Hybrid Summarization of Dakota Access Pipeline Protests (NoDAPL)

CS 5984/4984 Big Data Text Summarization Report

Xiaoyu Chen*, Haitao Wang, Maanav Mehrotra, Naman Chhikara, Di Sun
{xiaoyuch, wanght, maanav, namanchhikara, sdi1995} @vt.edu
Instructor: Dr. Edward A. Fox

Dept. of Computer Science, Virginia Tech
Blacksburg, Virginia 24061
December 2018

Outline

- **Introduction**
  - Automatic Summarization and NoDAPL Dataset
  - Related Work
  - Overview of the Proposed Framework
- Preprocessing of Data
  - Classification of Relevance
  - Topic Modelling by Latent Dirichlet Allocation (LDA) and LDA2Vec
- Hybrid Method
  - Latent Dirichlet Allocation based Extractive Summarization
  - Pointer-generator based Abstractive Summarization
  - Text Re-ranking based Hybrid Summarization
- Results and Evaluation
  - Compiled Summary
  - Extrinsic Evaluation
- Conclusion and Future Work

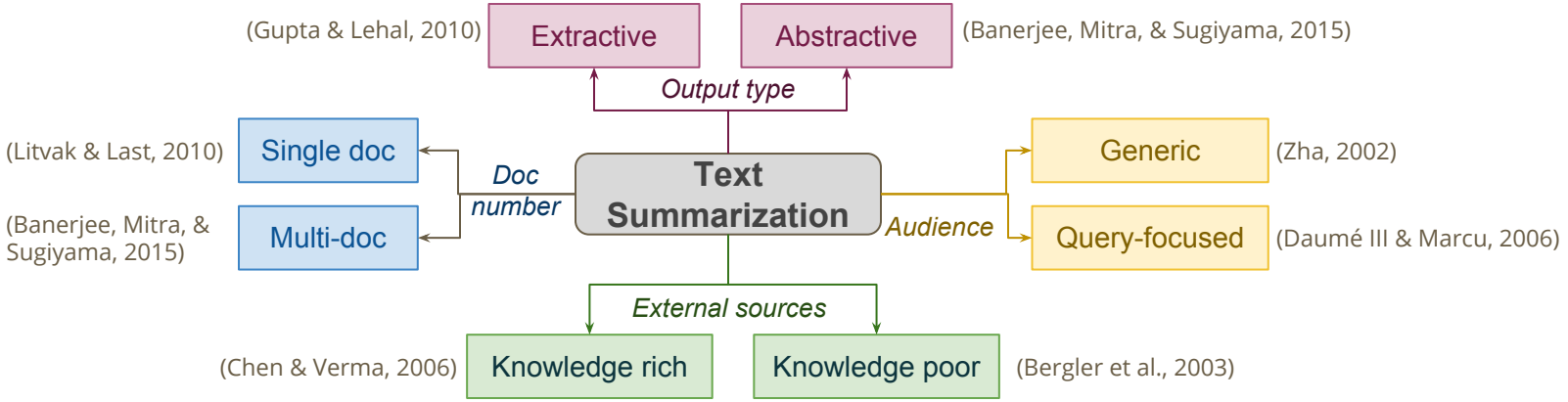
Introduction

- Automatic Summarization
  - Automatic summarization: has been investigated for more than 60 years since the publication of Luhn's seminal paper (Luhn, 1958)
  - Challenges: large proportion of noise texts (i.e., irrelevant documents/topics, noise sentences, etc.), highly redundant information, multiple latent topics
- Problem Statement
  - How to automatically / semi-automatically generate a good summary from a corpus collected from webpage with noisy, highly redundant, and large-scale text dataset?
  - Topic: NoDAPL - Protest movements against Dakota Access Pipeline construction in U.S.
  - Method: Automatic text summarization techniques with deep learning methodology.

Quality Quantity

Related Work

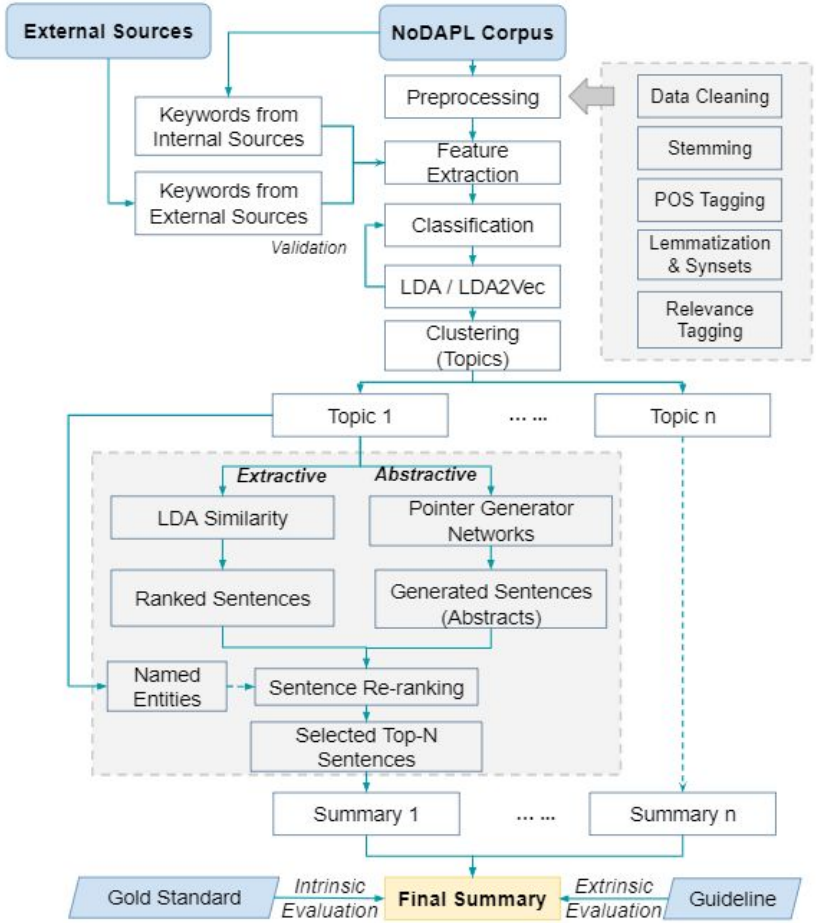
- Text Summarization Categories



- Deep Learning-based Automatic Text Summarization
  - Seq2Seq model (Khatri, Singh, & Parikh, 2018)
  - Pointer-generator network (See, Liu, & Manning, 2017)

Proposed Framework

- Proposed automatic text summarization framework with limited human effort
- Adopted both deep learning-based abstractive an LDA-based extractive summarization technologies to create hybrid method
- Advantages:
  - Without rely on deep understanding of the given events
  - Can be easily extended to other events
  - Not require large computation workload
- Disadvantages:
  - Highly depends on accuracy of extracting named entities and topics
  - Requires manually label 100 documents

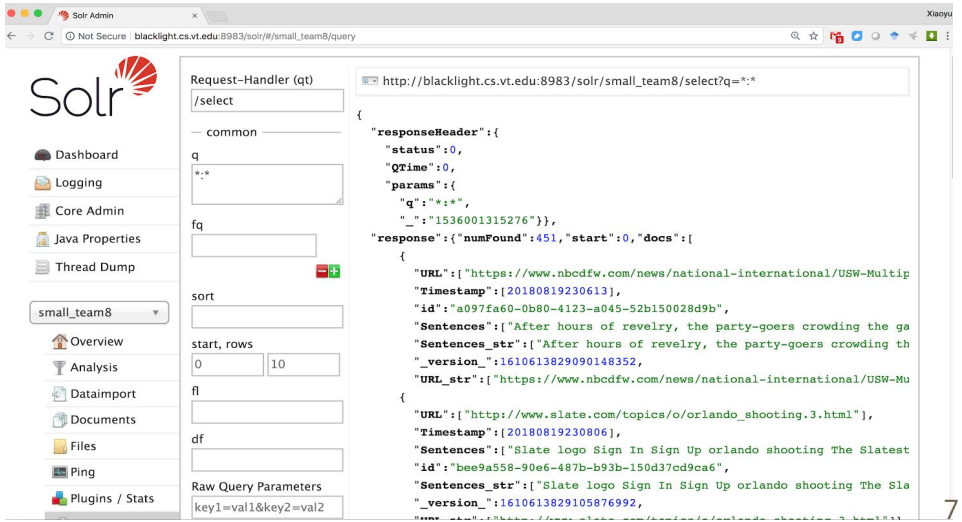


Outline

- Introduction
  - Automatic Summarization and NoDAPL Dataset
  - Related Work
  - Overview of the Proposed Framework
- **Preprocessing of Data**
  - Classification of Relevance
  - Topic Modelling by Latent Dirichlet Allocation (LDA) and LDA2Vec
- Hybrid Method
  - Latent Dirichlet Allocation based Extractive Summarization
  - Pointer-generator based Abstractive Summarization
  - Text Re-ranking based Hybrid Summarization
- Results and Evaluation
  - Compiled Summary
  - Extrinsic Evaluation
- Conclusion and Future Work

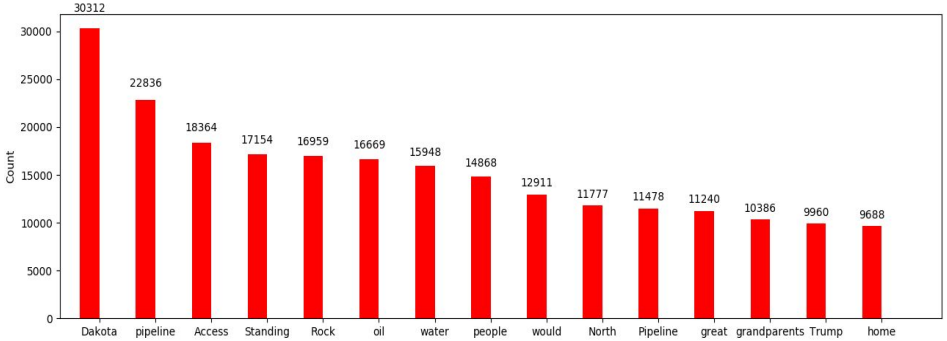
Formatting and Solr Indexing

- Converted WARC and CDX files into JSON formatted file
  - Total records (small dataset) : ~ **500**
  - Total records (big dataset) : ~ **1,1000**
  - Cleaned irrelevant content like HTML tags, javascript *etc.*
- Solr Indexing
  - Allowed us to query the data
  - Helped other teams for creating a gold standard



Stopwords Removal and POS Tagging

- Stopwords Removal
  - Used NLTK's default stopwords list
  - Made a custom stopwords list

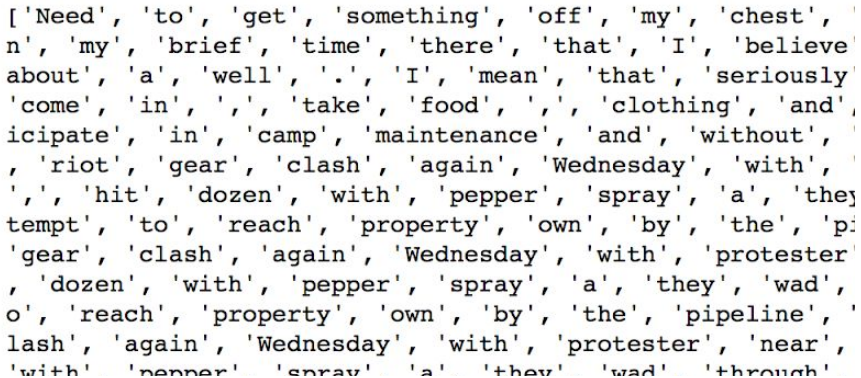


- POS Tagging
  - Used NLTK's default POS Tagger to tag tokens correctly
  - Note- Needs to be done without removing stopwords

```
[('Need', 'NN'), ('to', 'TO'), ('get', 'VB'), ('something', 'NN'),
that', 'IN'), ('I', 'PRP'), ('witnessed', 'VBD'), ('and', 'CC'), (
G'), ('in', 'IN'), ('my', 'PRP$'), ('brief', 'JJ'), ('time', 'NN')
believe', 'VBP'), ('many', 'JJ'), ('others', 'NNS'), ('have', 'VBP
'), ('up', 'RP'), ('about', 'RB'), ('as', 'RB'), ('well', 'RB'), (
'DT'), ('seriously', 'RB'), ('.', '.'), ('Plymouth', 'NNP'), ('roc
, 'PRP'), ('are', 'VBP'), ('coming', 'VBG'), ('in', 'IN'), (',', '
('clothing', 'NN'), ('and', 'CC'), ('occupying', 'VBG'), ('space',
, 'NN'), ('to', 'TO'), ('participate', 'VB'), ('in', 'IN'), ('camp
ithout', 'IN'), ('respect', 'NN'), ('of', 'IN'), ('tribal', 'JJ'),
S'), ('in', 'IN'), ('riot', 'NN'), ('gear', 'NN'), ('clashed', 'VB
, 'IN'), ('protesters', 'NNS'), ('near', 'IN'), ('the', 'DT'), ('D
```

Lemmatization

- Lemmatization
  - Lemmatization takes into consideration the morphological analysis of the words.
  - Analysed that it worked better than stemming
  - Used the WordNet library for lemmatization



POS-Tagging

- Manual Document Relevance Tagging and Internal Keywords Extraction
  - Initially tagged 50 documents as relevant and irrelevant for classifier.
  - Increased it to 100 documents.
  - Used it as our training set to classify all the documents as relevant and irrelevant.
- Using wikipedia as an external source
  - Two wikipedia articles (Dakota Access Pipeline,#NODAPL) used to extract keywords which were given as input to the classifier.

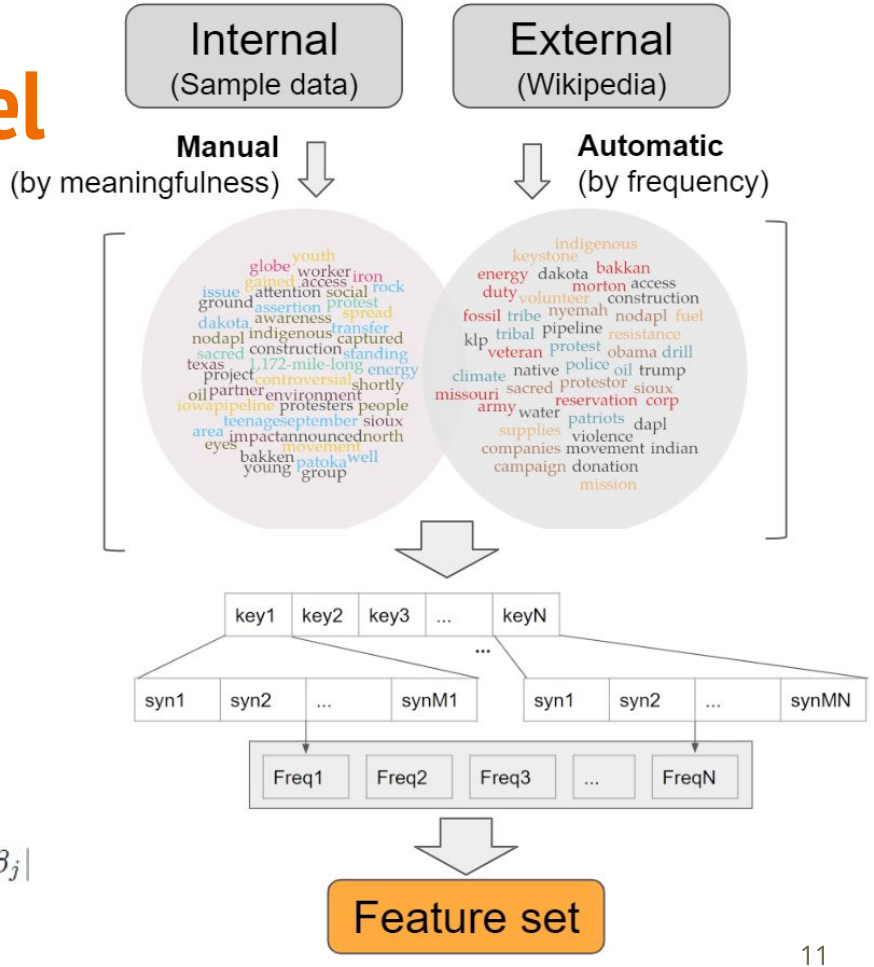
Document index ID	relevance (0 or 1)	keywords
63e65803-3546-4cab-a593-c78309e1356e	1	Indian Reservation, DAPL, Dakota Access pipeline, pipeline, Veterans, Sioux tribe, Standing Rock, Bakkan, fuel
208022ec-e5e4-40f1-8ac4-1c1c70807110	0	
37a24428-f7c5-4fb9-875d-985dd3e48916	1	Dakota Access pipeline, Oil, Donald Trump, U.S. Army Corps of Engineers
fe0fc10f-2e1a-4641-90de-6a278f1c992b	1	Dakota Access pipeline, Oil, drinking water
f3958d6f-41df-49a1-90b5-e7d727cb92f3	0	
be2be1cd-31ca-46ea-8afc-57b8c42a9f62	1	KLP, Dakota Access Pipeline, companies
4afcbd05-c48b-4c2f-bf4f-86b8c29ed449	1	Energy Transfer Partners, Drill, Missouri River, drinking water, Standing Rock Sioux Tribe
533214a9-7df2-4243-8bcc-71c1426906c3	1	Stop Trump, drinking water supplies, stop construction, Dakota Access pipeline
7347deed-dfc1-40dc-9f40-d60ea7db3b2d	0	
63be0e95-1c37-433f-b233-c985b9b6c298	1	Veterans, water protectors, Donation, Oil, Water source
0f44a3a3-f51b-4b10-9bdf-c9e0b6816dbc	0	

Feature Extraction and Model

- Feature Extraction
  - Calculated frequency of keywords and their synsets as features.
- Regularized Logistic regression-based Classification
  - Used 5-fold cross-validation (CV) to select the tuning parameter (λ) as well as to evaluate the model performance.

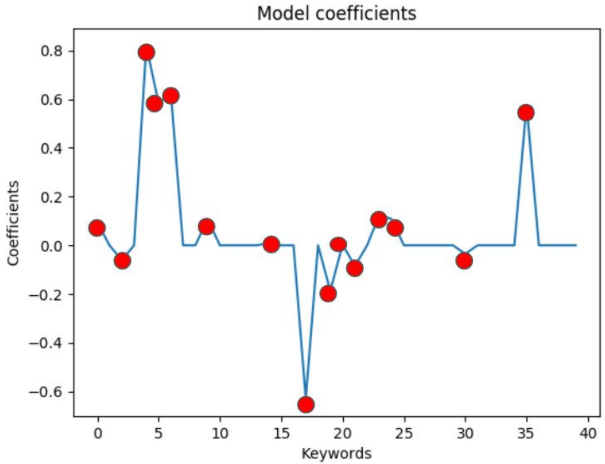
$$h(\boldsymbol{x}_i|\boldsymbol{\beta}) = \log(\frac{p(y_i=1)}{1-p(y_i=1)}) = \beta_0 + \sum_{j=1}^{p} x_{i,j}\beta_j + \epsilon_i$$

$$\min_{\boldsymbol{\beta}} -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(h(\boldsymbol{x}_i|\boldsymbol{\beta})) + (1-y_i)\log(1-h(\boldsymbol{x}_i|\boldsymbol{\beta}))] + \frac{\lambda}{n}\sum_{j=1}^{p}|\beta_j|$$



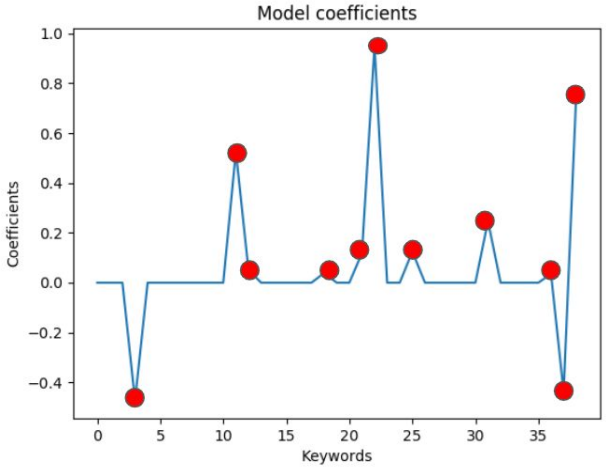
Classification Results

- LR1 using keywords from Wikipedia



Fold No.	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Training Accuracy	0.827	0.840	0.951	0.889	0.875	0.876
Testing Accuracy	0.900	0.750	0.750	0.600	0.810	0.762

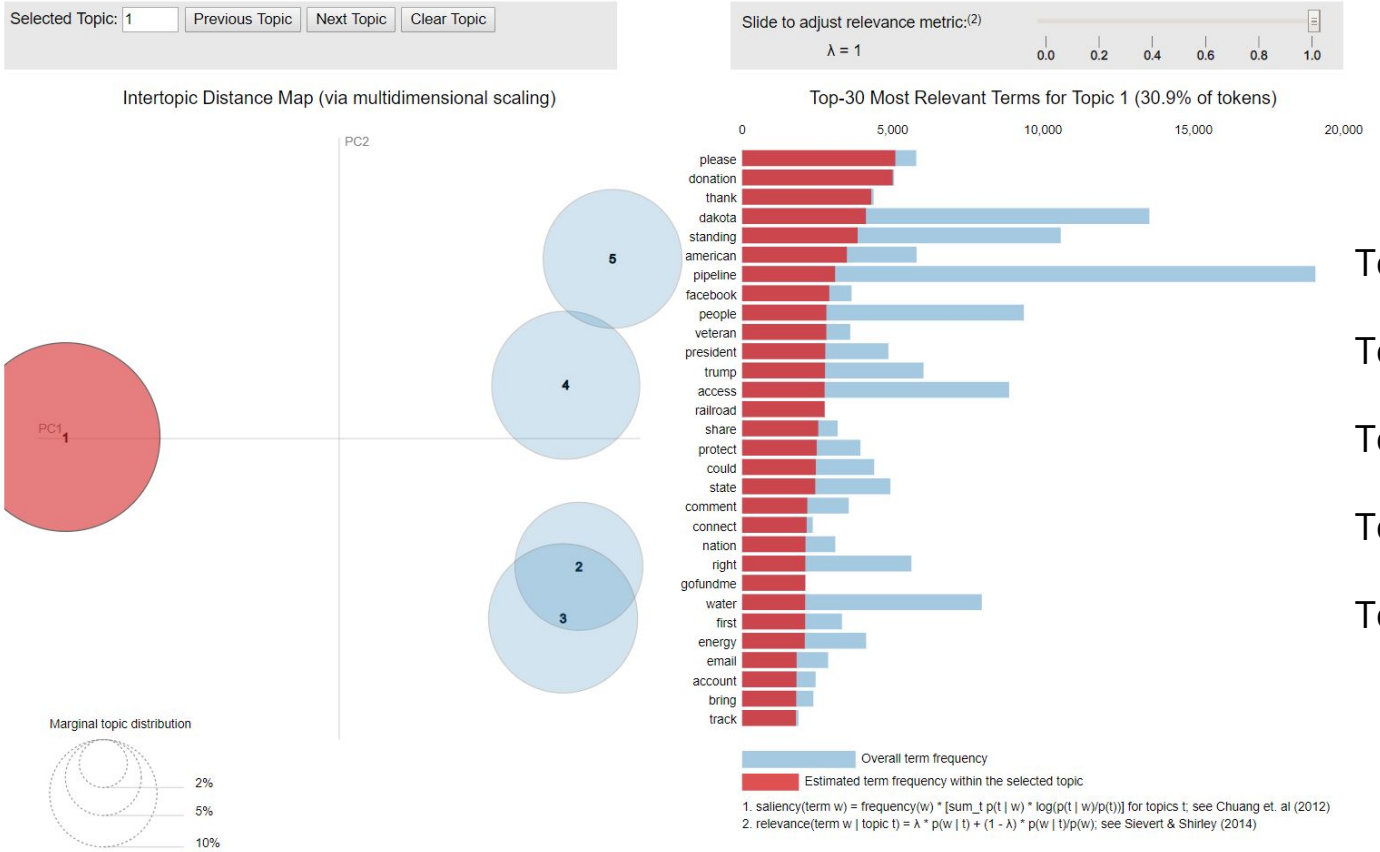
- LR2 using keywords from internal source



Fold No.	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average
Training Accuracy	0.876	0.988	0.876	0.876	0.962	0.916
Testing Accuracy	0.700	0.650	0.700	0.800	0.571	0.684

Topic Modelling - Using LDA and LDA2Vec

- LDA Model:
  - Document within a collection: is modeled as a finite mixture of topics
  - Each topic: is modeled as an infinite mixture distribution over an underlying set of topics probabilities.
- Used LDA model provided by Gensim
- Evaluation with small corpus
  - Performed LDA analysis on small corpus both before and after classification
  - Purpose: to evaluate the testing classification performance on corpus
- LDA2Vec:
  - A deep learning variant of LDA topic modelling developed recently by Moody (2016)
  - The topics found by LDA were consistently better than the topics from LDA2Vec



Topic 1: Donation Petition

Topic 2: Government Activities

Topic 3: Protest Preparation

Topic 4: Details of Protest

Topic 5: Concerns of Pipeline

Outline

- Introduction
  - Automatic Summarization and NoDAPL Dataset
  - Related Work
  - Overview of the Proposed Framework
- Preprocessing of Data
  - Classification of Relevance
  - Topic Modelling by Latent Dirichlet Allocation (LDA) and LDA2Vec
- **Hybrid Method**
  - Latent Dirichlet Allocation based Extractive Summarization
  - Pointer-generator based Abstractive Summarization
  - Text Re-ranking based Hybrid Summarization
- Results and Evaluation
  - Compiled Summary
  - Extrinsic Evaluation
- Conclusion and Future Work

Extractive Summary - TF-IDF based Ranking

- Sentence Extraction Using TF-IDF based Ranking
  - Created a counting vector (bag of words)- using sklearn's CountVectorizer
  - Built TF-IDF matrix- using sklearn's TF-IDF Transformer
  - Scoring each sentence
    - We only considered and added the tf-idf values where the underlying token was a noun. This total was then divided by the summation of all the document tf-idf values.
    - We added an additional value to a given sentence if it had any word that was included in the title of the document. This value was equal to the count of all words in a sentence found in the title divided by the total number of words in the title. This "heading similarity score" was then multiplied by an arbitrary constant (0.1) and added to the tf-idf value.
    - We then applied a position weighting. Each sentence was ordered from 0 to 1 equally based on the sentence number in the document. This weighting was then multiplied by the value in point 2.

Extractive Summary - TF-IDF Text Ranking Result

This outdated fossil fuel infrastructure is not needed and railroads serving the Bakkan North Dakota oil fields have invested heavily in improved railroad oil tank car safety and have stated that they have more than enough railroad capacity to transport the Bakkan oil. We know that President Elect Trump has a serious conflict of interest by owning large investments in DAPL and other fossil fuel assets and his energy team includes Harold Hamm billionaire founder of Continental Resources oil company and someone Mr Trump might name as his Secretary of Energy

Example: Summary for the first document.

Limitations:

- Requires large memory to store the frequent word dictionary and TF-IDF matrix
- Does not guarantee relevance of the top-ranked sentences

Extractive Summary - LDA based Ranking

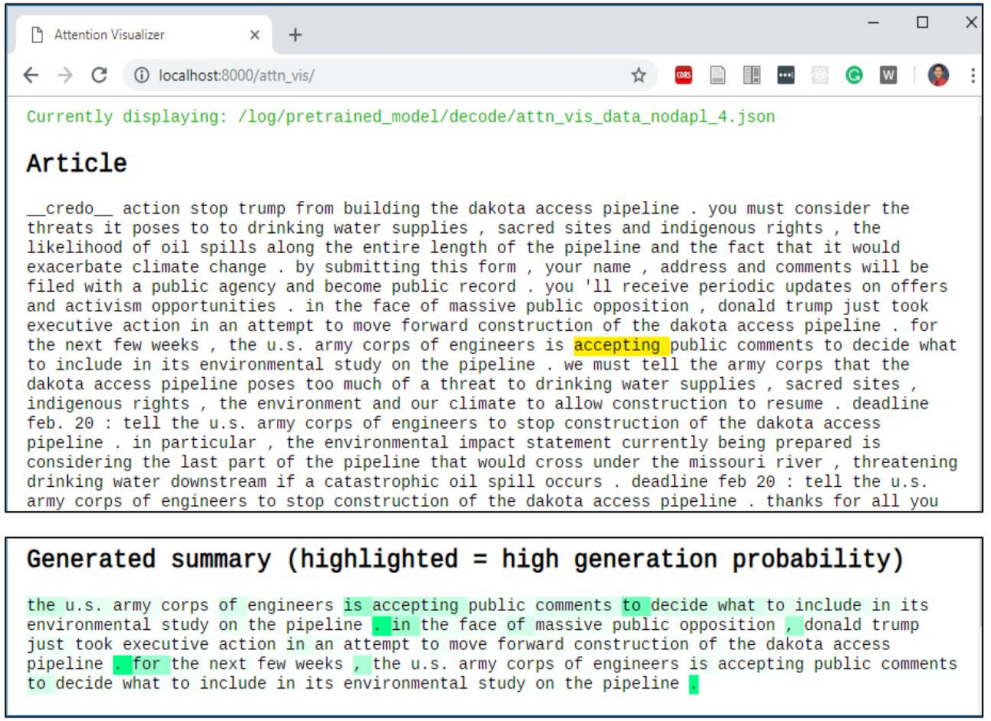
- LDA based Ranking
  - Use LDA model to rank sentences within each document for all topics
  - The rationale behind using LDA similarity queries as ranking score is
    - Given a topic, extracted sentences should be highly relevant to the topic
  - As a result, a list of sentences and the associated similarity measures (range: 0~1) were obtained for each document
  - The Top-N sentences were extracted by sorting the similarity scores
- Example Generated Summary

The permitting and construction process for the expansion of the existing clean energy facility has also been slowed because Dominion accountants, planners and decision makers convinced themselves that their stockholders would see more immediate profit from fracking, building pipelines, and converting the Cove Point facility into an export terminal than completing a facility that would supply a couple million of my fellow Virginians with 60-80 years worth of clean electricity. ...

Limitation: Failed to contain named entities.

Abstractive Summary

- Pointer Generator Network*
  - Generated 3-sentence abstract for each document
  - Used pre-trained with CNN/Daily Mail dataset
  - Used default hyperparameters with TensorFlow 1.2.1



Example PGN-generated abstract (in attention visualization)

* Abigail, *et.al* . "Get to the point: Summarization with pointer-generator networks." arXiv preprint arXiv:1704.04368 (2017).

Hybrid Automatic Summarization

- We proposed a hybrid automatic summarization approach to further improve the summarization performance in two considerations:
  - The extractive summarization only led to sentences which were highly relevant to corresponding topics but ignored the **important named entities**
  - The abstractive summarization produced high level summarization of documents but may not have **high relevance with topics**
- Therefore, we proposed to re-rank these sentences to generate better summarization according to two aspects:
  - The topics
  - The named entities

Named Entity Recognition

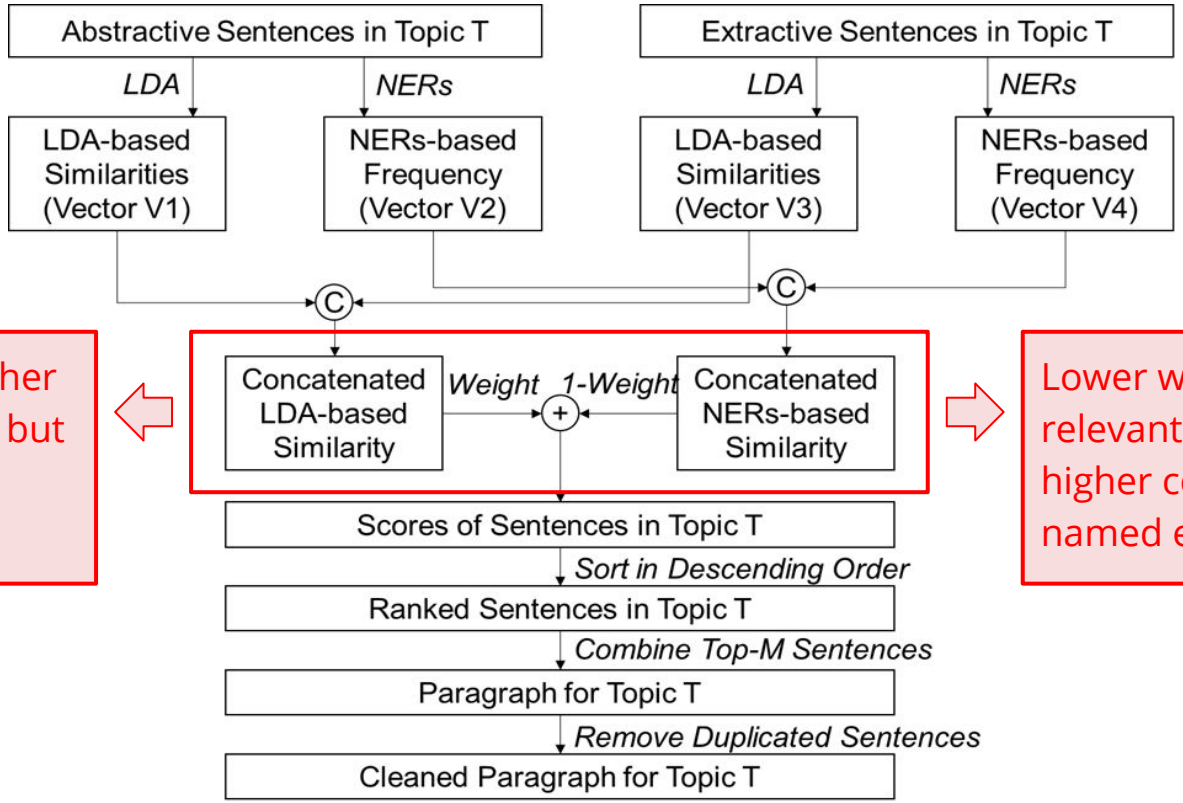
- Extracted information through named entity recognition using spaCy

```
Counter({('Wednesday', 'DATE'): 4, ('Dakota Access', 'ORG'): 4, ('dozens', 'CARDINAL'): 4, ('Plymouth', 'GPE'): 1, ('
North Dakota', 'NORP'): 1, ('four', 'CARDINAL'): 1, ('Energy Transfer Partners', 'ORG'): 1, ('Standing Rock', 'ORG'):
1, ('Thousands', 'CARDINAL'): 1, ('Standing Rock Sioux', 'GPE'): 1, ('Saturday', 'DATE'): 1, ('one', 'CARDINAL'): 1,
('5', 'CARDINAL'): 1, ('December', 'DATE'): 1, ("the Standing Rock Sioux Tribe's", 'ORG'): 1, ('the Army Corps', 'ORG
'): 1, ('US', 'GPE'): 1, ('Sunday', 'DATE'): 1, ('The US Army Corps of Engineers', 'ORG'): 1, ('Katy Perry', 'PERSON'
): 1, ('Orlando Bloom', 'PERSON'): 1, ("Lupita Nyong'o", 'PERSON'): 1, ('Whoopi Goldberg', 'PERSON'): 1, ('Mark Ruffa
lo', 'PERSON'): 1, ('Cancel', 'FAC'): 1, ('Cancel This', 'ORG'): 1, ('Cancel', 'GPE'): 1})
```

Type	DESCRIPTION
PERSON	People, including fictional
ORGANIZATION	Companies, agencies, institutions etc
GPE	Countries, cities, states
LOCATION	Non GPE locations, bodies of water etc

Type	DESCRIPTION
DATE	Absolute or relative dates or periods
TIME	Times smaller than a day
ORDINAL	First, second etc
CARDINAL	Numerals that do not fall under another type

Hybrid Summary



Higher weight: Higher relevant to Topics, but lower coverage of named entities

Lower weight: Lower relevant to Topics, but higher coverage of named entities

Outline

- Introduction
  - Automatic Summarization and NoDAPL Dataset
  - Related Work
  - Overview of the Proposed Framework
- Preprocessing of Data
  - Classification of Relevance
  - Topic Modelling by Latent Dirichlet Allocation (LDA) and LDA2Vec
- Hybrid Method
  - Latent Dirichlet Allocation based Extractive Summarization
  - Pointer-generator based Abstractive Summarization
  - Text Re-ranking based Hybrid Summarization
- **Results and Evaluation**
  - Compiled Summary
  - Extrinsic Evaluation
- Conclusion and Future Work

Hybrid Summarization Results

- Resulting in five paragraphs for five topics, respectively
- Summary on Topic 5 - Concerns about the Pipeline:

> *In particular, the environmental impact statement currently being prepared is considering the last part of the pipeline that would cross under the missouri river, threatening drinking water downstream if a catastrophic oil spill occurs. We must tell the army corps that the dakota access pipeline poses too much of a threat to drinking water supplies, sacred sites, indigenous rights, the environment and our climate to allow construction to resume. The dakota access pipeline would result in oil leaks in the missouri river watershed. Oil pipeline threatens standing rock sioux land, drinking water holy places. For the next few weeks, the u.s. Army corps of engineers to stop construction of the dakota access pipeline.*

Extrinsic Evaluation

- Task-based Measurement:
  - Developed guideline question as standards: 23 questions covering import information for NoDAPL
  - Built question-answer matches: mark questions that answered by summary sentences
- Evaluate Summarization Quality by Two Measures
  - Content relevance: **91.3%**
  - Question coverage: **69.6%**

Sentence ID	Sentences	Question ID
I-1	President obama says the dakota access...	T6-a
I-2	Thank you for temporarily halting the...	T4-i
I-3	Thank you for temporarily halting the...	T4-i
I-4	We know that president elect trump has...	T6-a
I-5	Trump might name as his secretary of...	
I-6	After the failed keystone pipeline, a ...	T2-e
I-7	The dakota access pipeline will be built...	T4-c
I-8	Veterans of the united states armed force...	T4-a
II-1	Members of the new orleans community ...	T4-a
II-2	Arlea ashcroft has been to the standing ...	T4-e
II-3	The wplc provided legal support on the...	T4-f
II-4	This year, all were in proud support of ...	T4-a
II-5	Everyone keeps asking how they can help ...	T4-b
II-6	By now, we all know how that turned out ...	T4-i
II-7	Sioux in the dakota borderlands and...	T4-e
II-8	Cambridge is standing in solidarity with...	T4-a
II-9	Oceti sakowin camp is one of eight stanl...	T4-a
II-10	The first amendment is under attack along...	T4-i
II-11	The standing rock nation film & music...	T4-e
...	...	...

Outline

- Introduction
  - Automatic Summarization and NoDAPL Dataset
  - Related Work
  - Overview of the Proposed Framework
- Preprocessing of Data
  - Classification of Relevance
  - Topic Modelling by Latent Dirichlet Allocation (LDA) and LDA2Vec
- Hybrid Method
  - Latent Dirichlet Allocation based Extractive Summarization
  - Pointer-generator based Abstractive Summarization
  - Text Re-ranking based Hybrid Summarization
- Results and Evaluation
  - Compiled Summary
  - Extrinsic Evaluation
- **Conclusion and Future Work**

Conclusion and Future Work

- Conclusion
  - A hybrid automatic summarization approach is proposed and tested on NoDAPL dataset with limited human effort
  - A summarization guideline according to our thorough understanding of NoDAPL events/topics was developed to extrinsically evaluate the generated summary
  - The content relevancy score and question coverage score indicates acceptable relevance and coverage of the generated summary
- Future Work
  - More text dataset for interesting topics can be tested by using the proposed hybrid text summarization approach
  - Topic modeling-supervised classification approach can be investigated to minimize human efforts in automatic summarization
  - Deep learning-based recommender system can be investigated for better sentence re-ranking.

Acknowledgement

We would like to thank the instructions and invaluable comments from Prof. Fox and GTA Mr. Liuqing Li. And special thanks to the creative ideas provided by our classmates!

NSF IIS-1619028

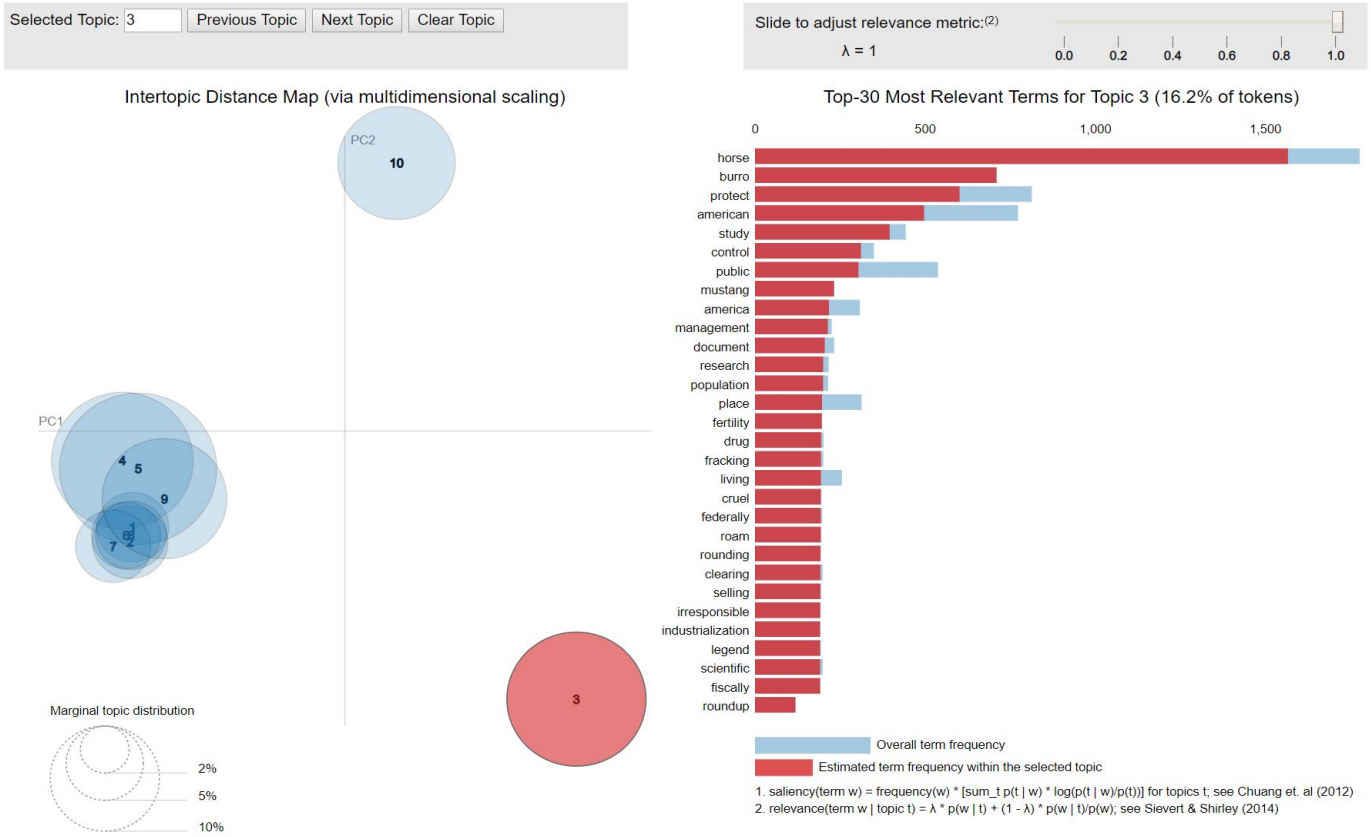
Thank you!

Questions?

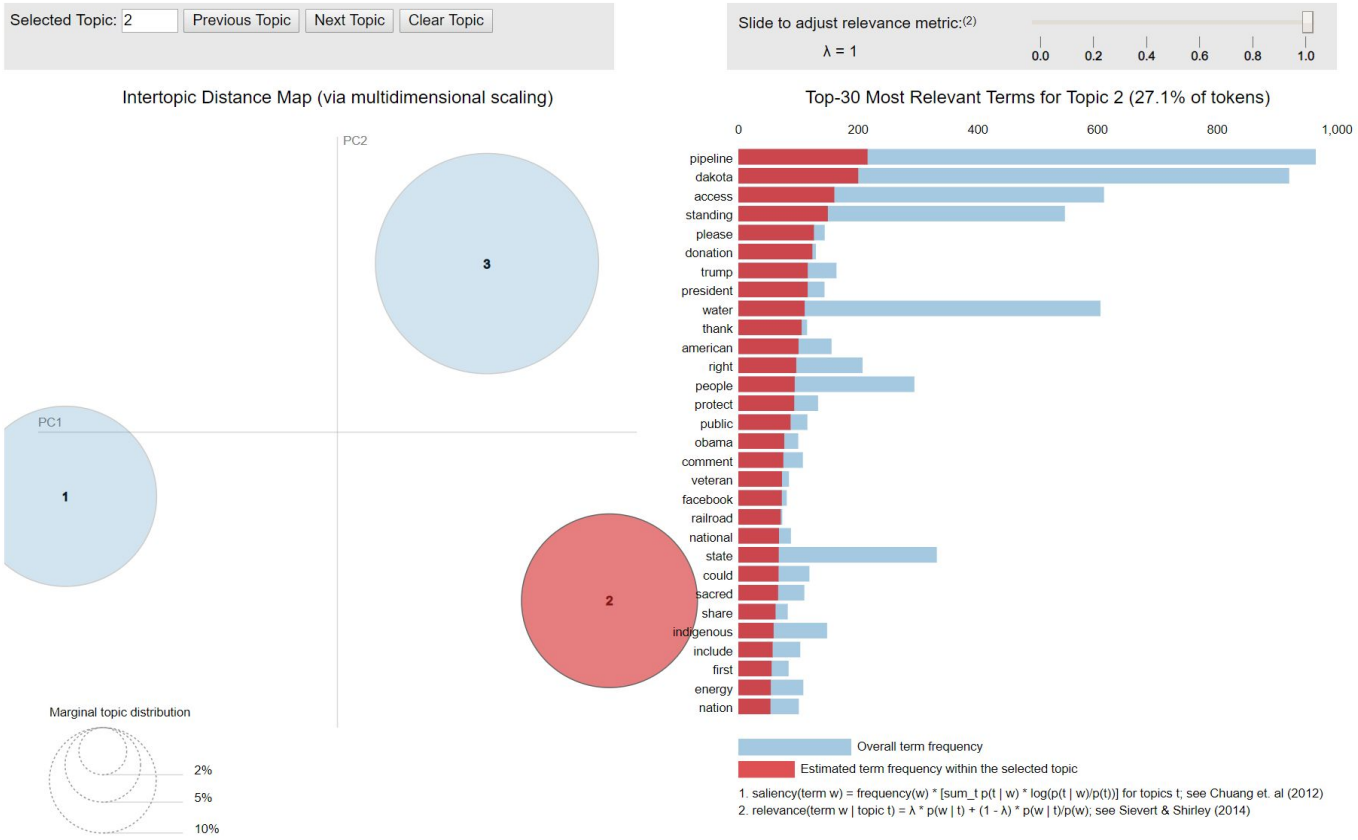
Backup

Related Work

- Based on Processing Techniques
  - Extractive summarization (Gupta & Lehal, 2010)
  - Abstractive summarization (Banerjee, Mitra, & Sugiyama, 2015)
- Based on Documents
  - Single document summarization (Litvak & Last, 2010)
  - Multi-document summarization (Banerjee, Mitra, & Sugiyama, 2015)
- Based on Audiences
  - Generic summarization (Zha, 2002)
  - Query-focused summarization (Daumé III & Marcu, 2006)
- Deep Learning-based Automatic Text Summarization
  - Seq2Seq model (Khatri, Singh, & Parikh, 2018)
  - Pointer-generator network (See, Liu, & Manning, 2017)



LDA analysis on original corpus without classification, the Topic 3 in red (i.e., circle in red) is irrelevant according to the frequent word list on the right hand side.



LDA analysis on the corpus after classification (relevant documents only). Three topics are found from the small corpus, each represents different relevant NoDAPL events/topics.

Abstractive Summary

- Beam Search Decoder
  - The framework we implemented relied on the encoder-decoder paradigm.
  - The encoder encoded the input sequence of words, while the decoder which uses a beam search algorithm transformed the probabilities over each word in the vocabulary and produced a final sequence of words.
  - The hyperparameters used for beam search decoder were
    - max_enc_steps=400
    - max_dec_steps=120
  - Coverage=1 eliminated repetition of same words

Topic Modelling- Using LDA2Vec

- LDA2Vec is a deep learning variant of LDA topic modelling developed recently by Moody (2016)
- LDA2Vec model mixed the best parts of LDA and word embedding method-word2vec into a single framework
- According to our analysis and results, traditional LDA outperformed LDA2Vec
- The topics found by LDA were consistently better than the topics from LDA2Vec
- Hence, we sticked with LDA for topic modelling and clustering