

Object Detection and Topic Modeling

Dr. Edward Fox
Virginia Tech, Blacksburg VA 24061
Dec. 1, 2022

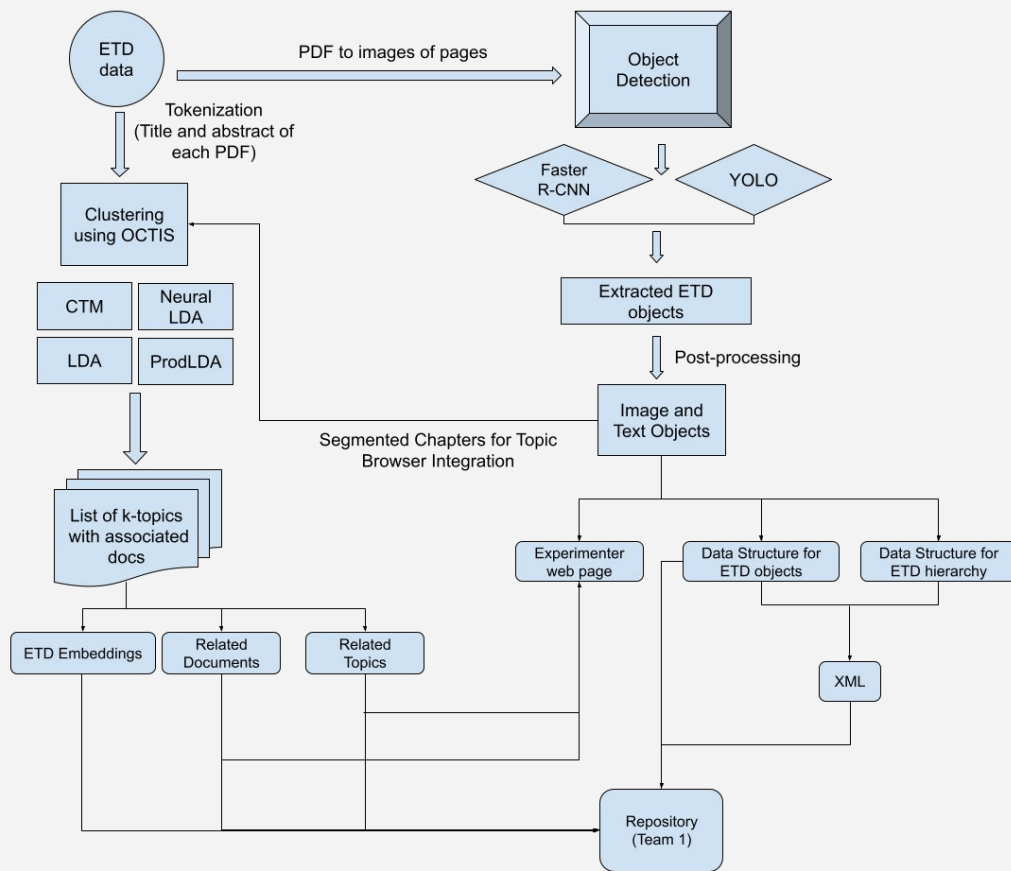
CS5604 Team 3

Nila Masrourisaadat, **Raj** Sahu, **Alan** Devera, **Nirmal** Amirthalingam, **Chenyu** Mao

Overview

- 5 teams were tasked to build an information storage and retrieval system from scratch to make Electronic Theses and Dissertations (ETDs) more accessible to the researchers, experimenters, and curators.
- Our team (Team 3) was tasked with detecting objects (Object Detection) within ETDs as well as determining topics (Topic Modeling) to store in a repository (Team 1) for better search and recommendation of ETDs (Team 2).

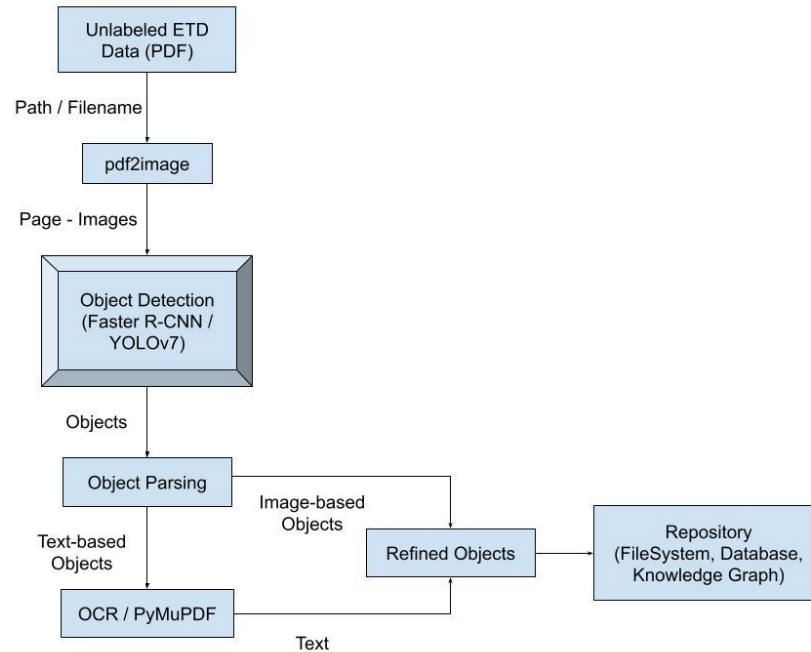
Overall Pipeline



Object Detection

- Object Detection overview
- ETD filtering rules
- Issues and challenges faced
- Services
- 5k ETDs
- Demo
- Future work

Object Detection Pipeline



Tools Used

- Pdf2image
- PyMuPdf
- YOLOv7
- Detectron2 (Faster R-CNN)
- Flask
- Docker
- VT CS Cloud

YOLOv7 & Detectron (Faster R-CNN)

YOLOv7:

- A real-time object detector that has greatly advanced the CV & ML world
- Fastest, and most accurate object detector to date
- Uses CNN to predict bounding boxes and class probabilities considering the entire image at one step

Detectron2 :

- A framework for object detection built on top of PyTorch
- Has support for object detection, activity recognition, semantic/instance segmentation
- Trains an object detection model on custom datasets using pre-trained weights
- All models in Detectron2 are pre-trained on the COCO dataset

***For our experiments, we would be using faster R-CNN models present in Detectron2

XML Schema

- Each PDF has its ETD ID as the root element
- 3 sub-elements - front, body, and back
- Image-based objects
 - Figures
 - Tables
 - Equations
 - Algorithms
- Image-based objects have captions / numbers associated with them
- The rest are text-based objects

object ▶ etd ▶ body ▶ chapter ▶ sections ▶

```
▼ object {1}
  ▼ etd {3}
    ▼ front {10}
      title : Document Title
      author : Author Name
      university : University
      degree : Degree Type
      committee : Committee
      date : Date or Month/Year
      abs_heading : Abstract
      abs_text : In this..
      toc_heading : Table of..
      toc_text : 1. Intro ...
    ▼ body {1}
      ▼ chapter {3}
        title : Chapter-1..
        page_no : 1
        ▼ sections {1}
          ▼ section {7}
            name : 1.1..
            ▶ paragraphs {1}
            ▶ figures {1}
            ▶ tables {1}
            ▶ equations {1}
            ▶ algorithms {1}
            ▶ footnotes {1}
      ▼ back {2}
        ref_heading : Ref..
        ref_text : ..
```


False positives

Problem:

- Last line of the paragraph in the previous page is detected as a chapter title

Solution:

- Checked the first letter of the title (capital)
- Avoided the text overflow for consecutive pages by checking the punctuation
- Check if it is a subset of table of contents

8

adjustments to society, or he does not, in varying amounts and degrees.

According to Burnham, there are three main factors which influence the adjustment of the child. They are the home, the school, and the child. Other factors also enter into the picture, such as the church, Boy and Girl Scout organizations, and other organizations.¹

False positives

Problem:

- Image-based objects (figures, tables) and their corresponding captions might not be on the same page

Solution:

- Determined the relationship between figure/table and their corresponding caption
- Paired images with their corresponding captions according to them being on top or bottom of images

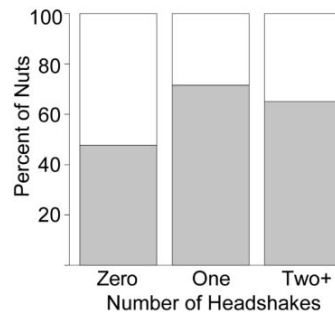


Figure 2.8. Outcome by assessment. The proportion of nuts eaten (□) or cached (■) based on the number of head flicks. Head flicks predicted a greater likelihood of caching nuts instead of eating them.

16

likelihood of caching ($Z = 3.13, p = 0.002$), with squirrels that did not head flick caching 48% of nuts, and squirrels that head flicked one or more times caching 69.8% of nuts

False positives

Problem:

- Chapter titles appear on multiple pages for the same chapter in some cases

Solution:

- Checked if two consequent chapter titles are the same

ETD Filtering Rules

Post-processing rules are as follows:

1. Linked image object with the previous caption object in the list
 - a. Determined the relationship between figure/table and their corresponding caption
 - b. Paired images with their corresponding captions according to them being on top or bottom of images
 - c. Checked the orientation of the pages in case of being horizontal and changed to vertical
2. Filtered chapters/sections titles
 - a. Checked the first letter of the title (capital)
 - b. Checked if the y-coordinate of the bounding box is within half of the page's height
 - c. Avoided the text overflow for consecutive pages by checking the punctuation
 - d. Checked if two consequent chapter titles are the same
 - e. Created null section tag in case of titles not being detected
3. Eliminated false positives based on matching with ToC
 - a. Extracted and saved all the detected chapter/section titles, ToC, page numbers
 - b. Filtered out the chapter and section titles that have been incorrectly detected
4. Refine chapter title detection by removing outliers
 - a. Find if the keyword “chapter” is in the list of all detected chapter titles
 - b. Difference in font size between chapter titles and other objects
 - c. Check the indentation level - left / center alignment

Team 3 Deliverables - Objects

Discussing with the other teams, and assessing the overall requirements of the class, we are providing the following outputs:

Output for a given ETD	Description/Contents
Page images	Each page of ETD is saved as a .jpg file
Detected images	Consists of sub-folders for each image-based object class (figure, table, equation, algorithm) saved as .jpg files
JSON object	Each object - text and image-based, is saved (unordered set as per the detection sequence)
Parsed XML	Detected objects in a tree structure with XML elements set to the clean text for text-based objects, and the image path for image-based objects

Team 3 - Services

We are providing 3 services that are containerized - to be used for workflow automation and our frontend webpage

- Generate page images given an ETD ID
- Generate outputs given an ETD ID - using YOLOv7
 - Detected images
 - JSON objects
 - XML
- Generate outputs given an ETD ID - using Faster R-CNN (Detectron2)
 - Detected images
 - JSON objects
 - XML

5k ETDs - Inference and Storing in DB

- Currently using the save detected object API (from team 1) to store objects in DB
- Finished running on 5000 ETDs
- Average inference time ~ 3 minutes (GPU)

Save detected object (POST) API -

`https://team-1-flask.discovery.cs.vt.edu/v1/objects/<etd_id>/<type_name>`

- `<etd_id>/page`
- `<etd_id>/text`
- `<etd_id>/image`
- `<etd_id>/xml`

5k ETDs - Inference and Storing in DB

Type	File	Metadata
page	<page_number>.jpg	-
text	-	{keys = type, text, page_num, bbox}
image	<etd_id>_<class>_<count>.jpg	{keys = type, path, page_num, bbox}
xml	<etd_id>.xml	-

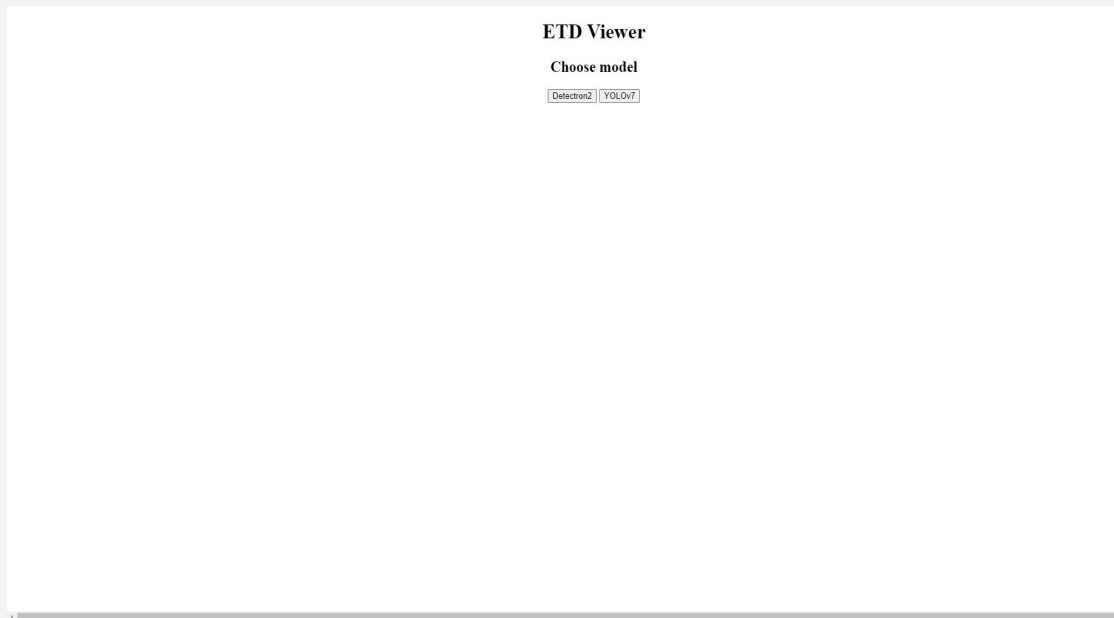
Get objects by ETD ID API

https://team-1-flask.discovery.cs.vt.edu/v1/etds/<etd_id>/objects?type=<type_name>

Experimenter UI Choosing a model


2 Models:

- YOLOv7
- Detectron2 (Faster R-CNN)



Experimenter UI Upload ETD

- Experimenter has the option to upload a pdf of the ETD that they want to perform object detection on



The screenshot shows a web interface titled "File Upload". Below the title, there is a text input field containing the placeholder text "Choose File" and "No file chosen". To the right of the input field is a "Submit" button.

Experimenter UI ETD Browser

ETD Browser

Introduction	>
Background	>
Modeling Independent Events	>
Modeling Temporal Evolutionary Events	>
Modeling Events from Multiple Data Sources	>
Conclusions and Future Work	>

Probabilistic Models of Topics and Social Events

CMU-ISR-16-113

University:

Carnegie Mellon University Pittsburgh, PA

Degree:

Committee:

Thesis Committee: Kathleen M. Carley Tom Mitchell Alexander J. Smola Huan Liu, Arizona State University Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Date:

Abstract

Abstract Structured probabilistic inference has shown to be useful in modeling complex latent structures of data. One successful way in which this technique has been applied is in the discovery of latent topical structures of text data, which is usually referred to as topic modeling. With the recent popularity of mobile devices and social networking, we can now easily acquire text data attached to meta information, such as geo-spatial coordinates and time stamps. This metadata can provide rich and accurate information that is helpful in answering many research questions related to spatial and temporal reasoning. However, such data must be treated differently from text data. For example, spatial data is usually organized in terms of a two-dimensional region while temporal information can exhibit periodicities. While some work existing in the topic modeling community that utilizes some of the meta information, these models largely focused on incorporating metadata into text analysis, rather than providing models that make full use of the joint distribution of meta information and text. In this thesis, I propose the event detection problem, which is a multi-dimensional latent clustering problem on spatial, temporal and topical data. I start with a simple parametric model to discover independent events using geo-tagged Twitter data. The model is then improved toward two directions. First, I augmented the model using Recurrent Chinese Restaurant Process (RCRP) to discover events that are dynamic in nature. Second, I studied a model that can detect events using data from multiple media sources. I studied the characteristics of different media in terms of reported event times and linguistic patterns. The approaches studied in

- Sidebar to navigate to different chapters and sections
- Images are linked to their respective captions

Introduction	>
Background	>
Modeling Independent Events	>
Modeling Temporal Evolutionary Events	>
Modeling Events from Multiple Data Sources	>
Conclusions and Future Work	>

model Third, there is a which contains information about the topical content of the language model, documents, Table 3.1 gives a summary of all notation that will be used as we describe the model in this section.

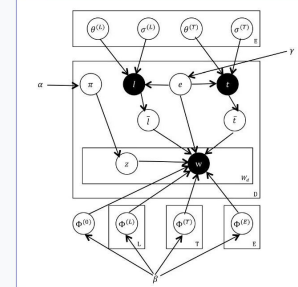


Figure 3.1: Illustrations of the model in plate notations

3.3.1 Event Model

An important observation incorporated into our model is that events are in many ways natural extensions of topics: events have a topical focus but also include a spatial and temporal region in which they are likely to occur. We thus assume events are defined by three things: first, each θ_{ij} event has a geographical center as well as a geographical variance controlled by a diagonal σ_{ij} covariance matrix with each value defined by σ_{ij} . The location of a document that belongs to an event is assumed to be drawn from a two dimensional Gaussian distribution governed by these

See the video demonstration using the file ETDViewerDemo.mp4

Milestones/Timeline

- Work done till IR-2
 - Implemented basic XML schema for detected objects
 - Developed post-processing rules/filters for YOLOv7
 - Front-end wireframes for the Experimenter Web Page
- Work done for IR-3
 - Added support for Faster R-CNN
 - Finalized working models and parser logic for YOLOv7 and Faster R-CNN (base models)
 - Deployed both models on the cloud server (containerized)
 - Team 1 deliverables
 - Working prototype of the Experimenter Web Page using Flask that supports both models
- Work done in November and December
 - Experiment with a bigger subset of the ETD dataset
 - Run inference on 5k ETDs using team 1's API to store into their DB
 - Dockerized services for object detection and enabled CI/CD features for the frontend

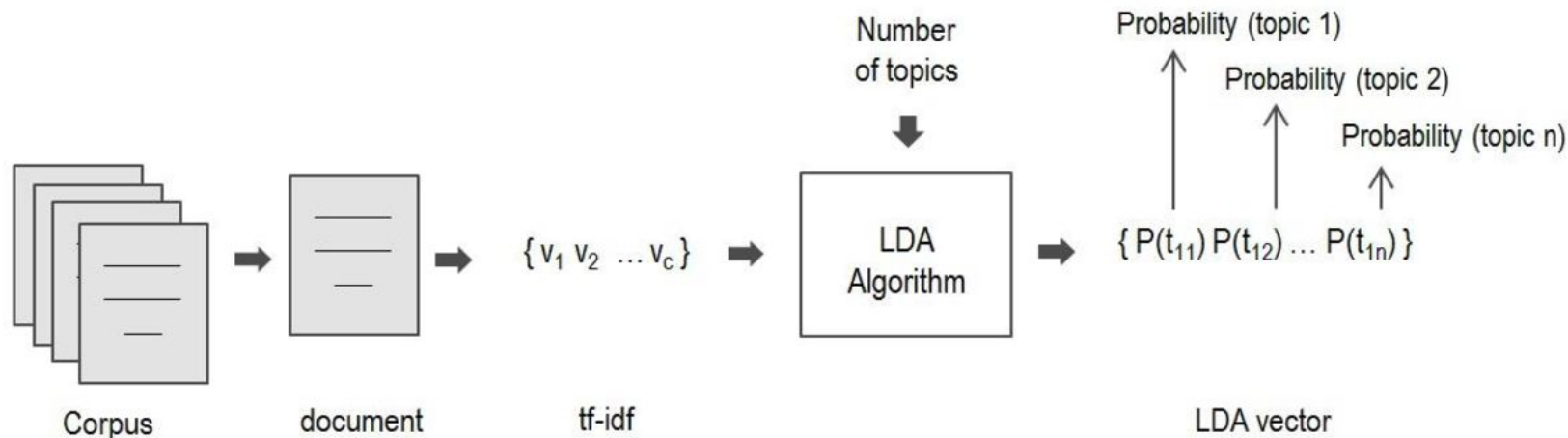
Future Work - Object Detection

- Integrate Flask app / Experimenter UI with the frontend and workflow services
- Update our pipeline to support “Add an ETD” once team 1 provides the API for saving metadata
- Set up workflow automation - team 5
- Add support for scanned documents - modify existing pipeline
- Add more post processing rules based on the results of the 5k dataset
- Improve UI layout and add more functionality

Topic Modeling

- Background
- Pipeline
- Chapters
- Experimenter pages + Demo
- Services
- Milestones
- Wrap-Up Tasks

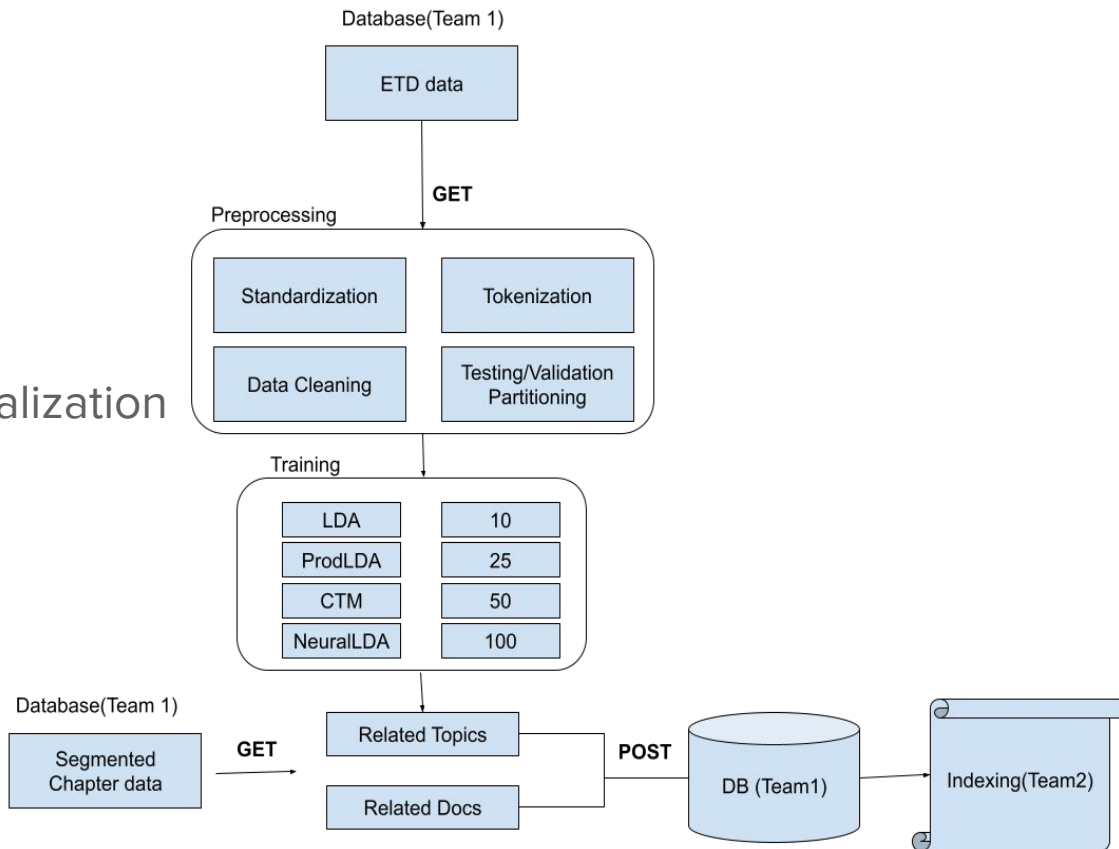
Background - **Vector Distance**



- Input ETD -> Octis -> Vectors
- Vectors [0.31, 0.47, 0.02, ... , 0.21]
- Shape : (1, 50)
- Similarity = Euclidean Distance of two vectors

TM - Pipeline

- Dataset preprocessing
- Topic Model training
- Inference Learning
- Topic and Documents Visualization





TM - Deliverables

Output	Description
Related Topics & Related Docs	For a given ETD id or Chapter id, find the most similar topics and documents respectively.
Topic Browser (Experimenter)	Allow a user to browse all topics, click on one and read associated documents. (Documents by Topic - Chapters by Topic)

TM - Chapters - Initial Experiment

- Extract all **<para>** tags to form a chapter from an ETD's XML.

```
object ▶ eld ▶ body ▶ chapter ▶ 4 ▶ sections ▶ section ▶ 3 ▶ equations ▶ equation ▶ 1 ▶  
▼ object {1}  
  ▼ eld {3}  
    ▶ front {10}  
    ▼ body {1}  
      ▼ chapter {6}  
        ▶ 0 {2}  
        ▶ 1 {2}  
        ▶ 2 {2}  
        ▶ 3 {2}  
        ▼ 4 {2}  
          title : Chapter 5\nModeling Events from Multiple Data\nSources  
          ▼ sections {1}  
            ▼ section {12}  
              ▶ 0 {7}  
              ▶ 1 {7}  
              ▶ 2 {7}  
              ▼ 3 {7}  
                name : 5.4 Initializations  
                ▼ paragraphs {1}  
                  para : A naïve approach to initialize the latent variables is to use  
uniformly generated random variables to serve as the initial values  
of  $z_i$ ,  $q$  and  $s$ . However, if those initial values are bad, it is  
likely that\nthe algorithm would take a long time to reach  
equilibrium. Instead, we can sample the initial  $z_i$ ,  $q_i$  values and  
using parts of Equation 5.3 and Equation 5.4 that do not require  
our knowledge\nof cluster index  $s$ . For location index  $i$ , its  
initial value is sampled purely based on its location proximity to  
the Gaussian centers. For word category variable  $q$ , its initial  
values are determined by both the values of previous words in the  
current document and prior  $\eta\delta$ .  
                figures :   
                tables :   
                ▼ equations {1}  
                  ▼ equation {2}  
                    ▶ 0 {2}  
                      path : detections/image-objects/equation_CMU-ISR-16-113_1.jpg  
                      eq_no : (5.11)  
                    ▼ 1 {2}
```

TM - Chapters - Initial Experiment - **Dataset**

Old Table schema :

id	Title	Abstract	Author	Year	University
1	ETD Title 1	ETD Abstract1	A1,A2	1900	Virginia Tech
2	ETD Title 2				
3	ETD Title 3				

TM - Chapters - Initial Experiment - **Dataset**

New Table schema :

id	Type	Parent	Title	Abstract	Author	Year	University
1	ETD	NULL	ETD Title 1	ETD Abstract 1	A1,A2	1900	Virginia Tech
2	ETD	NULL	ETD Title 2				
3	ETD	NULL	ETD Title 3				
4	Chapter	1	Chap Title 1	Chap Abstract 1			
5	Chapter	1	Chap Title 2	Chap Abstract 2			

TM - Chapters - Initial Experiment - **Issues**

Dataset size \rightarrow 10x (Assuming each ETD has 10 Chapters)

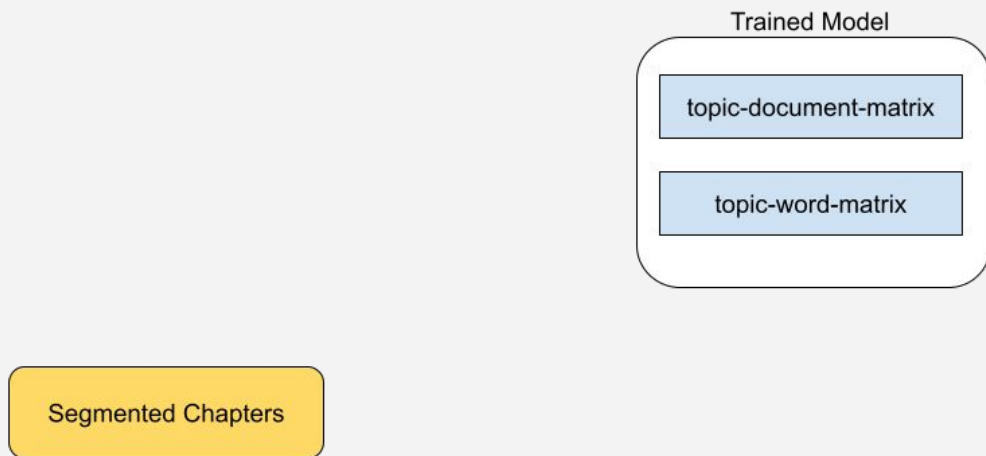


For same memory, reduce Overall number of ETDs and the vocabulary size.

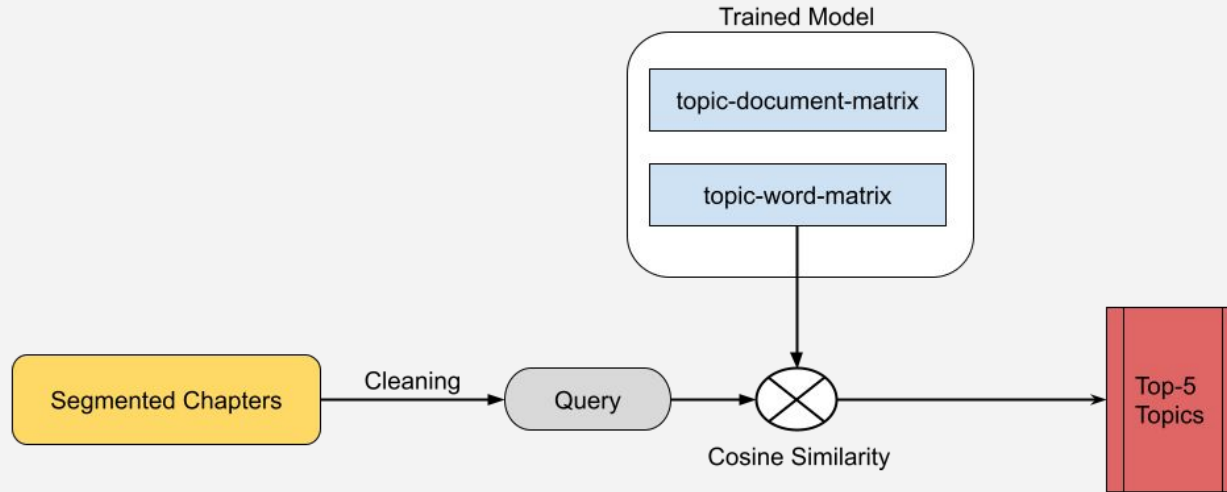


Reduced quality of topic categorization and search

TM - Chapters - **Inference**



TM - Chapters - Inference - **Related Topics**



TM - Chapters - Inference - Related Topics

```
# convert word list into N-hot encoding where total array size is length of vocabulary
terms_idx=[word2idx.get(w, None) for w in words_list ]
query_vector=np.zeros(len(word2idx))
count = 0
for idx in terms_idx:
    if idx is not None:
        query_vector[idx]=1.0
        count+=1

### Related Topics
# Using cosine similarity, find the most similar looking vocabulary array
# compared to our N-hot encoded words_list
query_vector =np.expand_dims(query_vector, axis=0) # convert to 2d for cosine similarity
distances=cosine_similarity(query_vector,model['topic-word-matrix'])
matching_topics_idx = np.argsort(-distances)
matching_topics_idx = np.squeeze(matching_topics_idx) # decreasing order of most similar topic ids

logger("Related Topics : ",matching_topics_idx[:topk],2)
```

TM - Chapters - Inference - Related Topics

```
# convert word list into N-hot encoding where total array size is length of vocabulary
terms_idx=[word2idx.get(w, None) for w in words_list ]
query_vector=np.zeros(len(word2idx))
count = 0
for idx in terms_idx:
    if idx is not None:
        query_vector[idx]=1.0
        count+=1
```

Cleaned
chapter
text

```
### Related Topics
# Using cosine similarity, find the most similar looking vocabulary array
# compared to our N-hot encoded words_list
query_vector =np.expand_dims(query_vector, axis=0) # convert to 2d for cosine similarity
distances=cosine_similarity(query_vector,model['topic-word-matrix'])
matching_topics_idx = np.argsort(-distances)
matching_topics_idx = np.squeeze(matching_topics_idx) # decreasing order of most similar topic ids

logger("Related Topics : ",matching_topics_idx[:topk],2)
```

TM - Chapters - Inference - Related Topics

```
# convert word list into N-hot encoding where total array size is length of vocabulary
terms_idx=[word2idx.get(w, None) for w in words_list ]
query_vector=np.zeros(len(word2idx))
count = 0
for idx in terms_idx:
    if idx is not None:
        query_vector[idx]=1.0
        count+=1

### Related Topics
# Using cosine similarity, find
# compared to our N-hot encoded
query_vector = np.expand_dims(query_vector, axis=-1) # convert to 2d for cosine similarity
distances=cosine_similarity(query_vector,model['topic-word-matrix'])
matching_topics_idx = np.argsort(-distances)
matching_topics_idx = np.squeeze(matching_topics_idx) # decreasing order of most similar topic ids

logger("Related Topics : ",matching_topics_idx[:topk],2)
```

Cleaned chapter text

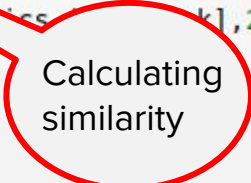
N-hot encoding

TM - Chapters - Inference - Related Topics

```
# convert word list into N-hot encoding where total array size is length of vocabulary
terms_idx=[word2idx.get(w,None) for w in words_list ]
query_vector=np.zeros(len(word2idx))
count = 0
for idx in terms_idx:
    if idx is not None:
        query_vector[idx]=1.0
        count+=1

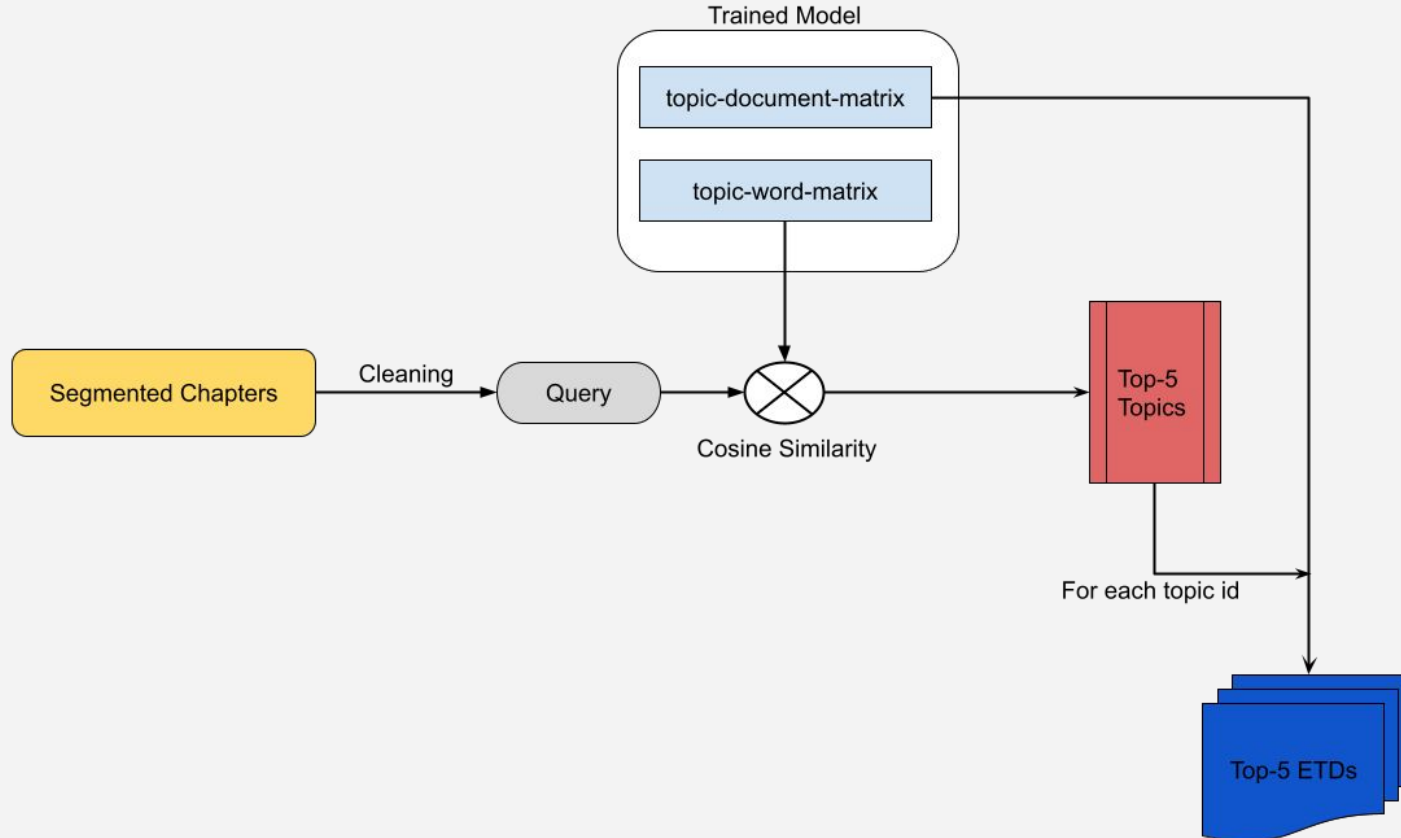
### Related Topics
# Using cosine similarity, find the most similar looking vocabulary array
# compared to our N-hot encoded words_list
query_vector =np.expand_dims(query_vector, axis=0) # convert to 2d for cosine similarity
distances=cosine_similarity(query_vector,model['topic-word-matrix'])
matching_topics_idx = np.argsort(-distances)
matching_topics_idx = np.squeeze(matching_topics_idx) # decreasing order of most similar topic ids

logger("Related Topics : ",matching_topics_idx[:k],2)
```



Calculating
similarity

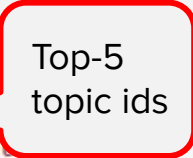
TM - Chapters - Inference - **Related ETDs**



TM - Chapters - Inference - Related ETDs

```
### Related etds
# iterate on all related topics, and fetch top-5 etds for each. Sort and find out overall top-5
topic_document_matrix = model['topic-document-matrix']
list_of_competitors = []
for topic_id in matching_topics_idx[:topk]:
    local_toppers_id = (-topic_document_matrix[topic_id]).argsort()[:topk].tolist()
    for etd_id in local_toppers_id:
        probability_value = topic_document_matrix[topic_id][etd_id]
        list_of_competitors.append([-probability_value, etd_id]) # appending -ve so that sorting gives the desired
list_of_competitors.sort()
matching_etd_idx = [etd_id for _, etd_id in list_of_competitors[:topk]]

logger("Related ETDs : ", matching_etd_idx, 2)
```



Top-5
topic ids

TM - Chapters - Inference - Related ETDs

```
### Related etds
# iterate on all related topics, and fetch top-5 etds for each. Sort and find out overall top-5
topic_document_matrix = model['topic-document-matrix']
list_of_competitors = []
for topic_id in matching_topics_idx[:topk]:
    local_toppers_id = (-topic_document_matrix[topic_id]).argsort()[:topk].tolist()
    for etd_id in local_toppers_id:
        probability_value = topic_document_matrix[topic_id][etd_id]
        list_of_competitors.append([-probability_value, etd_id]) # appending -ve so that sorting gives the desired
list_of_competitors.sort()
matching_etd_idx = [etd_id for _, etd_id in list_of_competitors[:topk]]

logger("Related ETDs : ", matching_etd_idx, 2)
```

Top-5
topic ids

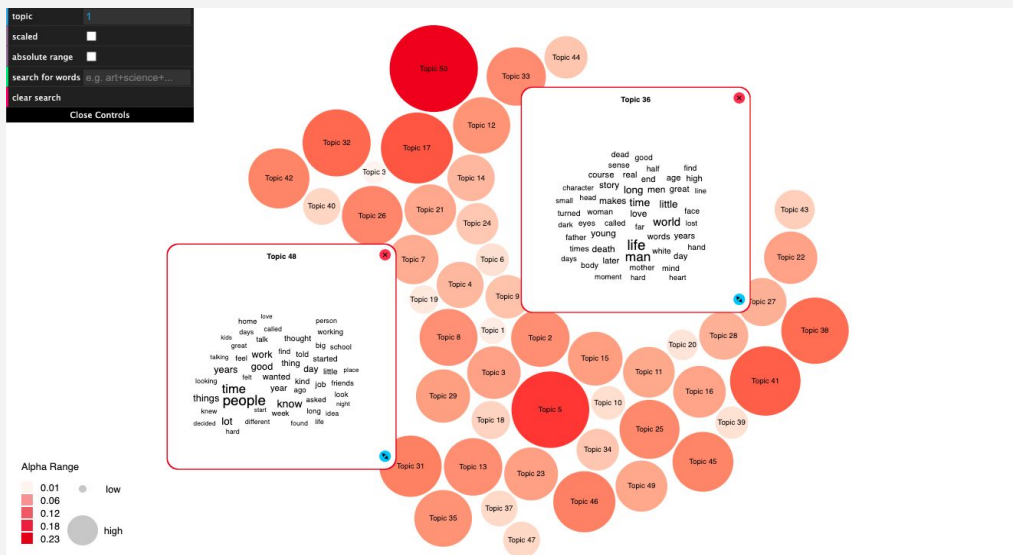
Top-5
etds /
topic

TM - Chapters - Inference - **Issues**

- Quality of output \propto Accuracy of Segmentation

Experimenter pages - Topic Bubbles

- Intuitive and Interactive visualization



Source : <https://github.com/sihwapark/topic-bubbles>

Experimenter pages - Topic Bubbles - Data Preparation

tw.json

topic_weights {
 alpha : [a_1, a_2, \dots, a_t],
 tw : [
 {words : [w_1, w_2, \dots, w_n],
 weights : [f_1, f_2, \dots, f_n]}
]
}

```
{ } twjson ×  
data > { } twjson > [ ] tw > { } 10 > [ ] words > [ ] 4  
1  
{"alpha": [0.0101420622464994, 0.017100462670444124, 0.0008927297556203249, 2.03  
031168090006726046, 0.009469457636100524, 0.020072152130933697, 0.0102480484275  
08002771946273159, 0.01977457554572692, 0.028379838167254347, 0.005906691396775  
013452092207977499, 0.023793897641807473, 0.034959134174428794, 0.0533110490593  
02913397060921672, 0.016660212380001222, 0.0030083769846931496, 0.0178545951124  
04462018221470354, 0.020642846951878197, 0.04462018221470354, 2.853474104722499  
030817520331002997, 0.015278315634999899, 0.016892566699957198], "tw": [{"words"  
"weights": [0.05728691, 0.04783485, 0.024015272, 0.017979702, 0.015738446, 0.01508  
"formation", "basin", "sea", "area", "lake", "low"], "weights": [0.013550137, 0.0118  
006021857]}, {"words": ["de", "worker", "et", "al", "la", "se", "labor", "el", "san", "  
0.02019842, 0.015976142, 0.015934184]}, {"words": ["channel", "event", "mass", "ms"  
040764313, 0.029591456, 0.02631898, 0.017177572, 0.015882565, 0.011919755, 0.01039  
"flux"], "weights": [0.12555389, 0.061532244, 0.05500461, 0.051334985, 0.017509744  
"white", "african", "race", "racial", "neighborhood", "pain", "ethnic", "color"], "v  
022940919, 0.022179365, 0.021536887]}, {"words": ["design", "system", "technology"  
02909869, 0.022343569, 0.020622479, 0.020355182, 0.01673883, 0.015476359, 0.014236  
"nanoparticle", "size", "high", "membrane", "study"], "weights": [0.021880854, 0.01  
007127685]}, {"words": ["urban", "environmental", "area", "city", "project", "texas  
0.013023239, 0.011130254, 0.0109691955, 0.010738383, 0.010430599, 0.010293547]}, {"  
"sequence"], "weights": [0.017287858, 0.017024837, 0.015502495, 0.014121457, 0.012  
"decision", "public", "government", "law", "military", "federal", "security", "nati  
007731809, 0.007072392, 0.006820462, 0.006808712]}, {"words": ["force", "motion", "  
023803955, 0.02174654, 0.012270515, 0.011152936, 0.010530921, 0.009603769, 0.00914  
"disease", "effect", "study", "bone"], "weights": [0.017075978, 0.016369948, 0.0124  
{"words": ["material", "film", "device", "high", "layer", "thin", "property", "metal  
010643125, 0.008210302, 0.007767597, 0.0076875775, 0.007193972]}, {"words": ["musi  
"weights": [0.040076546, 0.02662525, 0.020523332, 0.016882133, 0.014848347, 0.0109  
"study", "education", "program", "college", "high", "academic", "learning"], "weig  
0.008601092, 0.008502335]}, {"words": ["state", "phase", "transition", "energy", "s  
015189217, 0.01504769, 0.013477285, 0.013052032, 0.012639166, 0.012250538, 0.01189  
"experience", "support", "youth"], "weights": [0.046095956, 0.028611887, 0.0247263  
["language", "word", "english", "identity", "read", "reading", "speech", "study", "e
```

Experimenter pages - Topic Bubbles - Data Preparation

meta.csv

author	title
name ₁	etd ₁
name ₂	etd ₂
name ₃	etd ₃

```
meta.csv X
data > meta.csv > meta.csv
1 "Titus, Monica Joy","Plasma Diagnostics and Plasma-Surface Interactions in Inductively Coupled Plasmas"
2 "Condie, Tyson","Declarative Systems"
3 "Horberg, Elizabeth Jane","Portrait of the Rugged Individualist: The Nonverbal Pride Display Communicates Support for Meritocracy"
4 "Nelson Mondragon, John Alexander","Essays in Empirical Macroeconomics"
5 "Swift, Timothy Alan","Control and Trajectory Generation of a Wearable Mobility Exoskeleton for Spinal Cord Injury Patients"
6 "Leveille Buchanan, Nicole Therese","Errors as a Productive Context for Classroom Discussions: A Longitudinal Analysis of Normalization"
7 "Nair, Pradeep","Interrogating the Role of Spatial Organization in Receptor Function: Eph-Ephrin Signaling in Breast Cancer"
8 "HSU, KEVIN KAI-WIN","Hydrodynamic Exchange in Estuarine Perimeter Habitats"
9 "Robinson, Stefanie","An Environmental and Economic Trade-off Analysis of Manufacturing Process Chains to Inform Decision Making"
10 "GHANI, TAREK FOUAD","Essays on Markets and Institutions in Emerging Economies"
11 "Sheppard, Colin John Ritter","A Nexus Between Two Disruptions: A Multiscale Analysis of Transportation Electrification to Foster Sustainable Development"
12 "Sandjono, Sandra","Tracing Patterns of Textiles in Ancient Java (8th-15th century)"
13 "Patel, Anuj Ashwin","Microtopographical control of cell adhesion, organization, and proliferation in a cardiac tissue engineered scaffold"
14 "Elias, Renee Roy","Grocery Stores: Neighborhood Retail or Urban Panacea? Exploring the Intersections of Federal Policy, Community Development, and the Food Justice Movement"
```

dt.json

```
{ dtjson X
```

```
data > { dtjson > [ ]x
```

```
289900, 164/49, 288302, 398850, 30413, 383000, 208131], "t1":["410619, 1/0934, 384343, 29/30/, 19882, 2361, 9934/, 633/, 191323, 2/
```

```
177007, 206084, 303117, 313515, 51429, 73321], "t18":["23962, 92239, 212471, 452293, 368111, 71799, 355923, 154378, 459683, 6446, 306
```

```
267425, 316489, 118486, 227518, 354882], "t19":["103072, 9661, 330945, 241670, 29104, 446749, 201787, 290970, 34767, 337431, 70769, 4
```

```
425958, 471929, 410599, 42573], "t20":["173357, 169526, 158163, 226211, 113785, 478865, 371457, 233710, 7760, 216364, 32187, 13053, 1
```

```
22759, 85276, 482572], "t21":["360317, 389039, 338164, 202290, 386849, 358126, 400440, 279568, 152092, 326433, 298367, 170066, 388089
```

```
306381, 442178], "t22":["111837, 231056, 160122, 320808, 212467, 29974, 356372, 169502, 333081, 472919, 277012, 336304, 83104, 423437
```

```
399965], "t23":["247561, 446617, 320647, 109797, 469264, 63170, 16511, 151666, 152921, 14155, 457564, 488658, 475136, 431934, 211016,
```

```
"t24":["276345, 457379, 294430, 344917, 159669, 81271, 47626, 264425, 296000, 436045, 368101, 487155, 368117, 154017, 190077, 385870
```

```
[451248, 346581, 43/05/, 399623, 9035/, 211842, 45/940, 163098, 459304, 418300, 2169/9, 4493/3, 14804/, 18989, 399605, 390915, 2/1/
```

```
407842, 343672, 333713, 17202, 207684, 323406, 275302, 446772, 288508, 409729, 153606, 169096, 285390, 220086, 411715, 321760, 9195
```

```
39967, 217155, 384301, 115681, 486498, 384858, 322043, 456400, 240340, 315388, 442499, 348920, 116451, 71223, 156481, 167930, 34930
```

```
201242, 393240, 312057, 83536, 296696, 353099, 231308, 65366, 12286, 401861, 484674, 64694, 229562, 310403, 30644, 230618, 17931], "
```

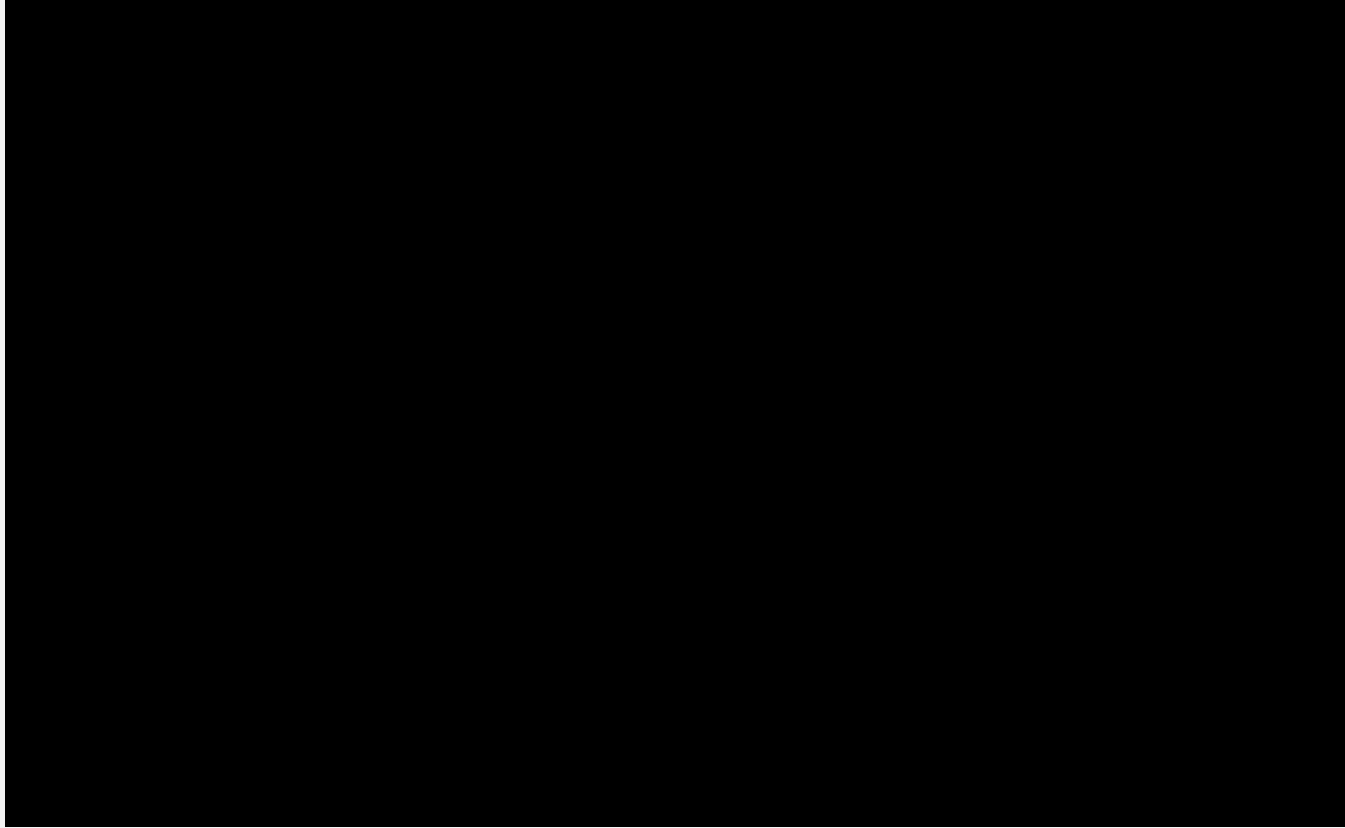
```
243323, 283340, 466906, 287001, 93782, 190513, 148550, 456835, 353682, 232216, 300349, 190329, 109300, 1959, 348225, 338180], "t30":
```

```
426431, 23649, 37540, 8192, 329944, 65146, 402883, 195761, 285669, 169074, 92753, 197199, 19233, 401929, 334382], "t31":["224725, 410
```

```
416417, 365024, 7579, 34239, 167345, 111574, 219437, 274815, 94809, 13180, 279534, 182895, 387046], "t32":["32238, 386070, 27182, 162
```

```
341102, 38307, 120007, 75103, 100016, 171071, 171705, 300000, 300000, 100000, 111007, 300007, 100000, 470000, 100000, 100000,
```

Experimenter pages - Topic Bubbles - **Demo**



TM - Services

We are providing 2 services that will be containerized - to be used for workflow automation and our frontend webpage

- Given a dataset, train TM model to generate vectors.
- Generate related topics and documents for an ETD and all its chapters.

Ps : The Frontend flask application (hosted on Cloud CS) is not linked with service yet.

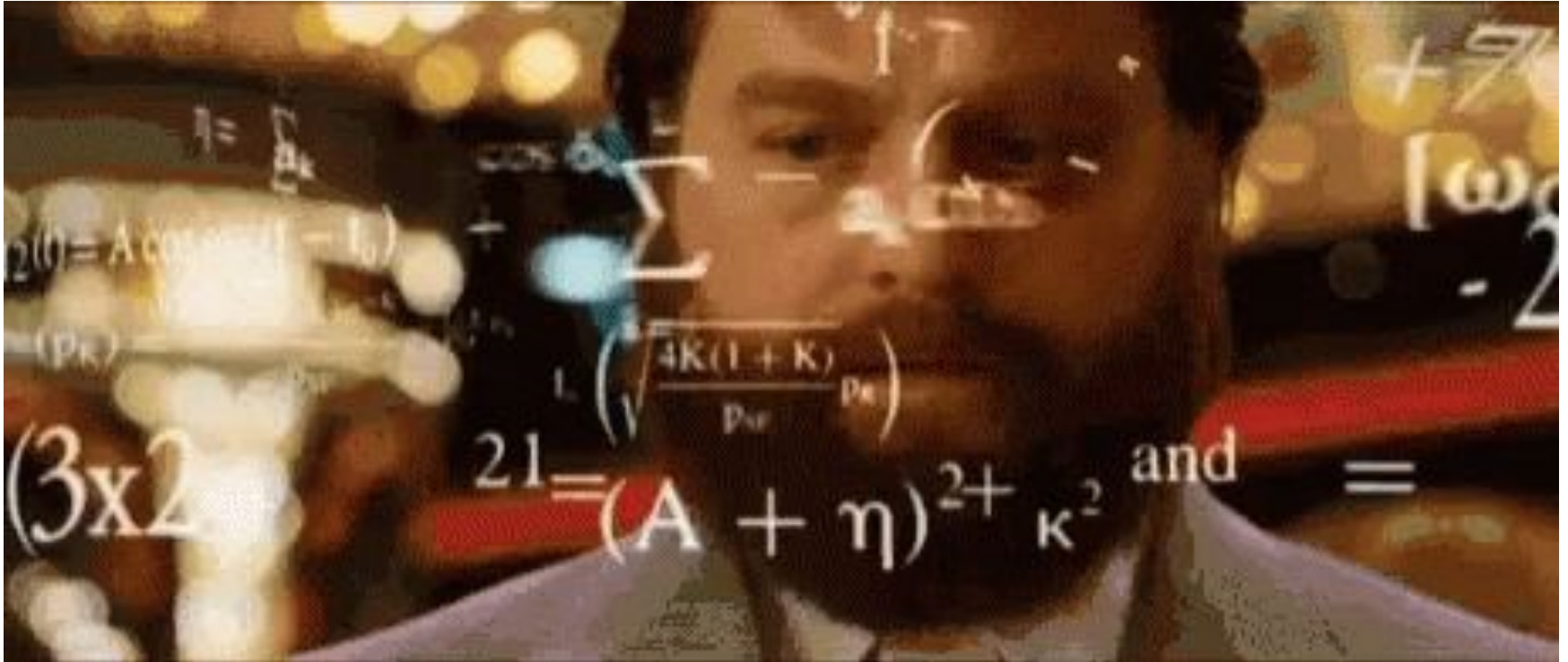
Milestones/Timeline

- Work done till IR-2
 - Train on complete 500k ETD dataset and integrate with UI
 - Perform similarity search (topic+document) through FaisNN
 - Display top-5 topics when topic search query is used
- Work done for IR-3
 - Chapter dataset generation, pre-processing
 - Integrating chapters into topic model
 - Preprocessing chapter data for OCTIS
 - Code refactoring and optimization for reducing load time
- **Work done for Final Report**
 - Topic Bubble Experimenter UI
 - Generate and save Related topics+Docs data for 5k dataset for all ETDs and Chapters

Wrap-Up Tasks

- Set up workflow automation - team 5
- Integrate Flask app / Experimenter UI with the frontend and workflow services.
- Build the Bubble UI for all the models (CTM, LDA, NeuralLDA and ProdLDA)

QUESTIONS?



BACKUP

Object Detection

- What has been accomplished
- Deliverables
- ETD filtering rules
- Issues and challenges faced
- Milestones
- Upcoming tasks and goals

What Has Been Accomplished

- The basic XML schema has been implemented for all object types
- Retrieved a trained YOLOv7 model to run inferences
- Post-processing rules for YOLOv7 detections
- Experimented on 10 randomly sampled PDFs for testing
- Front-end wireframe (Experimenter and page after search)
- Docker container with GPU access enabled

What Has Been Accomplished

- Set up Docker containers for YOLOv7 and Detectron2 frameworks
- Implemented YOLOv7 and Faster R-CNN algorithms for ETDs
- Implemented post-processing rules for detected objects
- Converted the unordered set of detections to XML
- Experimented on multiple randomly sampled ETDs
- Experimenter Web Page UI
- Team 1 deliverables - DB tasks (read/write)

Issues / Test Cases - OD

- Detections are not always in the top-bottom order for pages which is required to create the XML tree

Front sub-element

object ▶ eld ▶ front ▶ toc_text

```
▼ object {1}
  ▼ etd {3}
    ▼ front {10}
      Title : Probabilistic Models of Topics and Social Events
      author : CMU-ISR-16-113
      university : Carnegie Mellon University
      degree : value
      committee : Thesis Committee: Kathleen M. Carley, Tom Mitchell, Alexander J. Smola, Huan Liu, Arizona State University
      date : value
      abs_heading : Abstract
      abs_text : Structured probabilistic inference has shown to be useful in modeling complex latent structures of data. One successful way in which this technique has been applied is in the discovery of latent topical structures of text data, which is usually referred to as topic modeling. With the recent popularity of mobile devices and social networking, we can now easily acquire text data attached to meta information, such as geo-spatial coordinates and time stamps. This metadata can provide rich and accurate information that is helpful in answering many research questions related to spatial and temporal reasoning. However, such data must be treated differently from text data. For example, spatial data is usually organized in terms of a two dimensional region while temporal information can exhibit periodicities. While some work existing in the topic modeling community that utilizes some of the meta information, these models largely focused on incorporating metadata into text analysis, rather than providing models that make full use of the joint distribution of meta-information and text. In this thesis, I propose the event detection problem, which is a multi-dimensional latent clustering problem on spatial, temporal and topical data. I start with a simple parametric model to discover independent events using geo-tagged twitter data. The model is then improved toward two directions. First, I augmented the model using Recurrent Chinese Restaurant Process (RCRP) to discover events that are dynamic in nature. Second, I studied a model that can detect events using data from multiple media sources. I studied the characteristics of different media in terms of reported event times and linguistic patterns. The approaches studied in this thesis are largely based on Bayesian non-parametric methods to deal with streaming data and unpredictable number of clusters. The research will not only serve the event detection problem itself but also shed light into a more general structured clustering problem in spatial, temporal and textual data.
      toc_heading : Contents
      toc_text : List of Figures
      List of Tables
      1 Introduction
      2 Background
      3 Event Detections
      4 Bayesian Non-parametrics
      5 Modeling Independent Events
      6 Introduction
      7 Events Extraction from Text
      8 Events Extractions from Space and Time
      9 Graphical Models and Sampling Techniques
      9 Model
      10 Event Model
      10 Document Model
      11 Spatial and Temporal Boundaries
      12 Generative Model
      12 Language Model
      13 Model Inference
      14 E Step
      14 M step
      16 Prediction
      18 Experimental Results
      20 Data Set
      20 Qualitative Analysis of Events
      20 Quantitative Analysis
      22 Perplexity analysis
      24 Prediction of location and time
      24 Discussion
      25 Modeling Temporal Evolutionary Events
      27 Introduction
      28 Topic Modeling
      28 Non-parametric Bayesian
      28 Statistical Model
      30 Sequential Monte Carlo
      30 Integrating Variables
      31 Scalable Inference
      33 Laplace Approximation to Marginal Likelihood
      36 Sample Cluster Index
      38 Algorithm
      38 Sample Region
      38 Index
      38 SMC Updates
      39 Experimental Results
      39 Qualitative Results
      40 Numerical Results
      40 Discussion
      42
```

Body sub-element

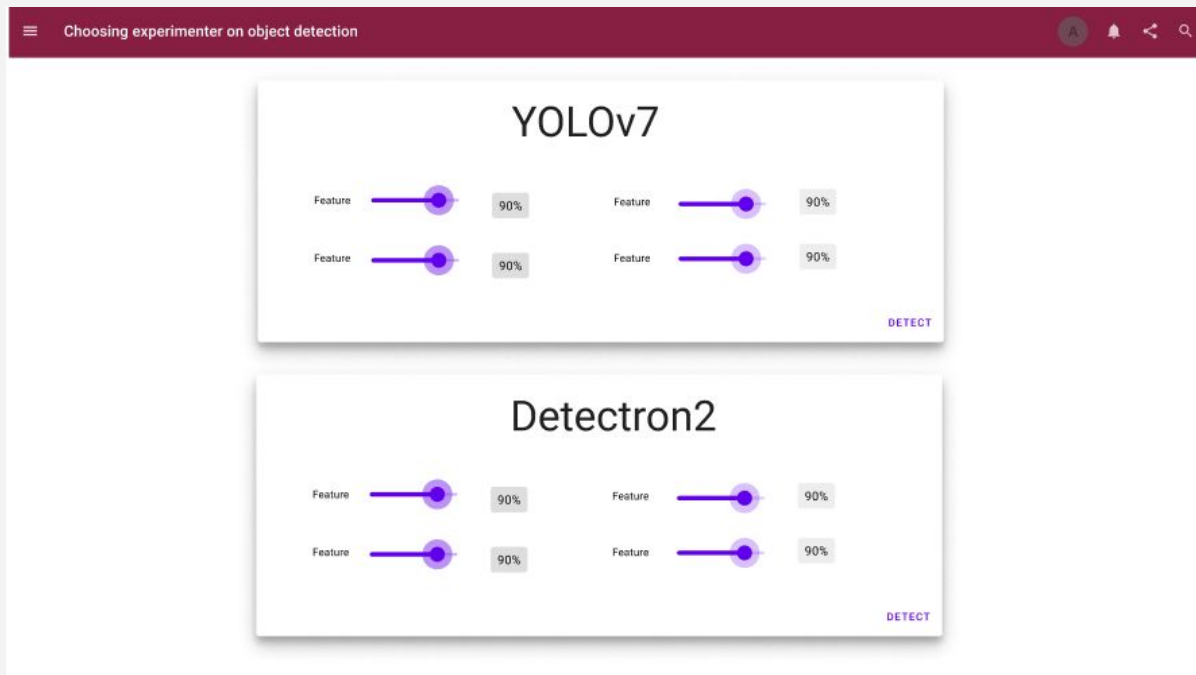
```
object ▶ eld ▶ body ▶ chapter ▶ 4 ▶ sections ▶ section ▶ 3 ▶ equations ▶ equation ▶ 1 ▶
▼ object {1}
  ▼ etd {3}
    ▶ front {10}
    ▼ body {1}
      ▼ chapter {6}
        ▶ 0 {2}
        ▶ 1 {2}
        ▶ 2 {2}
        ▶ 3 {2}
        ▼ 4 {2}
          title : Chapter 5\ Modeling Events from Multiple Data\ nSources
        ▼ sections {1}
          ▼ section {12}
            ▶ 0 {7}
            ▶ 1 {7}
            ▶ 2 {7}
            ▼ 3 {7}
              name : 5.4 Initializations
            ▼ paragraphs {1}
              para : A naive approach to initialize the latent variables is to use
                uniformly generated random variables to serve as the initial values
                of  $z, q$  and  $s$ . However, if those initial values are bad, it is
                likely that the algorithm would take a long time to reach
                equilibrium. Instead, we can sample the initial  $z, q, s$  values and
                using parts of Equation 5.3 and Equation 5.4 that do not require
                our knowledge of cluster index  $s$ . For location index  $z$ , its
                initial value is sampled purely based on its location proximity to
                the Gaussian centers. For word category variable  $q$ , its initial
                values are determined by both the values of previous words in the
                current document and prior  $\eta_0$ .
              figures : {value}
              tables : {value}
            ▼ equations {1}
              ▼ equation {2}
                ▼ 0 {2}
                  path : detections/image-objects/equation_CMU-ISR-16-113_1.jpg
                  eq_no : (5.11)
                ▼ 1 {2}
```

Back sub-element

```
object ▶ eld ▶ back ▶
▼ object {1}
  ▼ etd {3}
    ▶ front {10}
    body : {value}
    ▼ back {2}
      ref_heading : Bibliography
      ref_text : [1] Amr Ahmed and Eric P Xing. Dynamic non-parametric mixture models
        and the recurrent\chinese restaurant process: with applications to
        evolutionary clustering. In SDM, pages 219-230. SIAM, 2008. 1, 2.3,
        4.1, 4.2.1, 4.2.2, 4.2.3, 4.2.3, 4.3, 4.3, 4.4.2, 5.2.2\ n[2] Amr Ahmed,
        Qirong Ho, Choon H Teo, Jacob Eisenstein, Eric P Xing, and Alex J
        Smola. \Online inference for the infinite topic-cluster model:
        Storylines from streaming text. In\ nInternational conference on
        Artificial Intelligence and\ nStatistics, pages 101-109, 2011. \ n4.2.4,
        5.2.2, 5.2.3, 5.4\ n[3] Amr Ahmed, Mohamed Aly, Joseph Gonzalez, Shravan
        Narayanamurthy, and Alexander\ nInternational conference on Web\ nSmola.
        Scalable inference in latent variable models. In\ nsearch and data
        mining\ n(WSDM), pages 123-132, 2012. 4.2.1\ n[4] Amr Ahmed, Liangjie
        Hong, and Alexander J Smola. Nested chinese restaurant
        franchise\ nICML\ nprocess: Applications to user tracking and document
        modeling. In (3), pages 1426- \ n1434, 2013. 2.3, 4.2.2, 5.2.2,
        6.5.1\ nExchangeability and related\ n[5] David J Aldous. topics.
        Springer, 1985. 2.3\ nForeign\ n[6] Lisa Anderson. Demystifying the arab
        spring. Affairs, 90(3):2-7, 2011. 1, 3.5, \ n3.5.2\ n[7] Christophe
        Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An
        intro- \ nMachine\ nduction to mcmc for machine learning. learning, 50(1-
        2):5-43, 2003. 3.4\ n[8] Matthew J Beal, Zoubin Ghahramani, and Carl E
        Rasmussen. The infinite hidden markov\ nAdvances in neural information
        processing\ nmodel. In systems, pages 577-584, 2001. 2.3\ nIEEE\ n[9]
        George Bebis and Michael Georgiopoulos. Feed-forward neural networks.
        Potentials, \ n13(4):27-31, 1994. 6.5.3\ n[10] Edward Benson, Aria
        Haghighi, and Regina Barzilay. Event discovery in social
        media\ nProceedings of the 49th Annual Meeting of the Association for
        Computational\ nfeeds. In\ nLinguistics: Human Language Technologies-
        Volume\ n1, pages 389-398. Association for\ nComputational linguistics,
        2011. 2.1, 3.1, 3.2.1
```


Front End - Object Detection

- End product - Experimenter web page



Front End - Object Detection

- End product - Parsed ETD web page

The screenshot displays a web page with a dark red header. The header contains a hamburger menu icon, the title "The Syntax and Semantics of Do So Anaphora", and icons for notifications, sharing, and search. On the left side, there is a "Table of Contents" sidebar with a list of items, each preceded by a heart icon. The item "1.1 Verbal anaphora in English" is highlighted in pink. The main content area on the right shows the "Abstract" section, which discusses the use of "do so" anaphora in English and its semantic and syntactic properties. Below the abstract, there is a "Chapter 1 Introduction" section, followed by "1.1 Verbal anaphora in English", which includes examples of verbal anaphors and their use in sentences.

The Syntax and Semantics of Do So Anaphora

Table of Contents

- Abstract
- 1. Introduction
- 1.1 Verbal anaphora in English**
- 1.2 Do so anaphora
- 1.2.1 Part I: Do so and verb phrase ellipsis
- 1.2.2 Semantic Restriction on the use of do so
- 1.3 Preliminaries
- 1.3.1 The category of so
- 1.3.1 The category of do

Abstract

Do so anaphora is a fairly widely used in English, but has received relatively little treatment in the literature (especially when compared with verb phrase ellipsis). There are, however, two aspects of this anaphor that have gained prominence: i) its use as a test for constituency within the verb phrase, and ii) the semantic restriction it places on its antecedent. Though these two properties have been the most prominent, their analyses have not been uncontroversial. In this dissertation, I investigate these properties and give them a more complete analysis. The first part of the dissertation is devoted to a discussion of the use of do so as a test for constituency in the verb phrase, and the second part is devoted to understanding the semantic restriction that do so places on its antecedent. The behavior of do so anaphora has been used to argue both hierarchical structure (Lakoff and Ross 1976) and flat structure within the verb phrase (Culicover and Jackendoff 2005). In chapter 2, however, I argue that do so does not have any bearing on the debate about the internal structure of the verb phrase. The arguments put forth by these authors are predicated on do so being a surface anaphor in terms of Hankamer and Sag (1976). Instead I argue that do so is in fact a deep anaphor and that its purported surface anaphor properties fall out from independent semantic and pragmatic properties of the anaphor. As a deep anaphor, do so does not replace any structure in the verb phrase, but rather forms a verb phrase in its own right from the beginning of the derivation. Therefore, the use of do so to argue for or against hierarchical structure in the verb phrase has been misguided. I approach the semantic restriction that do so places on its antecedent from two angles. In chapter 3, I review the previous analyses of this restriction, and test their claims against a corpus of over 1000 naturally occurring examples extracted from the American National Corpus. None of the previous analyses are supported by the data, and I present a novel analysis that utilize three semantic parameters (agentivity, aktionsart, stativity) to predict which antecedents are possible with do so. One striking property of the counter examples found in the corpus is that they instantiate particular syntactic structures. The majority of them contain do so in a nonfinite form (usually in the infinitive), and in others, the antecedent is contained in a relative clause modifying the subject of do so. In chapter 4, I present experimental evidence that shows that these two syntactic environments lessen the effects of the restriction that do so normally places on its antecedent. I attribute this amelioration of the semantic restriction to the unavailability of verb phrase ellipsis in these 1 syntactic environments. The analysis falls out from the nonmonotonic interaction of the two restrictions: the syntactic restrictions on ellipsis force the use of do so to the detriment of the semantic restriction that do so normally places on its antecedent. I then situate this amelioration effect into the typology of coercion effects in general and argue that do so displays a novel type of coercion: subtractive coercion.

Chapter 1 Introduction

1.1 Verbal anaphora in English

In English, we have various strategies for avoiding the repetition of identical verb phrases. If we would like to express that both Steve and John have eaten an apple, it is not necessary to utter a sentence as in (1) where both conjoined clauses contain full VPs. In fact, this sounds quite unnatural.

(1) Steve has eaten an apple, and John has eaten an apple, too.

Instead, we have a number of verbal anaphors that can be used in the second clause to express that the same type of event has occurred as that expressed in the first clause. These verbal anaphors include Verb Phrase Ellipsis (2a), do it anaphora (2b), do that anaphora (2c), and do so anaphora (2d). In each case, the anaphor stands in for a full verb phrase, often referred to as the target of anaphora.

(2) Steve has eaten an apple, and . . .

a. John has, too.

Upcoming Tasks and Goals - OD

- Link image-based objects to their captions based on a distance metric (like Euclidean)
 - Figures - figure captions
 - Tables - table captions
 - Equations - equation numbers
- Recognizing the right chapter and section titles (delimiters)
- Fix errors and debug
- Deal with subsections and sections as they constitute the same object / class
- Building the web pages (Experimenter + Parsed ETD)

Issues/Test Cases - OD

False positives:

1. Chapter/section titles being incorrectly detected
 - a. Paragraph's last line
 - b. New chapter tag for the same chapter title
 - c. Chapter and paragraph is being created but not the section
2. Images/tables being linked to wrong caption
3. Chapter/section titles detected don't match with the titles in the ToC

Experimenter UI - YOLOv7 Detection

Introduction	»
Background	»
Modeling Independent Events	»
Modeling Temporal Evolutionary Events	»
Modeling Events from Multiple Data Sources	»
Conclusions and Future Work	»

model Third, there is a which contains information about the topical content of the language model, documents. Table 3.1 gives a summary of all notation that will be used as we describe the model in this section.

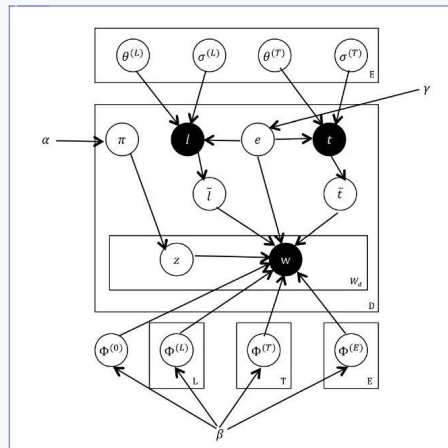


Figure 3.1: Illustrations of the model in plate notations

3.3.1 Event Model

An important observation incorporated into our model is that events are in many ways natural extensions of topics; events have a topical focus but also include a spatial and temporal region in which they are likely to occur. We thus assume events are defined by three things. First, each $\theta^{(L)}$ event has a geographical center as well as a geographical variance controlled by a diagonal $\sigma^{(L)}$ covariance matrix with each value defined by γ . The location of a document that belongs to e event is assumed to be drawn from a two dimensional Gaussian distribution governed by these

Experimenter UI - Faster R-CNN Detection

ETD Browser

- Chapter 1 Introduction »
- Chapter 2 Background »
- Chapter 3 Modeling Independent Events »
- Chapter 4 Modeling Temporal Evolutionary Events »
- Chapter 5 Modeling Events from Multiple Data Sources »
- Chapter 6 Conclusions and Future Work »

Probabilistic Models of Topics and Social Events

University:

Degree:

Committee:

Date:

Abstract

Abstract Structured probabilistic inference has shown to be useful in modeling complex latent structures of data. One successful way in which this technique has been applied is in the discovery of latent topical structures of text data, which is usually referred to as topic modeling. With the recent popularity of mobile devices and social networking, we can now easily acquire text data attached to meta information, such as geo-spatial coordinates and time stamps. This metadata can provide rich and accurate information that is helpful in answering many research questions related to spatial and temporal reasoning. However, such data must be treated differently from text data. For example, spatial data is usually organized in terms of a two-dimensional region while temporal information can exhibit periodicities. While some work existing in the topic modeling community that utilizes some of the meta information, these models largely focused on incorporating metadata into text analysis, rather than providing models that make full use of the joint distribution of meta-information and text. In this thesis, I propose the event detection problem, which is a multi-dimensional latent clustering problem on spatial, temporal and topical data. I start with a simple parametric model to discover independent events using geo-tagged Twitter data. The model is then improved toward two directions. First, I augmented the model using Recurrent Chinese Restaurant Process (RCRP) to discover events that are dynamic in nature. Second, I studied a model that can detect events using data from multiple media sources. I studied the characteristics of different media in terms of reported event times and linguistic patterns. The approaches studied in this thesis are largely based on Bayesian non-parametric methods to deal with streaming data and unpredictable number of clusters. The research will not only serve the event detection problem itself but also shed light into a more general structured clustering problem in spatial, temporal and textual data.

Structured probabilistic inference has shown to be useful in modeling complex latent structures of data. One successful way in which this technique has been applied is in the discovery of latent topical structures of text data,

UI improvements

- Filter Layout
 - Year range
 - Author names
 - University
 - Faculty
 - Committee

The screenshot displays the ETD Topic Browser interface. At the top, there are search filters for 'Start Year', 'End Year', and 'University'. Below these, a 'Topic Words' section lists terms like 'student', 'school', 'study', 'teacher', 'education', 'college', 'program', 'high', 'child', and 'academic'. A 'Similar Topics' section provides a list of related terms. The main content area is a table with two columns: 'TITLE' and 'ABSTRACT'. It lists several research topics, each with a title link and a brief abstract. A red arrow points from the 'Filter Layout' section of the slide to the filter area of the interface.

ETD Topic Browser

Start Year: End Year: University: (Please select a University) [View Documents](#)

Topic Words: 'student', 'school', 'study', 'teacher', 'education', 'college', 'program', 'high', 'child', 'academic'

Similar Topics:
[study](#) [research](#) [process](#) [decision](#) [program](#) [practice](#) [work](#) [university](#) [development](#) [organization](#)
[region](#) [fourth](#) [some](#) [datum](#) [track](#) [ice](#) [surface](#) [independent](#) [extent](#) [sea](#)
[mark](#) [chapter](#) [music](#) [century](#) [text](#) [character](#) [musical](#) [write](#) [form](#) [study](#)

TITLE	ABSTRACT
Texas Elementary Educators Professional Reading Practices	The purposes of this study were (1) to survey the amount of time spent by elementary educators in reading professional literature; (2) to survey elementary educators purposes for reading professional literature; (3) to Show more
Exploring Charter High School English Teachers' Perceptions of College-Ready Literacy	Students entering college underprepared for the standards and expectations of post-secondary schooling is a persistent academic problem. More specifically, students lack college-ready literacy skills, an important skill to Show more
Evaluating a Doctoral Program in College and University Teaching: A Single Case Study	This study assessed alumni of the College and University Teaching Program at the University of North Texas and how they perceived the training they received. Three hundred sixty alumni holding a college and university Show more
Four Case Studies: The Reading Attitudes and Practices of Teachers and Students in Second Grade	The purpose of this study was to investigate the nature of reading attitudes and practices in second-grade teachers and students. Surveys and interviews were conducted with second-grade teachers and students in Northwest Ohio. Show more
An Analysis of Enrollment Patterns in Required General Education Courses by Technical-Occupational Students in an Urban Community College	This study was concerned with the enrollment patterns in required general education courses by technical/occupational students in an urban community college. The purposes of this study were to (1) examine Show more
Pre-Service Teachers Attitudes and Assumptions of Culturally Responsive Teaching Practices	Previous literature has expanded on the importance of including culturally relevant teaching practices into classrooms in order to reach student populations related to culturally and linguistically diverse learners. Although Show more

Any time
Since 2022
Since 2021
Since 2018
Custom range...

Sort by relevance
Sort by date

Any type
Review articles

☐ include patents
☒ include citations

☒ Create alert