the poor performance. Similarly, TF*IDF also performed somewhat disappointingly. Several example topic network graphs were developed to explain the reasons behind this performance variability.

For the higher scores in performance measures such as tag set precision, recall, and F_1, the role of the gold standard topics were critical. In some cases, all of the metric values were close to zero because the gold standard had only a small number of topics, which also were too abstract and broad to represent the document comprehensively. Therefore, whenever available, instead of relying entirely on the existing gold standard, it is recommended to have human indexers tag (sample) documents from the text corpus of interest and examine the Inter-Indexer Consistency as a means to find optimal parameters or to measure the performance of a tagging system in a more realistic setting.

Chapter 5. Graphical User Interface

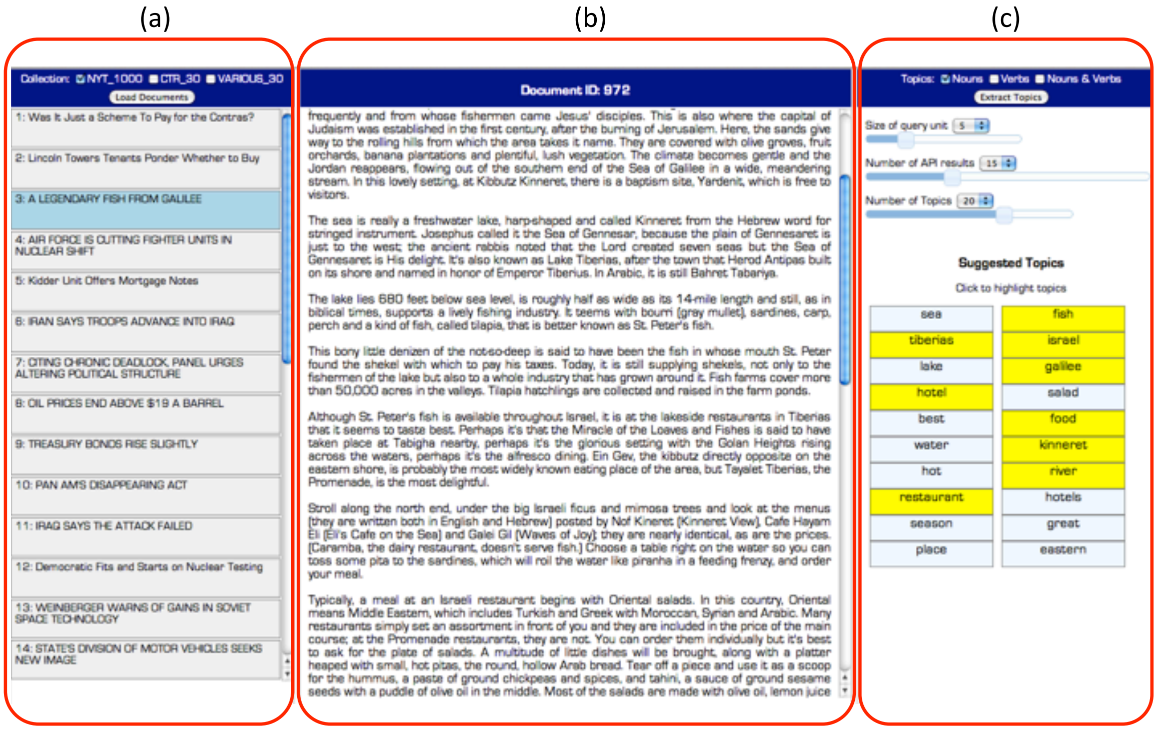
Chapters 3 and 4 detail the design, development, and evaluation of the Xpantrac system. Initially, we built a system that operates in command-line mode, which processes text documents continuously with minimal human intervention. Since it was a command-line tool, its utilization was limited to those who are familiar with executing commands on a computer's terminal window. However, general users might want to easily and accurately index their documents with topic tags in a semi-automatic way, interacting with a graphical user interface (GUI). Please consider the following scenario:

> *Rachel is a librarian working at a children's library. This library received about 100 short stories, each of which was written by young writers who recently started their literary career. To make these stories accessible online, Rachel decides to organize them based on the topic tags. So, she opens a Web browser and enters a URL of the Xpantrac UI. After loading documents that contain 100 stories, she selects each document to briefly view it, and then extracts suggested topic tags using the UI. After selecting several suggested tags from the Xpantrac UI, and also coming up with additional tags by herself, she enters them as the topic tags representing a story. A library patron, Jason, accesses the library homepage at home, clicks a tag "Christmas", which lists 5 stories about Christmas. He selects a story that might be appropriate for his 4-year daughter, and reads the story to her.*

To support interactive tagging tasks, a prototype of the GUI for Xpantrac has been built as a web-based tool, mainly using JavaScript, JQuery, and Apache SOLR. The initial version of the tool was devised in order to interface with three document collections (NYT_1000, CTR_30, or VARIOUS_30). A usability study for the GUI was subsequently conducted to examine its usefulness, which was based on user satisfaction reports. This chapter provides details about the UI design, development, and the analysis of user feedback collected from the usability study, followed by a discussion and a chapter summary.

5.1 Design of the User Interface

The graphical user interface (GUI) of the Xpantrac system is depicted in Figure 36. As shown below, it has three divided sections (from left to right), namely: the Collection Pane, the Document Pane, and the Topics Pane. In terms of its operation, users first load a document collection of interest in the Collection Pane (a). The titles of the documents from the loaded corpus are visualized as rectangular bars, each of which can be mouse-clicked. The content of the mouse-clicked document is then presented in the Document Pane (d) where users can read detailed content. The Topics Pane (c) allows users to adjust system parameters (e.g., the number of API returns and the number of topics). In addition, users can change the topic specifications; for example, the UI can suggest topic tags so that they appear only as nouns, only as verbs, or as both nouns and verbs. In general, topic suggestions in noun form would be the most effective. However, providing topic tags in verb format, or both noun and verb format, might better help the users in identifying the activities appearing in a document.



Collection: ☑ NYT_1000 □ CTR_30 □ VARIOUS_30

1: Was It Just a Scheme To Pay for the Contras?

2: Lincoln Towers Tenants Ponder Whether to Buy

3: A LEGENDARY FISH FROM GALILEE

4: AIR FORCE IS CUTTING FIGHTER UNITS IN NUCLEAR SHIFT

5: Kidder Unit Offers Mortgage Notes

6: IRAN SAYS TROOPS ADVANCE INTO IRAQ

7: CITING CHRONIC DEADLOCK, PANEL URGES ALTERING POLITICAL STRUCTURE

8: OIL PRICES END ABOVE $19 A BARREL

9: TREASURY BONDS RISE SLIGHTLY

10: PAN AM'S DISAPPEARING ACT

11: IRAQ SAYS THE ATTACK FAILED

12: Democratic Fits and Starts on Nuclear Testing

13: WEINBERGER WARNS OF GAINS IN SOVIET SPACE TECHNOLOGY

14: STATE'S DIVISION OF MOTOR VEHICLES SEEKS NEW IMAGE

Document ID: 972

frequently and from whose fishermen came Jesus' disciples. This is also where the capital of Judaism was established in the first century, after the burning of Jerusalem. Here, the sands give way to the rolling hills from which the area takes it name. They are covered with olive groves, fruit orchards, banana plantations and plentiful, lush vegetation. The climate becomes gentle and the Jordan reappears, flowing out of the southern end of the Sea of Galilee in a wide, meandering stream. In this lovely setting, at Kibbutz Kinneret, there is a baptism site, Yardenit, which is free to visitors.

The sea is really a freshwater lake, harp-shaped and called Kinneret from the Hebrew word for stringed instrument. Josephus called it the Sea of Gennesar, because the plain of Gennesaret is just to the west; the ancient rabbis noted that the Lord created seven seas but the Sea of Gennesaret is His delight. It's also known as Lake Tiberias, after the town that Herod Antipas built on its shore and named in honor of Emperor Tiberius. In Arabic, it is still Bahret Tabariya.

The lake lies 680 feet below sea level, is roughly half as wide as its 14-mile length and still, as in biblical times, supports a lively fishing industry. It teems with bourri (gray mullet), sardines, carp, perch and a kind of fish, called tilapia, that is better known as St. Peter's fish.

This bony little denizen of the not-so-deep is said to have been the fish in whose mouth St. Peter found the shekel with which to pay his taxes. Today, it is still supplying shekels, not only to the fishermen of the lake but also to a whole industry that has grown around it. Fish farms cover more than 50,000 acres in the valleys. Tilapia hatchlings are collected and raised in the farm ponds.

Although St. Peter's fish is available throughout Israel, it is at the lakeside restaurants in Tiberias that it seems to taste best. Perhaps it's that the Miracle of the Loaves and Fishes is said to have taken place at Tabigha nearby, perhaps it's the glorious setting with the Golan Heights rising across the waters, perhaps it's the alfresco dining. Ein Gev, the kibbutz directly opposite on the eastern shore, is probably the most widely known eating place of the area, but Tayalet Tiberias, the Promenade, is the most delightful.

Stroll along the north end, under the big Israeli ficus and mimosa trees and look at the menus (they are written both in English and Hebrew) posted by Nof Kineret (Kinneret View), Cafe Hayam Eli (Eli's Cafe on the Sea) and Galei Gil (Waves of Joy); they are nearly identical, as are the prices. (Caramba, the dairy restaurant, doesn't serve fish.) Choose a table right on the water so you can toss some pita to the sardines, which will roil the water like piranha in a feeding frenzy, and order your meal.

Typically, a meal at an Israeli restaurant begins with Oriental salads. In this country, Oriental means Middle Eastern, which includes Turkish and Greek with Moroccan, Syrian and Arabic. Many restaurants simply set an assortment in front of you and they are included in the price of the main course; at the Promenade restaurants, they are not. You can order them individually but it's best to ask for the plate of salads. A multitude of little dishes will be brought, along with a platter heaped with small, hot pitas, the round, hollow Arab bread. Tear off a piece and use it as a scoop for the hummus, a paste of ground chickpeas and spices, and tahini, a sauce of ground sesame seeds with a puddle of olive oil in the middle. Most of the salads are made with olive oil, lemon juice

Topics: ☑ Nouns □ Verbs □ Nouns & Verbs

Size of query unit 5

Number of API results 15

Number of Topics 20

sea	fish
tiberias	israel
lake	galilee
hotel	salad
best	food
water	kinneret
hot	river
restaurant	hotels
season	great
place	eastern

Figure 36. The overall user interface of the Xpantrac system: (a) The Collection Pane, (b) The Document Pane, and (c) The Topics Pane.

The topic extraction process allows users to go from (a) to (b) to (c). Moving sequentially from left to right is the equivalent of going from the larger scope of the collection-level, to the document-level, and finally to the smaller scope of the extracted topic-level (Figure 37).

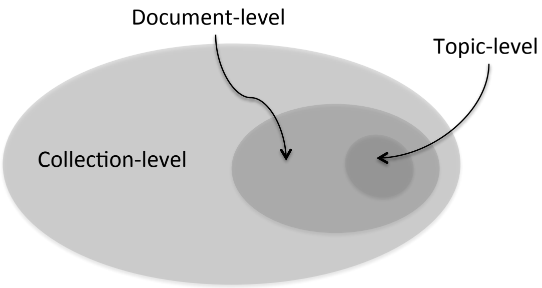
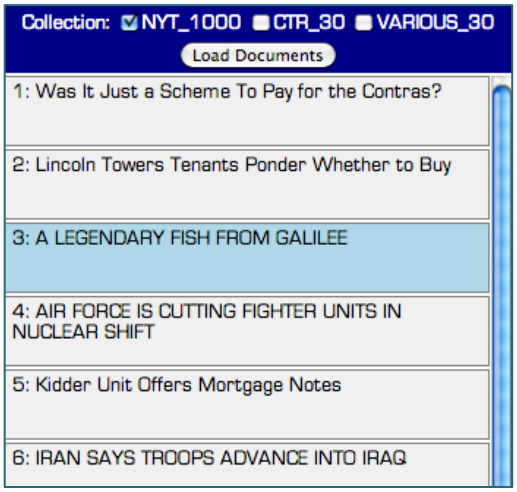


Figure 37. The topic extraction process: From macro to micro.

5.1.1 Components

5.1.1.1 The Collection Pane

For manual tagging tasks, users start by loading documents in the Collection Pane. First, they identify their collection of choice, NYT_1000, CTR_30, or VARIOUS_30, and then click the "Load Documents" button shown in the top portion of Figure 38. (The figure shows that the NYT_1000 collection is checked by default.) Once a document collection is loaded, the titles of the documents are displayed in stacked rectangular boxes. Users click one of the boxes to view the content of the article. Figure 38 shows that the article, "A Legendary Fish from Galilee," has been selected. The background color of the box changes to sky blue when it is clicked. To load another document collection, users simply check the other collection and press the "Load Document" button.



1: Was It Just a Scheme To Pay for the Contras?

2: Lincoln Towers Tenants Ponder Whether to Buy

3: A LEGENDARY FISH FROM GALILEE

4: AIR FORCE IS CUTTING FIGHTER UNITS IN NUCLEAR SHIFT

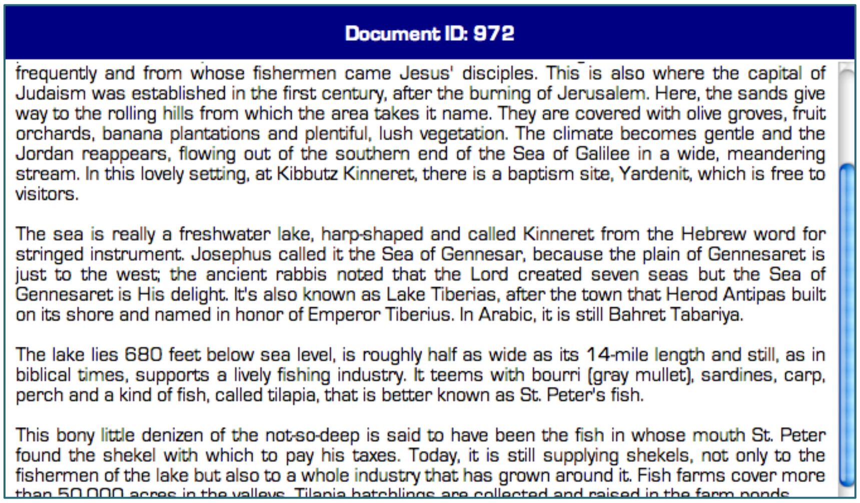
5: Kidder Unit Offers Mortgage Notes

6: IRAN SAYS TROOPS ADVANCE INTO IRAQ

Figure 38. The Collection Pane. The titles of the loaded documents are shown as stacked boxes.

5.1.1.2 The Document Pane

As a result of selecting the article, “3: A Legendary Fish from Galilee” in the Collection Pane (the rectangular box with sky blue color in Figure 38), appropriate documents then are presented in the Document Pane, located in the center of the UI (Figure 36 (b)). The document ID is displayed on top of the Pane, and a vertical scroll bar is automatically created for the articles that are longer than the height of the Pane (Figure 39). Users read or skim through the article looking for relevant topic tags.



Document ID: 972

frequently and from whose fishermen came Jesus' disciples. This is also where the capital of Judaism was established in the first century, after the burning of Jerusalem. Here, the sands give way to the rolling hills from which the area takes it name. They are covered with olive groves, fruit orchards, banana plantations and plentiful, lush vegetation. The climate becomes gentle and the Jordan reappears, flowing out of the southern end of the Sea of Galilee in a wide, meandering stream. In this lovely setting, at Kibbutz Kinneret, there is a baptism site, Yardenit, which is free to visitors.

The sea is really a freshwater lake, harp-shaped and called Kinneret from the Hebrew word for stringed instrument. Josephus called it the Sea of Gennesar, because the plain of Gennesaret is just to the west; the ancient rabbis noted that the Lord created seven seas but the Sea of Gennesaret is His delight. It's also known as Lake Tiberias, after the town that Herod Antipas built on its shore and named in honor of Emperor Tiberius. In Arabic, it is still Bahret Tabariya.

The lake lies 680 feet below sea level, is roughly half as wide as its 14-mile length and still, as in biblical times, supports a lively fishing industry. It teems with bourri (gray mullet), sardines, carp, perch and a kind of fish, called tilapia, that is better known as St. Peter's fish.

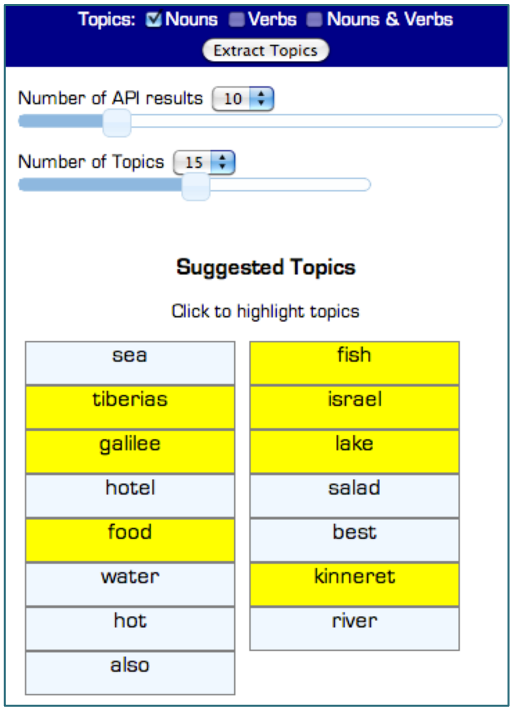
This bony little denizen of the not-so-deep is said to have been the fish in whose mouth St. Peter found the shekel with which to pay his taxes. Today, it is still supplying shekels, not only to the fishermen of the lake but also to a whole industry that has grown around it. Fish farms cover more than 50,000 acres in the valleys. Tilapia hatchlings are collected and raised in the farm ponds

Figure 39. The Document Pane. Users view the content of the document.

5.1.1.3 The Topics Pane

The Topics Pane is presented in Figure 40, which is located in the right side of the UI (Figure 36 (c)). As noted at the start of Section 5.1, it gives users some options for how the results can be displayed. For example, the system can suggest topic tags in noun form, verb form, or both. Right below those options, there are two slider bars, each with an associated drop-down list. Both the sliders and the drop-down lists can be used to change the settings.

Similarly, both the slider bar and drop-down list can be used to set the parameter, “Number of API results,” which controls the amount of information to process for extracting topics. Its parameter values are one of [1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50]. The larger the value, the more information is processed, presumably producing a more robust set of topics; it also, however, takes somewhat longer to process more information. Both the slider bar and the drop-down list that correspond to the parameter, “Number of Topics,” can be used to set the number of topic tags to be displayed on the UI. Its value can be selected from [5, 10, 15, 20, 25, 30]. The suggested topics are presented in descending order of their relevance to the document. The background color of each topic box toggles between “Alice blue” and “yellow” whenever it is clicked. Users can configure the parameters until the system suggests a list of useful topic tags, and then select (make yellow) the ones that interest them the most, to highlight for later reference.


Topics: ☑ Nouns □ Verbs □ Nouns & Verbs

Extract Topics

Number of API results 10

Number of Topics 15

Suggested Topics

Click to highlight topics

sea	fish
tiberias	israel
galilee	lake
hotel	salad
food	best
water	kinneret
hot	river
also	

Figure 40. The Topics Pane. Users adjust parameters to extract a useful set of suggested topics.

5.1.2 Implementation Details

We introduce the technical terms and their descriptions relevant to the implementation of the prototype UI in Section 5.1.2.1, to help readers becoming familiar with the subject. In its original design, Xpantrac system accesses a commercial search engine API to query segments of input texts, and to retrieve their relevant webpage descriptions as explained in Chapter 3. However, we can emulate the behavior of this commercial search engine using SOLR if the documents, from which to extract topics, are known. For this, the documents as well as their relevant webpage descriptions should be prepared in advance using a commercial search engine. Figure 41 in Section 5.1.2.2 illustrates this process of expanding a single document to its relevant webpage descriptions. Both the documents and the relevant descriptions in XML format then are indexed using SOLR. In other words, we expand the documents with a commercial search engine once, store the expanded information locally in SOLR, and then repeatedly use them for topic extraction purposes.

Once SOLR is ready to serve both the documents and their webpage descriptions, the UI can communicate with the API of SOLR using the REST API calls, as illustrated in Figure 42, Figure 43, and Figure 45. As a result, JSON files are returned from SOLR (Figure 44, Figure 46), and they are processed by the UI to identify topic suggestions, and displayed on the Topics Pane of the UI.

5.1.2.1 Technical Terms and Descriptions

The successful implementation of the prototype user interface of Xpantrac required a specific combination of multiple technical components. The relevant technical terms and their descriptions are listed in Table 40. Regarding the technologies, JQuery UI and JavaScript were used to establish the three-pane structure (see Figure 36), as well as to functionalize the buttons, sliders, and check boxes of the prototype UI. The prototype UI communicates with the REST API of the SOLR indexing engine to retrieve needed records in JSON, which are processed by functions implemented in the prototype UI using JavaScript.

Table 40. Technical terms and their descriptions.

Term	Description
Scripting vs. Programming	Traditionally, it was explained that scripts are interpreted, and programs are compiled (to bytecode or machine code). However, there is no clear distinction between them. It is rather a matter of how the language is used. A so-called scripting language (e.g., Python) can be used to implement sophisticated programs, and a programming language (e.g., C++) can be used to script a small object.
Java vs. JavaScript	Java is a full-fledged Object-Oriented Programming (OOP) language, which is capable of creating stand-alone applications that run in a virtual machine or in a browser. Client-side JavaScript resides inside HTML and can run only on a browser to enhance the interactivity of the webpages (Oracle, 2013).
JQuery, JQuery UI	JQuery is the most popular JavaScript library, which allows scripting of the client-side HTML to be simple. It is used to handle events, create animations, develop Ajax applications, and to navigate the document (The jQuery Foundation, 2013a). JQuery UI is a set of user interface widgets, interactions, effects, and themes, which are built on the basis of JQuery (The jQuery Foundation, 2013b).
JSON vs. XML	Both JavaScript Object Notation (JSON) and Extensible Markup Language

	(XML) are data formats for open data sharing. JSON has slightly simpler syntax and is more readable, but it can store only traditional data types such as text and numbers. XML, in contrast, can store any data types and extend the attributes of the stored data in exchange for a slight degradation in readability (Mikoluk, 2013).
REST API	The Representational State Transfer (REST) Application Programming Interface (API) facilitates the exchange of data on the Web by sending standard HTTP requests such as GET, POST, PUT, and DELETE. It allows communication between machines (i.e., Web apps) and webservers that serve resources (Fielding & Taylor, 2000; Fischer, 2013).
SOLR	Apache SOLR is a popular information retrieval system that provides fast access and retrieval of its (potentially massive) indexed resources using the REST-like API (The Apache Software Foundation, 2011).

5.1.2.2 SOLR

Chapter 3 detailed how Xpantrac uses a search engine API to expand a given text document to multiple (likely) relevant documents (i.e., merged webpage descriptions), after which they are processed to extract topic tags. In the prototype UI, however, SOLR replaces the role of a search engine because it can index the documents in the data sets (Figure 41 (a)). It also is capable of indexing all the webpage descriptions (Figure 41 (b)) retrieved using a search engine API for those documents. By reusing this type of indexed data, SOLR is able to emulate the search engine for our data sets (see Figure 42).

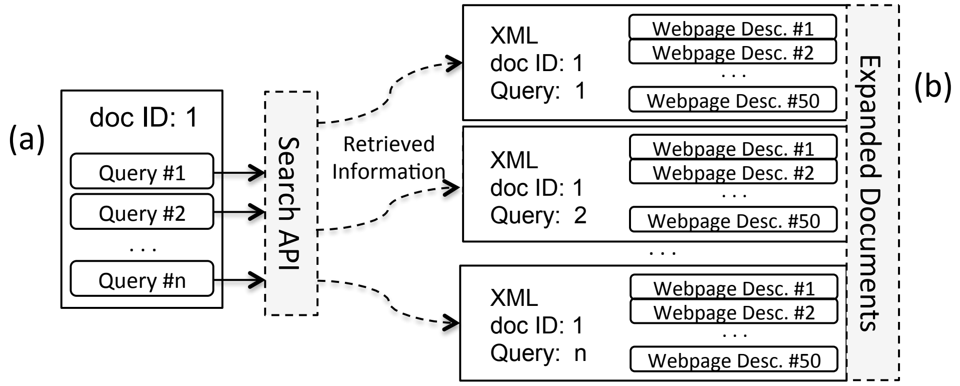


Figure 41. Process of expanding a single input document (a) to multiple expanded documents in XML format (b), using a search API.

Replacing the search engine with SOLR has an important benefit. For the usability evaluation of the prototype UI, users might repeatedly extract topic tags from the same documents every time they change the parameter settings of the UI. In the case of the SOLR indexing system, instead of querying the search engine API with the same queries multiple times, a user can query the search engine just once, index the returned results using SOLR, and query SOLR as many times as needed. Although it is an efficient and economic system, this approach is available only when the documents for the usability study are known in advance.

To deploy the Xpantrac UI for general use, a search engine API should be used in order to successfully expand any new text documents from users. As an alternative, a SOLR system could be constructed with the following two constraints:

1. It should index a massive number of documents, so that any documents from users could be expanded based on indexed information.
2. It should return the most relevant portion of the matching documents, for a query.

Satisfying the first constraint is important considering that the performance of Xpantrac was very low when the Yahoo! News API was used in the expansion process. Specifically, the Yahoo! News API did not always return the relevant webpage descriptions for the given query words, which led to an incomplete expansion of the input text and, therefore, the extraction of low quality topic tags. In contrast, SOLR would likely produce higher quality topic tags with less bias under Constraint 1. With any given set of query terms, the search engine API used by Xpantrac was designed to return multiple relevant webpage description, typically the first paragraph (or less) of the matching webpages. To simulate this outcome, SOLR can be designed to accommodate Constraint 2, above.

If that is done, Xpantrac might be applied to extracting topics for new documents related to a popular topic. For example, in the IDEAL project, which inherits data from the CTRnet project, SOLR might work with a large collection of webpages related to recent hurricanes. That collection might be used for expansion, using SOLR, instead of using a search engine API applied to the whole WWW.

5.1.2.3 Retrieving Information from SOLR

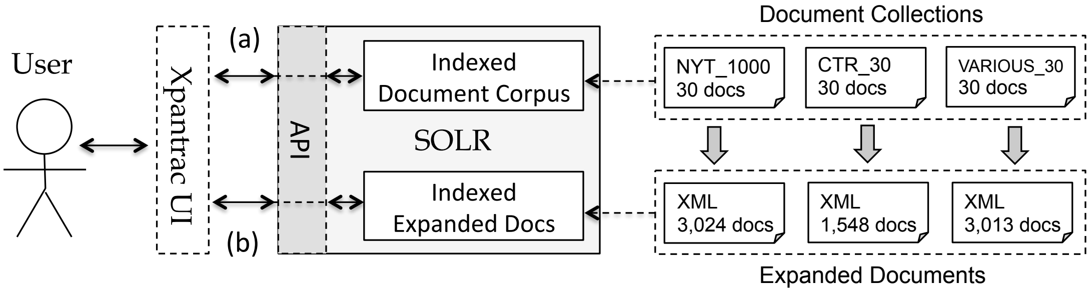


Figure 42. A diagram showing the indexing of documents and expanded documents using SOLR to emulate a search engine API.

Figure 42 illustrates that both of the document collections (located top right, inside a dotted line box) as well as their expanded documents in XML format (bottom right, inside a dotted line box) are indexed by SOLR. 30 documents were indexed from each document collection. Since the length of the documents in those 3 collections was different, each collection generated a different number of expanded documents in XML format. For example, 30 documents in NYT_1000 were expanded into 3,024 XML documents consisted with webpage descriptions. Likewise, the entire CTR_30 collection was expanded to 1,548 XML documents, and VARIOUS_30 was expanded to 3,013 XML documents.

Figure 42 also shows how the Xpantrac UI communicates with the SOLR API to access both the indexed documents (through (a)) and the indexed expanded documents (through (b)). An example API query to retrieve a document from an indexed document corpus is depicted in Figure 43. Notice the letters in bold font requesting a record, wherein "web_doc_content" is "1" and "web_doc_id" is "ctr_1." The values of "web_doc_content" are either 0 or 1; it is used to distinguish whether the record is the expanded webpage descriptions (value is 0) or text documents (value is 1). Also, "wt=json" requests to return the record in JSON format.

```
http://server_hostname:port_name/solr/collection1/select?q=web_doc_cont
ent%3A1%20AND%20web_doc_id%3Actr_1&wt=json&json.wrf=?
```

Figure 43. A SOLR query to retrieve a text document in JSON format.

Based on the communication pathway depicted in Figure 42 (a), the query in Figure 43 is sent to the SOLR API, which returns a record in JSON format as shown in Figure 44. The values for "web_doc_id" and "web_doc_content" match the ones specified in Figure 43. The text content of the key, "web_text," is to be presented in the Document Pane in Figure 36 (b) and in Figure 39.

```
{
  "responseHeader":{
    "status":0,
    "QTime":2,
    "params":{
      "indent":"true",
      "q":"web_doc_id:ctr_1 AND web_doc_content:1",
      "wt":"json"}},
  "response":{"numFound":1,"start":0,"docs":[
      {
        "id":"4603",
        "web_doc_id":"ctr_1",
        "web_title":"Haiti braces for Isaac's deluge",
        "web_text":"Haiti braces for Isaac’s deluge\n \nThe threat of a
        direct hit by Isaac on Southeast Florida might be declining but
        the Keys remain in the firing line. Rain and gusts could affect
        much of the state.\n\nRelated Content\nAn early look: What’s
        open and closed as Isaac approaches\nTrack Tropical Storm
        Isaac\n\nBY CURTIS MORGAN\nCMORGAN@MIAMIHERALD.COM\n\nPORT-AU-
        PRINCE -- Still weak Tropical Storm Isaac churned toward
        Hispaniola on Friday morning, a crossing that could blunt the
        impact on South Florida but...
        (continued…)
        Though storms in the past have dramatically weakened or
        dissolved over Haiti and Cuba, Feltgen said forecasters don’t
        see that happening with Isaac. In fact, Isaac is expected to
        strengthen in the Gulf, where it could threaten anywhere from
        the Florida Panhandle to Texas next week.",
        "web_doc_content":"1",
        "corpus":"ctr",
        "_version_":1448515558585663488}]
}}
```

Figure 44. A JSON document retrieved as a result of the SOLR query in Figure 43.

To extract the topic tags of a text document, the expanded documents of the given text document should be retrieved from SOLR. For this, another API query is used (Figure 45). This time, the value of “web_doc_content” is “0,” which means that the records to be retrieved do not correspond to the content of a text document; instead, they represent “expanded information” of a given text document. The value of “web_doc_id” is the same “ctr_1.”

```
http:// server_hostname:port_name
/solr/collection1/select?q=web_doc_content%3A0+AND+web_doc_id%3Actr_1&r
ows=1000&json.wrf=?&wt=json
```

Figure 45. A SOLR query to retrieve all the expanded documents in JSON format.

Based on the communication path depicted in Figure 42 (b), the query in Figure 45 is sent to the SOLR API, which then returns records in JSON format as shown in Figure 46. As the value of the “numFound” key denotes, a total of 68 records are retrieved, two of which are displayed. From this we know that the document with ID “ctr_1” was divided into 68 query units and each unit has been expanded to produce a single new record.

The values for “web_doc_id” and “web_doc_content” in each of two records match the ones specified in Figure 45. However, the value of the “web_query_id” key distinguishes the returned expanded document records. For example, the first record shown in Figure 46 has the value for “web_query_id” of “1,” whereas the second record in the same figure shows “68” as the value of “web_query_id.”

The value of the “web_results” key contains a series of text strings separated by a delimiter “|”. There are a total of 50 such strings, which is the maximum number of webpage descriptions provided by a search engine API per query. Users of the Xpantrac UI can adjust the value of the parameter “Num API Results” on top of the Topics Pane in Figure 36 (c). Thus, if a user set this parameter value to “20,” for example, only 20 text strings out of 50 strings would be collected from each of 68 records, which then are processed to extract topic tags. The produced tags are displayed on the middle and bottom part of Figure 36 (c) and in Figure 40.

```
{
  "responseHeader":{
    "status":0,
    "QTime":1,
    "params":{
      "indent":"true",
      "q":"web_doc_id:ctr_1 AND web_doc_content:0",
      "wt":"json",
      "rows":"1000"}},
  "response":{"numFound":68,"start":0,"docs":[
      {
        "id":"3055",
        "web_doc_id":"ctr_1",
        "web_query_id":"1",
        "web_query":"haiti braces isaacs deluge threat",
        "web_results":"The threat of a direct hit by Isaac on Southeast
        Florida might be declining but the Keys remain in the firing
        line Rain and gusts could affect much of the|The threat of...
        (continued…)
        Tracking Emily Haiti Braces for Deluge from Tropical Storm
        Emily Update at 8 a m Hurricane Irene Major East Coast Threat
        Congrats Courtney Station Team",
        "web_api":"bing",
        "web_doc_content":"0",
        "_version_":1448241280383451136},

        --(*66 more JSON documents with "web_query_id" from 2 to 67)--

      {
        "id":"3122",
        "web_doc_id":"ctr_1",
        "web_query_id":"68",
        "web_query":"florida panhandle texas week",
        "web_results":"Panhandle weather forecast and weather
        conditions Today s and tonight s Panhandle weather forecast
        plus Doppler radar from weather com|Vacation Rentals in Florida
        Panhandle starting from 672 per week View TripAdvisor s 35 655
        unbiased reviews 219 311 photos and great deals
        (continued...)
        Leaders to blue slip the S 744 amnesty bill that the Senate
        passed last week|could leave up to 15 inches of snow in some
        sections of Texas and Kansas where a storm last week dumped
        Alabama the Florida Panhandle and Georgia Home",
        "web_api":"bing",
        "web_doc_content":"0",
        "_version_":1448241280779812864}]
}}
```

Figure 46. The result of the query shown in Figure 45. Two JSON records (i.e., the first and the last one) are displayed from a total of 68 documents.

5.2 Usability Study

The goal of the usability study was to determine the usefulness of the Xpantrac UI as indicated by the perceived satisfaction of users when they were conducting tagging tasks supported by the Xpantrac UI. The context of the usability study required users to read newspaper articles that were presented on the UI, and then identify relevant topic tags. There was no control that the users had to select topics suggested by the UI, or that they had to run the UI at a certain point in time during the process. Instead, the experimental setting was purposefully designed to have the users decide whether to use the provided UI or not in a flexible and natural environment.

Specifically, the following questions helped to guide the usability study:

- How do people conduct tagging tasks supported by our prototype UI?
- How useful and satisfactory was the prototype UI for users?
- What was the quality of the topics indexed in an interactive semi-automatic way?

5.2.1 Study Design and Experiment

Participants for this study were recruited through an online listserv; they were situated in a variety of remote locations. Therefore, all communication between the study participants and the researchers had to be mediated by email messages, including submission of their signed informed consent forms in PDF format. In addition, the prototype UI for the study was made to be accessible as a Web-based application. All of the questionnaires were developed using *Qualtrics*, which is a sophisticated Web-based survey UI for aiding researchers in data collection and analysis (2002). The expected time to complete all tasks, including the questionnaire, was designed to be less than an hour to minimize the participants' mental fatigue.

5.2.1.1 Participant Recruitment

The participants were mainly recruited from the JESSE listserv[11] by posting a message regarding the usability study. Fortunately, a significant number of people showed an interest in taking part in the study. In fact, some of the JESSE members who were faculty members and data scientists in various universities around the country also forwarded the recruitment message to their students, thereby increasing the participation.

The rationale for selecting JESSE is because it is one of the largest online listserv discussion groups used by people with backgrounds in the Library and Information Sciences. Therefore, it would be likely that JESSE subscribers had prior experience with tagging or had studied related subjects, e.g., metadata preparation. In short, we expected that participants drawn from the JESSE population would be able to provide quality feedback about our prototype UI.

5.2.1.2 Study Procedures and Tasks

The experimental progression of the usability study is depicted in Figure 47. After participants submitted their signed informed consent forms to the researcher as email attachments, they completed an online demographic questionnaire, which took 5-8 minutes. Once they indicated completion of the questionnaire, they were sent a link to the task questionnaire, which consisted of the following items: a download link for a tutorial of the UI, an exercise task, and the three main tasks, each of which involved reading a news article, entering several topic tags, and answering questions regarding their actions. The expected time to complete the tasks and to fill in their responses to the questionnaire was typically less than 40 minutes. As a final step, the participants completed an exit questionnaire (5-10 minutes).

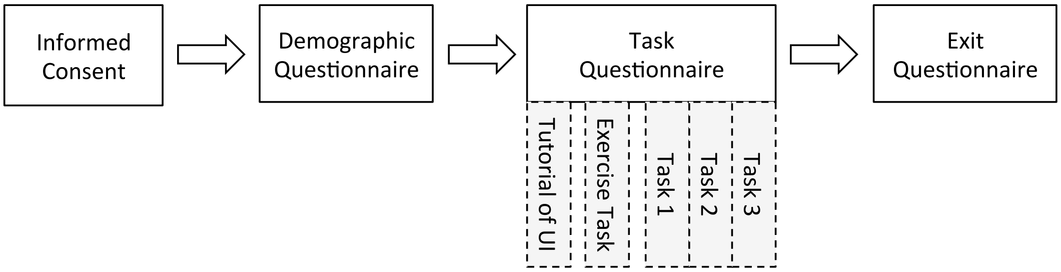


Figure 47. The full progression of the usability study.

The task questionnaire involved the following protocol. After entering their participant ID (provided by the researcher), the participants read a short tutorial about the prototype GUI system

[11] http://web.utk.edu/~gwhitney/jesse.html

after downloading a PDF document. Then, they worked on an optional exercise task by following the step-by-step instructions. The participants then assigned topic tags for the news articles in the main tasks, which are numbered from 1 to 3. There were no specific instructions on how to read the news articles; in other words, they could choose to read the articles thoroughly or just skim through them. In addition, participants were free to adjust the parameter settings according to their liking, thereby producing topic suggestions until they had a useful set of topics. They then were questioned about the reasons for their actions, decisions, and their strategies for identifying quality topic tags.

It is known that the topics tagged manually by one human indexer tend to be dissimilar from those tagged by another human indexer (Furnas, Landauer, Gomez, & Dumais, 1987). Therefore, we grouped the participants into five teams of two persons each, and had each team perform the tagging for the same set of three documents. Each of the three documents was randomly selected from one of the three following document collections: NYT_1000, CTR_30, or VARIOUS_30. The details for the teams, team members, and the document numbers from the collections are shown in Table 41. To receive adequate feedback from users, a total of five task questionnaires were developed for the five teams, and each questionnaire cited a different set of three articles to be used for the task.

Participants were encouraged to ask questions about the prototype UI or the tasks. For example, one participant had a technical problem with running the prototype UI on a browser. Another participant asked for clarification about a certain task. In such cases, the researchers replied to the participants with care, in order not to have the outputs/answers be biased by the researcher's response.

Table 41. The teams, team members, and the documents in the task surveys.

Team	Participant ID	NYT_1000 Doc ID	CTR_30 Doc ID	VARIOUS_30 Doc ID
Team A	P1, P3	660	1	1
Team B	P7, P8	1584	16	18
Team C	P11, P13	4981	19	21
Team D	P14, P20	9583	26	27
Team E	P16, P17	13786	28	28

5.2.1.3 Data Collection

All the responses from participants were collected using an online survey system, *Qualtrics*, which was provided free of charge (*Qualtrics*, 2002). As mentioned, three questionnaires were used in this study (Table 42):

Table 42. The questionnaires and their descriptions. Five different instances of the task questionnaires were used for the five corresponding teams.

Questionnaire	Description	Type
Demographic	Includes questions regarding general demographic information of the participants, and the details of their prior tagging experience.	Single type
Task	Used to collect the topic tags identified by the participants, as well as to obtain feedback regarding the UI usage and parameter settings.	5 types
Exit	Collects feedback from the users regarding the usability, e.g., usefulness and satisfaction, issues and problems, and feature suggestions for the UI.	Single type

Note: All the questionnaires used in this study are included in the Appendix.

5.2.2 Analyses and Findings

5.2.2.1 Participants Demographic

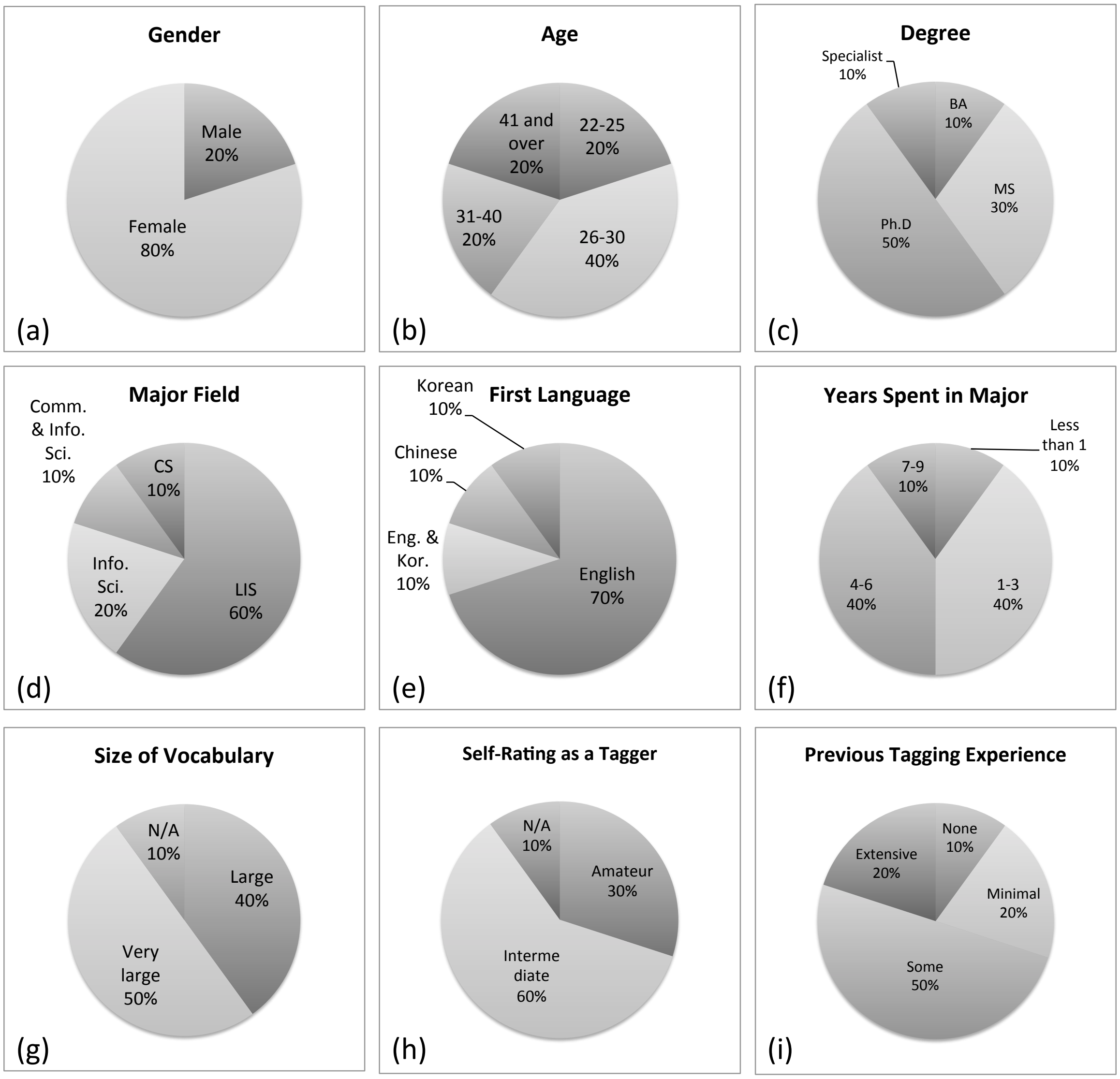


Figure 48. Participants' demographic information: (a) gender; (b) age; (c) degree; (d) major field of study; (e) first language; (f) years in their major field; (g) size of vocabulary (self-rating); (h) level of expertise as a tagger (self-rating); and (i) prior tagging experience.

Figure 48 presents a summary of the demographic information for the 10 participants. As shown below, 8 out of 10 people were female. Nearly half (4) ranged in age from 26-30, while the remaining six were spread evenly over the other age ranges, 22-25, 31-40, and 41+. Half of the participants were in Ph.D. programs, while three participants were in MS programs. There was one person who answered as "specialist" for her degree program.

Since the participants were recruited from JESSE, a popular listserv for people with an interest in library science, information science, and related fields, nine out of ten participants had a background in these areas; one person was studying computer science. English was the first language for eight people, including the one who answered her first language as "English and Korean." The majority of people had experience in their fields ranging from 1-3 years or 4-6 years.

Except for one person who did not respond to this particular indicator, participants rated themselves as having an either "large" or "very large" vocabulary size. More than half of the participants rated that they were an "intermediate" topic tagger, and about one third rated themselves as an "amateur" tagger. Half of the participants had "some" prior tagging experience, whereas two had "extensive" experience, and the other two had "minimal" experience (only one had no prior tagging experience). The list below encapsulates the participants' tagging experiences:

- Assigned tags and keywords to blogs, and social media posts
- Tagged images in Tumblr[12]
- Tagged and classified resources for a professional association's knowledge center
- Tagged books using a system provided by the university library
- Tagged movies and books to improve findability

Some participants mentioned other specific tagging domains as follows:

- Blogs: news, personal
- Social media
- Medicine
- Health science
- Photography
- Entertainment
- Books, textbooks
- Scholarly articles
- Corporate real estate

Based on the demographic information detailed above in Figure 48, and especially for (d) major field (90% in LIS and other information-related fields), (e) first language (80% English), (g) vocabulary size (90% either large or very large), and (i) prior tagging experience (90% had some level of experience)—the participants could be considered to be "quality topic taggers" appropriate for this usability study.

5.2.2.2 Usefulness and Satisfaction of the UI

To measure the usability of the Xpantrac GUI, which was used as an interactive semi-automatic topic suggestion tool, a usability questionnaire was developed with six questions regarding the usability and user satisfaction for the system (Davis, 1989; Lewis, 1995; Lund, 2001):

[12] https://www.tumblr.com/

(1) The system was easy to use

(2) Using the system was an enjoyable experience

(3) The system was useful in finding the indicative topics

(4) The system decreased the time to complete the tasks

(5) The system distracted me from focusing on the tasks

(6) Overall, I am satisfied with this system

Those questions were answered using a five-point Likert scale: -2: strongly disagree, -1: disagree, 0: neutral, 1: agree, and 2: strongly agree (Table 43). In summary, the participants agreed with or strongly agreed with questions (1), (2), (3), and importantly (6). Seven participants either strongly agreed or agreed, and one participant answered neutral, for (4). The participants also almost disagreed (-0.9) that the system was distracting from focusing on their tasks (5). Overall, the UI had consistently positive responses.

Regarding (4), however, there were two participants (P14 and P16) who strongly disagreed or disagreed, respectively. To understand the potential reasons for this outcome, their comments collected during the tagging tasks were examined. Results revealed that P14 went through almost every sentence for two out of three tasks for three reasons. P14 thought that: (1) the tags suggested by the UI did not provide enough context to understand the article, (2) setting a higher number for "the number of API return" parameter resulted in non-relevant tag suggestions in some cases, and (3) the article had multiple medical terms and jargon with which P14 was unfamiliar. It also should be noted that P14 responded as "neutral" for overall satisfaction with the UI (see last column in Table 43). Thus, it appears that P14 was not satisfied with the UI, conducted the tasks without consulting the suggested topics, and generally felt that the UI did not help decrease the time to complete the required tasks.

Although P16 also disagreed with (4), his/her reasons were different from those of P14. Despite the fact that P16 went through almost every sentence for two out of three articles (similar to P14), P16 had time to read the documents, and considered them to be interesting. This individual just wanted to be able to provide the best tags. For the other article, P16 skimmed through it due to less personal interest in the topic. In summary, P16 appeared to rely more on the manual way of tagging than referencing the topic suggestions from the UI for the reasons noted above. Thus, P16 did not need the UI and disagreed that the UI actually helped decrease task completion time.

Table 43. Participant responses for usability in Likert scale (-2: strongly disagree, -1: disagree, 0: neutral, 1: agree, and 2: strongly agree).

Participant ID	(1) Easy to use	(2) Enjoyable experience	(3) Useful to find indicative	(4) System decreased time	(5) System distracting	(6) Overall satisfaction
P1	2	2	2	1	-1	2
P3	0	0	0	1	-1	0
P7	2	1	1	1	1	2
P8	1	2	1	0	-2	1
P11	1	1	2	2	-2	2
P13	2	1	1	1	-1	1
P14	0	1	0	-2	-1	0
P20	2	1	2	1	-1	1
P16	0	0	0	-1	0	0
P17	2	2	2	2	-1	2
Mean	1.20	1.10	1.10	0.60	-0.90	1.10

Study participants also were required to state their opinions regarding each pane of the system UI while they were conducting the three tasks. Since the 10 participants were asked to comment on each of the 3 panes (Collection, Document, and Topics), 30 responses were expected. In summary, there were 14 positive feedbacks, 10 critical feedbacks, 4 “on-the-fence” feedbacks that included both positive and negative comments, and 2 cases for which no answers were provided. In one instance the participant responded with “none,” which meant “no problem with the pane;” this was counted as positive feedback.

Below are selected positive feedback comments:

- Easy to use.
- There were no problems with the pane.
- The Collection Pane was easy to use.
- The Document Pane worked well. It was also in a convenient location next to the Topics Pane.
- The Topics Pane was great. It was easy to use, and made the tasks more fun.

The critical feedback comments were organized into three categories: feature suggestions, problems, and confusions (see Table 44). The participants provided the feature suggestions voluntarily since the questionnaire only asked if there were any “problems” with the UI. In particular, features such as highlighting/underlining words in the Document Pane, automatically including such highlighted/underlined topics in the Topics Pane, changing the font size in the Document Pane, and sorting the topics in alphabetical order, might have affected the perceived usability of and satisfaction with the UI. Since the suggested features are not difficult to implement in the UI using JavaScript and JQuery, they will be incorporated in our next version of the UI.

Table 44. Descriptive critical feedback for each pane in the UI.

	Collection Pane	Document Pane	Topics Pane
Feature Suggestions	• Documents searchable by number or title • Show document publication dates • Documents sorted by dates	• Highlighting or underlining words in the document, which are included in the suggested topic list • Changing font size	• Mouse-over description presentation for parameter names on the UI • Sort topics by alphabetical order
Problems	• Font was too small • Unable to easily switch to another collection	• Cannot scroll-down to the last lines of the document • Document headings ran into their content without spaces • Bigger font needed	• Adjusting the parameter, “number of API results,” did not produce significantly different set of topics • Suggested topics are only in single words
Confusions	• Not sure why 30 docs are listed when I have to select only one	N/A	• Unable to select topic types in “verbs” or “nouns & verbs”

The participants also reported several problems they encountered while performing the tasks. One of the complaints was that participants could not scroll-down to the last lines of a document presented in the Document Pane. However, the size of the box in which the document content was displayed was statically set. Therefore, the last lines of the document were not viewable if the browser size was adjusted by the participants to be smaller than the size of the Document Pane. Designing the Document Pane to be automatically resized as the browser size changed might have prevented this issue.

The font size of the documents was an issue, especially because one of the participants had a slight vision problem. Another complaint was that the suggested topics were only in single words; however, this was an intentional choice in this study. Our next version of the UI might include an option to choose any combination from only single words, bigrams (i.e., two word phrases), and trigrams (i.e., three word phrases). One person mentioned that adjusting the parameters did not produce a significantly different set of topics. Even though parameters of the UI were changed, it is possible that the produced topics might not have varied significantly. The prototype system was designed to produce the main topics of a document in a robust way. In cases when the parameter setting for the “number of topics” was low—for example, “10,” when the system consistently produced a dozen main topics—then the users might feel that the topics extracted did not vary much for different parameter settings.

Several participants noted some confusing aspects of the UI. For instance, one participant was not sure why 30 documents were presented when only 3 of them were used in her tasks. The reason was that the other documents (i.e., those NOT part of her tasks) were included in other participants' tasks. Another reason was to show how the Collection Pane would perform in a real setting where multiple documents should be listed in the pane displaying their titles. An added confusion was that a couple of participants attempted to select the topic types as "verbs" or "nouns & verbs," which were disabled (i.e., made not-clickable and grayed-out) for this study. This confusion might have been prevented by giving participants clearer instructions for interacting with the Topics Pane, or by removing the disabled options from the UI prior to the study.

5.2.2.3 Quality of Indexed Topic Tags

Following the same methodology of evaluating topics in Chapter 4, the quality of topic tags was analyzed in two ways. The first was to measure precision, recall, and F1 values for the topics assigned to five sample documents from NYT_1000. The second was to compute the Rolling's inter-indexer consistency (IIC) (1981) regarding the topics assigned to the sample documents from CTR_30 and VARIOUS_30.

Precision, Recall, and F1 of Topics from NYT_1000

To reduce the subjectivity that might occur during topic-tagging tasks, each of five sample documents from the NYT_1000 was indexed by two participants. The precision, recall, and F1 values were computed using the existing gold standard topics as shown in Table 45. Compared to the average F1 values of the topics for the NYT_1000 generated by Xpantrac (command mode, automatic), OpenCalais, or TF*IDF from Chapter 4, which were 0.240, 0.226, and 0.173, respectively, the mean F1 value of 0.374 was much higher.

Table 45. The precision, recall, and F1 values for topics indexed by each participant for 5 documents from the NYT_1000.

Document ID	Participant ID	P	R	F1
660	P1	0.176	0.300	0.222
660	P3	0.143	0.100	0.118

1584	P7	0.500	0.294	0.370
1584	P8	0.438	0.412	0.424
4981	P11	0.800	0.571	0.667
4981	P13	0.222	0.286	0.250
9583	P14	0.500	0.333	0.400
9583	P20	0.714	0.333	0.455
13786	P16	0.385	0.556	0.455
13786	P17	0.429	0.333	0.375
Mean		0.431	0.352	0.374

In addition to the topic quality analysis for individual participants (Table 45), we also performed analyses both on the intersection and union sets of topics from the participant pairs (i.e., two participants working on the same document). The analysis results for both topic sets are displayed in Table 46 and Table 47, respectively.

Table 46. The precision, recall, and F1 for the intersection of topic sets indexed by the participant pairs for documents from the NYT_1000.

Document ID	Participant pair	P	R	F1
660	(P1, P3)	0.143	0.100	0.118
1584	(P7, P8)	0.500	0.176	0.261
4981	(P11, P13)	1.000	0.286	0.444
9583	(P14, P20)	0.833	0.333	0.476
13786	(P16, P17)	0.500	0.222	0.308
Mean		0.595	0.224	0.321

The intersection topic sets showed that their mean precision (0.595) was much higher than their mean recall (0.224) value (see the bottom row in Table 46). In the case of the union topic sets in Table 47, the mean recall (0.468) was higher than the mean precision (0.367). The union of topic sets resulted in the highest mean F1 value (0.4) amongst the individual, intersection, and union topic sets.

Table 47. The precision, recall, and F1 for the union of topic sets indexed by the participant pairs for documents from the NYT_1000.

Document ID	Participant pair	P	R	F1
660	(P1, P3)	0.176	0.300	0.222
1584	(P7, P8)	0.471	0.471	0.471
4981	(P11, P13)	0.333	0.571	0.421
9583	(P14, P20)	0.455	0.333	0.385
13786	(P16, P17)	0.400	0.667	0.500
Mean		0.367	0.468	0.400

Inter-Indexer Consistency: CTR_30

Instead of computing the IIC between participant pairs in this study only, we calculated the two groups of IIC values from participants in this study, as well as from the indexers (i.e., H1-H3) in our previous study who worked on CTR_30 (see Chapter 4). The results are shown in the fourth and fifth columns of Table 48. The reason for this approach was to examine whether there was any significant difference in IIC if the topic indexers from the previous study (i.e., manual tagging) worked with the indexers in this study (i.e., manual + UI tagging), instead of with their own members.

Among the IIC values in Table 48, the lowest value was 0.069 between H3 and P8 for Doc #16. Taking a closer look, one can see that this result was due to a large number of topics tagged by P8 (total 24 topics) and a small number of common topic words between the topics tagged by P8 and H3 (only 1 word). Since this finding was not in error, it was not discarded from the data table.

Table 48. The means of IIC values for two groups: group one (4th column) is when one of "manual only" indexers (i.e., one of H1-H3) works with the indexers with UI; group two (5th column) is when one of "manual only" indexers works with the other "manual only" indexers.

Doc #1	P1	P3	Mean (P1, P3)	Mean (H1-3)	H1	H2	H3
H1	0.500	0.400	0.450	0.561		0.533	0.588
H2	0.381	0.500	0.440	0.552	0.533		0.571
H3	0.522	0.286	0.404	0.580	0.588	0.571	
Doc #16	P7	P8	Mean (P7, P8)	Mean (H1-3)	H1	H2	H3
H1	0.364	0.333	0.348	0.450		0.500	0.400
H2	0.625	0.200	0.413	0.472	0.500		0.444
H3	0.267	0.069	0.168	0.422	0.400	0.444	
Doc #19	P11	P13	Mean (P11, P13)	Mean (H1-3)	H1	H2	H3
H1	0.211	0.438	0.324	0.368		0.320	0.417
H2	0.625	0.345	0.485	0.493	0.320		0.667
H3	0.667	0.429	0.548	0.542	0.417	0.667	
Doc #26	P14	P20	Mean (P14, P20)	Mean (H1-3)	H1	H2	H3
H1	0.444	0.211	0.327	0.450		0.400	0.500
H2	0.533	0.500	0.517	0.508	0.400		0.615
H3	0.625	0.588	0.607	0.558	0.500	0.615	
Doc #28	P16	P17	Mean (P16, P17)	Mean (H1-3)	H1	H2	H3
H1	0.444	0.353	0.399	0.330		0.160	0.500
H2	0.240	0.583	0.412	0.265	0.160		0.370
H3	0.400	0.526	0.463	0.435	0.500	0.370	
Mean	0.456	0.384	0.420	0.466	0.432	0.458	0.507
Std. Dev.	0.147	0.149	0.105	0.092	0.124	0.152	0.100

Once the two groups of IIC means were developed, a t-test was used to determine if they were significantly different. Based on the Quantile-Quantile (QQ) plot[13], each group of IIC values was found to have a normal distribution. Therefore, the following hypotheses were developed:

$$H_0 : \mu_x = \mu_y$$
$$H_1 : \mu_x \neq \mu_y$$

where H_0 is the null hypothesis, H_1 is an alternate hypothesis, μ_x is the mean of the values in the 4th column, and μ_y is the mean of the values in the 5th column of Table 48.

Table 49 shows the result of the t-test using JMP8 statistics software[14]. The t value was 2.04841 for alpha of 0.05, and the absolute value of t Ratio was 1.26235. Since this t Ratio was less than the t value, and thus not in the rejection range, we could not reject H_0. Therefore, it turned out that the mean IIC values of H1-H3 between working with the new participants (P1, P3,…) and working with their own members were not significantly different.

Table 49. The result of t-test comparing two groups of mean values in Table 48 (t = 2.04841 for alpha = 0.05).

t Ratio	-1.26235
DF	28
Prob > |t|	0.2172

Inter-Indexer Consistency: VARIOUS_30

Once again, two groups of IIC values were developed from the participants in this study, as well as from the indexers (i.e., H11-H15) in our previous study who worked on VARIOUS_30 (see Chapter 4). The results are displayed in the fourth and fifth columns of Table 50. The reason for this approach was to examine whether there was any significant difference in IIC values if the topic indexers from the previous study (H11-H15) worked with the indexers in this study (P1, P3,…), instead of with the rest of their own members.

It should be noted that during the study, P7 accidentally provided topic tags for another document instead of using Doc #18. Thus, the IIC values between P7 and H11-H15 were discarded from Table 50. The discarded data was marked with an asterisk (*).

[13] QQ-plot at http://www.stat.tamu.edu/~suhasini/teaching651/JMP_commands.pdf

[14] JMP8 tutorial at http://www.jmp.com/support/downloads/pdf/jmp8/jmp_user_guide.pdf

Table 50. The means of IIC values for two groups: group one (4th column) is when one of "manual only" indexers (i.e., H11-H15) works with "manual + UI" indexers; group two is when one of "manual only" indexers works with the other "manual only" indexers.

Doc #1	P1	P3	Mean(P1, P3)	Mean(H11-15)	H11	H12	H13	H14	H15
H11	0.143	0.333	0.238	0.376		0.429	0.375	0.500	0.200
H12	0.467	0.286	0.376	0.514	0.429		0.556	0.571	0.500
H13	0.375	0.375	0.375	0.393	0.375	0.556		0.500	0.143
H14	0.214	0.333	0.274	0.443	0.500	0.571	0.500		0.200
H15	0.231	0.200	0.215	0.261	0.200	0.500	0.143	0.200	
Doc #18	P7	P8	Mean(P7, P8)	Mean(H11-15)	H11	H12	H13	H14	H15
H11	*	0.343	0.343	0.369		0.444	0.444	0.471	0.118
H12	*	0.424	0.424	0.378	0.444		0.400	0.533	0.133
H13	*	0.333	0.333	0.357	0.444	0.400		0.417	0.167
H14	*	0.375	0.375	0.355	0.471	0.533	0.417		0.000
H15	*	0.000	0.000	0.104	0.118	0.133	0.167	0.000	
Doc #21	P11	P13	Mean(P11, P13)	Mean(H11-15)	H11	H12	H13	H14	H15
H11	0.200	0.381	0.290	0.285		0.222	0.216	0.526	0.174
H12	0.583	0.480	0.532	0.416	0.222		0.390	0.609	0.444
H13	0.372	0.545	0.459	0.334	0.216	0.390		0.381	0.348
H14	0.640	0.615	0.628	0.468	0.526	0.609	0.381		0.357
H15	0.276	0.467	0.371	0.331	0.174	0.444	0.348	0.357	
Doc #27	P14	P20	Mean(P14, P20)	Mean(H11-15)	H11	H12	H13	H14	H15
H11	0.444	0.516	0.480	0.473		0.471	0.389	0.533	0.500
H12	0.353	0.267	0.310	0.375	0.471		0.286	0.429	0.316
H13	0.389	0.490	0.439	0.333	0.389	0.286		0.182	0.474
H14	0.400	0.357	0.379	0.374	0.533	0.429	0.182		0.353
H15	0.500	0.485	0.492	0.411	0.500	0.316	0.474	0.353	
Doc #28	P16	P17	Mean(P16, P17)	Mean(H11-15)	H11	H12	H13	H14	H15
H11	0.400	0.414	0.407	0.459		0.737	0.343	0.357	0.400
H12	0.588	0.538	0.563	0.527	0.737		0.438	0.480	0.455
H13	0.545	0.571	0.558	0.474	0.343	0.438		0.537	0.579
H14	0.538	0.514	0.526	0.440	0.357	0.480	0.537		0.387
H15	0.696	0.625	0.660	0.455	0.400	0.455	0.579	0.387	
Mean	0.418	0.411	0.402	0.388	0.392	0.442	0.378	0.416	0.312
Std.Dev.	0.154	0.140	0.143	0.090	0.149	0.135	0.127	0.149	0.160

Note: Discarded data (due to a human error) is marked with *.

Once the two groups of IIC means were developed, the t-test was applied. The Quantile-Quantile (QQ) plot also showed normal distributions for both groups of data. Therefore, the following hypotheses were developed:

$$H_0 : \mu_x = \mu_y$$

$$H_1 : \mu_x \neq \mu_y$$

where H_0 is the null hypothesis, H_1 is an alternate hypothesis, μ_x is the mean of the values in the 4th column, and μ_y is the mean of the values in the 5th column of Table 50.

Table 51 presents the result of the t-test. The t value was 2.01063 for alpha of 0.05, and the absolute value of the t Ratio was 0.407157. Since this t Ratio was less than the t value, and not in the rejection range, we could not reject H_0. Therefore, the means of IIC values, μ_x and μ_y, were found to be not significantly different.

Table 51. The result of t-test comparing the two groups of mean values in Table 50 (t = 2.01063 for alpha = 0.05).

t Ratio	0.407157
DF	48
Prob > |t|	0.6857

5.2.2.4 Participant Interaction with the UI

Participants were able to interact with the UI in three ways: by changing the two parameters, by extracting topic suggestions, and by comparing them with their own set of topics. Table 52 presents details corresponding to their change in the parameters, and their reasons for doing so. The setting change for the "number of API returns" parameter is noted in the second column of the table, while the change for the "number of topics" parameter is noted in the third column. These two columns were developed by aggregating the responses for Question 1-3, 2-3, and 3-3 in the main task questionnaires. The fourth column, "Reason", was developed based on the responses for Questions 1-4, 2-4, and 3-4 in the main task questionnaires.

Findings showed that participants changed the settings (i.e., they increased or decreased parameter values) approximately 70% of the time, which means that the participants interacted with the UI trying to produce quality topic suggestions. Three participants (P1, P3, P8) increased the number of topics either to experiment with or to have more topic suggestions for later comparison to their own manual topics. In two cases (P14, P20) participants decreased the number of topics to reduce non-relevant topic suggestions. Additionally, P17 reported having carefully changed the outputs from the UI to the new parameter settings. Comments from P14 and P17 are provided below since their feedback contrasted with each other.

P14: Critical feedback

- *I expected that choosing the higher number of API results and topics would result in more accurate tag suggestions, but it actually ended up producing a lot of redundant tags.*
- *From Task 1, I learned that choosing the higher number of topics is not necessarily helpful and may result in a number of redundant tags.*
- *Similar to Task 2. I was trying to avoid getting a number of redundant suggestions.*

P17: Positive feedback

- *I wanted more choices. The higher API setting reduced redundant words. For example, setting it to 15 produced kidnappers and kidnapping. Setting API to 25 produced only kidnapping, which is the one I chose. I originally set the topic to 15, which gave me a good selection, but I wanted a few more. Setting the topic to 25 gave me more options that were very useful.*
- *I used a lower setting for API this time because I thought there would be a less chance of similar words. I used a setting 20 because I wanted more suggestions. It was great.*
- *I selected a higher API and number of topics because of the article's many different key terms. Xpantrac suggested every topic that I was thinking.*

Table 52. The parameter settings of the UI for the three tasks and reasons for change.

Participant ID	Change for No. API Returns	Change for No. Topics	Reason
1	Increase	Increase	To get the highest number of topics to compare with his/her own topics
3	Stay high	Increase	To have as much information as possible
7	Default	Default	Not specified
8	Increase	Increase	To experiment with the parameters
11	Default	Default	The same as those of the practice exercise
13	Default	Default	The same as those of the practice exercise
14	Stay high	Decrease	To reduce redundant topic suggestions
20	Default	Decrease	To experiment, to decrease unimportant tags
16	Default	Slight increase	To try a few more topics, not much attention
17	Slight decrease	Slight increase	To carefully adjust the parameters for optimal topic suggestions

5.2.2.5 Strategies for Identifying Topics

The participants described the process associated with their strategy for finding a set of quality topics (Table 53). Although the specifics varied slightly between participants, the majority of people read the article, came up with candidate topics by identifying frequent words, and then reviewed the topic suggestions provided by the UI in order to identify more topics that they might have missed. In the case of participants 13, 16, and 20, they also considered the effectiveness of their topics to be search query terms to distinguish documents.

In particular, five participants (P1, P3, P11, P14, and P17) directly mentioned their use of the UI. Two participants (P13 and P20) considered whether their identified topics could be used when searching for information. As shown in P17's positive feedback details presented previously, P17 controlled the parameters of the UI based on the document characteristics. Depending on the characteristics of the terms and topics in the document, participants adjusted both parameters of the UI, the number of API returns, and the number of topics.

Table 53. The participants' strategies to identify quality topics.

Participant ID	Strategy for Quality Topics
1	• Record tags while reading the article • Write down more tags after reading the article • Compare manual tags with the ones suggested by the UI o Include any missing tags from the topic suggestions o Exclude any tags that should be removed
3	• Read the texts thoroughly, but not repeatedly in order not to waste time • Generate many candidate terms • Review the suggested topics from the UI o Combine some of the suggested topics o Include some of the topics
7	• Good comprehension, vocabulary, and indexing skills are required depending on different levels of readings.
8	• Select concepts from the article that seemed most relevant and significant
11	• Look at the major overall subject of the article • Skim the paragraphs for paragraph topics • Crosscheck with the tags suggested by the UI.
13	• Look for words that appeared often or words that people would use when searching for information pertaining to each of the articles.
14	• Skim the article first to get an overall idea of content • Use the suggested topics by the UI to see the key points
20	• Use words and short phrases in the text for clarity • Think about how I might use my terms in a search
16	• Look for words that are used repeatedly, excluding stopwords • Make sure these words are unique to the text, and will distinguish them from other documents in the corpus

17	• If a term was used many times in different tenses such as Task 1, I would select a higher API • If there appeared to be many integral topics such as in Task 3, I selected a higher number of topics

5.3 Discussion

As shown in Table 53, the participants listed a number of strategies for identifying a set of quality topics in an interactive semi-automatic approach, where the UI suggested topic tags. For example, they used the suggested topics from the UI to crosscheck with their own topics, or to see the key points of the article. Except for one participant who was very interested in the content of the task articles and decided to read them thoroughly, all participants interacted with the UI in varying degrees (Table 52). Most importantly, the participants, nearly all of whom had backgrounds in LIS or information-related fields and were familiar with tagging tasks, agreed that the prototype UI was useful for tagging and rated it as being satisfactory overall (Table 43).

For the usability study, each participant conducted tasks involving topic identification for three news articles, accompanied by the topic suggestions from the UI. However, reading just three newspaper articles (e.g., from *The New York Times*) is not mentally taxing and would not take too long. Additionally, it also could be pleasurable if the articles were personally interesting to the reader. However, if there were 50, 100, or more articles that required tag identification, conducting this task manually would soon become almost impossible. Getting support from the UI would become a necessity. (In fact, one participant from a prior manual tagging study mentioned that it was extremely tiring to identify tags for more than 30 news webpages.) Therefore, using Xpantrac with the UI could be a valuable tool for alleviating a human indexer's mental workload for topic tagging tasks. Moreover, in cases when a significant number of articles had to be processed, Xpantrac used in command line mode could batch process these documents. A user could try out different parameter settings quite easily by using Xpantrac's user interface for several articles until a set of quality topics was extracted. The identified optimal parameters could be used in the batch mode tool for the highest quality topics processed via Xpantrac's automated function.

5.4 Summary

This chapter described the design, development, and the usability study of the graphical user interface (GUI) of Xpantrac, which is capable of suggesting a list of topics identified from an input text. The first part of this chapter detailed the components of the UI, which include (from left to right) the Collection Pane, the Document Pane, and the Topics Pane. Following the design of the UI, technical details also were provided, such as descriptions of the terms used, preparation for Apache SOLR to be used in place of a commercial search engine API, and the REST API queries and returned data structure, which showed communication details between the UI and SOLR.

The bulk of this chapter reviewed the results of the usability study using the prototype UI. As described, prospective study participants were recruited via the JESSE listserv with the expectation that they would have backgrounds in information-related fields (e.g., Library and Information Science), as well as be familiar with tagging tasks. Our recruiting approach was successful in that 90% of the participants had information-related backgrounds, and with one exception, all had prior tagging experience.

This usability study was structured around three questions. The first question (*How do people conduct tagging tasks supported by our prototype UI?*) used participant feedback collected at both the micro-level and macro-level to determine results. In terms of micro-level feedback, participants provided comments about each task, the specific parameter settings they selected, and their rationale for those decisions. From their comments organized in Table 52, it was found that 70% of the participants were actively interacting with the UI, adjusting the parameters and extracting topic suggestions iteratively to have satisfactory topic suggestions. The macro-level feedback they provided related to their overall strategies for identifying quality topics in the given setting (manual and access to the UI). The details were explained in Sections 5.2.2.4 and 5.2.2.5, respectively.

The second question investigated how useful and satisfactory it was to use the prototype UI. Based on participant feedback analyses (Section 5.2.2.2), we confirmed that participants agreed or strongly agreed that the Xpantrac UI was easy to use, enjoyable for some, useful for identifying indicative topics, and satisfactory overall. They also disagreed with the statement that the UI distracted them from focusing on their main tasks.

The third question (*What was the quality of the topics indexed using an interactive semi-automatic way?*) used a comparative process to assess the usefulness of the topics identified by the participants. In other words, participants could come up with their own topics manually, as well as use the suggested topics from the UI. Results indicated that the quality of the two topic sets (interactive semi-automatic versus manual) was not significantly different.

By incorporating the features suggested for the UI, the Xpantrac UI could be improved further as an easy-to-use tool with a richer set of features. Such modifications could enhance its usefulness for anyone seeking to have appropriate topic suggestions during topic tagging tasks.

Chapter 6. Conclusions and Future Work

6.1 Conclusions

6.1.1 Summary of the Chapters and Tests for Hypotheses

This dissertation was driven by two primary research motivations, detailed in Chapter 1. The first was prompted by the need for a flexible and effective topic tagging approach that could be used to tag textual resources in a specific domain, in this instance, disaster events. The second rationale for this study was inspired by the realm of Cognitive Informatics about incorporating the Web (considered to be a universal knowledge source) into the process of topic extraction. The specific challenge was to devise a system requiring minimal setup that would locate suitably descriptive topics for individual documents in a large disaster event archive. To address this two-part challenge, several research questions and corresponding hypotheses were developed. Based on study findings, a number of research contributions have been made. These include a survey of topic identification studies, the development of a novel approach for topic identification, the development of a prototype system, the generation of data sets indexed with human-assigned topic tags, the development of scripts for data pre-processing and analyses of the results, the generation of resources for the usability study, as well as a set of findings based on a thorough evaluation of the efficacy and usability of the proposed system.

Chapter 2, the review of salient literature, included an assessment of topic identification studies in the field of Cognitive Informatics (Blei et al., 2003; Massey, 2011; Y. Wang, 2002; Y. Wang et al., 2011), which as noted above, provided an impetus for this doctoral research. In addition, the Vector Space Model was presented in detail since this model from the Information Retrieval field played a major role in the proposed system. Three external knowledge sources, Wikipedia, WordNet, and the Web, specific resource items of them, and approaches for utilizing them, were described as well. Details concerning corpus-based statistical approaches such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003; *OpenCalais*, 2013), and Latent Semantic Analysis (LSA) (Deerwester et al., 1990) were introduced. Chapter 2 concluded with a discussion of the evaluation metrics used for measuring the quality and consistency of identified topics.

The details of the system design, which included the components of the proposed system, workflow diagrams based on the expansion and extraction processes, and the identification of optimal parameters for the system, were all presented in Chapter 3. The proposed system, which was named *Xpantrac,* which represents a combination of the words "expand" and "extract," is comprised of an expansion component and an extraction component. The idea of expanding an input text into its relevant information originated from topic identification studies found in Cognitive Informatics. In the same way that one reads and connects textual information with existing relevant knowledge residing in long-term memory, Xpantrac expands texts by segmenting the words into groups; the relevant information from each group of words is then retrieved from the Web by querying a search engine. Important to note is that the amount of retrieved information is much greater than the corresponding input text. During the extraction part of Xpantrac, the Vector Space Model (Salton et al., 1975; Salton & Buckley, 1988) was applied to identify a list of significant terms as topic tags.

The details of the experimental procedures for testing Hypothesis 1, shown below, are described in Section 3.3.

Hypothesis 1:

> *Optimal parameters can be identified by examining the cosine similarity between the topics generated by our system utilizing different combinations of parameters and the topics assigned by multiple human indexers. The average Inter-Indexer Consistency (IIC) values between our system with optimal parameters and human indexers will be greater or equal to the average of those between human indexers only.*

Two text collections, CTR_30 and VARIOUS_30, were developed in order to identify the optimal parameters of the Xpantrac system. CTR_30 contained a total of 30 homogeneous documents having to do with a hurricane event. The other collection, VARIOUS_30, included 30 heterogeneous documents randomly selected from 9 different archives of natural disasters, man-made disasters, and health emergencies. These two collections were amassed so that a robust set of optimal parameters could be identified based on the varying characteristics of these document collections. In order to reduce the inevitable subjectivity associated with human topic tagging, several people were recruited to be topic indexers for the documents. The topic tags generated by the study participants then were compared to the topic tags generated by Xpantrac using varying combinations of parameters. This process resulted in two optimal parameter settings, M39 and M43, which performed well for both data sets.

The average IIC values between Xpantrac using the two optimal settings and multiple human indexers then were computed using both the CTR_30 and VARIOUS_30 data sets. The analysis confirmed that the average IIC values between Xpantrac with the optimal settings were not significantly different when compared to those that the human indexers generated. In fact, in some cases (e.g., Xpantrac with M39 and M43 for CTR_30) Xpantrac outperformed the human indexers based on average IIC values. Therefore, Hypothesis 1 was proven. As a comparison, we replaced Xpantrac with OpenCalais (*OpenCalais*, 2013; Sandhaus, 2008), which is a popular NLP API, and then computed the average IIC values between OpenCalais and human indexers. Results confirmed that the OpenCalais IIC values were much lower than those generated by the human indexers (see Section 3.3.4.1).

Chapter 4 described the evaluation studies of Xpantrac. Section 4.1 details the evaluation strategy, the experimental settings, and the baseline approach (i.e., TF*IDF) using OpenCalais as a comparative system. Utilizing both CTR_30 and VARIOUS_30, topic quality ratings according to both a five-point Likert scale (Likert, 1932; Shwartz, 2010) and the Inter-Rater Consistency (IRC) were computed to test Hypothesis 2, below:

Hypothesis 2:

> *The average relevance rating for the topics generated by our system using a five-point Likert scale will be significantly higher than the average relevance rating for those extracted using the baseline, TF * IDF, for both the CTR_30 and VARIOUS_30 data sets.*

Testing this hypothesis was rather straightforward. A total of four settings were developed. The first setting included the topics extracted by Xpantrac using the M39 parameter configuration. The second included the topics generated by Xpantrac in the M43 configuration. Additionally, two "derived" topic settings also were developed. One was an *intersection* set of topics between the Xpantrac M39 and M43 settings, and the other was the *union* set of topics. As a baseline setting, the topics generated by TF*IDF were used. Human raters assigned a rating value from -2 to 2, depending on the relevance of each topic tag to its corresponding document in both CTR_30 and VARIOUS_30. To reduce the subjectivity in the relevance ratings, three people assigned ratings to the topics.

As described in Section 4.2, analytical results showed that the topic quality from all the settings of Xpantrac confirmed a significant difference in quality in comparison to the baseline approach when ANOVA was applied ($p < 0.0001$) to the relevance ratings for topics in CTR_30. In the case of VARIOUS_30, the quality gap between the topics generated by Xpantrac and the baseline approach was closer in comparison to results associated with CTR_30. However, the quality rating of topics using the three Xpantrac settings (M39, M43, and M39_AND_M43) showed a significant difference when compared to the quality rating for the baseline. Moreover, the M39_OR_M43 setting did not show a significant ratings difference in comparison to the baseline, although the mean rating of M39_OR_M43 (0.740) was much higher than that of the baseline (0.539). Considering that M39_OR_M43 was an "additionally derived" topic setting from M39 and M43, and the analyses results for M39 and M43 proved that their relevance ratings were significantly higher than that of the baseline, it would be reasonable to say that Hypothesis 2 also was proven.

To test Hypothesis 3 (below), a larger text collection, namely the NYT_1000, was used to compare the ability of Xpantrac, the baseline, and OpenCalais to identify document topic tags. In addition, three different kinds of search engine APIs, Bing Azure Search, Yahoo! Web Search, and Yahoo! News Search, were incorporated in the expansion process of Xpantrac so that their performance could be compared.

Hypothesis 3:

> *The average F1 score for the topics generated by our system using 1000 randomly-selected New York Times articles will be significantly higher than or equal to the average F1 score for those extracted using either the baseline or OpenCalais natural language processing API.*

The articles in the New York Times corpus (Furnas et al., 1987; Sandhaus, 2008), of which NYT_1000 is a subset, were indexed with a list of metadata such as topics, locations, people, etc. These metadata were merged and used as the gold standard for the evaluation. The precision, recall, and F1 measures were used as evaluation metrics. Regarding the performance of the APIs, analytical results revealed that the performance of Xpantrac was the lowest when the Yahoo! News API was used in the expansion process. This outcome was principally associated with the incomplete expansion of the input text. The other two APIs performed equally well, with little differences noted between the two. This finding was not surprising because the Yahoo! Web Search API was developed based on the Bing Azure API technology (Iivonen, 1995; Shwartz, 2010).

In terms of precision outcomes, Xpantrac with M39, M43, and M39_AND_M43 performed significantly better in comparison to either OpenCalais or the baseline TF*IDF, excluding the case with the Yahoo! News API. In terms of recall scores, there were no significant differences among OpenCalais, M39 with Yahoo! Web Search, M43 with Yahoo! Web Search, and M39_OR_M43. Moreover Xpantrac with (a) M39 and M43 with Bing Azure Search, (b) M39_AND_M43 with the Yahoo! Web, and (c) M39_AND_M43 with Bing Azure Search showed significantly lower performance in comparison to OpenCalais. When precision and recall measures were combined as F_1 scores, only the baseline TF*IDF was statistically lower than either Xpantrac or OpenCalais (excluding Xpantrac using the Yahoo! News Search API). All the remaining settings—i.e., Xpantrac with M39, M43, M39_AND_M43, M39_OR_M43, and OpenCalais—did not show any statistical differences among their F_1 values. Therefore, results confirmed that Xpantrac performed as well as the widely-used OpenCalais API in terms of F_1, thereby proving Hypothesis 3.

Chapter 5 detailed the design, development, and usability of the user interface (UI) of Xpantrac, which consisted of three panes (Section 5.1). To extract topics using the UI, users were asked first to select a document corpus and load the documents in the Collection Pane (left side). The content of a selected document then is displayed in the Document Pane (middle). The Topics Pane (right side) then displays a set of topic suggestions after a user adjusts parameters by controlling the sliders or drop-down lists and pressing the button to extract the topics. Finally, the user is able to read the document presented in the Document Pane, and generate relevant topic tags, while at the same time consulting the topic suggestions in the Topics Pane.

A usability study was conducted to evaluate the usefulness and satisfaction regarding this UI. After performing the tasks using the interactive semi-automatic approach, users then provided feedback via a series of usability-related questions that assessed ease of use, how enjoyable the experience was, the ease of finding indicative topics, and overall satisfaction. The mean value of user responses according to the Likert scale was above 1.0, which denotes "agrees with" the questions. Therefore, users tended to "more than agree" with the usefulness of the UI; they also were satisfied with it. In terms of whether the UI system helped decrease task completion time, users responded with 0.6, which was between "neutral (0)" and "agree (1)." Regarding the question of whether the UI was distracting them from focusing on the main tasks, the mean value of responses was -0.9 (-1 was "disagree"). Thus, these findings proved the first part of Hypothesis 4.

Hypothesis 4:

> *The usability and user satisfaction for our prototype user interface will exceed a "1: Useful and Satisfactory" rating according to a five-point Likert scale from -2 to 2.*
> *In addition, the topic tags generated by human indexers, consulting the suggested tags from our prototype system, will have the higher average F1 score when compared with the F1 scores of those generated by either the baseline or OpenCalais.*

The precision, recall, and F1 values of the topic tags assigned by the five teams of human indexers also were computed using several documents in NYT_1000. The best mean F1 score (0.4) was computed when the union sets of the topic tags between the team members were used as the representative topics. Compared to the average F1 values of the topics for the NYT_1000 generated by Xpantrac command line mode, OpenCalais, and the baseline TF*IDF (see Chapter 4), which were 0.240, 0.226, and 0.173, respectively, the mean F1 value of 0.4 was much higher. Thus, this finding proved the second part of Hypothesis 4.

In summary, all of the hypotheses listed in Chapter 1 were proven. Correspondingly, this study's principal hypothesis, shown below, was proven as well:

Principal hypothesis:

> *The proposed approach, which uses a Web search engine, can produce tags of similar quality to human tags without depending on a large text corpus or a large training set. The tags will be comparable in quality to those obtained from a state-of-the-art topic identification application programming interface (API). It is also hypothesized that the front-end user interface for the system will be usable and satisfactory for suggesting topic tags during an interactive semi-automatic tagging task.*

6.1.2 Responding to the Research Questions

Table 54 reveals the relationships among the research questions, their corresponding hypotheses, and the related dissertation chapters. In particular, the design, development, and evaluation of the proposed Xpantrac system, which is capable of generating human-comparable topic tags, are explained in Chapters 3 and 4. The procedures for identifying a set of optimal parameters also are presented in Chapter 3. Using the two small text collections, CTR_30 and VARIOUS_30, each with only 30 documents indexed with topics by human indexers, the optimal parameters were identified using the cosine similarity values. Xpantrac with this optimal configuration showed human-comparable performance (Xpantrac even outperformed in some cases) when examined with the IIC metric. With the same parameter setting, Xpantrac also outperformed the OpenCalais API with a significantly higher precision using NYT_1000, although the F1 value did not show a significant difference. Thus, we would say that a topic identification system, which requires only minimal human intervention, could be developed. In addition, this system could generate human-comparable topics. Thus, Research Question 1 (Q1 and Q1.1) is answered.

Table 54. Relationships among the research questions, hypotheses, and chapters.

No.	Research Question	Hypothesis	Chapter
1	Q1. Can one design a semi-automatic approach (that requires only minimal human intervention) for generating human-comparable topic tags without relying on a prepared large text corpus? Q1.1. If such a tagging system were built, what would be its optimal parameters?	Hypothesis 1	Entire dissertation, Chapters 3 & 4
2	Q2. What is an effective evaluation approach for comparing the quality of topics generated by the prototype system with other methods? Q2.1. What is the measured quality of topics generated by our system or other methods?	Hypotheses 1 & 2	Chapters 3 & 4
3	Q3. Can one apply the optimal parameters, developed in response to Q1, to our tagging system in order to produce quality topics using a larger data set? Q3.1. Does this (proposed) approach outperform other approaches?	Hypothesis 3	Chapter 4
4	Q4. If we build a front-end user interface of our system, what is an effective approach and how useful will it be in supporting an interactive semi-automatic topic tagging tasks?	Hypothesis 4	Chapter 5

To reiterate, there were two metrics used for measuring the quality of topics. The first was the Inter-Indexer Consistency (IIC), and the second was the relevance rating in the Likert scale. Using IIC, we were able to evaluate the Xpantrac-generated topics in relation to the topics generated by other approaches (e.g., TF*IDF, OpenCalais, and human indexers). By obtaining a relevance rating using the Likert scale, which IIC could not effectively address, this study broadened its evaluation metrics. Chapters 3 and 4 provides details of evaluation results using both metrics. As detailed in those chapters, Research Question 2 (Q2 and Q2.1) was answered as well.

In order to answer Research Question 3, a larger text collection, the NYT_1000, was developed. A detailed analysis of the evaluation using this collection is presented in Chapter 4, which confirms that Xpantrac with optimal parameters performed analogously (in F1 measure) to that of the widely-used natural language processing API, OpenCalais. It did not, however, outperform it. In contrast, Xpantrac did outperform the baseline approach, TF*IDF. Research Question 3 (Q3 and Q3.1) is addressed in Chapter 4.

Chapter 5 discusses Research Question 4 (Q4). Although the usability study had limitations and constraints, user feedback indicated a generally positive experience with the Xpantrac UI. Nonetheless, several critical comments were noted, which will be addressed in our next update of the Xpantrac UI.

In summary, all the research questions developed for this study, and detailed in Chapter 1, were addressed by the proven hypotheses. However, additional studies should be designed to advance the proposed approach and prototypes based on the lessons and inspirations gained from this dissertation research.

6.2 Limitations of the Study

Although this study was designed to produce optimal results, it was inevitably constrained by issues of time and available human resources (e.g., human topic indexers). It also was limited by the use of a gold standard in the evaluation of the Xpantrac system. The limitations of this study can be grouped into three categories:

1. The generalizability of Xpantrac, especially in terms of an optimal set of parameters. Using a greater number of document collections with increased content variability would likely improve the generalizability of the system.

2. Evaluation of the topics. Topics as bigrams or trigrams were tokenized into unigrams, since only the unigram topics with exact term matching (e.g., IIC) were used to compute evaluation metric values throughout the study. Using bigram or trigram topics, in addition to considering synonyms for term matching, might enhance the performance of Xpantrac in more realistic settings.

3. The weaknesses in the gold standard due to the inherent subjective nature of tagging, and their impact. The variability in quality of the gold standard also might have limited the study.

The first point, the generalizability of Xpantrac, merits further investigation. One of the design goals of Xpantrac was ensuring minimal human intervention in the process of identifying topics. Typically, human intervention in a system of this kind involves identifying system parameters. In order to identify the optimal parameters for Xpantrac, two document collections with different characteristics (i.e., homogeneous and heterogeneous) were developed by multiple human topic indexers. An assumption associated with this study was that any parameters that work well for both collections might work well for other text collections—providing that the context of the documents in the collection was in the disaster domain. Building and applying more document collections about various types of disasters or significant community events might be effective in identifying the parameters that are robust and work well for documents with diverse content.

Evaluating topics generated by algorithms or even by human indexers is well known to be difficult due to its subjective nature (Furnas et al., 1987; S. Yang et al., 2011). This issue also limited the current study. As detailed herein, all the topics in multi-word phrases from human indexers, OpenCalais, and the gold standard (e.g., New York Times corpus), were tokenized into unigrams, after which duplicate terms and stopwords (e.g., of, for, in, the, etc. appearing in the phrases) were removed before the evaluation metrics were applied.

Computing the evaluation metric values for the quality of topics based on exact-term matching represents another potential limitation of this study. In other words, the consistency of topics only at the terminological level, not at the conceptual level, was considered (Iivonen, 1995). It is possible that two alphabetically different terms have close conceptual similarity with each other or might even be synonyms. However, the exact-term matching approach does not take this possibility into consideration. Since it is difficult to accurately identify the corresponding synonym or concept of a term due to the difficulty in identifying the context, our evaluation relied on exact matching. In an effort to reduce potential problems associated with the exact matching, the topic tags were lemmatized into their root forms before they were compared with each other. As long as the context of the term usage is detected and a relevant synonym can be identified, considering the synonyms and the conceptual similarity of terms might lead to a more realistic evaluation of topics.

The third point is the weaknesses in the gold standard due to the subjectivity in its development. In order to evaluate the performance of an information retrieval system, metrics such as precision, recall, and F1 are often computed based on the matching between the retrieved information and a gold standard, which we trust as the *true answer*. This approach is beneficial and, in some cases, necessary, considering that it allows the evaluation of a system at a large scale, as well as the comparison between the performance of systems, provided that there exists a large gold standard. However, there is no guarantee that this *true answer* is actually the truth. Due to the cost and complexity associated with multiple human taggers, each document in the gold standard is often tagged by a single human indexer, or by a software tool, of which humans examine the results. Therefore, there is a chance that one person's subjective tags are considered to be the *true answer*, and this becomes a weakness in the gold standard.

The variability in quality of the gold standard was observed during the study. The gold standard for the NYT_1000 data set was prepared by merging tags, which were assigned by single human taggers or by a software tool and human examiners to various metadata categories such as people, location, (high level) topics, and events. From a close look, it is found that the quality of such gold standard tags varies. As an example of low quality tags, there exist only two gold standard tags, "news" and "media", when in fact several more tags, such as "mortgage", "stock", "oil", "company", and "crisis", might have been made more sense as the topic tags for the NYT article with ID 51133 (Figure 35). Such high level topic tags, "news" and "media", in the example above might impact the computed performance of the system negatively. However, it would not be feasible to examine the validity of the gold standard in its entirety, thus we accepted the gold standard of NYT_1000 as the *true answer* in this study.

6.3 Future Work

There are four directions that this study could take in terms of future research. The first involves the expansion process of Xpantrac, in particular, replacing the commercial search engine API with a custom API. The second corresponds to the extraction process of Xpantrac UI. Specifically, two additional features could be implemented: topic extraction in bigrams and trigrams, as well as the topic forms in *verbs* or *nouns and verbs*. These features would provide more realistic and detailed topic identification. The third direction for future research is to implement the various additional features that were suggested by participants from the usability study of the user interface (UI). Finally, a logical fourth avenue for further studies involves an additional round of usability study for the UI once all the updates have been applied.

(1) Updating the Expansion Process

An expansion process for Xpantrac would be to develop a custom API to replace the present search engine API in order to retrieve a large amount of relevant information for an input text. The Crisis, Tragedy, and Recovery network project (S. Yang et al., 2011) has been archiving terabytes of webpages related to diverse disaster events in collaboration with the Internet Archive[15]. One potential approach for developing the custom API would be to index those archived webpages using Apache SOLR, as we demonstrated in our prototype UI.

There are, however, conditions to be satisfied to develop an effective custom API. The first condition is that the number of indexed webpages should be very large. Recall that the quality of the generated topics was low compared to the quality of topics using OpenCalais or even TF*IDF when the Yahoo! News Search API was used in the expansion process of Xpantrac. The reason for the low output was that not all the query units in the input text could be expanded with the Yahoo! News Search API. Another important point would be to have a balanced number of webpages in terms of their domains of coverage. For example, if the indexed documents included various earthquake events, the custom API would perform well for an expansion of news articles about earthquakes, thereby providing enough relevant information for the query units. In contrast, the custom API might not work well for another news article about floods or hurricanes due to a shortage of relevant information regarding the events. Therefore, two criteria should ideally be fulfilled in order to create an effective custom API: a greater amount of content and more balanced content of the indexed webpages.

An important detail for the development of a custom API is to separately index several sentences of the first paragraph of the webpages in order to emulate the behavior of the commercial search engine APIs. For example, search engine APIs such as Bing Azure Search API[16] and Yahoo! Web Search API[17] return webpage descriptions that include several sentences selected from the first paragraph of the matching webpage. Therefore, indexing several initial sentences, or even

[15] https://archive.org/

[16] http://datamarket.azure.com/dataset/bing/search

[17] http://developer.yahoo.com/boss/search/

the entire first paragraph, would make the custom API behave as if it were a commercial search API. Another approach for avoiding this "selective indexing" is to index the entire text of a webpage. Since the custom API would return the entire text of relevant webpages, a new set of optimal parameters should be identified—again based on the process explained in Chapter 3.

(2) Updating Extraction Process

Two additional features could be implemented in an expansion process. The first involves topic extraction in bigrams and trigrams (we will call them *n-grams*), which was suggested by the participants during the usability study. Considering that the topics extracted by human indexers, using OpenCalais, and even with the gold standard included some of these n-grams, it makes sense for Xpantrac to extract topics in n-grams as well. In an initial approach, for example, existing methods could be used to detect n-grams from the expanded information, after which significant n-grams might be extracted as topic tags.

The second feature to be implemented is the extraction of topics in *verb* form, as well as in *noun and verb* form. This feature is already included in the Xpantrac command-line mode; however, it is not part of the UI at present. The underlying premise for updating the extraction process is that topic tags suggested as verbs might better identify the main activities appearing in news articles. For example, news articles about the response activities of emergency organizations during a hurricane event might include topics in verbs such as "rescue," "deploy," "recover," "save," "supply," "damage," "submerge," and so on.

(3) Implementing Additional Features

The participants in the usability study suggested the incorporation of a number of system modifications. In particular, they expressed their preference to be able to control the font size, especially in the Document Pane where the content of the document is displayed. In fact, one of the participants emphasized that the ability to change the font size would be very helpful for people with vision problems (like her). Other participants advocated a means to organize suggested topics in the Topics Pane in alphabetical order. This modification would enable users to quickly compare the topics generated by the human topic indexer with the suggested topics by Xpantrac. Two additional changes were suggested: (1) improved filtering of noise topics, and (2) the ability to automatically resize the panes in the UI following any changes in the browser window size.

With respect to improved filtering and redundant noise reduction, lemmatization of the words would be essential. However, considering that there does not currently exist an appropriate lemmatizer built for JavaScript used in the UI, one idea is to lemmatize terms when indexing the news articles in the custom API. Then, those terms could be used for expansion and extraction of the topic tags. As indicated, some participants complained that they could not fully scroll-down the displayed article because of an issue with the pane size. This problem could be rectified by ensuring that the pane size changed automatically and proportionally to any size change in the browser window. Ideally, all of these changes should be incorporated on a step-by-step basis, and should be accompanied by usability studies for added features.

(4) Usability Studies with Updated UI

Regarding the updates from (1) to (3) mentioned above, the effectiveness of such added features should be examined through usability studies. For (1), a new set of optimal parameters should be identified. Then, the topic tags generated by Xpantrac using the custom API could be compared with the topics that were previously generated by Xpantrac, as well as by the human taggers using the two document collections, CTR_30 and VARIOUS_30. As a topic quality metric, Rolling's IIC (Massey, 2011; Rolling, 1981; Y. Wang, 2002; Y. Wang et al., 2011) can be utilized. In addition to topic quality analysis, detailed analyses of the workings of the custom API also should be conducted. For an expansion of news articles pertaining to various disaster events, the custom API should return the relevant information in a balanced fashion. In other words, if the beginning part of an article is expanded appropriately—but the middle or concluding sections are not similarly expanded—the retrieved information is not balanced, resulting in the extraction of topics with low quality.

For (2), the new feature of the n-gram topic suggestion implemented in the UI could be evaluated using metrics such as Modified R-precision (Kim, Baldwin, & Kan, 2010). The Modified R-precision metric is based on the partial matching between n-grams, and the component weight as well. Considering that the two data sets tagged by the human indexers (CTR_30 and VARIOUS_30), as well as the gold standard (the New York Times corpus) contained several n-grams as topics, it would be straightforward to compute the Modified R-precision values. An interesting follow-up study would be to evaluate the topics in *verb form*, or in *nouns and verbs* form, as suggested by the UI. The relevance of each such topic for describing activities in texts could be evaluated using the 5-point Likert scale as previously described (Section 4.2).

All of the feedback from the participants in these usability studies will be carefully considered prior to any further expansions to the Xpantrac UI.

6.4 Implications

Compared to other corpus-based statistical approaches such as LDA or LSA, and machine-learning-based approaches, Xpantrac does not require much human intervention, for example, the development of training sets or preparation of a large corpus. In this way, it will reduce the time and effort in setting up the tool because the users only have to specify the locations of input documents. Since this robust approach is not dependent on the training sets or corpus, document collections of any size could be processed. From a small collection of personal documents to a large news article corpus, Xpantrac can process them without much variation in its effectiveness. By utilizing the Xpantrac UI tool, users could interactively tag the documents by referencing the topic suggestions provided by the UI tool.

As we have seen in Section 4.4 (i.e., Evaluation Using a Larger Data Set), Xpantrac will be scalable to larger collections. It can autonomously process individual documents without considering the properties of an entire collection when identifying the topic tags. Therefore, the quality of extracted topic tags will not be affected by the size of the collection. The main issue of processing larger collections with Xpantrac will be the processing time. Xpantrac identifies topic tags by retrieving related information from the Web by using search engine APIs. Therefore, the network speed and response time of the used search API might affect the topic identification speed. One approach to overcome this would be to incorporate parallel processing using multiple machines. The documents in a collection might be divided and loaded in different machines. Then, each machine runs Xpantrac and extracts topic tags. Identified tags from each machine then are stored in a shared database as the result.

Section 4.2 (i.e., Topic Quality Rating) and Section 4.3 (i.e., Inter-Rater Consistency) described the evaluation of Xpantrac using a homogeneous data set (CTR_30) as well as a heterogeneous data set (VARIOUS_30), which showed a significantly better performance of Xpantrac over the baseline approach, TF*IDF. Another important point was that the relevance rating of the topics extracted by Xpantrac was fairly consistent in the range 0.5 – 1.0 for both data sets although the VARIOUS_30 data set contained documents from diverse events. From this result, it would make sense to presume that the text documents in various other domains can be successfully processed with Xpantrac.

Xpantrac was designed and optimized for extracting topic tags from webpages, and then it was found that the tags could be extracted in good quality from the New York Times articles as well. However, more studies would be needed to examine the performance of Xpantrac when it is applied to documents that are much longer than the NYT articles.

6.5 Summary

This chapter provides a summary of each chapter, while at the same time proving the hypotheses that were developed prior to the onset of this dissertation research. Responses to the proposed research questions are reviewed in connection with the proven hypotheses. This study features a number of inherent limitations, described in Section 6.2, as well as a discussion of four suggestions for future research in Section 6.3, including suggested modifications for the prototype system. It concludes with implications of the findings in Section 6.4.

Appendix A. Usability Study

The resources for the usability study of the Xpantrac UI are presented in this appendix. They include the study abstract, a participant recruitment letter, an informed consent form, an IRB approval letter, and all types of the questionnaires including the demographic, different types of the task, and the exit questionnaires.

A.1 Study Abstract

Title: Investigating the Usability of a Topic Tagging System

The investigators have been developing a web-based topic tagging system, which would help human users' tagging tasks by suggesting a list of significant terms given a text document. In this study, we will evaluate the usability of the prototype tagging system. For the usability evaluation, the study participants perform tagging tasks for the provided text documents, using our prototype system and their own intelligence. Documents are selected from the New York Times corpus as well as from other online articles.

The metrics from the Information Retrieval field, such as precision, recall, and F1, of the resulting topics are computed using the gold standard topic sets, which have been developed by aggregating manually indexed tags for the documents prior to this study. The study results will be valuable for understanding the usefulness of the tagging system, as well as improving the user interface.

A.2 Participant Recruitment Letter

Title: Investigating the Usability of a Topic Tagging System

The recruiting criteria includes that the participant is not a minor, and s/he is either a graduate or an undergrad student. Participants will identify a list of topic tags for online news and articles by using a prototype topic tagging system and their intelligence simultaneously. Although we do not measure the task completion time, we encourage the participants to complete the task as accurately as possible without wasting time.

The task procedures are listed below:

1. Print and sign the informed consent form emailed by the researcher. Then, send the signed consent form to the researcher by cell phone photographs/scanned PDF/images/mail (select one).
2. Complete the demographic questionnaire.
3. Complete the task questionnaire by identifying relevant topic tags for 3 news articles, based on reading the articles and selecting the suggested tags from the system.
4. As a last step, complete a short exit questionnaire.

You will assign tags for 3 articles. One article is from the *New York Times*, and the other two are from disaster and health related news webpages. It will take approximately less than 1 hour to complete the experiment. You will be compensated for your participation. Out of three tasks in the task questionnaire, for each completed task, you will receive a $10 gift card. Therefore, if you complete all three tasks, you will receive $30 gift card.

If you are interested in this study or have a question, please don't hesitate to contact me at:

Email: seungwon@vt.edu

Phone:(540) 230-8983

Department of Computer Science

114 McBryde Hall, Mail Code 0106

Virginia Tech, Blacksburg, VA 24061

Thank you,
Sincerely, Seungwon Yang

Ph.D. student

A.3 Informed Consent

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Informed consent form for participants of Research Involving Human Subject

Title of the Project: Investigating the Usability of a Topic Tagging System

Investigators: **Seungwon Yang**

Faculty investigators: Dr. Edward A. Fox

Participant number: ____________________________

I. PURPOSE OF THIS RESEARCH

The investigators of this study are working on developing a web-based topic tagging system, which produces topics given an electronic textual document. The purpose of this study is to examine the usability of the prototype system. The responses to questionnaires will be used to improve the system's user interface, algorithms, and interaction methods.

II. PROCEDURES

All the communications, for example, sending links to questionnaires, web-based system, and a tutorial document, are conducted by email. Your answers for the questionnaires are collected using an online survey system.

Upon receiving your signed informed consent (this form) by, e.g., cellphone pictures/scanned PDF/images/mail, we will send you a link to a demographic questionnaire for you to complete. Please make sure your participant ID is present in all questionnaires. After completing the demographic questionnaire, please email the researcher at seungwon@vt.edu to notify the completion. Then, a link to the task questionnaire and a link to an online prototype tool will be sent to you by email.

You will become familiar with the steps of doing the task by reading the instructions, which are found in the top portion of the task questionnaire. You also will become familiar with how the topic tagging system works by following the tutorial.

The tasks are to open a text document using the tagging system, prepare a list of representative topical words for the opened document using both the tagging system and your intelligence, and enter those topics as the answer in the task questionnaire. Please use only the tagging system and your intelligence to answer the questionnaire. You will work on a total of 3 online newspaper articles, and the task will take approximately less than 1 hour. After completing the task questionnaire, please email the researcher at seungwon@vt.edu to notify the completion. Then, the researcher will send you a link to an exit questionnaire. Please notify the researcher after completing the exit questionnaire as well.

(Note: the values 3 and 1 may change slightly as a result of a pilot study.)

You can stop and resume work anytime during your participation. There is no time limit for the task completion, but we encourage you to finish the tasks in 2 days at most. Once you complete the tasks and questionnaires, you will receive a gift card by email as compensation.

III. RISKS

You are involved with reading textual documents on a computer screen, and interacting with the system's user interface to perform the tasks and to answer the questionnaire. The physical components of the tasks previously described may result in eye, wrist, or finger fatigue.

If you experience any fatigue during the task, you are free to rest and may continue when you are ready. If the fatigue becomes uncomfortable, you will be allowed to leave the study with no penalty.

IV. BENEFITS OF THIS RESEARCH

You will experience the topic tagging task, using both the provided web-based system and your intelligence to achieve the best results. In addition, you will learn about the usability testing process by going through the experiment.

V. EXTENT OF ANONYMITY AND CONFIDENTIALITY

This research will assure your confidentiality. We will have your signatures on the Informed Consent document. Photographs with a cell phone, scanned PDFs, images of the signed document, or a paper form are all acceptable. This document will be kept in the computer of a researcher, Seungwon Yang, until his dissertation and related publications are published. Your name will not be released in any publication. After publishing the dissertation and after the final acceptance of related papers, Seungwon Yang will destroy the documents (in electronic forms) from his computer.

Only the researcher, Seungwon Yang, will collect the data. No one other than the researchers will have access to the data. All responses will be coded so as not to include the name of the participant. Only your participant number will identify you during analysis and any written reports of the research.

This study is being conducted solely for research and development purposes, and the resulting data and interpretations also will be the part of the researcher's academic work. Consistent with these academic purposes, any results would be freely publishable. Again, to protect your identity, personal names will not be used in any published works. We are willing to share drafts of reports with you before submitting them for publication.

VI. COMPENSATION

A $30 gift card is provided upon completion of all three tasks in the experiment. However, the compensation is also prorated, so you can receive $10 gift card for each completed task.

VII. FREEDOM TO WITHDRAW

Your participation is voluntary and the decision on whether you wish to participate or not is strictly your own. You may discontinue participation at any time. However, the compensation will be given only to those who complete the experiment.

VIII. APPROVAL OF RESEARCH
This research project has been approved by the Institutional Review Board for Research Involving Human Subjects at Virginia Polytechnic Institute and State University and by the Department of Computer Science (College of Engineering).

IX. PARTICIPANT'S RESPONSIBILITIES

Upon signing this form, I voluntarily agree to participate in this study. I have no restrictions to my participation in this study. As a participant, I may withdraw from this experiment at any time without penalty. I agree to abide by the rules of this project.

X. PARTICIPANT'S PERMISSION

I have read and understand the Informed Consent and conditions of this study. All of my questions have been answered. I agree to participate in this experiment.

______________________________ ______________________
Participant's Signature Date

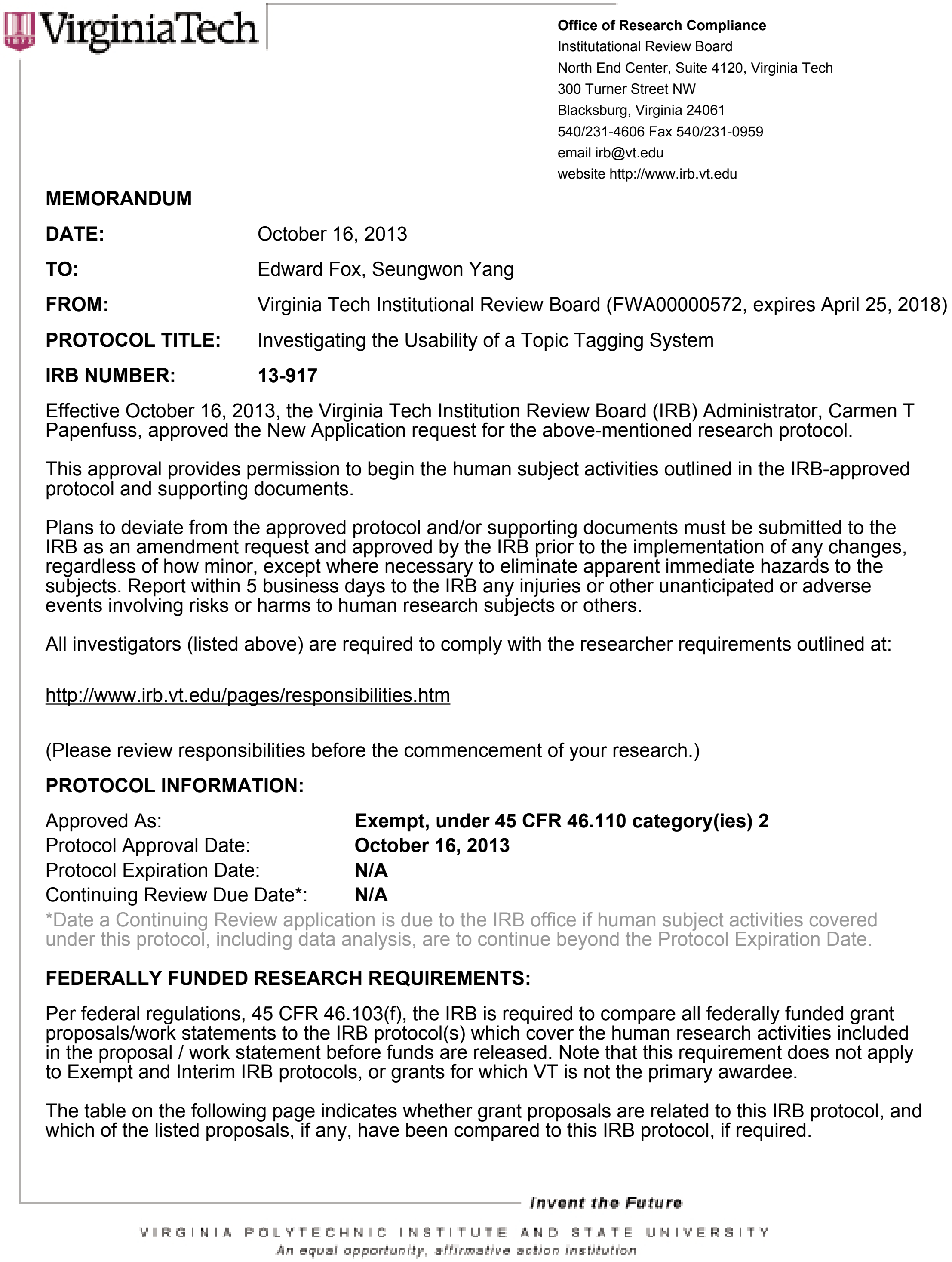
Should I have any questions about the research or its conduct, I may contact:

Faculty Advisor: Edward A. Fox
Professor, Department of Computer Science, Virginia Tech
Email: fox@vt.edu Phone: (540) 231-5113
Mailing address:
Dept. of Computer Science
114 McBryde Hall, Mail Code 0106
Virginia Tech, Blacksburg, VA 24061

Investigator: Seungwon Yang
Ph.D. student, Department of Computer Science, Virginia Tech
Email: seungwon@vt.edu Phone: (540) 230-8983
Mailing address:
Dept. of Computer Science
114 McBryde Hall, Mail Code 0106
Virginia Tech, Blacksburg, VA 24061

Dr. David M. Moore, Email : moored@vt.edu Phone: (540) 231-4991
Chair, IRB

A.4 IRB Approval Letter



VirginiaTech

Office of Research Compliance
Institutational Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE:	October 16, 2013
TO:	Edward Fox, Seungwon Yang
FROM:	Virginia Tech Institutional Review Board (FWA00000572, expires April 25, 2018)
PROTOCOL TITLE:	Investigating the Usability of a Topic Tagging System
IRB NUMBER:	**13-917**

Effective October 16, 2013, the Virginia Tech Institution Review Board (IRB) Administrator, Carmen T Papenfuss, approved the New Application request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Exempt, under 45 CFR 46.110 category(ies) 2**
Protocol Approval Date:	**October 16, 2013**
Protocol Expiration Date:	**N/A**
Continuing Review Due Date*:	**N/A**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Invent the Future

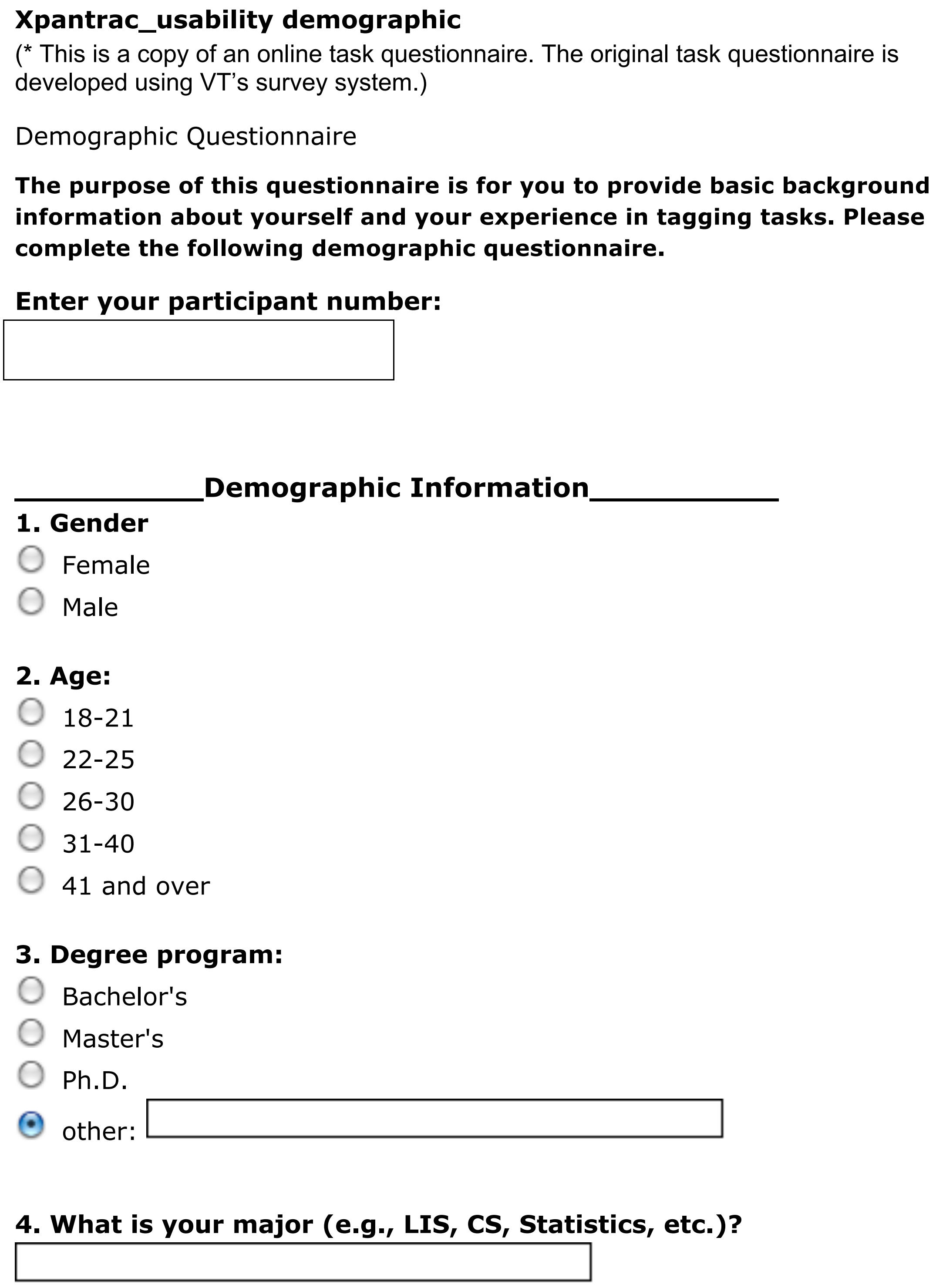
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
An equal opportunity, affirmative action institution

Date*	OSP Number	Sponsor	Grant Comparison Conducted?
10/14/2013	13130712	National Science Foundation	Not required (Exempt approval)

* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this IRB protocol is to cover any other grant proposals, please contact the IRB office (irbadmin@vt.edu) immediately.

A.5 Demographic Questionnaire



Xpantrac_usability demographic

(* This is a copy of an online task questionnaire. The original task questionnaire is developed using VT's survey system.)

Demographic Questionnaire

The purpose of this questionnaire is for you to provide basic background information about yourself and your experience in tagging tasks. Please complete the following demographic questionnaire.

Enter your participant number:

___________Demographic Information___________

1. Gender

○ Female
○ Male

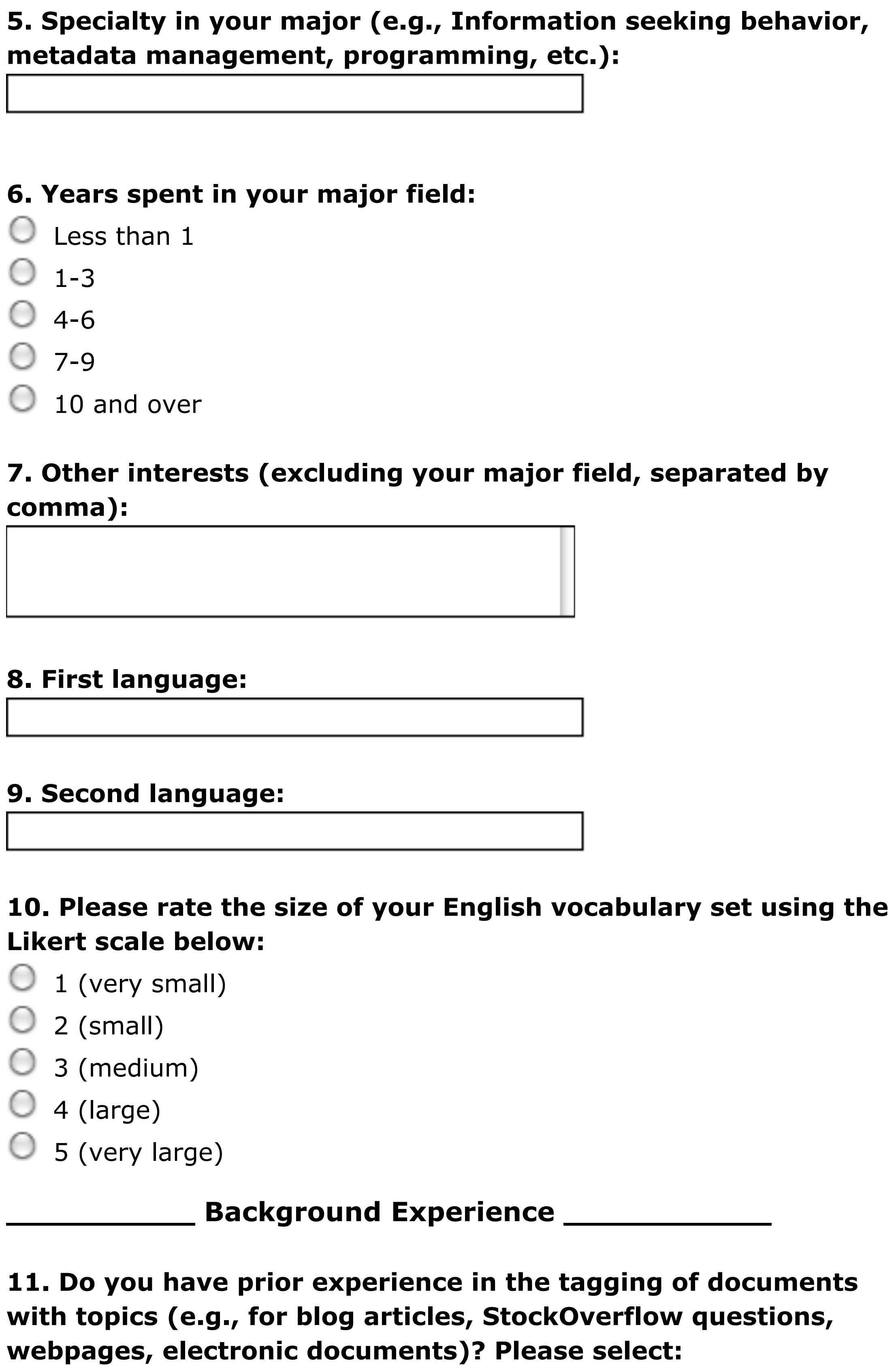
2. Age:

○ 18-21
○ 22-25
○ 26-30
○ 31-40
○ 41 and over

3. Degree program:

○ Bachelor's
○ Master's
○ Ph.D.
● other:

4. What is your major (e.g., LIS, CS, Statistics, etc.)?



5. Specialty in your major (e.g., Information seeking behavior, metadata management, programming, etc.):

6. Years spent in your major field:

○ Less than 1

○ 1-3

○ 4-6

○ 7-9

○ 10 and over

7. Other interests (excluding your major field, separated by comma):

8. First language:

9. Second language:

10. Please rate the size of your English vocabulary set using the Likert scale below:

○ 1 (very small)

○ 2 (small)

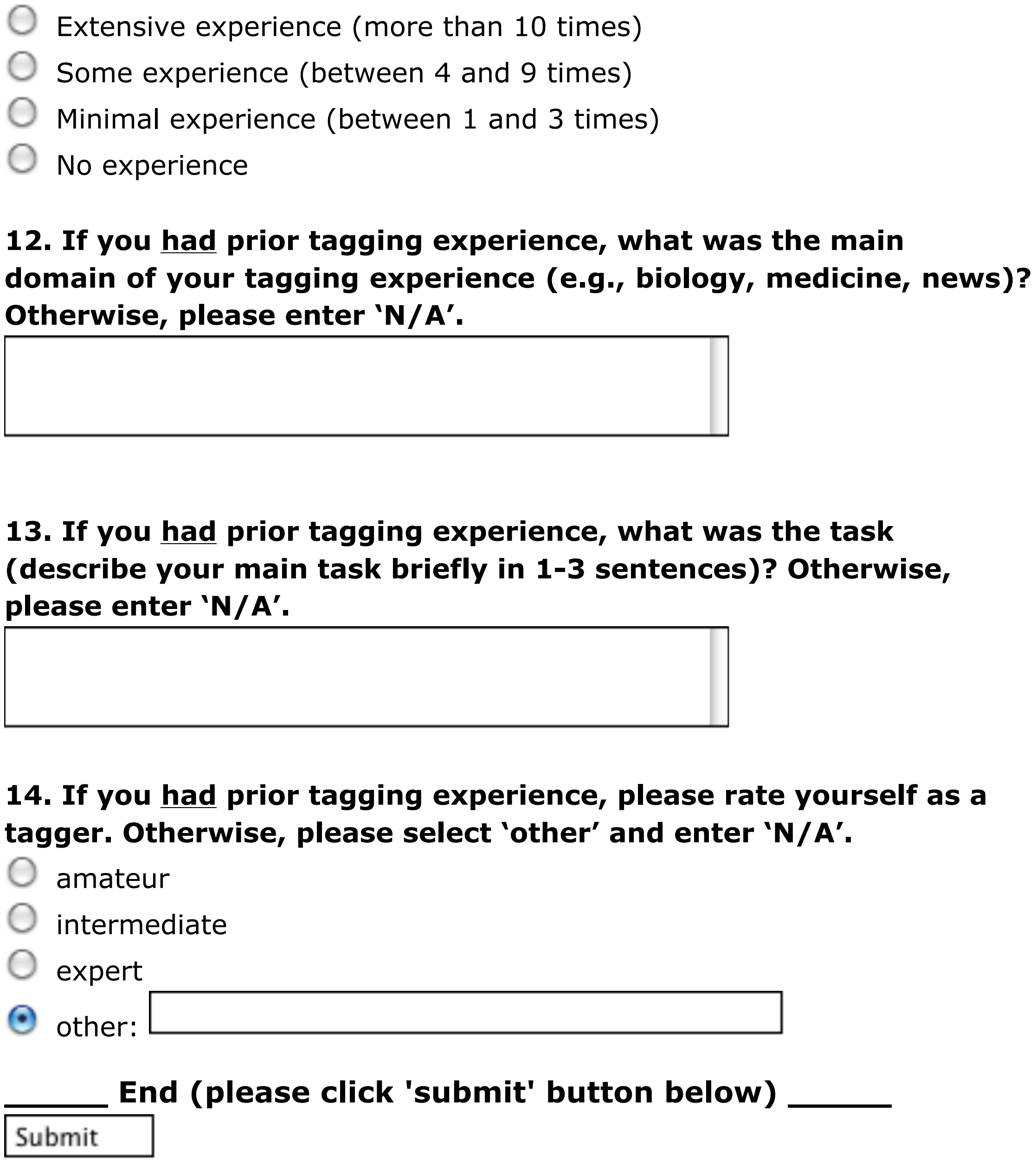
○ 3 (medium)

○ 4 (large)

○ 5 (very large)

___________ Background Experience ____________

11. Do you have prior experience in the tagging of documents with topics (e.g., for blog articles, StockOverflow questions, webpages, electronic documents)? Please select:



○ Extensive experience (more than 10 times)
○ Some experience (between 4 and 9 times)
○ Minimal experience (between 1 and 3 times)
○ No experience

12. If you had prior tagging experience, what was the main domain of your tagging experience (e.g., biology, medicine, news)? Otherwise, please enter 'N/A'.

13. If you had prior tagging experience, what was the task (describe your main task briefly in 1-3 sentences)? Otherwise, please enter 'N/A'.

14. If you had prior tagging experience, please rate yourself as a tagger. Otherwise, please select 'other' and enter 'N/A'.

○ amateur
○ intermediate
○ expert
● other:

_____ End (please click 'submit' button below) _____

A.6 Task Questionnaires

A.6.1 Type 1

Task Questionnaire TYPE 1

(The actual questionnaire will be provided to the participants as an online survey form.)

* Please enter your participant number:

A Tutorial of the Topic Tagging System

Please download the tutorial document from the following link here, and read it.

(Tutorial at: http://spare05.dlib.vt.edu/~xpantrac/XpantracTutorial.pdf)

Exercise Task

Before you start the main tasks, we ask that you do an exercise task to become familiar with the tagging system. Please follow the instructions below:

1. Open the web-based tagging system by visiting: http://spare05.dlib.vt.edu/~xpantrac/ui/xpantrac_proto.html
2. On the top of the Collection pane located on the left, press the "Load Document" button to load the documents from the pre-selected *NYT_1000* corpus.
3. The titles of the 30 documents are shown in gray boxes on the Collection pane.
4. Click to open the 26th document, "26. SHULTZ SAID TO WARN REAGAN OF CASEY'S VERSION"
5. You should read and come up with topics, based on your understanding of a document, and your intellectual ability to summarize that document with a list of topics (i.e., keywords or key phrases) that indicate the several topics covered by the document. Jot those down, to be entered into the text box in step 8 below.
6. The system has been built to suggest topics from documents. To supplement your own list from step 5 above, you can ask the system to generate a list of suggested topics, and then you can select ones you agree with from the system's topic suggestions. Before you can generate topic tags, a preparatory action is needed: You must guide the system regarding the number and breadth of topics it should produced. For example, set "Number of API results" to 10, and "Number of Topics" to 20 as well. You can choose different values. Remember that the higher number of API results means using more information to produce topics, but it takes a slightly longer time.
7. Click the "Extract Topics" button. You will see topic suggestions in two-columns. You can click each topic box that has a useful topic, to mark it as useful, for example, click "reagan", "president", "iran", "Shultz", "intelligence", etc.
8. Based on what you did in steps 5 and 7 above, you can enter your final topics (i.e., your own + those you selected that were suggested by the system) below:

Beginning of the Main Tasks

---- Task 1----

The Questions 1 and its sub questions are about the NYT_1000 corpus. Please load the NYT_1000 corpus by checking it and clicking “Load Documents” button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 1. Click “2. Lincoln Towers Tenants Ponder Whether to Buy” article in the Collection pane on the left. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below. You should find topics that are as accurate as possible, without wasting time. You are allowed to read the document and come up with your own tags, and also to use the tagging system’s suggestions.

Question 1-1. How thoroughly did you read the document?

(0) Did not read the document at all
(1) Skimmed through the document
(2) Went through almost every sentence
(3) Read more than once

Question 1-2. What was the reason for your answer in Question 1-1?

Question 1-3. What were your settings for the “Number of API results” and “Number of Topics” in the system? If you did not use the system, please enter “0” and “0”.

- Number of API results: ______
- Number of Topics: ______

Question 1-4. What was your rationale for the settings in Question 1-3?

---- Task 2 ----

The Questions 2 and its sub-questions are about the CTR_30 corpus. Please load the CTR_30 corpus by checking it and clicking “Load Documents” button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 2. Click “1. Haiti braces for Isaac’s deluge” article in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below. You should find topics as accurate as possible, without wasting time.

Question 2-1. How thoroughly did you read the document?

(0) Did not read the document at all
(1) Skimmed through the document
(2) Went through almost every sentence
(3) Read more than once

Question 2-2. What was the reason for your answer in Question 2-1?

Question 2-3. What were your settings for the “Number of API results” and “Number of Topics” in the system? If you did not use the system, please enter “0” and “0”.

- Number of API results: ______
- Number of Topics: ______

Question 2-4. What was your rationale for the settings in Question 2-3?

---- Task 3 ----

The Question 3 and its sub-questions are about the VARIOUS_30 corpus. Please load the VARIOUS_30 corpus by checking it and clicking “Load Documents” button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 3. Click “1. Hope fades after Guatemala earthquake” article in the Collection pane on the left. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below.

Question 3-1. How thoroughly did you read the document?

(0) Did not read the document at all
(1) Skimmed through the document
(2) Went through almost every sentence
(3) Read more than once

Question 3-2. What was the reason for your answer in Question 3-1?

Question 3-3. What were your settings for the “Number of API results” and “Number of Topics” in the system? If you did not use the system, please enter “0” and “0”.

- Number of API results: ______
- Number of Topics: ______

Question 3-4. What was your rationale for the settings in Question 3-3?

Question 4. Describe your overall strategy to come up with quality topic tags:

<u>End of the Main Tasks</u>

* Submit the task form by pressing the ‘Submit’ button on the bottom of the survey.

A.6.2 Type 2

Task Questionnaire TYPE 2

(The actual questionnaire will be provided to the participants as an online survey form.)

* Please enter your participant number:

A Tutorial of the Topic Tagging System

Please download the tutorial document from the following link here, and read it.

Exercise Task

Before you start the main tasks, we ask that you do an exercise task to become familiar with the tagging system. Please follow the instructions below:

9. Open the web-based tagging system by visiting: http://spare05.dlib.vt.edu/~xpantrac/ui/xpantrac_proto.html
10. On the top of the Collection pane located on the left, press the "Load Document" button to load the documents from the pre-selected *NYT_1000* corpus.
11. The titles of the 30 documents are shown in gray boxes on the Collection pane.
12. Click to open the 26th document, "26. SHULTZ SAID TO WARN REAGAN OF CASEY'S VERSION"
13. You should read and come up with topics, based on your understanding of a document, and your intellectual ability to summarize that document with a list of topics (i.e., keywords or key phrases) that indicate the several topics covered by the document. Jot those down, to be entered into the text box in step 8 below.
14. The system has been built to suggest topics from documents. To supplement your own list from step 5 above, you can ask the system to generate a list of suggested topics, and then you can select ones you agree with from the system's topic suggestions. Before you can generate topic tags, a preparatory action is needed: You must guide the system regarding the number and breadth of topics it should produced. For example, set "Number of API results" to 10, and "Number of Topics" to 20 as well. You can choose different values. Remember that the higher number of API results means using more information to produce topics, but it takes a slightly longer time.
15. Click the "Extract Topics" button. You will see topic suggestions in two-columns. You can click each topic box that has a useful topic, to mark it as useful, for example, click "reagan", "president", "iran", "Shultz", "intelligence", etc.
16. Based on what you did in steps 5 and 7 above, you can enter your final topics (i.e., your own + those you selected that were suggested by the system) below:

Beginning of the Main Tasks

---- NYT_1000 ----

The Questions 1 and its sub questions are about the NYT_1000 corpus. Please load the NYT_1000 corpus by checking it and clicking “Load Documents” button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 1. Click “3. A LEGENDARY FISH FROM GALILEE” in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below. You should find topics that are as accurate as possible, without wasting time. You are allowed to read the document and come up with your own tags, and also to use the tagging system’s suggestions.

Question 1-1. How thoroughly did you read the document?

(4) Did not read the document at all
(5) Skimmed through the document
(6) Went through almost every sentence
(7) Read more than once

Question 1-2. What was the reason for your answer in Question 1-1?

Question 1-3. What were your settings for the “Number of API results” and “Number of Topics” in the system? If you did not use the system, please enter “0” and “0”.

- Number of API results: ______
- Number of Topics: ______

Question 1-4. What was your rationale for the settings in Question 1-3?

---- CTR_30 document corpus ----

The Questions 2 and its sub-questions are about the CTR_30 corpus. Please load the CTR_30 corpus by checking it and clicking “Load Documents” button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 2. Click “2. Tropical Storm Isaac shifts west slightly, but be prepared” in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below. You should find topics as accurate as possible, without wasting time.

Question 2-1. How thoroughly did you read the document?

(4) Did not read the document at all
(5) Skimmed through the document
(6) Went through almost every sentence
(7) Read more than once

Question 2-2. What was the reason for your answer in Question 2-1?

Question 2-3. What were your settings for the “Number of API results” and “Number of Topics” in the system? If you did not use the system, please enter “0” and “0”.

- Number of API results: ______
- Number of Topics: ______

Question 2-4. What was your rationale for the settings in Question 2-3?

---- VARIOUS_30 document corpus ----

The Question 3 and its sub-questions are about the VARIOUS_30 corpus. Please load the VARIOUS_30 corpus by checking it and clicking “Load Documents” button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 3. Click “11. Troops movements, call-up of reservists signal Israeli ground operation in Gaza imminent” in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below.

Question 3-1. How thoroughly did you read the document?

(0) Did not read the document at all
(1) Skimmed through the document
(2) Went through almost every sentence
(3) Read more than once

Question 3-2. What was the reason for your answer in Question 3-1?

Question 3-3. What were your settings for the “Number of API results” and “Number of Topics” in the system? If you did not use the system, please enter “0” and “0”.

- Number of API results: ______
- Number of Topics: ______

Question 3-4. What was your rationale for the settings in Question 3-3?

Question 4. Describe your overall strategy to come up with quality topic tags:

End of the Main Tasks

* Submit the task form by pressing the ‘Submit’ button on the bottom of the survey.

A.6.3 Type 3

Task Questionnaire TYPE 3

(The actual questionnaire will be provided to the participants as an online survey form.)

* Please enter your participant number:

A Tutorial of the Topic Tagging System

Please download the tutorial document from the following link here, and read it.

Exercise Task

Before you start the main tasks, we ask that you do an exercise task to become familiar with the tagging system. Please follow the instructions below:

17. Open the web-based tagging system by visiting:
http://spare05.dlib.vt.edu/~xpantrac/ui/xpantrac_proto.html

18. On the top of the Collection pane located on the left, press the “Load Document” button to load the documents from the pre-selected *NYT_1000* corpus.
19. The titles of the 30 documents are shown in gray boxes on the Collection pane.
20. Click to open the 26th document, “26. SHULTZ SAID TO WARN REAGAN OF CASEY’S VERSION”
21. You should read and come up with topics, based on your understanding of a document, and your intellectual ability to summarize that document with a list of topics (i.e., keywords or key phrases) that indicate the several topics covered by the document. Jot those down, to be entered into the text box in step 8 below.
22. The system has been built to suggest topics from documents. To supplement your own list from step 5 above, you can ask the system to generate a list of suggested topics, and then you can select ones you agree with from the system’s topic suggestions. Before you can generate topic tags, a preparatory action is needed: You must guide the system regarding the number and breadth of topics it should produced. For example, set “Number of API results” to 10, and “Number of Topics” to 20 as well. You can choose different values. Remember that the higher number of API results means using more information to produce topics, but it takes a slightly longer time.
23. Click the “Extract Topics” button. You will see topic suggestions in two-columns. You can click each topic box that has a useful topic, to mark it as useful, for example, click “reagan”, “president”, “iran”, “Shultz”, “intelligence”, etc.
24. Based on what you did in steps 5 and 7 above, you can enter your final topics (i.e., your own + those you selected that were suggested by the system) below:

Beginning of the Main Tasks

---- NYT_1000 ----

The Questions 1 and its sub questions are about the NYT_1000 corpus. Please load the NYT_1000 corpus by checking it and clicking “Load Documents” button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 1. Click “4. AIR FORCE IS CUTTING FIGHTER UNITS IN NUCLEAR SHIFT ” in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below. You should find topics that are as accurate as possible, without wasting time. You are allowed to read the document and come up with your own tags, and also to use the tagging system’s suggestions.

Question 1-1. How thoroughly did you read the document?

(8) Did not read the document at all
(9) Skimmed through the document
(10) Went through almost every sentence
(11) Read more than once

Question 1-2. What was the reason for your answer in Question 1-1?

Question 1-3. What were your settings for the "Number of API results" and "Number of Topics" in the system? If you did not use the system, please enter "0" and "0".

- Number of API results: ______
- Number of Topics: ______

Question 1-4. What was your rationale for the settings in Question 1-3?

---- CTR_30 document corpus ----

The Questions 2 and its sub-questions are about the CTR_30 corpus. Please load the CTR_30 corpus by checking it and clicking "Load Documents" button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 2. Click "15. No title" article in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below. You should find topics as accurate as possible, without wasting time.

Question 2-1. How thoroughly did you read the document?

(8) Did not read the document at all
(9) Skimmed through the document
(10) Went through almost every sentence
(11) Read more than once

Question 2-2. What was the reason for your answer in Question 2-1?

Question 2-3. What were your settings for the "Number of API results" and "Number of Topics" in the system? If you did not use the system, please enter "0" and "0".

- Number of API results: ______
- Number of Topics: ______

Question 2-4. What was your rationale for the settings in Question 2-3?

---- VARIOUS_30 document corpus ----

The Question 3 and its sub-questions are about the VARIOUS_30 corpus. Please load the VARIOUS_30 corpus by checking it and clicking “Load Documents” button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 3. Click “18. Car bombs, aerial attacks pummel Syria” in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below.

Question 3-1. How thoroughly did you read the document?

(0) Did not read the document at all
(1) Skimmed through the document
(2) Went through almost every sentence
(3) Read more than once

Question 3-2. What was the reason for your answer in Question 3-1?

Question 3-3. What were your settings for the “Number of API results” and “Number of Topics” in the system? If you did not use the system, please enter “0” and “0”.

- Number of API results: ______
- Number of Topics: ______

Question 3-4. What was your rationale for the settings in Question 3-3?

Question 4. Describe your overall strategy to come up with quality topic tags:

End of the Main Tasks

* Submit the task form by pressing the ‘Submit’ button on the bottom of the survey.

A.6.4 Type 4

Task Questionnaire TYPE 4

(The actual questionnaire will be provided to the participants as an online survey form.)

* Please enter your participant number:

A Tutorial of the Topic Tagging System

Please download the tutorial document from the following link here, and read it.

Exercise Task

Before you start the main tasks, we ask that you do an exercise task to become familiar with the tagging system. Please follow the instructions below:

25. Open the web-based tagging system by visiting: http://spare05.dlib.vt.edu/~xpantrac/ui/xpantrac_proto.html
26. On the top of the Collection pane located on the left, press the “Load Document” button to load the documents from the pre-selected *NYT_1000* corpus.
27. The titles of the 30 documents are shown in gray boxes on the Collection pane.
28. Click to open the 26th document, “26. SHULTZ SAID TO WARN REAGAN OF CASEY’S VERSION”
29. You should read and come up with topics, based on your understanding of a document, and your intellectual ability to summarize that document with a list of topics (i.e., keywords or key phrases) that indicate the several topics covered by the document. Jot those down, to be entered into the text box in step 8 below.
30. The system has been built to suggest topics from documents. To supplement your own list from step 5 above, you can ask the system to generate a list of suggested topics, and then you can select ones you agree with from the system’s topic suggestions. Before you can generate topic tags, a preparatory action is needed: You must guide the system regarding the number and breadth of topics it should produced. For example, set “Number of API results” to 10, and “Number of Topics” to 20 as well. You can choose different values. Remember that the higher number of API results means using more information to produce topics, but it takes a slightly longer time.
31. Click the “Extract Topics” button. You will see topic suggestions in two-columns. You can click each topic box that has a useful topic, to mark it as useful, for example, click “reagan”, “president”, “iran”, “Shultz”, “intelligence”, etc.
32. Based on what you did in steps 5 and 7 above, you can enter your final topics (i.e., your own + those you selected that were suggested by the system) below:

Beginning of the Main Tasks

---- NYT_1000 ----

The Questions 1 and its sub questions are about the NYT_1000 corpus. Please load the NYT_1000 corpus by checking it and clicking "Load Documents" button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 1. Click "11. IRAQ SAYS THE ATTACK FAILED" in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below. You should find topics that are as accurate as possible, without wasting time. You are allowed to read the document and come up with your own tags, and also to use the tagging system's suggestions.

Question 1-1. How thoroughly did you read the document?

(12) Did not read the document at all
(13) Skimmed through the document
(14) Went through almost every sentence
(15) Read more than once

Question 1-2. What was the reason for your answer in Question 1-1?

Question 1-3. What were your settings for the "Number of API results" and "Number of Topics" in the system? If you did not use the system, please enter "0" and "0".

- Number of API results: ______
- Number of Topics: ______

Question 1-4. What was your rationale for the settings in Question 1-3?

---- CTR_30 document corpus ----

The Questions 2 and its sub-questions are about the CTR_30 corpus. Please load the CTR_30 corpus by checking it and clicking "Load Documents" button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 2. Click "18. After drenching New Orleans, Isaac threatens dam" in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below. You should find topics as accurate as possible, without wasting time.

Question 2-1. How thoroughly did you read the document?

(12) Did not read the document at all
(13) Skimmed through the document
(14) Went through almost every sentence
(15) Read more than once

Question 2-2. What was the reason for your answer in Question 2-1?

Question 2-3. What were your settings for the "Number of API results" and "Number of Topics" in the system? If you did not use the system, please enter "0" and "0".

- Number of API results: ______
- Number of Topics: ______

Question 2-4. What was your rationale for the settings in Question 2-3?

---- VARIOUS_30 document corpus ----

The Question 3 and its sub-questions are about the VARIOUS_30 corpus. Please load the VARIOUS_30 corpus by checking it and clicking "Load Documents" button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 3. Click "21. Diabetes is a chronic (lifelong) disease marked by high levels of sugar in the blood" in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below.

Question 3-1. How thoroughly did you read the document?

(0) Did not read the document at all
(1) Skimmed through the document
(2) Went through almost every sentence
(3) Read more than once

Question 3-2. What was the reason for your answer in Question 3-1?

Question 3-3. What were your settings for the “Number of API results” and “Number of Topics” in the system? If you did not use the system, please enter “0” and “0”.

- Number of API results: ______
- Number of Topics: ______

Question 3-4. What was your rationale for the settings in Question 3-3?

Question 4. Describe your overall strategy to come up with quality topic tags:

End of the Main Tasks

* Submit the task form by pressing the ‘Submit’ button on the bottom of the survey.

A.6.5 Type 5

Task Questionnaire TYPE 5

(The actual questionnaire will be provided to the participants as an online survey form.)

* Please enter your participant number:

A Tutorial of the Topic Tagging System

Please download the tutorial document from the following link here, and read it.

Exercise Task

Before you start the main tasks, we ask that you do an exercise task to become familiar with the tagging system. Please follow the instructions below:

33. Open the web-based tagging system by visiting: http://spare05.dlib.vt.edu/~xpantrac/ui/xpantrac_proto.html
34. On the top of the Collection pane located on the left, press the “Load Document” button to load the documents from the pre-selected *NYT_1000* corpus.
35. The titles of the 30 documents are shown in gray boxes on the Collection pane.
36. Click to open the 26th document, “26. SHULTZ SAID TO WARN REAGAN OF CASEY’S VERSION”
37. You should read and come up with topics, based on your understanding of a document, and your intellectual ability to summarize that document with a list of topics (i.e., keywords or key phrases) that indicate the several topics covered by the document. Jot those down, to be entered into the text box in step 8 below.
38. The system has been built to suggest topics from documents. To supplement your own list from step 5 above, you can ask the system to generate a list of suggested topics, and then you can select ones you agree with from the system’s topic suggestions. Before you can generate topic tags, a preparatory action is needed: You must guide the system regarding the number and breadth of topics it should produced. For example, set “Number of API results” to 10, and “Number of Topics” to 20 as well. You can choose different values. Remember that the higher number of API results means using more information to produce topics, but it takes a slightly longer time.
39. Click the “Extract Topics” button. You will see topic suggestions in two-columns. You can click each topic box that has a useful topic, to mark it as useful, for example, click “reagan”, “president”, “iran”, “Shultz”, “intelligence”, etc.
40. Based on what you did in steps 5 and 7 above, you can enter your final topics (i.e., your own + those you selected that were suggested by the system) below:

Beginning of the Main Tasks

---- NYT_1000 ----

The Questions 1 and its sub questions are about the NYT_1000 corpus. Please load the NYT_1000 corpus by checking it and clicking “Load Documents” button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 1. Click “20. REASON UNCLEAR” in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below. You should find topics that are as accurate as possible, without wasting time. You are allowed to read the document and come up with your own tags, and also to use the tagging system’s suggestions.

Question 1-1. How thoroughly did you read the document?

(16) Did not read the document at all
(17) Skimmed through the document
(18) Went through almost every sentence
(19) Read more than once

Question 1-2. What was the reason for your answer in Question 1-1?

Question 1-3. What were your settings for the “Number of API results” and “Number of Topics” in the system? If you did not use the system, please enter “0” and “0”.

- Number of API results: ______
- Number of Topics: ______

Question 1-4. What was your rationale for the settings in Question 1-3?

---- CTR_30 document corpus ----

The Questions 2 and its sub-questions are about the CTR_30 corpus. Please load the CTR_30 corpus by checking it and clicking “Load Documents” button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 2. Click “25. Hurricane Isaac relief items stolen from St. Bernard Parish school” in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below. You should find topics as accurate as possible, without wasting time.

Question 2-1. How thoroughly did you read the document?

(16) Did not read the document at all
(17) Skimmed through the document
(18) Went through almost every sentence
(19) Read more than once

Question 2-2. What was the reason for your answer in Question 2-1?

Question 2-3. What were your settings for the “Number of API results” and “Number of Topics” in the system? If you did not use the system, please enter “0” and “0”.

- Number of API results: ______
- Number of Topics: ______

Question 2-4. What was your rationale for the settings in Question 2-3?

---- VARIOUS_30 document corpus ----

The Question 3 and its sub-questions are about the VARIOUS_30 corpus. Please load the VARIOUS_30 corpus by checking it and clicking “Load Documents” button in the top portion of the Collection pane. Now you should see a total of 30 gray boxes on the Collection pane.

Question 3. Click “27. Heart Attach Prevention” in the Collection pane. Then, enter a list of topic tags that you think are relevant to the document in the text box provided below.

Question 3-1. How thoroughly did you read the document?

(20) Did not read the document at all
(21) Skimmed through the document
(22) Went through almost every sentence
(23) Read more than once

Question 3-2. What was the reason for your answer in Question 3-1?

Question 3-3. What were your settings for the “Number of API results” and “Number of Topics” in the system? If you did not use the system, please enter “0” and “0”.

- Number of API results: ______
- Number of Topics: ______

Question 3-4. What was your rationale for the settings in Question 3-3?

Question 4. Describe your overall strategy to come up with quality topic tags:

End of the Main Tasks

* Submit the task form by pressing the ‘Submit’ button on the bottom of the survey.

A.7 Exit Questionnaire

Exit Questionnaire

(The actual questionnaire will be provided to the participants as an online survey form.)

* Please enter your participant number:

Q.1. Please rate how much you agree with, “The system was easy to use”:

-2 (strongly disagree)

-1 (disagree)

0 (neutral)

1 (agree)

2 (strongly agree)

Q.2. Please rate how much you agree with, “Using the topic tagging system was an enjoyable experience”:

-2 (strongly disagree)

-1 (disagree)

0 (neutral)

1 (agree)

2 (strongly agree)

Q.3. Please rate how much you agree with, “The system was useful in finding the indicative topics:

-2 (strongly disagree)

-1 (disagree)

0 (neutral)

1 (agree)

2 (strongly agree)

Q.4. Please rate how much you agree with, “Using the system decreased the time to complete the tasks”:

-2 (strongly disagree)

-1 (disagree)

0 (neutral)

1 (agree)

2 (strongly agree)

Q.5. Please rate how much you agree with, “Using the system distracted me from focusing on the tasks”:

-2 (strongly disagree)

-1 (disagree)

0 (neutral)

1 (agree)

2 (strongly agree)

Q.6. (If any) What was the main problem you noticed with the “Collection pane” on the left, when conducting the tasks?

Q.7. (If any) What was the main problem you noticed with the “Document pane” on the center, when conducting the tasks?

Q.8. (If any) What was the main problem you notices with the “Topics pane” on the right, when conducting the tasks?

Q. 9. Please rate how much you agree with, “Overall, I was satisfied with this topic tagging system”:

-2 (strongly disagree)

-1 (disagree)

0 (neutral)

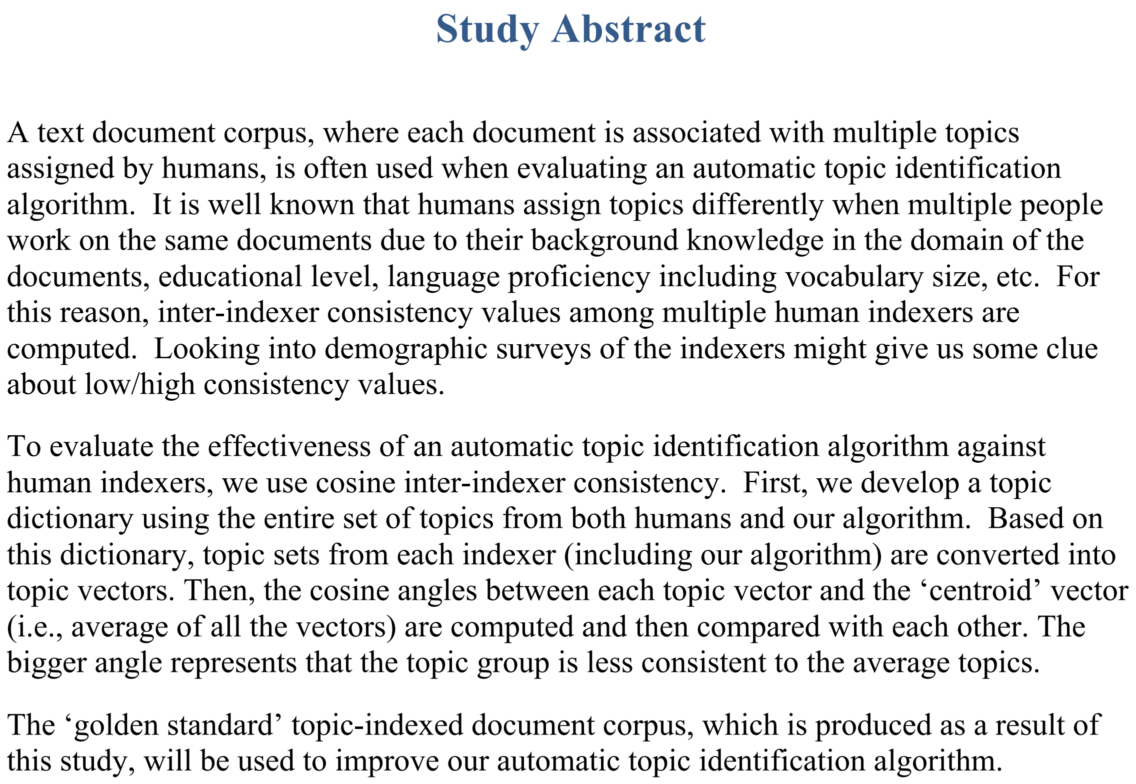
1 (agree)

2 (strongly agree)

Appendix B. Manual Topic Tagging Study

This appendix contains resources used for a manual topic tagging study, which involved multiple human topic indexers. The type of the documents used are the same as the ones in Appendix A: the study abstract, a participant recruitment letter, an informed consent form, an IRB approval letter, and a demographic, and a task questionnaire.

B.1 Study Abstract



Study Abstract

A text document corpus, where each document is associated with multiple topics assigned by humans, is often used when evaluating an automatic topic identification algorithm. It is well known that humans assign topics differently when multiple people work on the same documents due to their background knowledge in the domain of the documents, educational level, language proficiency including vocabulary size, etc. For this reason, inter-indexer consistency values among multiple human indexers are computed. Looking into demographic surveys of the indexers might give us some clue about low/high consistency values.

To evaluate the effectiveness of an automatic topic identification algorithm against human indexers, we use cosine inter-indexer consistency. First, we develop a topic dictionary using the entire set of topics from both humans and our algorithm. Based on this dictionary, topic sets from each indexer (including our algorithm) are converted into topic vectors. Then, the cosine angles between each topic vector and the 'centroid' vector (i.e., average of all the vectors) are computed and then compared with each other. The bigger angle represents that the topic group is less consistent to the average topics.

The 'golden standard' topic-indexed document corpus, which is produced as a result of this study, will be used to improve our automatic topic identification algorithm.

B.2 Participant Recruitment Letter

Title: A Study of Multiple Topic Assignment to Disaster Webpage Content and Inter-Indexer Consistency Computation

We would like to recruit participants, who were trained in the Library and Information Sciences field, and whose first language is English. We will keep recruiting people until we have five completed results. The recruiting criteria are that the participant is not a minor and English is her/his first language. Participants individually index a total of 30 textual documents with topical keywords that match the theme of documents. Each document is a news article related to a hurricane disaster. A couple of example documents are here:

http://spare05.dlib.vt.edu/~seungwon/ctr_30/1.pdf

http://spare05.dlib.vt.edu/~seungwon/ctr_30/5.pdf

The study involves the steps listed below:

- Step 1: Read, sign the informed consent form (if you agree), and send it to researchers by FAX/scanned PDF/images/mail.
- Step 2: Complete an online demographic survey and submit the form.
- Step 3: In the main task survey form, read each document (click a link to open), think of multiple topics that may describe/explain the main theme of the document, and add those topics into an empty text box under each document link.
- Step 4: Do step 3 for all 30 documents, answer an optional question (if desired), and submit the completed form.

It will take approximately 3 hours to complete the task. You can stop and resume work anytime; however, we encourage finishing the task in at most 2 days. A $30 Amazon gift card will be provided online to you upon completion of the study without partial compensation. If you are interested in this study or have a question, please don't hesitate to contact me at:

Email: seungwon@vt.edu
Phone:(540) 230-8983

FAX: (540) 231-6075
Mailing address:

Dept. of Computer Science

114 McBryde Hall, Mail Code 0106

Virginia Tech, Blacksburg, VA 24061

Thank you,
Sincerely, Seungwon Yang

Ph.D. Candidate

B.3 Informed Consent

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Informed consent form for participants of Research Involving Human Subject

Title of the Project: A Study of Multiple Topic Assignment to Disaster Webpage Content and Inter-Indexer Consistency Computation

Investigators: **Seungwon Yang**

Faculty investigators: Dr. Edward A. Fox

Participant number: _________________

I. PURPOSE OF THIS RESEARCH

The investigators of this study are working on developing an automatic topic identification method, which assigns topics given an electronic textual document. The first purpose of this research is to develop a 'golden standard' test set, which includes documents tagged (i.e., topic-assigned) by multiple human indexers. This 'golden standard' set is used to evaluate our automatic method by comparing results from both approaches.

The second purpose is to investigate inter-indexer consistency. It is well known that human indexers may assign topics differently from each other due to their different background knowledge in the domain of the document, vocabulary size, language proficiency, education level, etc. We would like to measure the consistency among human indexers, and also the consistency between the highest performing human indexers and our automatic approach. It is our hope that our automatic approach will help reduce the human intervention of assigning meaningful topics/tags to textual documents in the era of information inundation.

II. PROCEDURES

Since you were recruited online using the JESSE listserv discussion group, all the communications, sending (links of) forms are conducted by email. Your answers for the demographic questionnaire and the main task form are collected using an online survey system.

Upon receiving your signed informed consent (this form) by FAX/scanned PDF/images/mail, a link to an online demographic survey form is sent to you by email. Once the researchers receive completed demographic survey forms, we will send you a link to an online main task form. The main task form consists of instructions and 30 textual documents, each of which has an empty text box for adding topics.

You will get familiarized with the steps of doing the task by reading the instructions, which is in the top portion of the main task form. Then, you read each of the 30 text documents, find topics that may contribute to the main theme of the document, and add those into an empty text box under each document link. There is an optional question at the bottom of the online task form, which you may answer. In total, the task will take approximately 3 hours. You can stop and resume work anytime during your participation. There is no time limit in the task completion, but we encourage you to finish the tasks in 2 days at most. When you complete the task, you will receive a $30 gift card by online (e.g., Amazon gift card) for compensation.

III. RISKS

You are involved with reading textual documents on a computer screen and typing in several words/phrases as topics for each document. The physical components of the task previously described may result in eye, wrist, or finger fatigue.

If you experience any fatigue during the task, you are free to rest and may continue when you are ready. If the fatigue becomes uncomfortable, you will be allowed to leave the study with no penalty.

IV. BENEFITS OF THIS RESEARCH

You will experience the task of topic assignment, which is relevant to the field of Library and Information Sciences. In addition, you will receive a gift card of $30 upon completion of the study.

V. EXTENT OF ANONYMITY AND CONFIDENTIALITY

This research will assure your confidentiality. We will have your signatures on the Informed Consent document. Faxed, scanned PDFs, images of the signed document, or a paper form are all acceptable. This document will be kept in the computer of a researcher, Seungwon Yang, until his dissertation and related publications are published. Your name will not be released in any publications. After publishing the dissertation and acceptance of related papers, Seungwon Yang will destroy the documents (in electronic forms) from his computer.

Only the researcher, Seungwon Yang, will collect the data. No one other than the researchers will have access to the data. All responses will be coded so as not to include the name of the participant. Your two-digit participant number will only identify you during analysis and any written reports of the research.

This study is being conducted solely for research and development purposes, and the resulting data and interpretations also will be the part of the researcher's academic work. Consistent with these academic purposes, any results would be freely publishable. Again, to protect your identity, personal names will not be used in any published works. We are willing to share drafts of reports with you before submitting them for publication.

VI. COMPENSATION

A $30 gift card is provided upon completion of the study.

VII. FREEDOM TO WITHDRAW

Your participation is voluntary and the decision on whether you wish to participate or not is strictly your own. You may discontinue participation at any time. However, the compensation will be given to those who complete the experiment.

VIII. APPROVAL OF RESEARCH

This research project has been approved by the Institutional Review Board for Research Involving Human Subjects at Virginia Polytechnic Institute and State University and by the Department of Industrial and Systems Engineering (College of Engineering).

IX. PARTICIPANT’S RESPONSIBILITIES

Upon signing this form, I voluntarily agree to participate in this study. I have no restrictions to my participation in this study. As a participant, I may withdraw from this experiment at any time without penalty. I agree to abide by the rules of this project.

X. PARTICIPANT’S PERMISSION

I have read and understand the Informed Consent and conditions of this study. All of my questions have been answered. I agree to participate in this experiment.

______________________________ ______________________

Participant’s Signature Date

Should I have any questions about the research or its conduct, I may contact:

Faculty Advisor: Edward A. Fox

Professor, Department of Computer Science, Virginia Tech

Email: fox@vt.edu

Phone: (540) 231-5113 FAX: (540) 231-6075
Mailing address:

Dept. of Computer Science

114 McBryde Hall, Mail Code 0106

Virginia Tech, Blacksburg, VA 24061

Investigator: Seungwon Yang

Ph.D. student, Department of Computer Science, Virginia Tech

Email: seungwon@vt.edu

Phone: (540) 230-8983 FAX: (540) 231-6075
Mailing address:

Dept. of Computer Science

114 McBryde Hall, Mail Code 0106

Virginia Tech, Blacksburg, VA 24061

Dr. David M. Moore, Email : moored@vt.edu Phone: (540) 231-4991

Chair, IRB

B.4 Study Procedure

Manual Topic Assignment Study Procedure

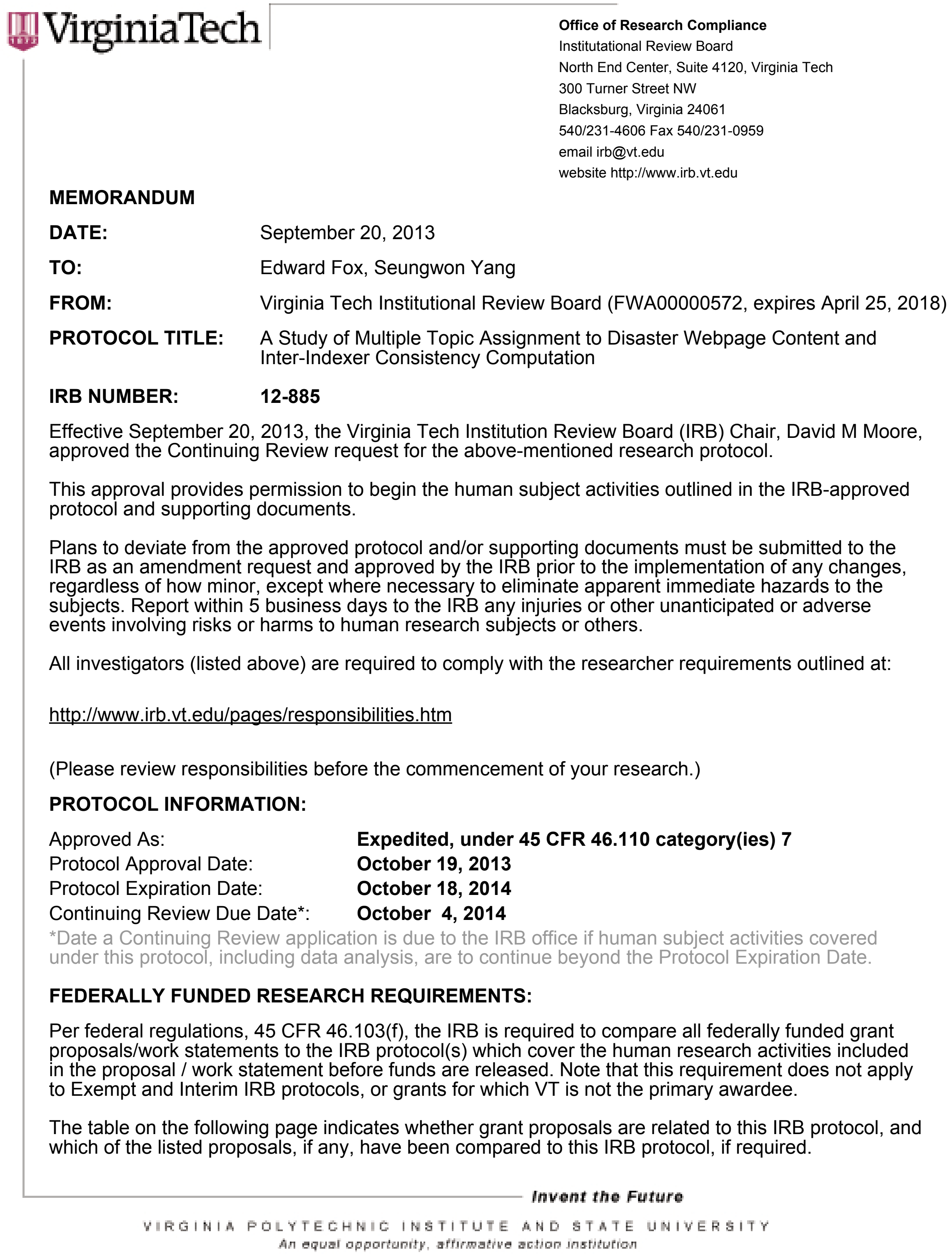
For Researchers

1. An informed consent form is emailed to participants as an attachment.
2. Upon receiving his/her signed informed consent form by FAX/scanned PDF/images/mail, an online demographic survey form is sent to the participant by email.
3. A link of an online main task form, which contains links to 30 textual documents with empty text boxes to enter topics, is sent to the participant by email.
4. Upon receiving the results from participants, $30 gift cards are sent to them.

For Participants

1. Print and sign the informed consent form. Then, send the signed consent form to the researchers by FAX/scanned PDF/images/mail (select one).
2. Complete an online demographic form. Then submit the completed form online.
3. Open each of 30 document links in the main task form, read the document, and add multiple topics in an empty box provided. You can also answer an optional question. Then submit the completed results online.

B.5 IRB Approval Letter



VirginiaTech

Office of Research Compliance
Institutational Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE:	September 20, 2013
TO:	Edward Fox, Seungwon Yang
FROM:	Virginia Tech Institutional Review Board (FWA00000572, expires April 25, 2018)
PROTOCOL TITLE:	A Study of Multiple Topic Assignment to Disaster Webpage Content and Inter-Indexer Consistency Computation
IRB NUMBER:	**12-885**

Effective September 20, 2013, the Virginia Tech Institution Review Board (IRB) Chair, David M Moore, approved the Continuing Review request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As:	**Expedited, under 45 CFR 46.110 category(ies) 7**
Protocol Approval Date:	**October 19, 2013**
Protocol Expiration Date:	**October 18, 2014**
Continuing Review Due Date*:	**October 4, 2014**

*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Invent the Future

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
An equal opportunity, affirmative action institution

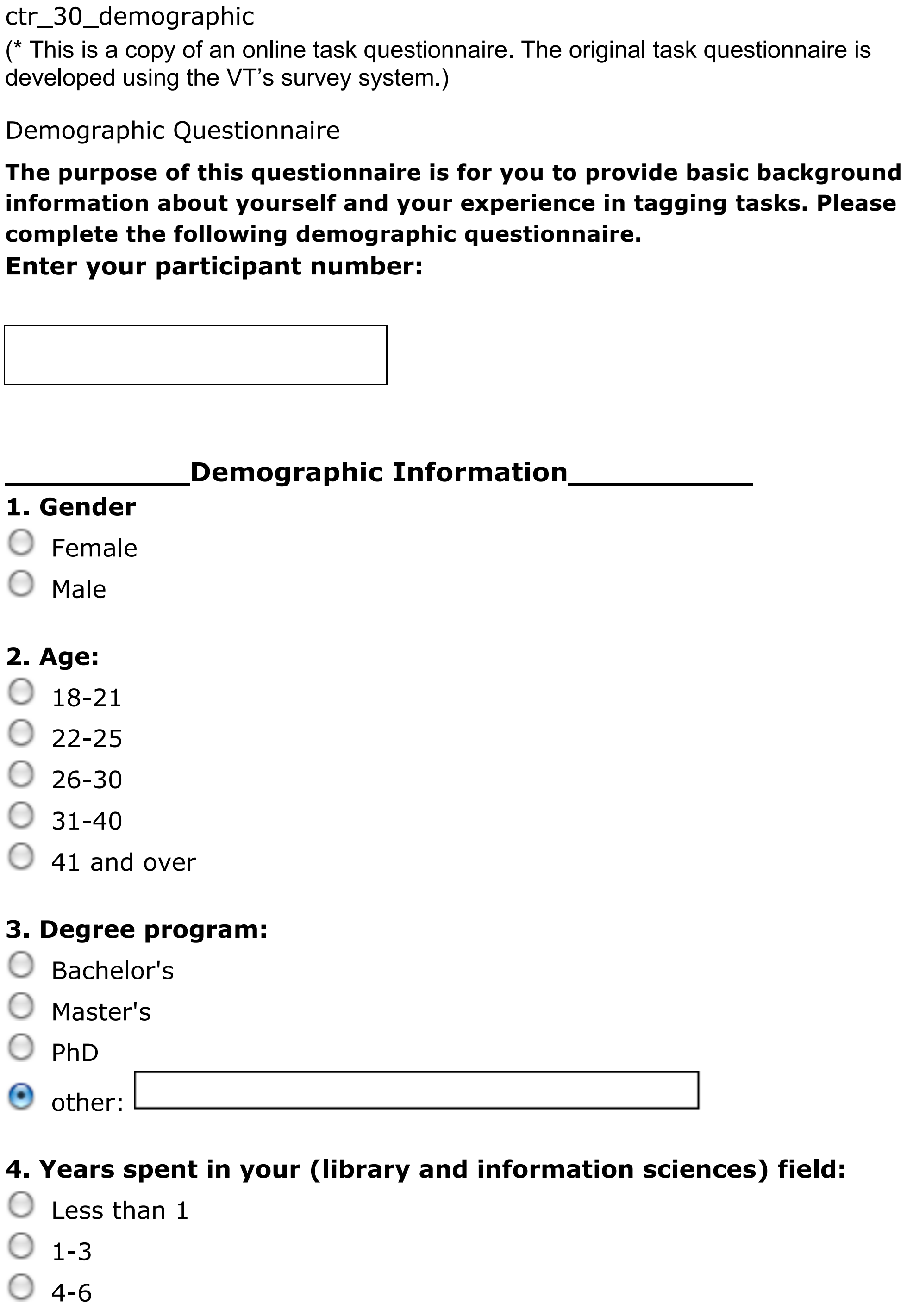
Date*	OSP Number	Sponsor	Grant Comparison Conducted?

* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this IRB protocol is to cover any other grant proposals, please contact the IRB office (irbadmin@vt.edu) immediately.

B.6 Demographic Questionnaire

Note: The demographic questionnaire below was used to develop CTR_30 data set. Then, the same form was used to develop VARIOUS_30 data set. Thus, we only present the form for CTR_30 here.



ctr_30_demographic

(* This is a copy of an online task questionnaire. The original task questionnaire is developed using the VT's survey system.)

Demographic Questionnaire

The purpose of this questionnaire is for you to provide basic background information about yourself and your experience in tagging tasks. Please complete the following demographic questionnaire.

Enter your participant number:

____________Demographic Information____________

1. Gender

○ Female

○ Male

2. Age:

○ 18-21

○ 22-25

○ 26-30

○ 31-40

○ 41 and over

3. Degree program:

○ Bachelor's

○ Master's

○ PhD

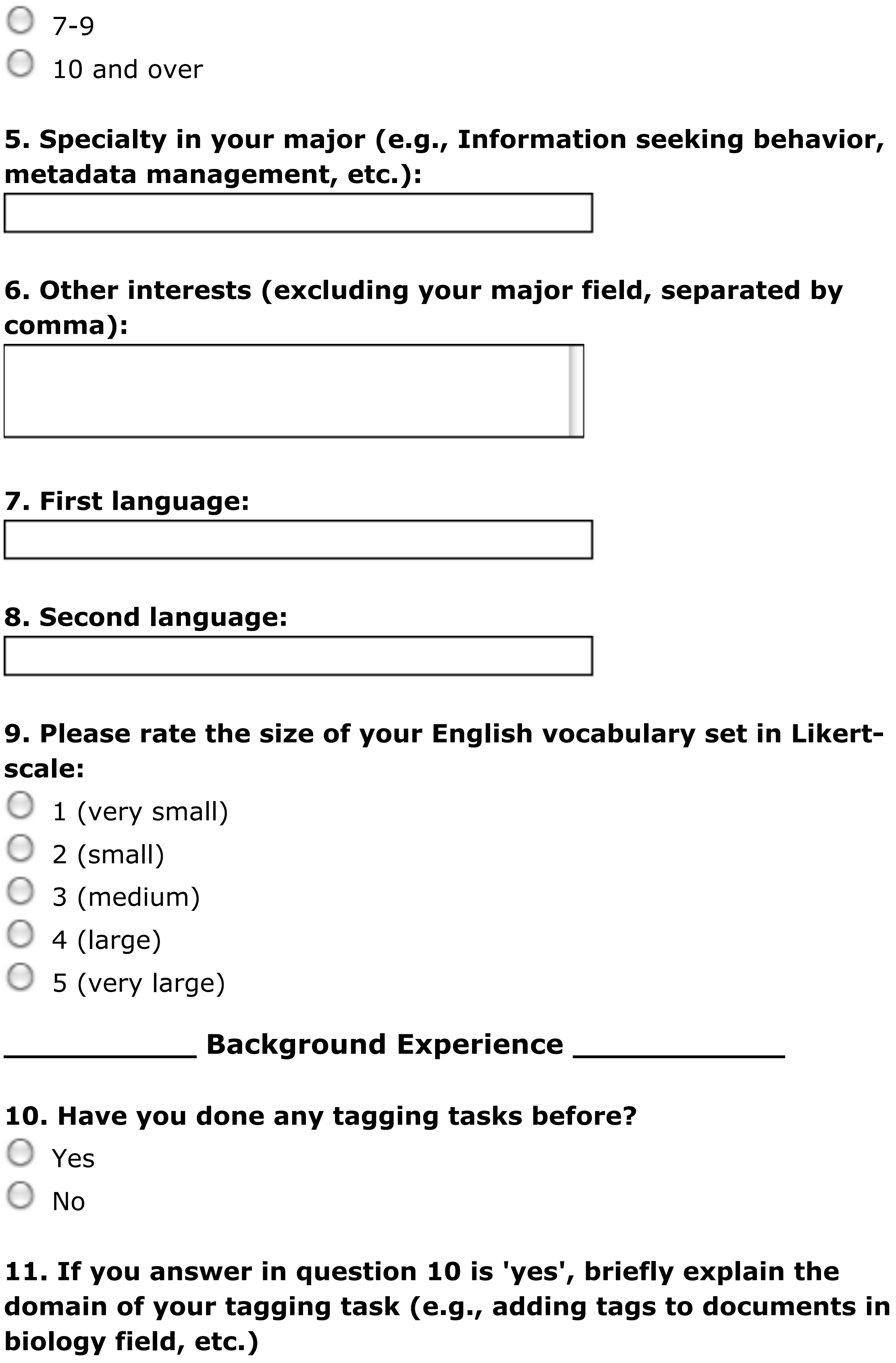
● other:

4. Years spent in your (library and information sciences) field:

○ Less than 1

○ 1-3

○ 4-6



○ 7-9

○ 10 and over

5. Specialty in your major (e.g., Information seeking behavior, metadata management, etc.):

6. Other interests (excluding your major field, separated by comma):

7. First language:

8. Second language:

9. Please rate the size of your English vocabulary set in Likert-scale:

○ 1 (very small)

○ 2 (small)

○ 3 (medium)

○ 4 (large)

○ 5 (very large)

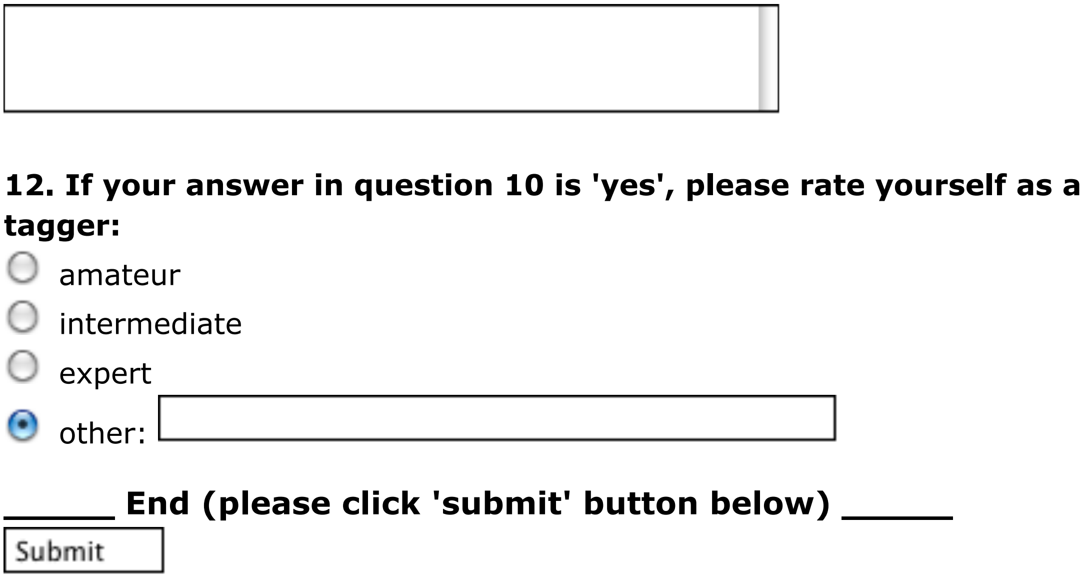
___________ Background Experience ____________

10. Have you done any tagging tasks before?

○ Yes

○ No

11. If you answer in question 10 is 'yes', briefly explain the domain of your tagging task (e.g., adding tags to documents in biology field, etc.)



12. If your answer in question 10 is 'yes', please rate yourself as a tagger:

○ amateur

○ intermediate

○ expert

● other:

______ End (please click 'submit' button below) ______

B.7 Task Questionnaire

Note: The task questionnaire below was used to develop CTR_30 data set with human-assigned tags. Then, the same form, with modifications for the links to documents in VARIOUS_30, was used to develop VARIOUS_30 data set. Thus, we only present the form for CTR_30 here.

ctr_30_topic_assignment

(* this document is a copy of an online task form. The original form is developed using the VT's survey system)

Main Tasks

Instructions

1. **Click a document link. A PDF document will open.**
2. **Read the document thoroughly to understand its main theme.**
3. **Find several topic words that might describe or explain the theme.**
4. **Add those topic words to an empty text box. Separate words by a comma. More relevant words should come first.**
5. **Do steps 1-4 for all 30 documents.**
6. **You may answer an optional question.**
7. **Submit the completed form by pressing the 'Submit' button at the bottom of the form.**

1. Click the link: Document_1

Read the document and add topics below separated by comma:

2. Click the link: Document_2

Read the document and add topics below separated by comma:

3. Click the link: Document_3

Read the document and add topics below separated by comma:

4. Click the link: Document_4

Read the document and add topics below separated by comma:

5. Click the link: Document_5

Read the document and add topics below separated by comma:

6. Click the link: Document_6

Read the document and add topics below separated by comma:

7. Click the link: Document_7

Read the document and add topics below separated by comma:

8. Click the link: Document_8

Read the document and add topics below separated by comma:

9. Click the link: Document_9

Read the document and add topics below separated by comma:

10. Click the link: Document_10

Read the document and add topics below separated by comma:

11. Click the link: Document_11

Read the document and add topics below separated by comma:

12. Click the link: Document_12

Read the document and add topics below separated by comma:

13. Click the link: Document_13

Read the document and add topics below separated by comma:

14. Click the link: Document_14

Read the document and add topics below separated by comma:

15. Click the link: Document_15

Read the document and add topics below separated by comma:

16. Click the link: Document_16

Read the document and add topics below separated by comma:

17. Click the link: Document_17

Read the document and add topics below separated by comma:

18. Click the link: Document_18

Read the document and add topics below separated by comma:

19. Click the link: Document_19

Read the document and add topics below separated by comma:

20. Click the link: Document_20

Read the document and add topics below separated by comma:

21. Click the link: Document_21

Read the document and add topics below separated by comma:

22. Click the link: Document_22

Read the document and add topics below separated by comma:

23. Click the link: Document_23

Read the document and add topics below separated by comma:

24. Click the link: Document_24

Read the document and add topics below separated by comma:

25. Click the link: Document_25

Read the document and add topics below separated by comma:

26. Click the link: Document_26

Read the document and add topics below separated by comma:

27. Click the link: Document_27

Read the document and add topics below separated by comma:

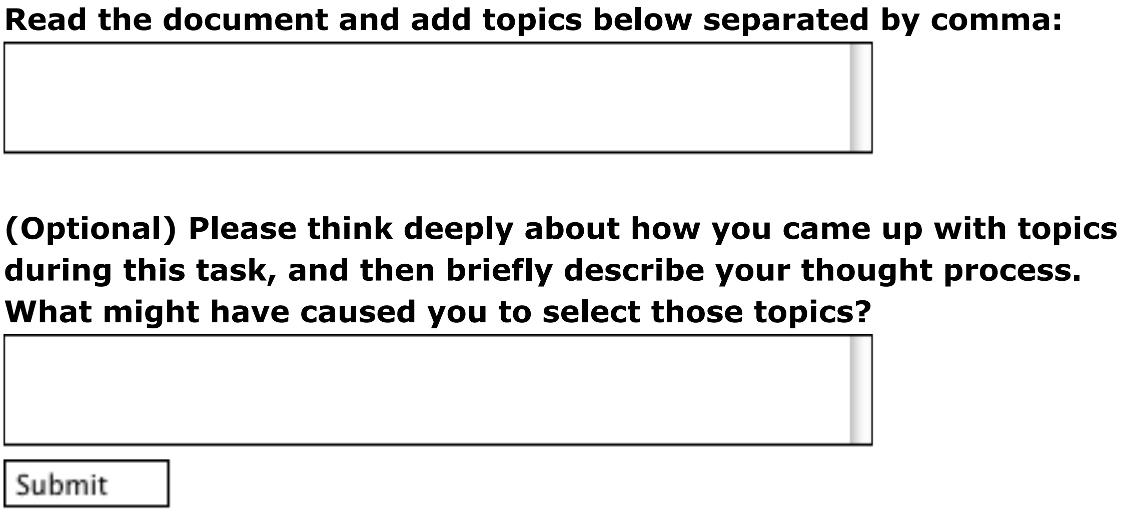
28. Click the link: Document_28

Read the document and add topics below separated by comma:

29. Click the link: Document_29

Read the document and add topics below separated by comma:

30. Click the link: Document_30



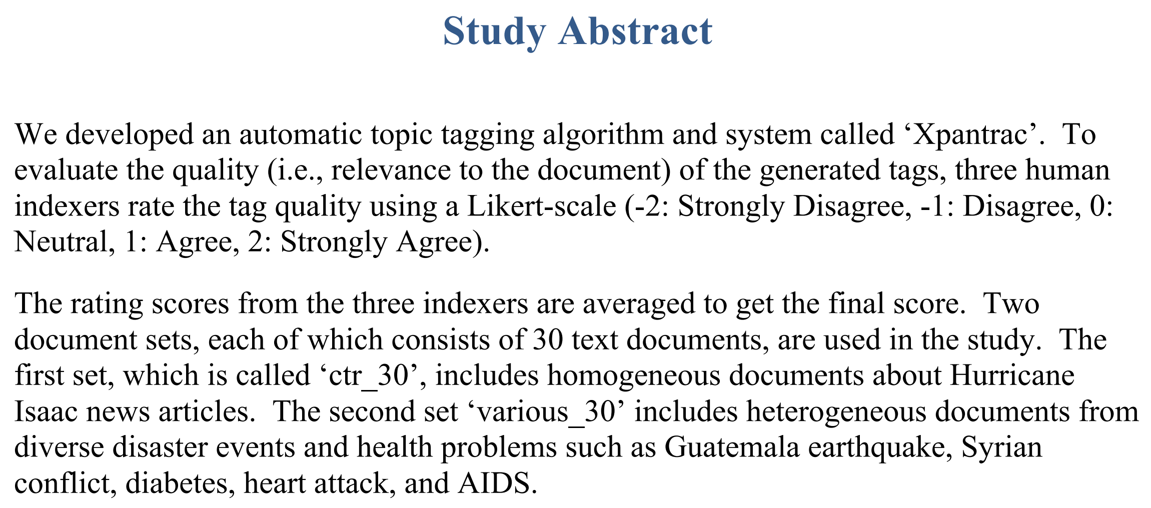
Read the document and add topics below separated by comma:

(Optional) Please think deeply about how you came up with topics during this task, and then briefly describe your thought process. What might have caused you to select those topics?

Appendix C. Topic Relevance Rating Study

Resources used for a topic relevance rating study, which involved multiple human topic raters, are presented. The type of the documents includes the study abstract, a participant recruitment letter, a description of study procedures, an informed consent form, an IRB approval letter, and a task questionnaire.

C.1 Study Abstract



Study Abstract

We developed an automatic topic tagging algorithm and system called ‘Xpantrac’. To evaluate the quality (i.e., relevance to the document) of the generated tags, three human indexers rate the tag quality using a Likert-scale (-2: Strongly Disagree, -1: Disagree, 0: Neutral, 1: Agree, 2: Strongly Agree).

The rating scores from the three indexers are averaged to get the final score. Two document sets, each of which consists of 30 text documents, are used in the study. The first set, which is called ‘ctr_30’, includes homogeneous documents about Hurricane Isaac news articles. The second set ‘various_30’ includes heterogeneous documents from diverse disaster events and health problems such as Guatemala earthquake, Syrian conflict, diabetes, heart attack, and AIDS.

C.2 Participant Recruitment Letter

Participant Recruitment Letter

Title: Investigating the Quality of Automatically Generated Tags for Disaster Webpages

The recruiting criteria include that the participant is not a minor and s/he is in a graduate program at a university. Participants individually rate tags attached for disaster webpage content using the Likert-scale (-2: Strongly Disagree, -1: Disagree, 0: Neutral, 1: Agree, 2: Strongly Agree).

There are two data sets with attached tags. The first one (i.e., ctr_30) consists of 30 documents with homogeneous content (i.e., single disaster event). The second one (i.e., various_30) consists of 30 text documents with heterogeneous content.

The procedures are listed below:

1. Print and sign the informed consent form. Then, send the signed consent form to the researcher by FAX/scanned PDF/images/mail (select one).
2. Click a link in an email from the researcher to open the first task form.
3. Enter your participant number.
4. Open a document link and read the document to understand the main theme/topics.
5. Rate each tag associated with the document by choosing from a Likert-scale (-2: Strongly Disagree, -1: Disagree, 0: Neutral, 1: Agree, 2: Strongly Agree).
6. Repeat steps 4 and 5 to complete the task for all 30 documents in the first task form.
7. You may answer an optional question at the bottom of the task form.
8. Submit the form by clicking the ‘submit’ button at the bottom of the form.
9. Click a link in an email from the researcher to open the second task form.
10. Repeat steps 3 to 8 for the second task form.

There are about 9-20 tags for each document. It will take approximately 3-4 hours to complete both data sets. You can stop and resume work anytime; however, we encourage finishing the task in at most 2 days. A total of $40 Amazon gift card will be provided online to you **upon completion** of both data sets. You can complete only one data set and receive $20, however, there is no partial compensation for a portion of a data set. If you are interested in this study or have a question, please don't hesitate to contact me at:

Email: seungwon@vt.edu Phone:(540) 230-8983
FAX: (540) 231-6075

Dept. of Computer Science
114 McBryde Hall, Mail Code 0106
Virginia Tech, Blacksburg, VA 24061

Thank you,
Sincerely, Seungwon Yang
Ph.D. Candidate

C.3 Informed Consent

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Informed consent form for participants of Research Involving Human Subject

Title of the Project: Investigating the Quality of Automatically Generated Tags for Disaster Webpages

Investigators: **Seungwon Yang**

Faculty investigators: Dr. Edward A. Fox

Participant number: _________________

I. PURPOSE OF THIS RESEARCH

The investigators of this study are working on developing an automatic topic identification method, which assigns topics given an electronic textual document. The purpose of this study is to rate the automatically generated tags, by three human indexers. The rating scores are used to evaluate our automatic topic tagging system.

II. PROCEDURES

All the communications, i.e., sending (links of) forms are conducted by email. Your answers for the main task forms are collected using an online survey system.

Upon receiving your signed informed consent (this form) by FAX/scanned PDF/images/mail, we will send you a link to the first task form. It consists of instructions, 30 textual documents, and about 9-21 automatically generated topic tags.

You will get familiarized with the steps of doing the task by reading the instructions, which is in the top portion of the task form. Then, you read each of the 30 text documents, and rate tags associated with each document by selecting a Likert-scale score (-2: Strongly Disagree, -1: Disagree, 0: Neutral, 1: Agree, 2: Strongly Agree). There is an optional question at the bottom of the online task form that you may answer. In total, the task will take approximately 2-4 hours. You can stop and resume work anytime during your participation. There is no time limit for the task completion, but we encourage you to finish the tasks in 2 days at most. If you complete the task before the end of our study, you will receive a $20 Amazon gift card by email as compensation.

Upon completion of the first task form, we will send you a link to the second task form. When you complete the second task form, you will receive another $20 Amazon gift card by email. Therefore, you can receive a maximum of $40 for two completed tasks.

III. RISKS

You are involved with reading textual documents on a computer screen and rating tag quality by clicking one of the score options. The physical components of the task previously described may result in eye, wrist, or finger fatigue.

If you experience any fatigue during the task, you are free to rest and may continue when you are ready. If the fatigue becomes uncomfortable, you will be allowed to leave the study with no penalty.

IV. BENEFITS OF THIS RESEARCH

You will experience the task of tag rating, which is relevant to the field of Computational Linguistics and Text Mining. In addition, you will receive a gift card of $20 for each completed task form ($40 for both completed tasks).

V. EXTENT OF ANONYMITY AND CONFIDENTIALITY

This research will assure your confidentiality. We will have your signatures on the Informed Consent document. Faxed, scanned PDFs, images of the signed document, or a paper form are all acceptable. This document will be kept in the computer of a researcher, Seungwon Yang, until his dissertation and related publications are published. Your name will not be released in any publications. After publishing the dissertation and final acceptance of related papers, Seungwon Yang will destroy the documents (in electronic forms) from his computer.

Only the researcher, Seungwon Yang, will collect the data. No one other than the researchers will have access to the data. All responses will be coded so as not to include the name of the participant. Your two-digit participant number only will identify you during analysis and any written reports of the research.

This study is being conducted solely for research and development purposes, and the resulting data and interpretations also will be the part of the researcher's academic work. Consistent with these academic purposes, any results would be freely publishable. Again, to protect your identity, personal names will not be used in any published works. We are willing to share drafts of reports with you before submitting them for publication.

VI. COMPENSATION

A $20 gift card is provided for each completed task form. Therefore, a maximum of $40 gift card can be provided for the two completed task forms.

VII. FREEDOM TO WITHDRAW

Your participation is voluntary and the decision on whether you wish to participate or not is strictly your own. You may discontinue participation at any time. However, the compensation will be given to those who complete the experiment.

VIII. APPROVAL OF RESEARCH

This research project has been approved by the Institutional Review Board for Research Involving Human Subjects at Virginia Polytechnic Institute and State University and by the Department of Computer Science (College of Engineering).

IX. PARTICIPANT'S RESPONSIBILITIES

Upon signing this form, I voluntarily agree to participate in this study. I have no restrictions to my participation in this study. As a participant, I may withdraw from this experiment at any time without penalty. I agree to abide by the rules of this project.

X. PARTICIPANT'S PERMISSION

I have read and understand the Informed Consent and conditions of this study. All of my questions have been answered. I agree to participate in this experiment.

_____________________________ Participant's Signature

_____________________ Date

Should I have any questions about the research or its conduct, I may contact:

Faculty Advisor: Edward A. Fox
Professor, Department of Computer Science, Virginia Tech
Email: fox@vt.edu
Phone: (540) 231-5113 FAX: (540) 231-6075
Mailing address:
Dept. of Computer Science
114 McBryde Hall, Mail Code 0106
Virginia Tech, Blacksburg, VA 24061

Investigator: Seungwon Yang
Ph.D. student, Department of Computer Science, Virginia Tech
Email: seungwon@vt.edu
Phone: (540) 230-8983 FAX: (540) 231-6075
Mailing address:
Dept. of Computer Science
114 McBryde Hall, Mail Code 0106
Virginia Tech, Blacksburg, VA 24061

Dr. David M. Moore, Email : moored@vt.edu Phone: (540) 231-4991
Chair, IRB

C.4 Study Procedure

Tag Quality Rating Study Procedure

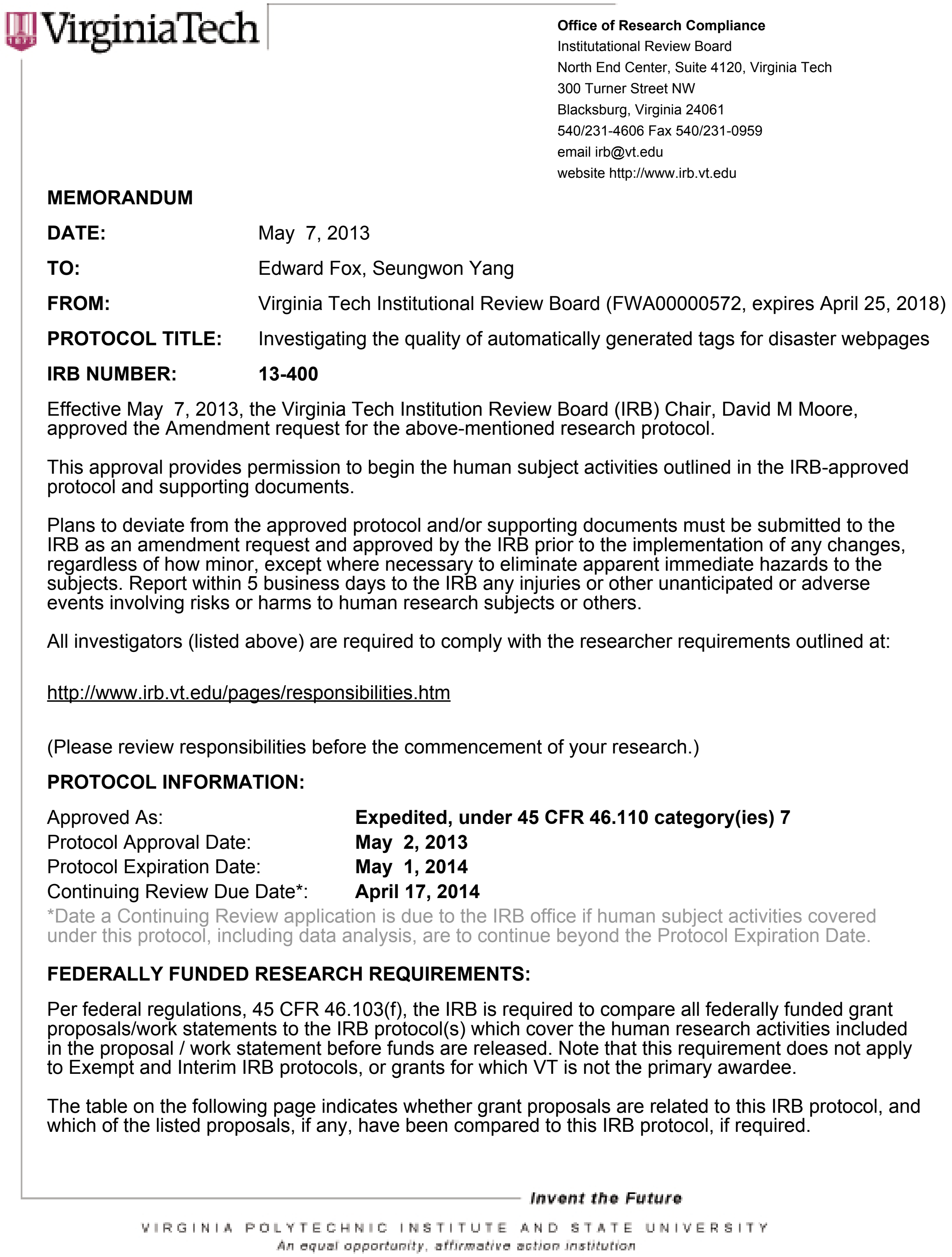
For Researchers

1. An informed consent form is emailed to participants as an attachment.
2. Upon receiving his/her signed informed consent form by FAX/scanned PDF/images/mail, a link of the first task form is sent to the participant by email.
3. Upon receiving the result of the first task, a link for the second task form is sent to the participant by email.
4. Whenever a participant completes a task form, a $20 Amazon gift card is sent to her/him by email (for two completed tasks, they receive a total of $40).

For Participants

1. Print and sign the informed consent form. Then, send the signed consent form to the researcher by FAX/scanned PDF/images/mail (select one).
2. Click a link to open the first task form provided by the researcher.
3. Type in your participant number.
4. Open a document link and read the document to understand the main theme/topics.
5. Rate each tag associated with the document by clicking one of Likert-scale scores (0: non-relevant, 1: weakly relevant, 2: relevant, 3: strongly relevant).
6. Repeat steps 4 and 5 to complete the task for all 30 documents.
7. You may answer an optional question at the bottom of the task form.
8. Submit the form by clicking the ‘submit’ button at the bottom of the form.
9. Repeat steps from 2 to 8 for the second task form.

C.5 IRB Approval Letter



VirginiaTech

Office of Research Compliance
Institutational Review Board
North End Center, Suite 4120, Virginia Tech
300 Turner Street NW
Blacksburg, Virginia 24061
540/231-4606 Fax 540/231-0959
email irb@vt.edu
website http://www.irb.vt.edu

MEMORANDUM

DATE: May 7, 2013

TO: Edward Fox, Seungwon Yang

FROM: Virginia Tech Institutional Review Board (FWA00000572, expires April 25, 2018)

PROTOCOL TITLE: Investigating the quality of automatically generated tags for disaster webpages

IRB NUMBER: **13-400**

Effective May 7, 2013, the Virginia Tech Institution Review Board (IRB) Chair, David M Moore, approved the Amendment request for the above-mentioned research protocol.

This approval provides permission to begin the human subject activities outlined in the IRB-approved protocol and supporting documents.

Plans to deviate from the approved protocol and/or supporting documents must be submitted to the IRB as an amendment request and approved by the IRB prior to the implementation of any changes, regardless of how minor, except where necessary to eliminate apparent immediate hazards to the subjects. Report within 5 business days to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.

All investigators (listed above) are required to comply with the researcher requirements outlined at:

http://www.irb.vt.edu/pages/responsibilities.htm

(Please review responsibilities before the commencement of your research.)

PROTOCOL INFORMATION:

Approved As: **Expedited, under 45 CFR 46.110 category(ies) 7**
Protocol Approval Date: **May 2, 2013**
Protocol Expiration Date: **May 1, 2014**
Continuing Review Due Date*: **April 17, 2014**
*Date a Continuing Review application is due to the IRB office if human subject activities covered under this protocol, including data analysis, are to continue beyond the Protocol Expiration Date.

FEDERALLY FUNDED RESEARCH REQUIREMENTS:

Per federal regulations, 45 CFR 46.103(f), the IRB is required to compare all federally funded grant proposals/work statements to the IRB protocol(s) which cover the human research activities included in the proposal / work statement before funds are released. Note that this requirement does not apply to Exempt and Interim IRB protocols, or grants for which VT is not the primary awardee.

The table on the following page indicates whether grant proposals are related to this IRB protocol, and which of the listed proposals, if any, have been compared to this IRB protocol, if required.

Invent the Future

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY
An equal opportunity, affirmative action institution

Date*	OSP Number	Sponsor	Grant Comparison Conducted?

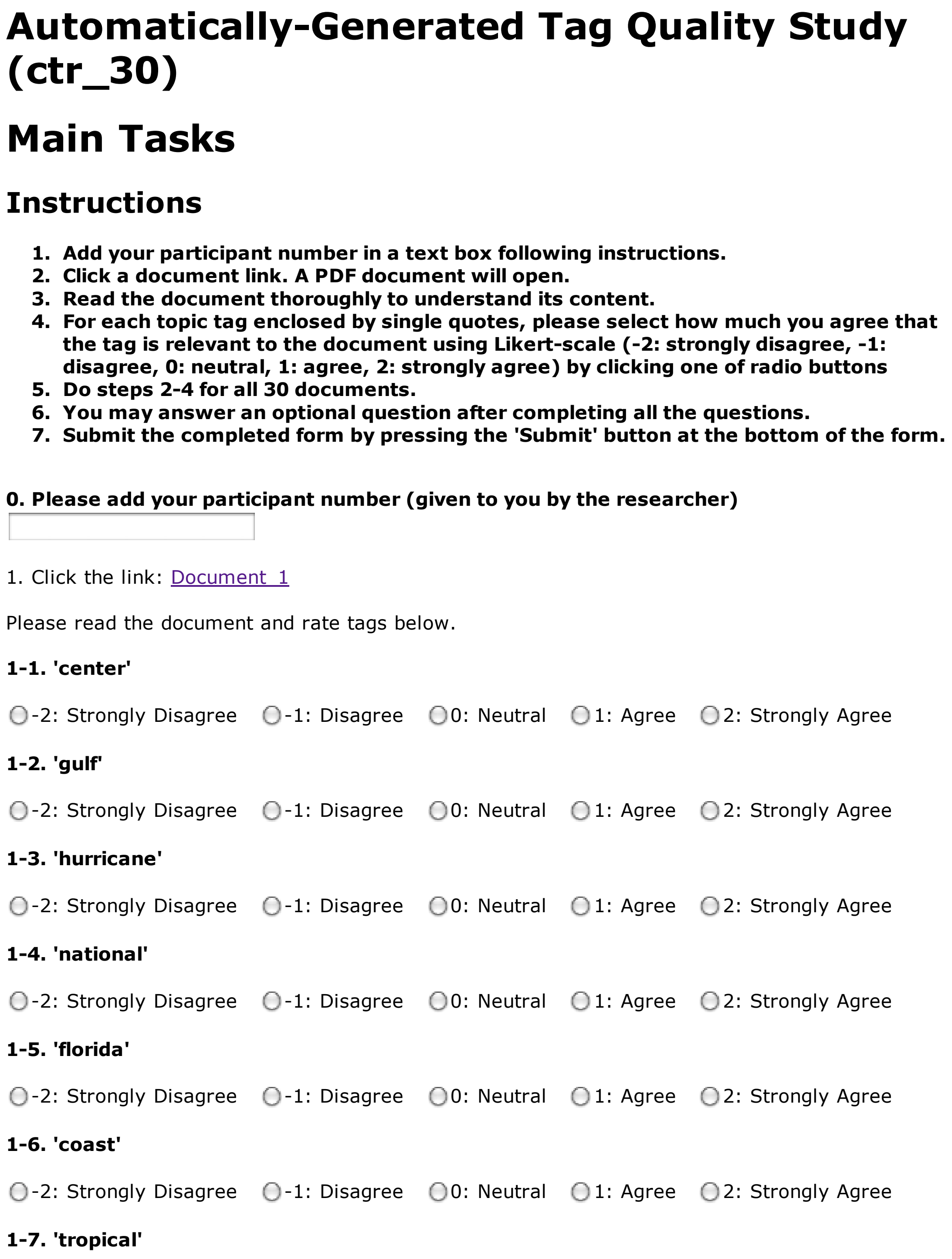
* Date this proposal number was compared, assessed as not requiring comparison, or comparison information was revised.

If this IRB protocol is to cover any other grant proposals, please contact the IRB office (irbadmin@vt.edu) immediately.

C.6 Task Questionnaire

Due to the length of the two task questionnaires (40 pages each), only the first3 pages are displayed here. Entire questionnaires can be downloaded from a repository at: https://github.com/seungwonyang/dissertation

C.6.1 CTR_30 Topic Rating



Automatically-Generated Tag Quality Study (ctr_30)

Main Tasks

Instructions

1. **Add your participant number in a text box following instructions.**
2. **Click a document link. A PDF document will open.**
3. **Read the document thoroughly to understand its content.**
4. **For each topic tag enclosed by single quotes, please select how much you agree that the tag is relevant to the document using Likert-scale (-2: strongly disagree, -1: disagree, 0: neutral, 1: agree, 2: strongly agree) by clicking one of radio buttons**
5. **Do steps 2-4 for all 30 documents.**
6. **You may answer an optional question after completing all the questions.**
7. **Submit the completed form by pressing the 'Submit' button at the bottom of the form.**

0. Please add your participant number (given to you by the researcher)

1. Click the link: Document_1

Please read the document and rate tags below.

1-1. 'center'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-2. 'gulf'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-3. 'hurricane'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-4. 'national'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

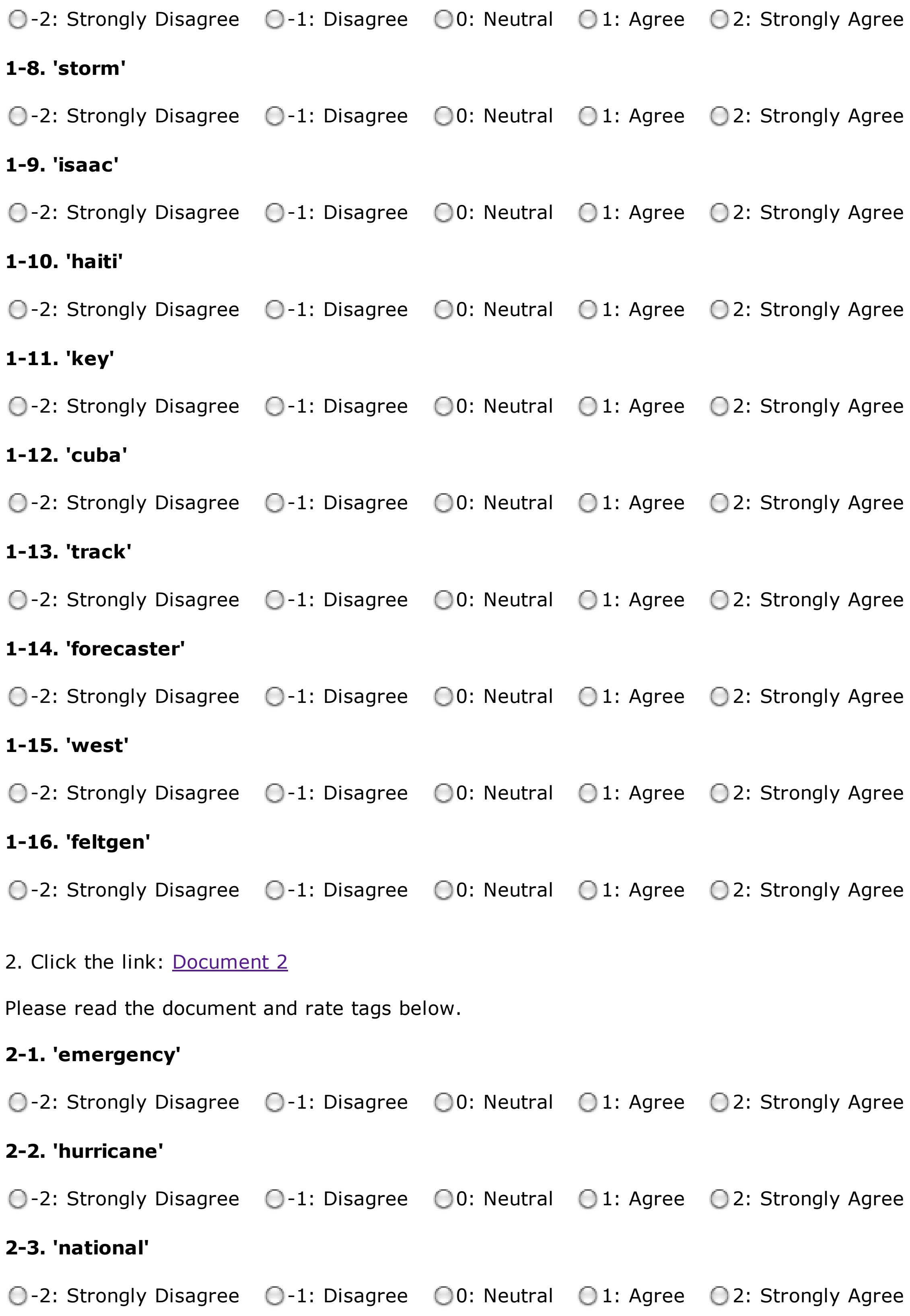
1-5. 'florida'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-6. 'coast'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-7. 'tropical'



○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-8. 'storm'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-9. 'isaac'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-10. 'haiti'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-11. 'key'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-12. 'cuba'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-13. 'track'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-14. 'forecaster'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-15. 'west'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-16. 'feltgen'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2. Click the link: Document 2

Please read the document and rate tags below.

2-1. 'emergency'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-2. 'hurricane'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-3. 'national'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-4. 'official'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-5. 'florida'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-6. 'tropical'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-7. 'scott'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-8. 'storm'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-9. 'isaac'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-10. 'sandy'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-11. 'republican'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-12. 'mph'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-13. 'harris'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-14. 'west'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-15. 'central'

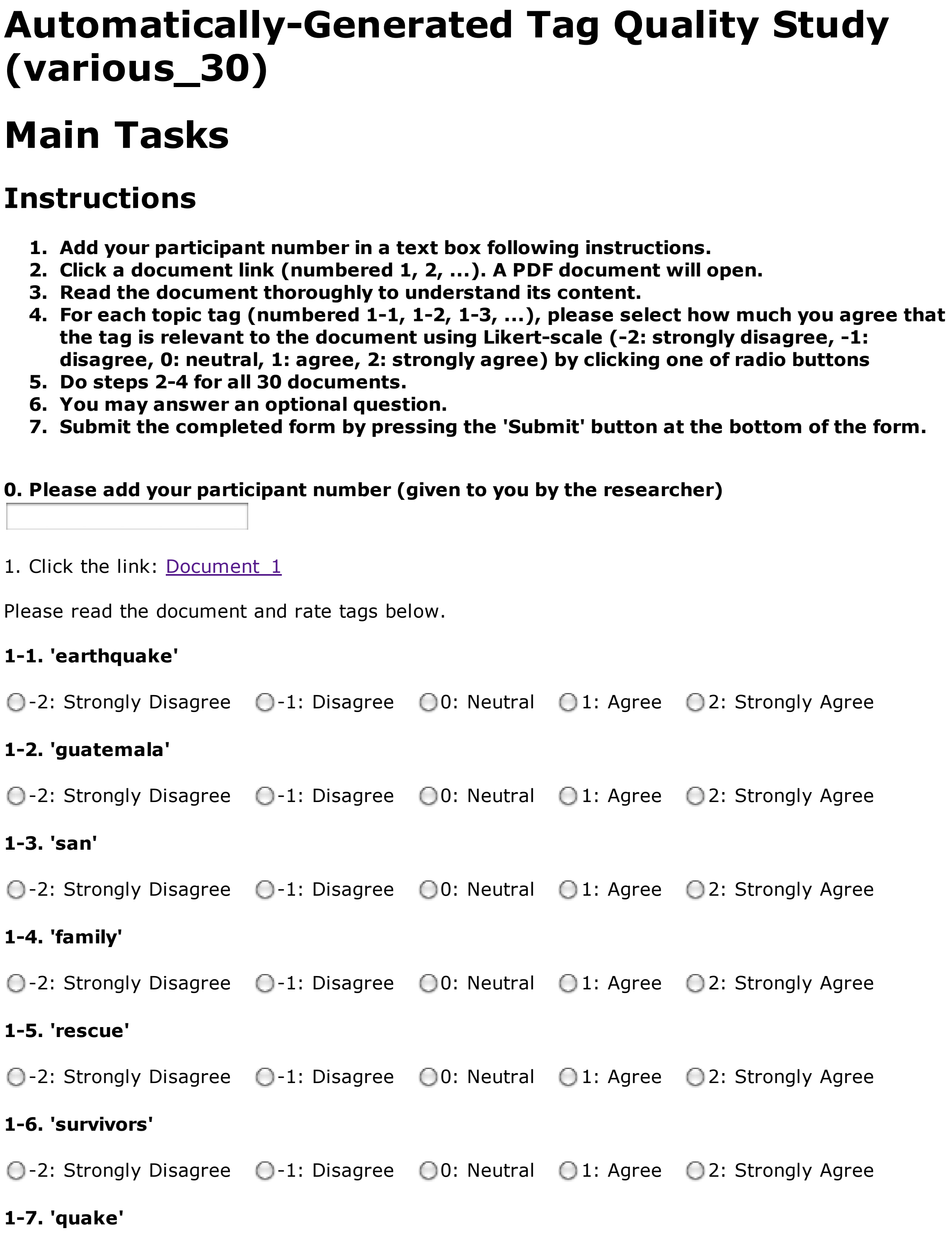
○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-16. 'convention'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-17. 'hispaniola'

C.6.2 VARIOUS_30 Topic Rating



Automatically-Generated Tag Quality Study (various_30)

Main Tasks

Instructions

1. **Add your participant number in a text box following instructions.**
2. **Click a document link (numbered 1, 2, ...). A PDF document will open.**
3. **Read the document thoroughly to understand its content.**
4. **For each topic tag (numbered 1-1, 1-2, 1-3, ...), please select how much you agree that the tag is relevant to the document using Likert-scale (-2: strongly disagree, -1: disagree, 0: neutral, 1: agree, 2: strongly agree) by clicking one of radio buttons**
5. **Do steps 2-4 for all 30 documents.**
6. **You may answer an optional question.**
7. **Submit the completed form by pressing the 'Submit' button at the bottom of the form.**

0. Please add your participant number (given to you by the researcher)

1. Click the link: Document_1

Please read the document and rate tags below.

1-1. 'earthquake'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-2. 'guatemala'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-3. 'san'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-4. 'family'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

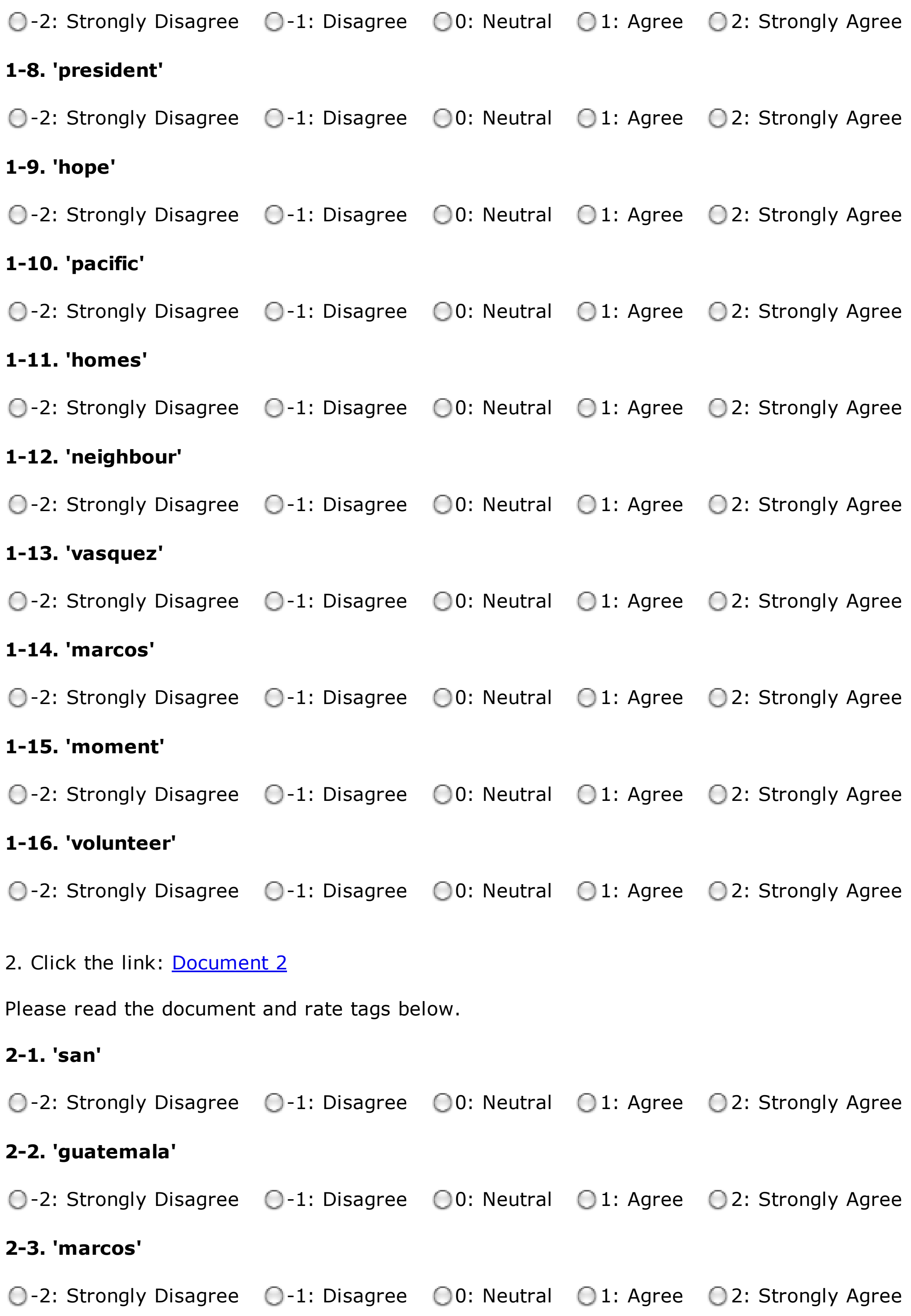
1-5. 'rescue'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-6. 'survivors'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-7. 'quake'



○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-8. 'president'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-9. 'hope'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-10. 'pacific'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-11. 'homes'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-12. 'neighbour'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-13. 'vasquez'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-14. 'marcos'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-15. 'moment'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

1-16. 'volunteer'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2. Click the link: Document 2

Please read the document and rate tags below.

2-1. 'san'

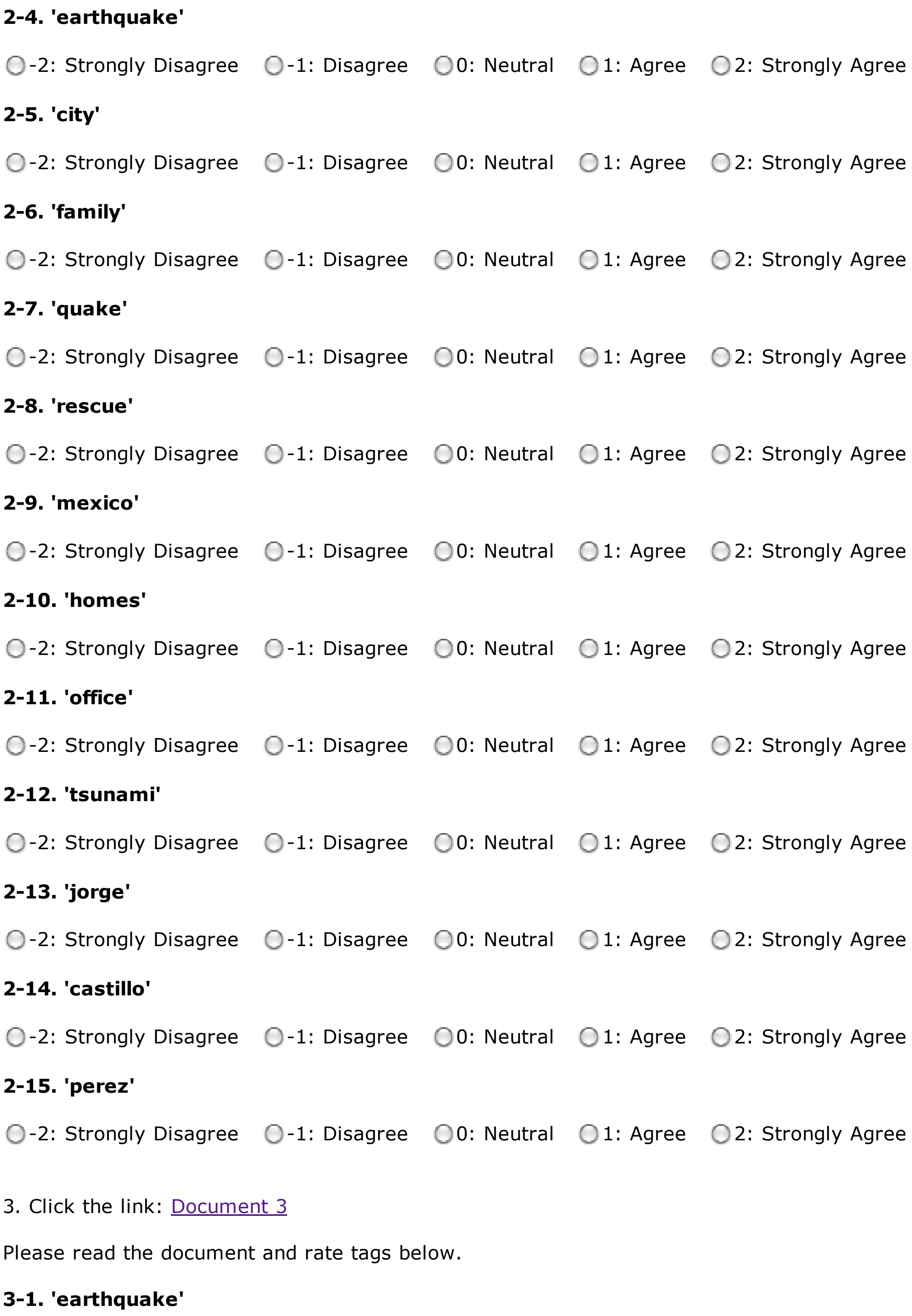
○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-2. 'guatemala'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-3. 'marcos'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree



2-4. 'earthquake'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-5. 'city'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-6. 'family'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-7. 'quake'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-8. 'rescue'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-9. 'mexico'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-10. 'homes'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-11. 'office'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-12. 'tsunami'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-13. 'jorge'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-14. 'castillo'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

2-15. 'perez'

○-2: Strongly Disagree ○-1: Disagree ○0: Neutral ○1: Agree ○2: Strongly Agree

3. Click the link: Document 3

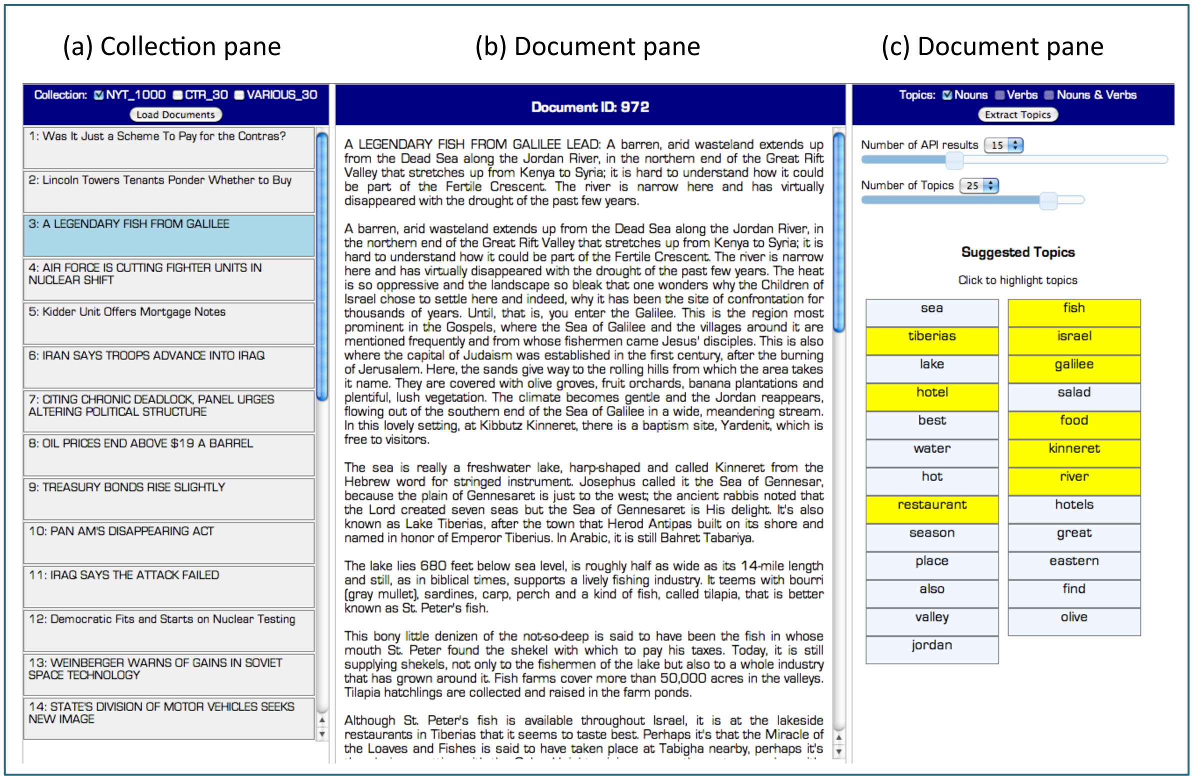
Please read the document and rate tags below.

3-1. 'earthquake'

Appendix D. Xpantrac User Interface Tutorial

1. Overall user interface

The user interface (UI) of the prototype Xpantrac (the name came from eXpansion + extraction algorithm) system is shown in Figure D.1. The interface has three divided sections, namely the Collection pane, the Document pane, and the Topics pane. Users will work on each pane in the order of (a), (b), and (c) to extract topic tags.


(a) Collection pane
(b) Document pane
(c) Document pane
Collection: ☑ NYT_1000 □ CTR_30 □ VARIOUS_30
Load Documents
Topics: ☑ Nouns □ Verbs □ Nouns & Verbs
Extract Topics

1: Was It Just a Scheme To Pay for the Contras?

2: Lincoln Towers Tenants Ponder Whether to Buy

3: A LEGENDARY FISH FROM GALILEE

4: AIR FORCE IS CUTTING FIGHTER UNITS IN NUCLEAR SHIFT

5: Kidder Unit Offers Mortgage Notes

6: IRAN SAYS TROOPS ADVANCE INTO IRAQ

7: CITING CHRONIC DEADLOCK, PANEL URGES ALTERING POLITICAL STRUCTURE

8: OIL PRICES END ABOVE $19 A BARREL

9: TREASURY BONDS RISE SLIGHTLY

10: PAN AM'S DISAPPEARING ACT

11: IRAQ SAYS THE ATTACK FAILED

12: Democratic Fits and Starts on Nuclear Testing

13: WEINBERGER WARNS OF GAINS IN SOVIET SPACE TECHNOLOGY

14: STATE'S DIVISION OF MOTOR VEHICLES SEEKS NEW IMAGE

Document ID: 972

A LEGENDARY FISH FROM GALILEE LEAD: A barren, arid wasteland extends up from the Dead Sea along the Jordan River, in the northern end of the Great Rift Valley that stretches up from Kenya to Syria; it is hard to understand how it could be part of the Fertile Crescent. The river is narrow here and has virtually disappeared with the drought of the past few years.

A barren, arid wasteland extends up from the Dead Sea along the Jordan River, in the northern end of the Great Rift Valley that stretches up from Kenya to Syria; it is hard to understand how it could be part of the Fertile Crescent. The river is narrow here and has virtually disappeared with the drought of the past few years. The heat is so oppressive and the landscape so bleak that one wonders why the Children of Israel chose to settle here and indeed, why it has been the site of confrontation for thousands of years. Until, that is, you enter the Galilee. This is the region most prominent in the Gospels, where the Sea of Galilee and the villages around it are mentioned frequently and from whose fishermen came Jesus' disciples. This is also where the capital of Judaism was established in the first century, after the burning of Jerusalem. Here, the sands give way to the rolling hills from which the area takes it name. They are covered with olive groves, fruit orchards, banana plantations and plentiful, lush vegetation. The climate becomes gentle and the Jordan reappears, flowing out of the southern end of the Sea of Galilee in a wide, meandering stream. In this lovely setting, at Kibbutz Kinneret, there is a baptism site, Yardenit, which is free to visitors.

The sea is really a freshwater lake, harp-shaped and called Kinneret from the Hebrew word for stringed instrument. Josephus called it the Sea of Gennesar, because the plain of Gennesaret is just to the west; the ancient rabbis noted that the Lord created seven seas but the Sea of Gennesaret is His delight. It's also known as Lake Tiberias, after the town that Herod Antipas built on its shore and named in honor of Emperor Tiberius. In Arabic, it is still Bahret Tabariya.

The lake lies 680 feet below sea level, is roughly half as wide as its 14-mile length and still, as in biblical times, supports a lively fishing industry. It teems with bourri (gray mullet), sardines, carp, perch and a kind of fish, called tilapia, that is better known as St. Peter's fish.

This bony little denizen of the not-so-deep is said to have been the fish in whose mouth St. Peter found the shekel with which to pay his taxes. Today, it is still supplying shekels, not only to the fishermen of the lake but also to a whole industry that has grown around it. Fish farms cover more than 50,000 acres in the valleys. Tilapia hatchlings are collected and raised in the farm ponds.

Although St. Peter's fish is available throughout Israel, it is at the lakeside restaurants in Tiberias that it seems to taste best. Perhaps it's that the Miracle of the Loaves and Fishes is said to have taken place at Tabigha nearby, perhaps it's

Number of API results 15

Number of Topics 25

Suggested Topics

Click to highlight topics

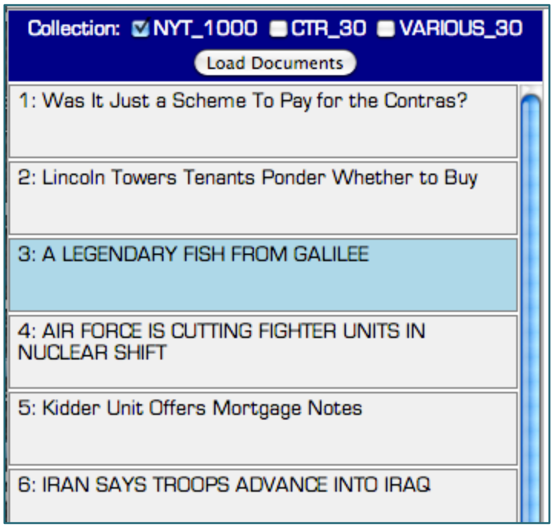
sea	fish
tiberias	israel
lake	galilee
hotel	salad
best	food
water	kinneret
hot	river
restaurant	hotels
season	great
place	eastern
also	find
valley	olive
jordan	

Figure D.1. The user interface of the Xpantrac system. (a) Collection pane, (b) Document pane, and (c) Topics pane.

2. Collection pane: **Load the documents**

Users first load a document collection by adding a check mark for one of the collections: NYT_1000 (*New York Times* articles), CTR_30 (webpages about the Hurricane Isaac disaster), and VARIOUS_30 (various disaster webpages), and clicking the "Load Documents" button shown on the top portion of Figure 2. Make sure the button is pressed.

We use 30 documents from each collection for this study. Once a document collection is loaded, the titles of the documents are shown in rectangular boxes as shown in Figure D.2. Users click the box to view the content of the article. The background color of the box changes to sky blue if it is clicked.



Collection: ☑ NYT_1000 □ CTR_30 □ VARIOUS_30

Load Documents

1: Was It Just a Scheme To Pay for the Contras?

2: Lincoln Towers Tenants Ponder Whether to Buy

3: A LEGENDARY FISH FROM GALILEE

4: AIR FORCE IS CUTTING FIGHTER UNITS IN NUCLEAR SHIFT

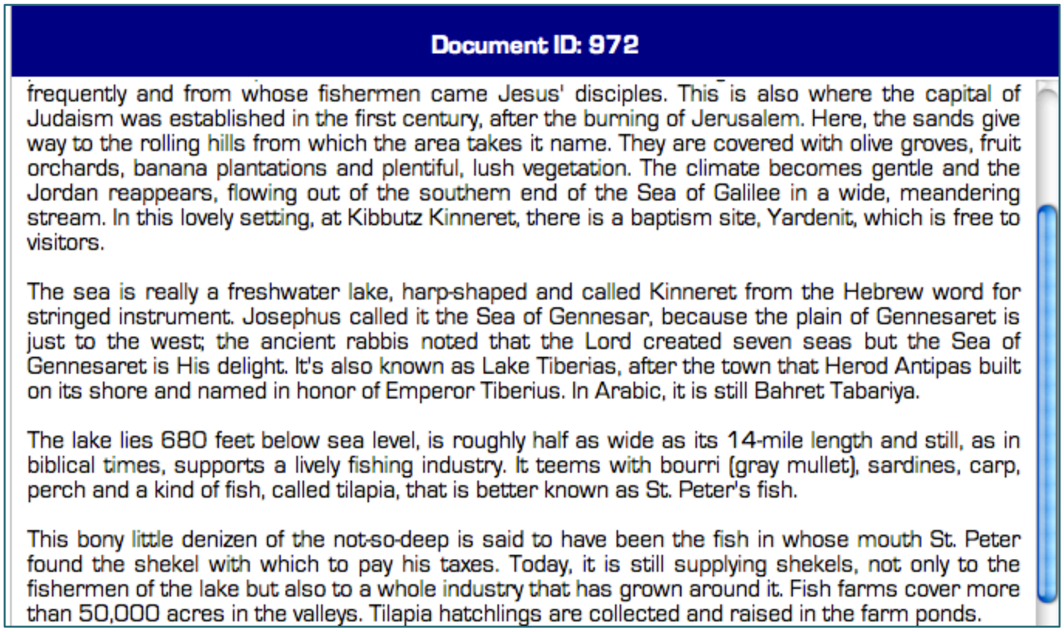
5: Kidder Unit Offers Mortgage Notes

6: IRAN SAYS TROOPS ADVANCE INTO IRAQ

Figure D.2. Collection pane. Users load documents to assign tags.

3. Document pane: **View the content of the selected document**

The content of the article is presented in the Document pane (Figure D.3), located in the center of the UI. A vertical scroll bar is provided for the longer articles.



Document ID: 972

frequently and from whose fishermen came Jesus' disciples. This is also where the capital of Judaism was established in the first century, after the burning of Jerusalem. Here, the sands give way to the rolling hills from which the area takes it name. They are covered with olive groves, fruit orchards, banana plantations and plentiful, lush vegetation. The climate becomes gentle and the Jordan reappears, flowing out of the southern end of the Sea of Galilee in a wide, meandering stream. In this lovely setting, at Kibbutz Kinneret, there is a baptism site, Yardenit, which is free to visitors.

The sea is really a freshwater lake, harp-shaped and called Kinneret from the Hebrew word for stringed instrument. Josephus called it the Sea of Gennesar, because the plain of Gennesaret is just to the west; the ancient rabbis noted that the Lord created seven seas but the Sea of Gennesaret is His delight. It's also known as Lake Tiberias, after the town that Herod Antipas built on its shore and named in honor of Emperor Tiberius. In Arabic, it is still Bahret Tabariya.

The lake lies 680 feet below sea level, is roughly half as wide as its 14-mile length and still, as in biblical times, supports a lively fishing industry. It teems with bourri (gray mullet), sardines, carp, perch and a kind of fish, called tilapia, that is better known as St. Peter's fish.

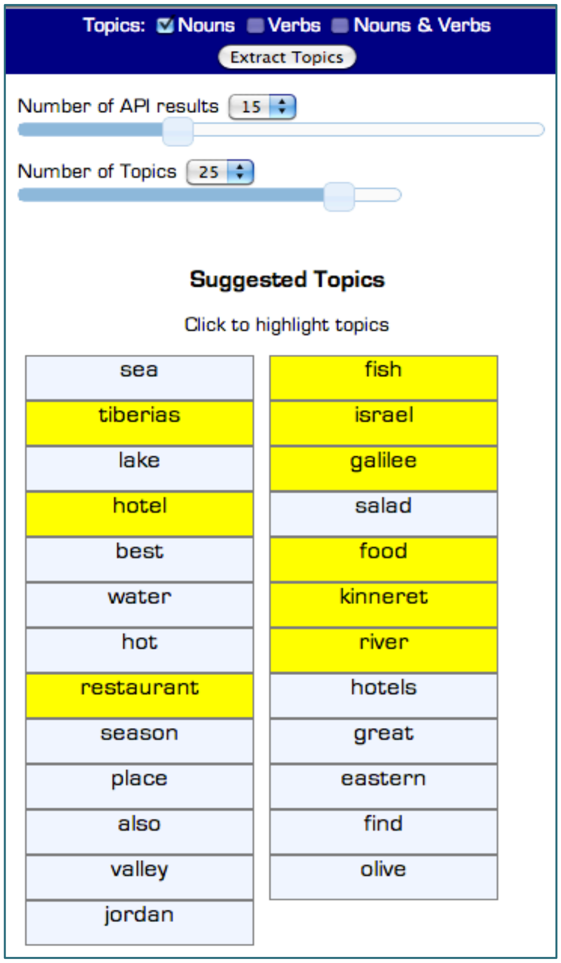
This bony little denizen of the not-so-deep is said to have been the fish in whose mouth St. Peter found the shekel with which to pay his taxes. Today, it is still supplying shekels, not only to the fishermen of the lake but also to a whole industry that has grown around it. Fish farms cover more than 50,000 acres in the valleys. Tilapia hatchlings are collected and raised in the farm ponds.

Figure D.3. Document pane. Users view the content of the document.

4. Topic pane: **Configure parameters and extract topic tags**

The Topic pane is located on the far right side of the UI. Users configure the parameters and setting before extracting a list of topic tags. First, there are options for whether the users would like to extract topics from (only) nouns, (only) verbs, or from both nouns and verbs. For this study, it is set to (only) nouns, so leave the choices.

There are two slider bars, "Number of API results" and "Number of Topics", to control parameters of the system (Figure D.4). The higher number of API results means the system uses more information in producing the topic suggestions; however, it also means that it takes longer time to extract topics. As the name suggests, "Number of Topics" setting controls how many topic suggestions to be presented. Topics shown on top of the two-column list should be more relevant to the document than the ones at the bottom (e.g., sea → fish → tiberias → Israel →…). Users can click the topic boxes to highlight them for later reference.



Topics: ☑ Nouns □ Verbs □ Nouns & Verbs

Extract Topics

Number of API results 15

Number of Topics 25

Suggested Topics

Click to highlight topics

sea	fish
tiberias	israel
lake	galilee
hotel	salad
best	food
water	kinneret
hot	river
restaurant	hotels
season	great
place	eastern
also	find
valley	olive
jordan	

Figure D.4. Topic pane. System parameters such as no. of API results and no. of topic suggestions can be changed. Users can mark topics of interest by clicking them.

Appendix E. Data Sets, Software, and Scripts

Documentations and download links for the data sets used (i.e., CTR_30, VARIOUS_30, document IDs for NYT_1000), manually-assigned and automatically-generated topic tags for the documents in those data sets, the produced software (i.e., Xpantrac command line mode, and it user interface), and the scripts (i.e., for text pre-processing, and computing evaluation metrics), are included in this appendix.

E.1 Data Sets

E.1.1 CTR_30 Data Set

- Download link for 30 documents with associated topic tags assigned by 3 human indexers
  - 30 homogeneous documents https://github.com/seungwonyang/dissertation/blob/master/dataset/CTR_30.zip
  - Tags assigned by human indexers https://github.com/seungwonyang/dissertation/blob/master/dataset/TagsBy3HumanIndexers_EXP1_CTR_30.csv
- Download link for a script to extract topics using the OpenCalais API. It stores the extracted topics into a database table, which should be developed in advance. https://github.com/seungwonyang/dissertation/blob/master/script/topic_extraction_opencalais.py
- Download link for a script to extract topics using a baseline, TF*IDF https://github.com/seungwonyang/dissertation/blob/master/script/topic_extraction_TFIDF_CTR30%2CVARIOUS_30.py

E.1.2 VARIOUS_30 Data Set

- Download link for 30 documents with associated topic tags assigned by 5 human indexers
  - 30 heterogeneous documents https://github.com/seungwonyang/dissertation/blob/master/dataset/VARIOUS_30.zip
  - Tags assigned by human indexers https://github.com/seungwonyang/dissertation/blob/master/dataset/TagsBy5HumanIndexers_EXP1_VARIOUS_30.csv
- Download link for a script to extract topics using the OpenCalais API. It stores the extracted topics into a database table, which should be developed in advance. https://github.com/seungwonyang/dissertation/blob/master/script/topic_extraction_opencalais.py

- Download link for a script to extract topics using a baseline, TF*IDF
  https://github.com/seungwonyang/dissertation/blob/master/script/topic_extraction_TFIDF_CTR30%2CVARIOUS_30.py

E.1.3 NYT_1000 Data Set

- The IDs of the documents in NYT_1000
  https://github.com/seungwonyang/dissertation/blob/master/dataset/NYT_1000_IDs_EXP3.txt
- Download link for the topic tags produced by Xpantrac with 3 search APIs (i.e., Bing Azure, Yahoo! Web, and Yahoo! News APIs) in SQL file format
  https://github.com/seungwonyang/dissertation/blob/master/dataset/NYT_1000_taggedTopics_EXP3.sql
- Download link for a script to extract topics from NYT_1000 using a baseline, TF*IDF
  https://github.com/seungwonyang/dissertation/blob/master/script/topic_extraction_TFIDF_NYT_1000.py

E.1.4 Documents Used for Usability Study

- Demographic Questionnaire
  https://github.com/seungwonyang/dissertation/blob/master/survey/DemographicsQuestionnaire_EXP4.pdf
- Task Questionnaires (Note: different documents were used for different survey types)
  - Type 1
    https://github.com/seungwonyang/dissertation/blob/master/survey/TaskQuestionnaire_EXP4_TYPE1.docx
  - Type 2
    https://github.com/seungwonyang/dissertation/blob/master/survey/TaskQuestionnaire_EXP4_TYPE2.docx
  - Type 3
    https://github.com/seungwonyang/dissertation/blob/master/survey/TaskQuestionnaire_EXP4_TYPE3.docx
  - Type 4
    https://github.com/seungwonyang/dissertation/blob/master/survey/TaskQuestionnaire_EXP4_TYPE4.docx
  - Type 5
    https://github.com/seungwonyang/dissertation/blob/master/survey/TaskQuestionnaire_EXP4_TYPE5.docx
  - Type 6
    https://github.com/seungwonyang/dissertation/blob/master/survey/TaskQuestionnaire_EXP4_TYPE6.docx
  - Type 7
    https://github.com/seungwonyang/dissertation/blob/master/survey/TaskQuestionnaire_EXP4_TYPE7.docx
- Exit Questionnaire
  https://github.com/seungwonyang/dissertation/blob/master/survey/ExitQuestionnaire_EXP4.pdf

- Xpantrac Tutorial
https://github.com/seungwonyang/dissertation/blob/master/survey/XpantracTutorial.pdf

E.2 Software

E.2.1 Xpantrac Command Line (Python)

- Xpantrac with Bing Azure API
https://github.com/seungwonyang/dissertation/blob/master/code/Xpantrac_extractTopics_bing.py
    - Stopwords files used along Xpantrac
    Stopwords.txt:
    https://github.com/seungwonyang/dissertation/blob/master/code/stopwords.txt
    Custom_stops.txt:
    https://github.com/seungwonyang/dissertation/blob/master/code/custom_stops.txt
- Xpantrac to build the cache (it expands input texts using a search API, and then stores the expanded information to a database table, which is accessed by Xpantrac_bing_DB.py)
https://github.com/seungwonyang/dissertation/blob/master/code/Xpantrac_bing_buildCache.py
- Xpantrac to access cached information in a database table
https://github.com/seungwonyang/dissertation/blob/master/code/Xpantrac_bing_DB.py
    - Chained part-of-speech (POS) taggers used with Xpantrac
    https://github.com/seungwonyang/dissertation/blob/master/code/pos_tagger.py

E.2.2 Xpantrac User Interface (JavaScript)

- Download link
https://github.com/seungwonyang/dissertation/blob/master/code/Xpantrac_UI.html

E.3 Scripts

E.3.1 For Computing IIC, Precision, Recall, and F1

- A script to compute precision, recall, and F1
https://github.com/seungwonyang/dissertation/blob/master/script/topic_tag_evaluator.py
- A script to compute the Rolling's Inter-Indexer Consistency (IIC)
    - IIC between human topics and Xpantrac topics

https://github.com/seungwonyang/dissertation/blob/master/script/compute_IIC_human_machine.py

- IIC between human topics and OpenCalais topics
  https://github.com/seungwonyang/dissertation/blob/master/script/compute_IIC_human_opencalais.py

References

Abdi, H. E., & Williams, L. J. (2010). Tukey's Honestly Significant Difference (HSD) Test. In N. Salkind (Ed.), *Encyclopedia of Research Design*. Thousand Oaks, CA: SAGE Publications Inc.

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., & Shadbolt, N. R. (2003). Automated Ontology-Based Knowledge Extraction from Web Documents. *IEEE Intelligent Systems*, 18(1), 14–21.

Antoniou, G., & Van Harmelen, F. (2004). *A Semantic Web Primer*. MIT Press.

Baaeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern Information Retrieval* (1st ed.). Addison Wesley. New York: ACM Press.

Baker, C. J. O., & Cheung, K.-H. (2007). *Semantic Web*. New York: Springer.

Banko, M., & Etzioni, O. (2007). Strategies for Lifelong Knowledge Extraction from the Web. In Proceedings of the 4th International Conference on Knowledge Capture (K-CAP'07), New York, New York, USA: ACM Press. 95-102. doi:10.1145/1298406.1298425

Bastian, M., Heymann, S., & Jacomy, M. (2009). Gephi: An Open Source Software for Exploring and Manipulating Networks. In Proceedings of the Third International ICWSM Conference. 361-362.

Becker, J., & Kuropka, D. (2003). Topic-based vector space model. In W. Abramowicz & G. Klein (Eds.). In Proceedings of the International Conference on Business Information Systems (BIS'03), Colorado Springs, CO, USA.

Bing Search API. (2013). *Bing Search API*. Retrieved September 21, 2013, from http://datamarket.azure.com

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. (J. Lafferty, Ed.), *Journal of Machine Learning Research*, 3, 993–1022.

Bradford, R. B. (2008). An Empirical Study of Required Dimensionality for Large-Scale Latent Semantic Indexing Applications. Proceedings of the 17th ACM conference on Information and Knowledge Management (CIKM'08), Napa Valley, CA: ACM Press. 153-162 doi:10.1145/1458082.1458105

Breitman, K. K., Casanova, M. A., & Truszkowski, W. (2007). *Semantic Web: Concepts, Technologies and Applications*. Springer.

Calais. (2013). *Calais*. Thomson Reuters. Retrieved August 18, 2013, from http://www.opencalais.com/

Coursey, K., Mihalcea, R., & Moen, W. (2009). Using Encyclopedic Knowledge for Automatic Topic Identification (pp. 210–218). In Proceedings of NODALIDA, Boulder, CO. Retrieved from http://dl.acm.org/citation.cfm?id=1596407

Danielsson, P. (2003). Automatic extraction of meaningful units from corpora: A corpus-driven approach using the word stroke. *International Journal of Corpus Linguistics*, *8*(1), 109–127. doi:10.1075/ijcl.8.1.06dan

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS quarterly*, *13*, 319–340. Retrieved from http://hcibib.org/perlman/question.cgi?form=PUEU

Deerwester, S. C., Dumais, S. T., Landauer, T. K., & Furnas, G. W. (1990). Indexing by latent semantic analysis. *JASIS*, *41*(6), 391–407.

DiCiuccio, R. (2010, May 18). Entity Extraction & Content API Evaluation. Retrieved October 6, 2013, from http://blog.viewchange.org/2010/05/entity-extraction-content-api-evaluation/#ee-evri

Du, L., Jin, H., de Vel, O., & Liu, N. (2008). A latent semantic indexing and WordNet based information retrieval model for digital forensics (pp. 70–75). In Proceedings of the Intelligence and Security Informatics, 2008. ISI 2008. IEEE International Conference on. doi:10.1109/ISI.2008.4565032

Dumais, S., & Chen, H. (2000). Hierarchical classification of Web content (pp. 256–263). In Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'00), New York, New York, USA: ACM Press. doi:10.1145/345508.345593

Enriching very large ontologies using the WWW. (2000). Enriching very large ontologies using the WWW. *In Proceedings of the ECAI 2000 Workshop on Ontology Learning*.

Fagan, M. (2010, January 1). Comparing NLP APIs for Entity Extraction. Retrieved October 6, 2013, from http://faganm.com/blog/2010/01/02/1009/

Fellbaum, C. (2010). WordNet. In R. Poli, M. Healy, & A. Kameas (Eds.), *Theory and Applications of Ontology: Computer Applications* (pp. 231–243). Dordrecht Heidelberg London New York: Springer Netherlands. doi:10.1007/978-90-481-8847-5_10

Fielding, R. T., & Taylor, R. N. (2000). Principled design of the modern Web architecture (pp. 407–416). In Proceedings of the International Conference on Software Engineering. doi:10.1109/ICSE.2000.870431

Fischer, L. (2013, January 22). A Beginner's Guide to HTTP and REST. *net.tutsplus.com*.

Retrieved July 22, 2013, from http://net.tutsplus.com/tutorials/other/a-beginners-introduction-to-http-and-rest/

Foltz, P. W., & Foltz, P. W. (1990). *Using latent semantic indexing for information filtering*. *ACM SIGOIS Bulletin* (Vol. 11, pp. 40–47). ACM. doi:10.1145/91478.91486

Furnas, G. W., Landauer, T. K., Gomez, L. M., & Dumais, S. T. (1987). The vocabulary problem in human-system communication. *Communications of the ACM*, *30*(11), 964–971. doi:10.1145/32206.32212

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, *33*(1), 1–53.

Gentner, D. (1989). The Mechanisms of Analogical Reasoning. In S. Vosniadou & A. Ortony (Eds.), *Similarity and Analogical Reasoning* (pp. 199–241). Cambridge University Press.

Gentner, D., & Markman, A. B. (1997). Structure mapping in analogy and similarity. *American Psychologist*, *52*(1), 45. doi:10.1037/0003-066X.52.1.45

Gentner, D., Rattermann, M. J., & Forbus, K. D. (1993). The Roles of Similarity in Transfer: Separating Retrievability From Inferential Soundness. *Cognitive Psychology*, *25*(4), 524–575. doi:10.1006/cogp.1993.1013

Google Custom Search. (2013). *Google Custom Search*. Retrieved August 27, 2013, from https://developers.google.com/custom-search/

Griffiths, T. L., Kemp, C. & Tenenbaum, J. B. (2008). Bayesian models of cognition. In Ron Sun (ed.), *The Cambridge handbook of computational psychology.* Cambridge University Press.

He, X., Ding, C. H. Q., Zha, H., & Simon, H. D. (2001). Automatic Topic Identification Using Webpage Clustering (pp. 195–202). In Proceedings of the IEEE International Conference on Data Mining (ICDM 2001), San Jose, CA.

Hearst, M. A., Dumais, S. T., Osman, E., Platt, J., & Scholkopf, B. (1998). Support Vector Machines. *IEEE Intelligent Systems*, *13*(4), 18–28. doi:10.1109/5254.708428

Hofmann, T. (1999). Probabilistic latent semantic indexing (pp. 50–57). In Proceedings of the 22nd annual international ACM SIGIR conference, New York, New York, USA: ACM Press. doi:10.1145/312624.312649

Iivonen, M. (1995). Consistency in the selection of search concepts and search terms. *Information Processing & Management*, *31*(2), 173–190. doi:10.1016/0306-4573(95)80034-Q

Kantrowitz, M., Mohit, B., & Mittal, V. O. (2000). Stemming and its effects on TFIDF ranking. *SIGIR 2000*, 357–359. doi:10.1145/345508.345650

Kim, S. N., Baldwin, T., & Kan, M.-Y. (2010). Evaluating N-gram based Evaluation Metrics for Automatic Keyphrase Extraction. *International Conference on Computational Linguistics*

(Coling 2010), 572–580.

Landauer, T., & Dumais, S. (2008). Latent Semantic Analysis. *Scholarpedia*, *3*(11), 4356. doi:10.4249/scholarpedia.4356

Lange, T. E., & Wharton, C. M. (1994). REMIND: Retrieval From Episodic Memory by Inferencing and Disambiguation. In J. A. Barnden & K. J. Holyoak (Eds.), *Advances in Connectionist and Neural Computation Theory, Volume 3: Analogy, Metaphor, and Reminding* (Vol. 3, pp. 29–94). Ablex Publishing Corporation.

Lassila, O., Swick, R. R., W3C. (2004). RDF/XML Syntax Specification (Revised). D. Beckett & B. McBride, Eds. *W3C Recommendation*. Retrieved August 27, 2013, from http://www.w3.org/TR/REC-rdf-syntax/

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, *7*(1), 57–78. doi:10.1080/10447319509526110

Likert, R. (1932). A technique for the measurement of attitudes. R. S. Woodworth, Ed. *Archives of psychology*, *22*(140), 5–55. Retrieved from http://www.voteview.com/Likert_1932.pdf

Liu, G. Z. (1997). Semantic Vector Space Model: Implementation and evaluation. *Journal of the American Society for Information Science and Technology*, *48*(5), 395–417. doi:10.1002/(SICI)1097-4571(199705)48:5<395::AID-ASI3>3.0.CO;2-Q

Liu, X., Wang, G. A., Johri, A., Zhou, M., & Fan, W. (2012). Harnessing global expertise: A comparative study of expertise profiling methods for online communities. *Information Systems Frontiers*. doi:10.1007/s10796-012-9385-6. pp. 1-13, Springer US.

Lund, A. M. (2001, October 12). Measuring Usability with the USE Questionnaire. *STC Usability SIG Newsletter*. 8(2). Retrieved November 12, 2013, from http://www.stcsig.org/usability/newsletter/0110_measuring_with_use.html

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Markman, A. B., & Gentner, D. (2001). Thinking. *Annual Review of Psychology*, *52*(1), 223–247. doi:10.1146/annurev.psych.52.1.223

Martin, O. S. (2011, October 17). A Wikipedia Literature Review. *arXiv.org*. Retrieved from http://arxiv.org/PS_cache/arxiv/pdf/1110/1110.5863v1.pdf

Massey, L. (2011). A cognitive informatics framework for language understanding (pp. 167–174). In Proceedings of the 10th IEEE International Conference on Cognitive Informatics, Banff, Alberta, Canada. doi:10.1109/COGINF.2011.6016137

Massey, L., & Wong, W. (2011). A Cognitive-Based Approach to Identify Topics in Text Using the Web as a Knowledge Source (pp. 61–78). IGI Global. doi:10.4018/978-1-60960-625-1.ch004

Matthew D Hoffman, D. M. B. F. B. (2010). Online Learning for Latent Dirichlet Allocation. *Neural Information Processing Systems (NIPS)*, 856–864.

Medelyan, O. (2009, July 16). *Human-competitive automatic topic indexing*. The University of Waikato, Himilton, New Zealand.

Medelyan, O., Frank, E., & Witten, I. H. (2009). Human-competitive tagging using automatic keyphrase extraction (Vol. 3). In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP'09), Association for Computational Linguistics.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*(2), 254. doi:10.1037/0033-295X.100.2.254

Mikoluk, K. (2013, August 16). JSON vs XML: How JSON Is Superior To XML. *www.udemy.com*. Retrieved September 22, 2013, from https://www.udemy.com/blog/json-vs-xml/

Miller, E. (2005). An Introduction to the Resource Description Framework. *Bulletin of the American Society for Information Science and Technology*, *25*(1), 15–19. doi:10.1002/bult.105

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41. doi:10.1145/219717.219748

Milne, D., Medelyan, O., & Witten, I. H. (2006). Mining Domain-Specific Thesauri from Wikipedia: A case study (pp. 442–448). In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI'06), Washington DC, USA. doi:10.1109/WI.2006.119

Ogiela, L., Tadeusiewicz, R., & Ogiela, M. R. (2007). Cognitive Informatics in Automatic Pattern Understanding (pp. 79–84). In Proceedings of the 6th IEEE International Conference on Cognitive Informatics, Lake Tahoe, CA. doi:10.1109/COGINF.2007.4341875

OpenCalais. (2013). *OpenCalais*. Retrieved September 5, 2013, from http://www.opencalais.com/

Oracle. (2013, September 22). How is JavaScript different from Java? Retrieved October 22, 2013, from http://www.java.com/en/download/faq/java_javascript.xml

Ozsoy, M. G., Alpaslan, F. N., & Cicekli, I. (2011). Text summarization using Latent Semantic Analysis. *Journal of Information Science*, *37*(4), 405–417.

Papineni, K. (2001). Why inverse document frequency? (pp. 1–8). In Proceedings of the Second meeting of the North American Chapter of the Association for Computational Linguistics, Morristown, NJ, USA: Association for Computational Linguistics.

doi:10.3115/1073336.1073340

Pease, A., & Fellbaum, C. (2010). Formal ontology as interlingua: the SUMO and WordNet linking project and global WordNet. In C.-R. Huang, N. Calzolari, A. Gangemi, A. Lenci, A. Oltramari, & L. Prevot (Eds.), *ebooks.cambridge.org* (pp. 25–35). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511676536.003

Perkins, J. (2013a, May 24). Part of Speech Tagging with NLTK Part 4 – Brill Tagger vs. Classifier Taggers | StreamHacker. *Stream Hacker*. Retrieved July 28, 2013, from http://streamhacker.com/2010/04/12/pos-tag-nltk-brill-classifier/

Perkins, J. (Ed.). (2013b). *Stream Hacker*. Retrieved July 28, 2013, from http://streamhacker.com

Qualtrics. (2002). *Qualtrics*. Retrieved September 25, 2013, from https://survey.vt.edu/survey/

Raghavan, V. V., & Wong, S. K. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for Information Science and Technology*, *37*(5), 279–287. doi:10.1002/(SICI)1097-4571(198609)37:5<279::AID-ASI1>3.0.CO;2-Q

Rogers, D. J., & Tanimoto, T. T. (1960). A Computer Program for Classifying Plants. *Science*, *132*(3434), 1115–1118. doi:10.1126/science.132.3434.1115

Rolling, L. (1981). Indexing consistency, quality and efficiency. *Information Processing & Management*, *17*(2), 69–76. doi:10.1016/0306-4573(81)90028-5

Salton, G. (1971). *The SMART Retrieval System—Experiments in Automatic Document Processing*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620. doi:10.1145/361219.361220

Sandhaus, E. (Ed.). (2008). *The New York Times Annotated Corpus*. Linguistic Data Consortium. Retrieved July 18, 2013, from http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2008T19

Schonhofen, P. (2008). Annotating Documents by Wikipedia Concepts. In Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WIIAT'08). 1, 461-467. doi:10.1109/WIIAT.2008.56

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. doi:10.1145/505282.505283

Shi, Z., & Shi, J. (2003). Perspectives on cognitive informatics. In Proceedings of the Second IEEE International Conference on Cognitive Informatics (ICCI'03). 129-133.

doi:10.1109/COGINF.2003.1225970

Shuda, W., Jiangping, L., & Riu, W. (2009). Research of Information Filtering Based on Vector Space Model. *Second International Workshop on Computer Science and Engineering, 2009 (WCSE '09)*, *1*, 42–46. doi:10.1109/WCSE.2009.618

Shwartz, B. (2010, August 24). Official: Yahoo's Results Now Come From Bing. (B. Schwartz, Ed.). Retrieved November 13, 2013, from http://searchengineland.com/yahoos-transition-to-bing-organic-results-complete-49228

Sjöbergh, J. (2003). Combining POS-taggers for improved accuracy on Swedish text. Proceedings of NODALIDA.

Sparck-Jones, K. (1972). A Statistical Interpretation of Term Specificity and Its Application in Retrieval. *Journal of Documentation*, *28*(1), 11–21. doi:10.1108/eb026526

Stein, B., & Meyer zu Eissen, S. (2011). Topic Identification: Framework and Application. In M. Tochtermann (Ed.). In Proceedings of the 4th International Conference on Knowledge Management (I-KNOW'04). 353-360.

The Apache Software Foundation. (2011, September 22). Apache Solr. Retrieved September 22, 2013, from http://lucene.apache.org/solr/

The jQuery Foundation. (2013a, September 22). jQuery. Retrieved September 22, 2013, from http://jquery.com/

The jQuery Foundation. (2013b, September 22). jQuery User Interface. Retrieved September 22, 2013, from http://jqueryui.com/

Thompson Reuters. (2013). *Thompson Reuters*. Retrieved August 18, 2013, from http://thomsonreuters.com/

Tiedemann, J., & Mur, J. (2008). Simple is best: experiments with different document segmentation strategies for passage retrieval. In Proceedings of the 2nd workshop on Information Retrieval for Question Answering (IRQA '08), Proceedings of the Association for Computational Linguistics (Coling 2008).

Tiun, S., Abdullah, R., & Kong, T. E. (2001). Automatic Topic Identification by Using Ontology Hierarchy (pp. 444–453). In Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'01), Springer-Verlag.

van Rijsbergen, C. J. (1979). *Information Retrieval* (2nd ed.). Butterworth-Heinemann Newton, MA, USA.

Vasishth, S., & Broe, M. (2010). Analysis of Variance (ANOVA). In *The Foundations of Statistics: A Simulation-based Approach* (pp. 97–126). Berlin, Heidelberg: Springer. doi:10.1007/978-3-642-16313-5_5

W3C. (2004, February 10). Resource Description Framework (RDF): concepts and abstract syntax. *www.w3.org*. World Wide Web Consortium. Retrieved August 27, 2013, from http://travesia.mcu.es/portalnb/jspui/handle/10421/2427

Wang, Y. (2002). On cognitive informatics. In Proceedings of the First IEEE International Conference on Cognitive Informatics (ICCI'02). 34-42. doi:10.1109/COGINF.2002.1039280

Wang, Y. (2007). Cognitive Informatics Foundations of Nature and Machine Intelligence. In Proceedings of the 6th IEEE International Conference on Cognitive Informatics. 3-12. doi:10.1109/COGINF.2007.4341867

Wang, Y., Berwick, R. C., Haykin, S., Pedrycz, W., Baciu, G., Bhavsar, V. C., et al. (2011). Cognitive Informatics in Year 10 and Beyond: summary of the plenary panel. In Proceedings of the 10th IEEE International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC'11), Banff, Alberta, Canada. 11-22. doi:10.1109/COGINF.2011.6016117

Wharton, C. M., Holyoak, K. J., Downing, P. E., & Lange, T. E. (1994). Below the Surface: Analogical Similarity and Retrieval Competition in Reminding. *Cognitive Psychology*, *26*(1), 64–101. doi:10.1006/cogp.1994.1003

Wong, S. K. M., & Raghavan, V. V. (1984). *Vector Space Model of Information Retrieval: A Reevaluation*. In Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'84). 167-185. British Computer Society. Swinton, UK.

Yahoo! Search BOSS API. (2013). Retrieved August 27, 2013, from http://developer.yahoo.com

Yang, S., Kavanaugh, A., Kozievitch, N. P., Li, L. T., Srinivasan, V., Sheetz, S. D., et al. (2011). CTRnet DL for Disaster Information Services. In Proceedings of the 11th International ACM/IEEE Joint Conference on Digital Libraries (JCDL'11), 437-438. New York, NY, USA: ACM. doi:10.1145/1998076.1998173

Yingxu Wang, & Ying Wang. (2006). Cognitive informatics models of the brain. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, *36*(2), 203–207. doi:10.1109/TSMCC.2006.871151

Yun, J., Jing, L., Yu, J., Huang, H., & Zhang, Y. (2011). Document Topic Extraction Based on Wikipedia Category. In Proceedings of the Fourth International Conference on Computational Sciences and Optimization (CSO), 852-856. Yunnan, China. doi:10.1109/CSO.2011.119

Zhang, W., Yoshida, T., & Tang, X. (2008). TFIDF, LSI and multi-word in information retrieval and text categorization. In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics (SMC'08), 108-113. Singapore. doi:10.1109/ICSMC.2008.4811259

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Systems With Applications*, *38*(3), 2758–2765. doi:10.1016/j.eswa.2010.08.066

Zhanguo, M., Jing, F., Xiangyi, H., Yanqin, S., & Liang, C. (2011). Improved Terms Weighting Algorithm of Text. In Proceedings of the International Conference on Network Computing and Information Security (NCIS). 2, 367-370. doi:10.1109/NCIS.2011.171

Zunde, P., & Dexter, M. E. (1969). Indexing consistency and quality. *Journal of the American Society for Information Science and Technology*, *20*(3), 259–267. doi:10.1002/asi.4630200313