Enhancing Digital History – Event Discovery via Topic Modeling and Change Detection

King-Ip Lin, Sabrina Peng
Department of Computer Science
Lyle School of Engineering
Southern Methodist University
Dallas, TX 75205, USA
{kdlin, shpeng}@smu.edu

Abstract

Digital history is the application of computer science techniques to historical data in order to uncover insights into events occurring during specific time periods from the past. This relatively new interdisciplinary field can help identify and record latent information about political, cultural, and economic trends that are not otherwise apparent from traditional historical analysis. This paper presents a method that uses topic modeling and breakpoint detection to observe how extracted topics come in and out of prominence over various time periods. We apply our techniques on British parliamentary speech data from the 19th century. Findings show that some of the events produced are cohesive in topic content (religion, transportation, economics, etc.) and time period (events are focused in the same year or month). Topic content identified should be further analyzed for specific events and undergo external validation to determine the quality and value of the findings to historians specializing in 19th century Britain.

1 Introduction

The field of digital history involves the application of computer science techniques to historical data. Research in this field is aimed at uncovering both obvious and latent information about specific time periods from the past, allowing for a deeper understanding of historical events.

Specifically, using natural language processing techniques on historic text data can be valuable in determining what factors are catalysts for change. Issues, ideas, and sentiments can suddenly become viral and become triggers for influential events.

In this paper, we present our work on detecting these factors by pinpointing which topics gain or lose prominence over certain time periods in history. We test our methods by applying them to a dataset of 19th century British parliamentary debates from the House of Commons. We define our task as one that discovers when political, cultural, and economic trends grow and/or shrink with respect to significant increased or decreased discussion of certain topics in parliamentary discourse.

We model the ideas by turning to standard (generative) topic models, such as LDA (Biel, Ng, & Jordan, 2013). These models are well studied and have been applied in a variety of fields, including the humanities (Günther & Quandt, 2016; Ramage, Rosen, Chuang, Manning, & McFarland, 2009; Thomas & Droge, 2022; Guldi, 2019). In many such models, there is an intuitive description for topics that makes it feasible for users to detect what ideas are being represented in their data.

The next thing we need to model is the change in prominence of topics. There is a class of topic models known as Dynamic Topic Models (Biel & Lafferty, 2006) that attempts to determine the evolution of the most prominent topics over time. While this is useful in a lot of applications, this information does not necessarily show us the scope of the change in each topic's prominence. For instance, while a topic that is generated in time t does not appear in time t-1, it is not easy to determine whether its prominence suffers just

The 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 69–78

a small drop or falls off a cliff. Thus, in this work, we took the approach of developing a single topic model for all documents across all time slots. We intentionally generate a larger number of topics and use measures to quantitively measure the importance of each topic for each time slot. Thus, we can generate a time series of importance for each topic. We then apply a changepoint/breakpoint detection algorithm on the time series to detect major changes in the time series and capture where certain topics enter/leave parliamentary debate.

Another issue we look at is the robustness of the results. Topic model algorithms generate different results for each run, an undesirable characteristic (Yong, Pan, Lu, Topkara, & Song, 2016). Methods have been proposed for combating this instability (Montyla, Claes, & Faroaq, 2018; Miller & McCoy, 2017) (Rieger, 2020). In our work, we incorporate methods to overcome the instability by running the model multiple times and using clustering techniques to combine the results and enhance stability.

2 Background

2.1 Topic Modeling and Latent Dirichlet Allocation

Topic modeling is a language modeling technique that represents a large corpus of documents via topics. In such models, like Probabilistic Latent Semantic Indexing (PLSI) (Hoffmann, 1999) and Latent Dirichlet Allocation (LDA) (Biel, Ng, & Jordan, 2013), a topic is represented by a probability distribution over the vocabulary of the corpus. Intuitively, a topic is defined by the words that are heavily associated with it.

We use the following notation for the rest of the paper:

- We have a corpus C of n documents, denoted by $C_1, \ldots C_n$
- The set of all distinct words that makes up all documents is denoted by the set W ($w_1, \ldots w_m$)
- Let k be the number of topics describing the corpus (provided by the user)

Given the above, the topic model is described by two sets of probability distributions, each represented by a set of vectors.

- Topic-word vector ($t_1, \ldots, t_k$): each vector corresponds to a topic, which is a probability distribution on W.
- Document-topic vector ($d_1, \ldots, d_n$): each vector corresponds to a probability distribution of topics 1..k. This represents the association of each topic to a given document.

The goal of the topic modeling is to find the set of vectors/distributions that maximizes the probability that the corpus is actually being represented by the corresponding model.

Among the most widely used topic models today is Latent Dirichlet Allocation (LDA). It assumes there is an underlying Dirichlet distribution governing the choice of the vectors. Two parameters that are associated with the Dirichlet distribution, α and β, are used to affect the likelihood of a certain probability distribution being picked.

Typically, users of LDA can examine the topics, and for each topic, extract the words that have high probability to describe them. Also, they can look at the document-topic vectors to cluster documents along the topics.

2.2 Dynamic Topic Model

While the basic topic model does not have a time dimension, there has been work done to incorporate the time dimension. Dynamic Topic Model (Biel & Lafferty, 2006; Wang, Biel, & Heckerman, 2008) is one such approach. For the discrete case (Biel & Lafferty, 2006), it assumes the topic-word vector at time t is conditional on the topic-word vector at time t-1. The method generates a set of topics for each time t, enabling the user to see the most prevalent topics at certain times. Other dynamic topic models have been

proposed, many of which are being applied in a large variety of applications (Xu, Chen, Dai, & Chen, 2017; Hida, Takeishi, & Hori, 2018; Rieger, Jentsch, & Rahnenführer, 2021). While these models incorporate the notion of topic changes over time, they mostly focus on the generation of topics at different time points, meaning extra efforts are needed to obtain what we are looking for – the gain or loss of topic prominence.

The work by Wang and Goutte (Yunli & Cyril, 2018) is similar to this work in the sense that they also generate time series and apply change point detection. However, they are still generating topics on a per time slot basis and calculate the "dissimilarity" of topics from 1 slot to the next. The topic-CD model proposed in (Lu, Guo, & Chen, 2022) is also similar, with the caveat that the model builds in a fixed number of change points.

2.3 Changepoint/Breakpoint Detection Algorithms for Time Series

Changepoint / breakpoint detection in time series (Troung, Oudre, & Vayatis , 2020) has been applied to many problems involving climate data (Reeves, Chen, Wang, Lund, & Lu, 2007) and bioinformatics (Vito M. R. Muggeo, 2011). In this paper, we utilize the "ruptures" package (Truong, 2018), which contains a variety of change point detection algorithms. After some research, we selected the Pelt ("Pruned Exact Linear Time") algorithm, which computes the segmentation of the time series that minimizes the constrained sum of approximation errors. The Pelt algorithm does not require a fixed number of change points to be detected, which is ideal in our case, as we are conducting unsupervised learning and do not know the number of true breakpoints. The algorithm uses pruning rules to keep or discard samples from the set of potential change points, resulting in a considerable speedup when compared to other algorithms and a computational complexity that is linear on average.

3 Our Approach

3.1 Problem Specification and Basic Algorithm

Our goal for this work is to, given a set of historic documents spanning a time period, determine when and how certain ideas rise to prominence or fade into non-existence over that period.

We assume there is a corpus C of documents (C_1, …, C_n). Each document has a time point (chosen from a set of time points $t_1 \leq \ldots \leq t_m$) associated with it. We assume m is much smaller than n. Notice that a timepoint can be a single instance in time (e.g. 1/1/2001, 12:00 am), or a period of time (e.g. March 1854 – June 1855). Our approach allows the user to choose any way of grouping the documents by time periods as they see fit.

We capture the notion of ideas by using topic models to represent them. Each topic can be represented by the words associated with it that have the highest probabilities. This provides a reasonable starting point for users to infer the ideas based on the words that are used to describe it.

Our approach consists of the following steps (for the rest of the paper, we use LDA as our topic model, but any topic model that generates topic-word and document-topic vectors can be used):

1. Run LDA on C, with k topics.
2. For each timepoint t_i, calculate and aggregate the document-topic vectors for all documents to form a vector denoting the importance of each topic at each timepoint.
3. For each topic, generate a time series based on the aggregated vector's value over the timepoints.
4. Apply breakpoint detection algorithms on the time series to detect when there is a sudden increase/decrease of weight of each topic.

Here we provide some additional details about each step:

- We want the number of topics k to cover the possible topics over all timepoints. Thus, we suggest setting k to a larger number than normal – i.e. larger than what one expects the number of topics to be over the timepoints.

- In step 2, we leave the option of how to aggregate the document-topic vector open. In this paper, we choose to simply add the document-topic vectors for all documents – essentially treating probabilities as "weights". We also choose not to normalize the results to get back to a probability distribution. One reason we take the raw sum is that we want to model not just the relative importance of the topics amongst themselves, but also the quantitative strength of the topic being mentioned. Other aggregation functions can be chosen if they can be justified.

- As mentioned in section 2, we use Pelt as our breakpoint detection algorithm.

3.2 Data used and simple example

To illustrate our methods, we use a data set of British parliamentary debates from 1803-1910. The dataset contains raw text and metadata for 10,979,009 sentences in speeches made by the legislators during parliamentary debates. In addition to the raw text of each sentence spoken, important metadata fields used in the event detection process include the date the sentence was spoken and the speech the sentence belongs to. As the dataset is large and analysis requires extensive computational resources, a subset of the dataset is created by performing stratified random sampling by speech month. For data cleaning and preparation, the raw text from each sentence is tokenized into words. We lowercase all words, strip out all punctuation, and filter out words that are less than 3 letters long. Then, all common English and dataset-specific (government-related) stop words are removed to retain more interesting terms. Finally, lemmatization is conducted to remove inflectional endings and retain the base form of each word.

In terms of segmenting the speech into documents, we consider each time a legislator speaks as a document to be fed into LDA.

Figures 1 and 2 show sample results from various steps of our methodology – LDA document-topic vector aggregation, time series generation, and breakpoint detection.

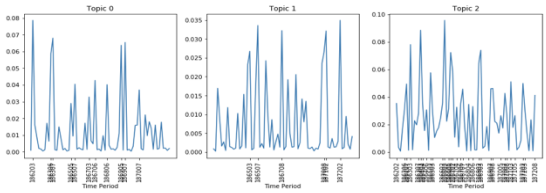


Figure 1. Examples of time series generated from aggregated LDA document-topic vectors.

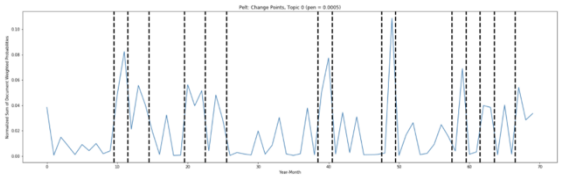


Figure 2. Example of breakpoints detected by the Pelt change point algorithm.

3.3 Enhancing robustness

Using LDA introduces the problem of instability. LDA is a non-deterministic algorithm that uses a stochastic process to update internal weights. Therefore, the results generated by LDA are not reproducible between different runs of the algorithm on the same dataset.

As stated in the introduction, there has been work on enhancing the stability of the method. Most methods try to run LDA on the same data set multiple times, and then aggregate the results. We follow a similar technique here. In our experiments below, we run our algorithms 10 times and aggregate the results for analysis.

However, compared to other methods, we have options on how we aggregate the topics generated over multiple runs. In our case, each

topic is associated with two items: the topic-word vector describing it, and the time series that is generated from that topic. Thus, we can aggregate the topics in one of two ways.

The first way is to cluster the topics based on the topic-word vector with agglomerative hierarchical clustering. Once the documents are clustered, the previously calculated changepoints of each time series for each topic are examined. Those points that appear with high frequency in the cluster will be returned as the breakpoints.

For clustering purposes, we need a similarity/distance metric between pairwise topic-word vectors. Our approaches rely on using selective terms from each topic. With a decent vocabulary size, each topic-word vector will have a lot of terms that have small (but non-zero) values. Since those terms are usually ignored by humans anyway, it makes some sense to ignore those terms when calculating similarity between topics. Thus, each topic is now represented by a subset of the vocabulary that is deemed "important" – for example, the set of words having high probabilities of belonging to the topic. After that, we apply Jaccard coefficient and Jensen-Shannon distance to calculate the similarity between topics. We apply two versions of the Jaccard coefficient, by considering only the top k words of each topic (denoted by Jaccard), or by considering all words in a topic that have a probability greater than a threshold p (denoted by Jaccard-p). We also apply Jensen-Shannon distance, which is the square root of Jensen-Shannon divergence. It measures the similarity between two probability distributions and is the symmetric version of Kullback-Leibler divergence.

The advantage of this method is that since the topic-word vector is the defining feature of the topic, it theoretically makes sense to cluster the topics in this way (as opposed to other stability methods). However, there is no guarantee that they share the same breakpoints, which may render some clusters useless.

Alternatively, we can cluster the topics based on the time series that are associated with each topic. We use both Euclidean and Manhattan distance as distance measures. Once the topics have been clustered, we examine the topics within a cluster and find words that have high probability among most of the topics and use them to represent the clusters.

We will then apply the changepoint detection algorithms to the sequences of the clusters to denote the breakpoints. For this method, the clustering usually places sequences with similar breakpoints together. The challenge is to find frequent words that are shared among the topics. Space limitations means that we will only discuss the result of our first approach.

4 Experimental Results

We create a subset of the data for use in our experiments by selecting 500 samples from each month of the dataset's representative time period using stratified random sampling. We set the number of clusters detected by agglomerative clustering to $N = 10$. We evaluate results for the distance metrics used in both methods on the basis of both cluster cohesion and topic distinctiveness. For each approach-metric combination, we analyze cluster tendency plots and the spread of topics across clusters. Cluster tendency plots used include VAT and iVAT, which reveals hidden cluster structures as dark blocks along the diagonal of the image representation. We also analyze topic annotations created by extracting the top documents from each cluster based on the aggregated probability of a document belonging to the cluster topics.

As mentioned in the previous section, we utilize the Jaccard coefficient, Jaccard-p coefficient, and Jensen-Shannon distance as distance metrics for agglomerative clustering. For the Jaccard approach, the sets of terms used to calculate the coefficient include the top terms from each topic with a topic-word probability 100x greater than the overall probability the term would appear in a random document. We further add rare words (those

occurring less than 10 times across all topics) and unique words (those that were completely unique to the given topic). For the Jaccard-p approach, the sets of terms used to calculate the coefficient include the top terms from a topic that have a topic-word probability of greater than 0.25%. For the Jensen-Shannon approach, the topic-word vectors used to calculate the distance include the top 1000 terms from each topic with the highest probabilities.

The Jaccard approach resulted in a less effective extraction and clustering of topics. Over 70% of the total topics were contained within one cluster, indicating one large generic cluster and many small specific ones. The topic annotations corroborate this finding – the top documents from the large cluster have a variety of topics, and the number of topics in the other clusters are too small.

The Jaccard-p approach seems to mitigate the original issues of using Jaccard due to its different word set composition and probability threshold. One larger cluster still exists, but the topics are more evenly spread across the identified cluster, as shown in Figure 3. Moreover, the VAT diagram (Figure 4) shows greater cluster distinctiveness.

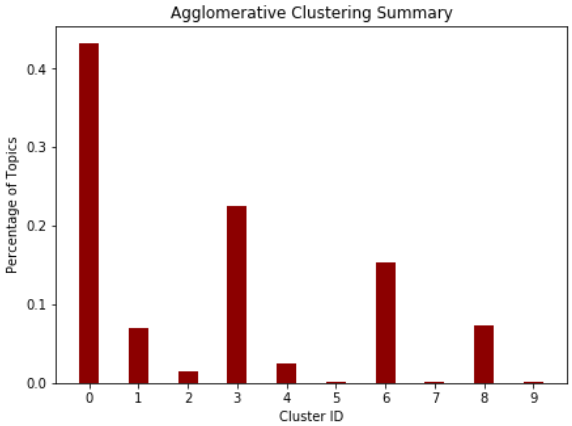


Figure 3. LDA topic distribution over Jaccard-p agglomerative clustering.

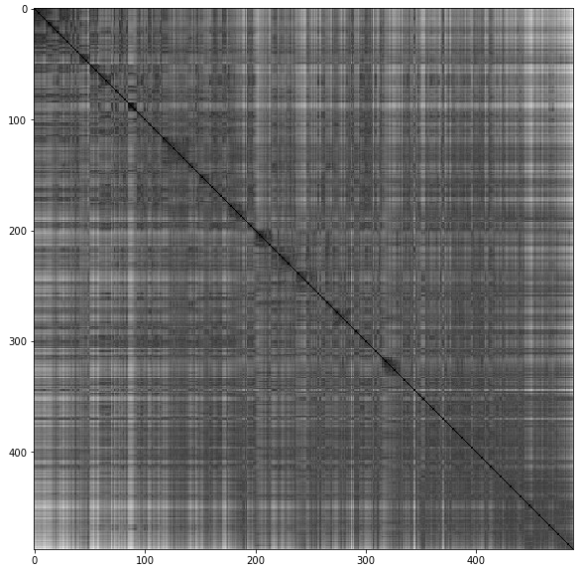


Figure 4. VAT diagram for Jaccard-p agglomerative clustering.

The topic annotations generated from the Jaccard-p approach indicate that certain clusters do exhibit topic cohesion. Relevant speeches for the clusters show thematic similarities, and topic-generated time series show similar trends in rise and fall across time intervals.

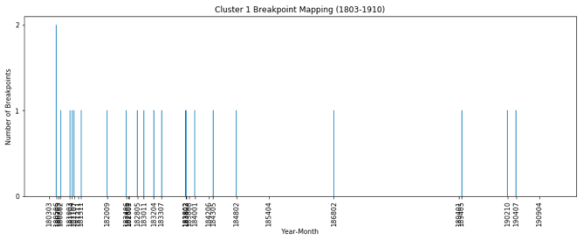
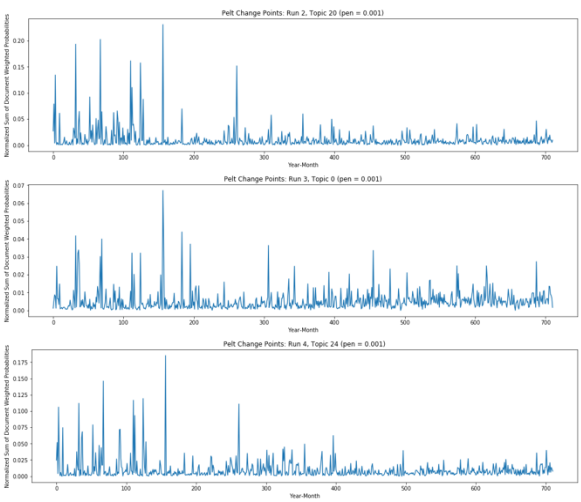


Figure 5. Breakpoints associated with topics in Jaccard-p cluster, mapped against time.



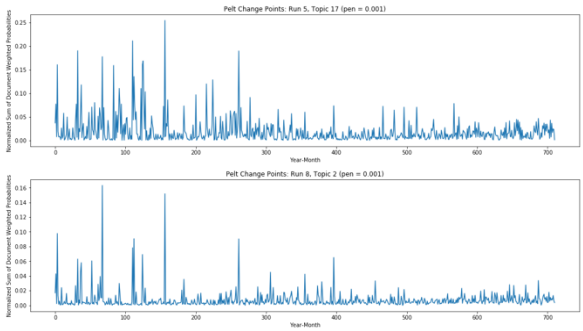


Figure 6. Subset of topic time series associated with a Jaccard-p cluster, indicating corresponding movement across many timepoints.

Figures 5 and 6 shows the information of one such cluster. The breakpoint mapping over time in Figure 5 shows that there are identifiable periods of time (spanning months or years) for the events or trends associated with this cluster. Figure 6 shows a selected subset of the time series of the LDA topics from this cluster. There are similarities in time series across multiple runs, showing that there are corresponding rises and falls in topic prominence over time. The similarities also show that the agglomerative clustering was effective in combating LDA instability.

The examined Jaccard-p cluster's topic annotations indicate topic cohesion. The top documents of this cluster cover discussions on treasury legal tender, the value of money used international trade, and the interest rates established by the Bank of England. These points of discussion are related in the areas of economics, finance, and trade. Common important terms extracted from the documents include "gold," "payment," "price," and "bank." One note is that clusters, including the one being examined, can include certain documents that are not as related to the common theme. For example, this cluster's fifth most important document relates to education, instead of economics. This indicates that we can continue to improve upon our approach to filter out unrelated documents.

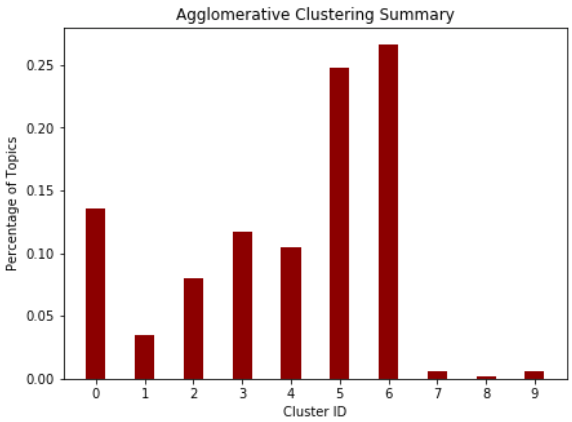


Figure 7. LDA topic distribution over Jensen-Shannon agglomerative clustering.

The Jensen-Shannon approach behaves somewhat better than the Jaccard-p approach in terms of topic distribution across clusters (Figure 7). The VAT diagram (Figure 8) shows internal cluster cohesion, and the topic cohesion is present for many clusters. In addition, we discovered a Jensen-Shannon cluster about finance and economics, containing similar content, documents, and breakpoints to the Jaccard-p cluster discussed earlier. This observation indicates that we can compare clusters across approaches.

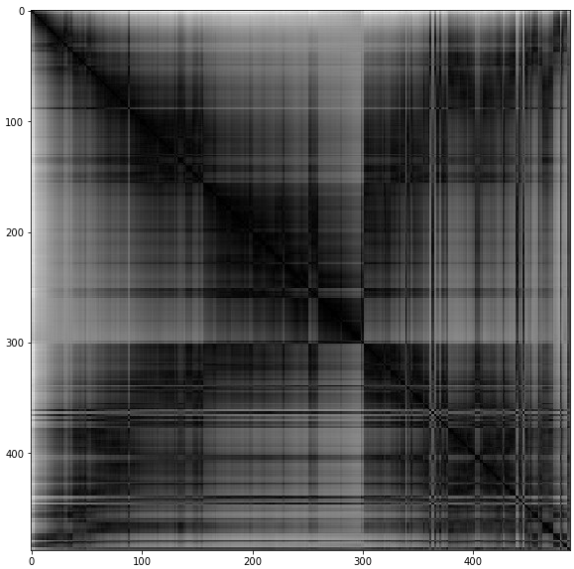


Figure 8. VAT diagram for Jensen-Shannon agglomerative clustering.

The Jensen-Shannon cluster chosen for examination here (Figures 9 and 10) highlights ideas that reoccur frequently across the century. The selected subset of LDA topic time

series from this cluster again show similarities between time series and robustness across runs of LDA.

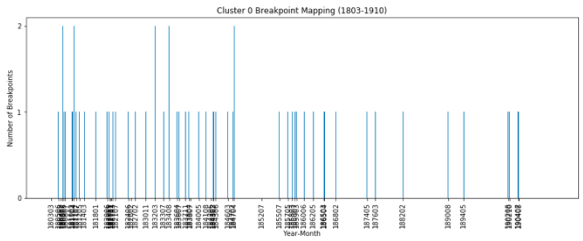


Figure 9. Breakpoints associated with topics in Jensen-Shannon cluster, mapped against time.

The topic annotations of the Jensen-Shannon cluster also indicate topic cohesion. The top documents of this cluster have a focus on educational systems, with additional commentary on government and political systems. Common important terms extracted from the documents include "school," "teacher," "election," and "representative." Future work can focus on distinguishing between these somewhat discrete topics – breaking down larger clusters into smaller ones on other criteria can yield more specific identifications of events and trends.

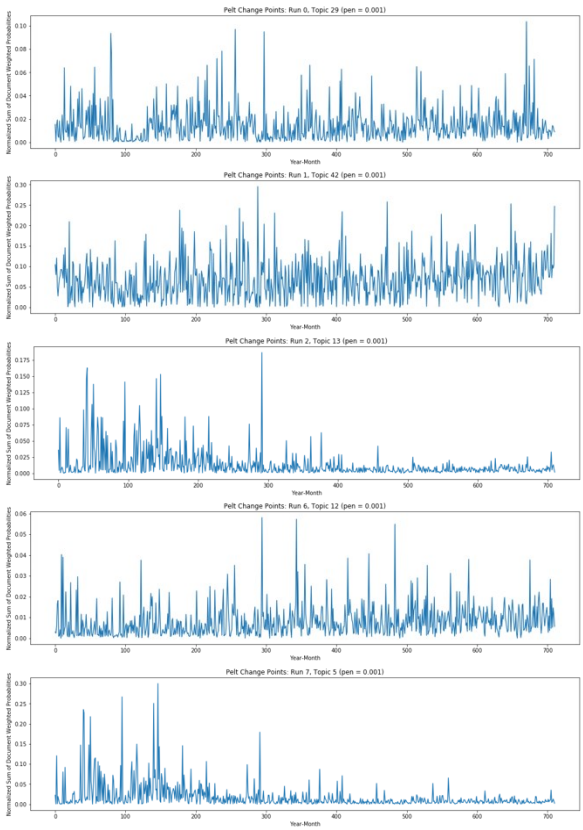


Figure 10. Subset of topic time series associated with a Jensen-Shannon cluster, indicating corresponding movement across many timepoints.

The difference in the word sets used for each approach contributed to the differences seen in the results. With the Jaccard approach, we saw less success with clustering and identification of topics, indicating that we can modify the Jaccard word set composition to be similar to those used in the Jaccard-p and Jensen-Shannon approaches for future experiments.

5 Conclusions and Future Work

The problem of event discovery using topic modeling and change detection is a challenging one. The two experimental methods we define in this paper yielded results with varying degrees of success. Our most reliable results came from the Jaccard-p and Jensen-Shannon approaches from Method 1, where generated LDA topics were clustered based on their document content. We were able to create clusters with distinct areas of discussion, such as finance or education, which we can continue to do analysis on to identify specific historical events.

Our first approach can be improved by increasing the number of samples analyzed per time interval or increasing the granularity of the time interval used. We plan to break down each cluster into smaller sub-clusters to examine more specific topic content – for example, our Jaccard-p cluster could be dissected to explore historical discussions on specific components of the British economic system.

We would also like to explore dynamic time warping technique to measure time series similarity. The simple distance metrics used in our approach suffer from a misalignment problem, where computations rely on a one-to-one mapping of corresponding observations in time series. Dynamic time warping solves the misalignment issue by exploring different warping paths and finding the optimal one that allows for matching of similar time series with different phases.

Finally, we aim to consult with historical experts specializing in the analyzed time period. These experts can provide external validation of the topics generated and insight into what potential changes can be made to our approaches to benefit future historical work.

Acknowledgments

The authors would like to thank the SMU Computer Science and History departments, with particular thanks to Dr. Jo Guldi, Dr. Eric Larson, and Steph Buongiorno.

References

David M. Biel and John D. Lafferty, Dynamic topic models, in *Proceedings of the 23rd international conference on Machine learning*, New York, NY, 2006.

David M. Biel, Andrew Y. Ng and Michael I. Jordan, Latent dirichlet allocation, *Journal of machine Learning research,* vol. 3, pp. 993-1022, 2013.

Jo Guldi, Parliament's debates about infrastructure: an exercise in using dynamic topic models to synthesize historical change., *Technology and Culture,* vol. 60, no. 1, pp. 1-33, 2019.

Elisabeth Günther and Thorsten Quandt, Word Counts and Topic Models, Automated text analysis methods for digital journalism research, *Digital Journalism,* vol. 4, no. 1, pp. 75-88, 2016.

Rem Hida, Naoya Takeishi, Takehisa Yairi and Koichi Hori, Dynamic and Static Topic Model for Analyzing Time-Series Document Collections, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Melbourne, Australia, 2018.

Thomas Hoffmann, Probabilistic latent semantic indexing, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, Berkeley, CA, 1999.

Xiaoling Lu, Yuxuan Guo, Jiayi Chen and Feifei Wang, Topic change point detection using a mixed Bayesian model, *Data Mining and Knowledge Discovery,* vol. 36, pp. 146-173, 2022.

John Miller and Kathleen McCoy, Topic Model Stability for Hierarchical Summarization, in *Proceedings of the Workshop on New Frontiers in Summarization*, 2017.

Mika V. Montyla, Maelick Claes and Umar Faroaq, Measuring LDA topic stability from clusters of replicated runs, *In Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM '18)*, 2018.

Vito M. R. Muggeo and Giada Adelifio, Efficient change point detection for genomic sequences of continuous measurements, *Bioinformatics,* vol. 27, no. 2, pp. 161-166, 2011

Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D. Manning and Daniel A. McFarland, Topic Modeling for the Social Sciences, in *NIPS*, 2009.

Jaxk Reeves, Jien Chen, Xiaolan L. Wang, Robert Lund and Qi Qi Lu, A Review and Comparison of Changepoint Detection Techniques for Climate Data, *Journal of Applied Meteorology and Climatology,* vol. 46, no. 6, pp. 900-915, 2007.

Jonas Rieger, ldaPrototype: A method in R to get a Prototype of multiple Latent Dirichlet Allocations, *The Journal of Open Source Software,* vol. 5, no. 51, 2020.

Jonas Rieger, Carsten Jentsch and Jorg Rahnenführer, RollingLDA: An Update Algorithm of Latent Dirichlet Allocation to Construct Consistent Time Series from Textual Data, in *EMNLP*, Punta Cana, Dominican Republic, 2021.

Charles Truong, Laurent Oudre, Nicolas Vayatis, ruptures: change point detection in Python, arvix preprint, 2018. [Online]. Available: https://arxiv.org/abs/1801.00826.

Charles Troung, Laurent Oudre and Nicolas Vayatis, Selective review of offline change point detection methods, *Signal Processing,* vol. 167, 2020.

Chong Wang, David M. Biel and David Heckerman, Continuous time dynamic topic models, in *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI)*, Helsinki, Finland, 2008.

Yunli Wang and Cyril Goutte, Real-time Change Point Detection using On-line Topic Models, in *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA. 2018

Zhengxing Xu, Ling Chen, Yimeng Dai and Gencai Chen, A Dynamic Topic Model and Matrix Factorization-Based Travel Recommendation Method Exploiting Ubiquitous Data, *IEEE Transactions on Multimedia,,* vol. 19, no. 8, pp. 1933-1945, August 2017.

Yi Yong, Shimei Pan, Jie Lu, Mercan Topkara and Yangqiu Song, The stability and usability of

statistical topic models, A*CM Trans. Interact. Intell. Syst,* vol. 6, no. 2, 2016.