Analyzing Presidential Speeches with Topic Modeling

Ole R. Villadsen
University of Tennessee—Knoxville
School of Information Sciences
IS590 Intro. to Digital Scholarship
Summer 2014, Dr. P. Bogen
Term Project

ABSTRACT

This paper discusses my course term project to assess the utility of a computational analytic technique called probabilistic topic modeling to identify latent topics or themes present in a large corpus of textual information. I set out to accomplish this goal by performing a topic modeling text analysis on a corpus of 622 key U.S. presidential speeches identified by the University of Virginia Miller Center and archived on their web site at http://millercenter.org/president/speeches.

The results of this project, together with a review of the available literature on topic modeling, suggest that this technique is an effective tool for mining large data sets to identify latent themes or topics. The results of the topic modeling analysis of the presidential speeches suggest that the technique accurately identified latent themes or discourses across different presidential speeches over time. The results also suggest that it is an effective tool for producing new insights into the history of presidential speeches, including finding similarities between speeches that otherwise might not be apparent.

Categories and Subject Descriptors

I.2.7 [Artificial Intelligence]: Natural Language Processing—*Text Analysis*

General Terms

Algorithms, Performance

Keywords

Topic Modeling, Latent Dirichlet Allocation, Data Mining, Digital Humanities

1. INTRODUCTION

Advances in computational data mining techniques are presenting new opportunities to extract meaningful information from a variety of text-based corpora. One such technique, probabilistic topic modeling, has demonstrated the ability to identify latent themes aka "topics" resident in a large collection of text documents, and to enable researchers to gain new insight into these documents upon close examination of these topics and their distribution within the corpus.

Probabilistic topic modeling is a method of text analysis that extracts a pre-determined number of word clusters—which can be thought of as topics, themes, or discourses—that reside in a large collection of text-based documents. In recent years researchers across many domains, ranging from the humanities to political science to commercial interests, have embraced topic modeling's capabilities to further their research.

A review of recent articles on probabilistic topic modeling suggests that it is an effective method for a researcher to extract latent themes or topics from a large corpus and to identify and explore new insights into the collection. The project to analyze 622 key presidential speeches reinforced several key themes that emerged from the literature on topic modeling, including that topic modeling requires considerable knowledge and effort to interpret the results for meaningful information.

2. WHAT IS TOPIC MODELING?

Topic modeling is an approach to mining a large collection of textual documents to extract a pre-determined number of hidden or latent "topics" from a body of text documents, or corpus. Discovering these hidden topics provides researchers greater insight into the documents and the ability "to summarize, visualize, explore, and theorize about a corpus." [1].

Topic modeling itself is about 15 years old and has been applied to numerous fields, including bioinformatics, comparative literature, history, political science, and social media [8]. Latent Dirichlet allocation (LDA), a form of topic modeling that was developed in 2002-2003 by David N. Blei among other colleagues, has been particularly popular with digital humanities and is the method of topic modeling used for this project.

2.1 What is a 'Topic'?

A critical aspect to understanding topic modeling and its utility begins with defining what is meant by a "topic" and how the model views the notion of a "topic." A topic can be effectively defined as a "recurring pattern of co-occurring words" [3, p.12], which is what topic modeling seeks to extract from a large body of texts. While we can think of a topic as an underlying "theme" that plays a role in generating a document, the tool itself regards a topic as a "collection of words that have different probabilities of appearance in passages discussing the topic" [14].

2.2 How Does Topic Modeling Work?

A key feature of probabilistic topic modeling is that it has no insight itself into the meanings of the words or their syntax. Each document is essentially treated as a "bag of words," with two important assumptions at work: 1) there are a fixed number of "topics" i.e. recurring patterns of co-occurring words existing within the documents, and 2) each document in the corpus contains the aforementioned topics to varying

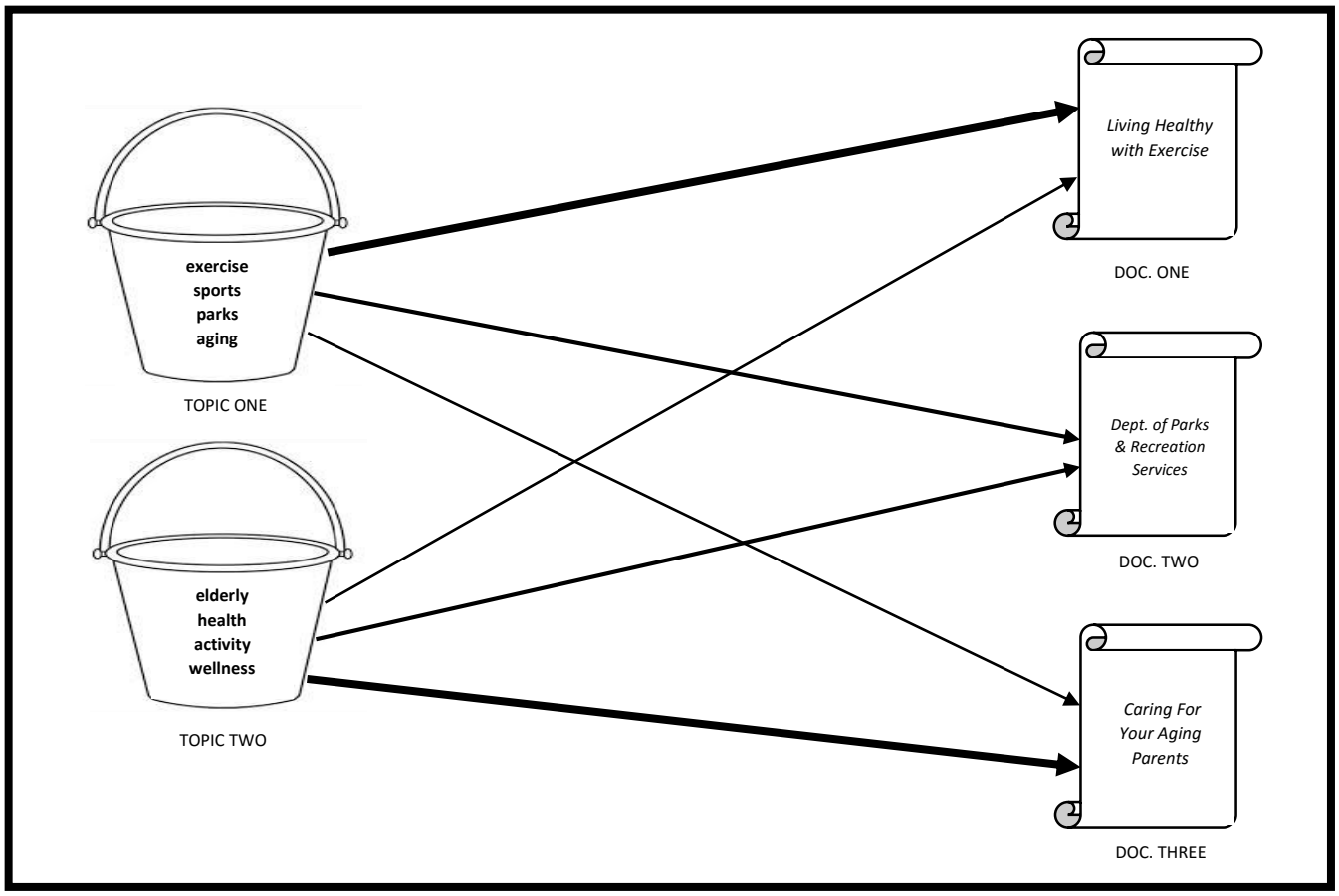

Figure 1. Topic modeling at work

An effective topic model analysis of a corpus would yield clusters of words aka "topics" that look something like "'navy, ship, captain' and 'tobacco, farm, crops.'" [3]. One might think of topic modeling as a tool that allows a researcher to look inside a corpus and do some reverse engineering to discover the topics or themes that the documents' authors might have been considering when they generated the documents within the collection. [14].

Figure 1 provides a very simple demonstration of how a topic model might extract two word clusters i.e. topics from three documents. Topic One words can be found primarily in Document One, and Topic Two words are most common in Document Three. Each topic is present in roughly the same amounts in Document Two.

degrees [1]. In practice, the researcher assigns a fixed number of topics to the topic modeling tool, which can have a significant impact on the results (see Section 3.3 below).

Provided a collection of documents and a fixed number of topics determined by the researcher that reside within this corpus, the topic modeling algorithms go to work to find the clusters of words that tend to reside together within the documents by determining the probability that a word in each document belongs to a specific word cluster or topic [1]. Put very simply (and jumping over the "occult statistics" that provide the mathematical muscle behind the topic modeling algorithms), the model begins with randomly

assigning words to each of the predetermined number of word clusters. It then examines each word against each word cluster, reassigning the words to different word clusters in such a way as to increase the frequency of the word within the word cluster (aka topic) and the presence of the word cluster within a document.

As this process unfolds, words will increasingly have a greater presence within word clusters where they are already present, and the word clusters will have a greater presence within documents where they already reside [14]. The topic modeling algorithms continue to perform these computations until the model reaches a state of equilibrium consistent with the collection. While this may sound simple enough, keep in mind that the model performs up to several billion calculations depending on the size of the corpus to achieve the highest probabilities.

2.3 Preparing for Using Topic Modeling

There are several steps a researcher must take before embarking on a project to topic model a collection of documents. First, there must be a sufficiently large number of documents for the model to work properly, at least several hundred. The documents must also be prepared before loading into the topic modeling tool, which means tokenizing the text and removing stopwords [3, 9]. This process may also necessitate web scraping or other techniques to harvest documents that, for example, reside as hyperlinks on web pages.

Once the text has been prepared, there are a few topic modeling tools that use LDA from which to choose. One such tool—MAchine Learning for LanguagE Toolkit (MALLET)—is particularly popular within digital humanities; it can be accessed through the command line as well as a Java graphical user interface [7]. Paper Machines, a plugin for Zotero bibliographic management software, also uses MALLET for its topic modeling features. *Gensim*, which comes from "generate similarities," is another open source vector space modeling and topic modeling toolkit that is available for use by the public. *Gensim* is based on the programming language Python that uses a variational Bayes algorithm for LDA.

3. APPLYING TOPIC MODELING TO TEXT ANALYSIS

3.1 Topic Modeling in Practice

Topic modeling has been used increasingly over the past several years to analyze large collections of text documents in the humanities and social sciences to derive new insights into the collection that would otherwise be unavailable. Indeed, interest in text modeling has grown so much so that two peer-reviewed journals devoted to humanities and social sciences, *Journal of Digital Humanities* and *Poetics*, devoted entire issues to the use of topic modeling in 2012 and 2013, respectively.

3.1.1 Case Studies in Topic Modeling

Cameron Blevins' use of topic modeling to analyze Martha Ballard's diary helped open the door to the use of this technique to examine humanities' texts [8, 2]. Topic modeling has also penetrated other domains, including political science, history, and media analysis. Examples of research studies based on topic modeling include:

- Analyzing the controversy over US federal funding of the arts during the 1980s and 1990s by examining 7,958 newspaper articles [4].
- Exploring the journal of the Modern Language Association, PMLA, to provide new insight into the history of literary studies [6].
- Exploring the public discourse over science from 1980 through 2012 through topic modeling about 15,000 newspaper articles [5].
- Examining 118,000 speeches from the Congressional Record to identify the agenda of the US Senate from 1997-2004 [10].

Topic modeling has even extended into commercial applications. For example, one group of researchers investigated how to use topic modeling to identify customer sentiment in Twitter. The researchers investigated how to use LDA techniques to identify Twitter posts that contained useful information for exploring customer sentiment [13].

3.2 Benefits of Topic Modeling

Topic modeling is a cost-effective way to explore a large collection of documents to extract prevailing themes or topics that reside in the corpus. It is also a tool that enables researchers to identify and explore new insights into the corpus that may otherwise not be evident. Most researchers agree that topic modeling is mainly useful for examining a very large collection i.e. hundreds or thousands of documents that are too large for a single person to read and analyze effectively.

3.2.1 Cost Effective

In this respect, topic modeling is able to analyze and identify key topics in a large collection of documents produced over a long period of time more effectively than, for example, reading each individual document and manually coding its content. Indeed, Quinn et al. in 2010 used a topic modeling analysis of the Congressional Record to show that from a cost perspective, topic modeling offers many advantages over other methods of analyzing large collections of text, such as reading, human coding, and dictionary-based coding [10].

3.2.2 New and Unexpected Insights

Beyond the practical utility of analyzing large corpora, topic modeling also provides a way to uncover new insights within the documents under examination. Following his topic modeling analysis of the discourse over science, Evans [5] found that the model's results yielded subjects of discussion

that a prior qualitative analysis of the material failed to identify in addition to those that he already expected to see, such as "race" and "evolution." Evans [5] subsequently labeled these new unexpected word clusters as "cultural history" and "presidential politics."

Topic modeling also allowed Evans [5] to examine individual case studies, such as the topic "evolution," and examine their importance by looking across a broader range of topics over time for the necessary comparison data. Finally, Evans [5] was able to compare two different topics over time, such as "evolution" and "race," to determine their relationship and correlation [5].

Researchers also have identified topic modeling's advantages in analyzing humanities literature, including poetry—an arguably more difficult form of text to topic model given the heavy emphasis on figurative rather than direct language. As a result, analyzing poetry with probabilistic topic modeling arguably may fail to yield the kind of concrete word clusters present in other forms of corpora from which topics can be readily identified. However, Rhody [11] and Underwood [14] uncovered that the seemingly opaque topics that arise from topic modeling poetry also provide fertile ground for analysis and new discoveries after investigating their origin within the documents where these topics are most prevalent.

In this respect, such opaque word clusters aka topics can be viewed as a form of discourse or rhetoric rather than concrete topics based on a theme, and discovering the documents in which they reside and over what time periods can yield new insights about the collection. For example, Rhody [11] found that a topic consisting of "death, lie, heart, dead, long, world, blood, earth…" was found not only in poems about death but also pieces about struggles and loss in general. Underwood [14] found that another opaque topic that came out of a topic modeling exercise on poetry—"thy, where, over, still, when, oh, deep, bright…."—correlated with poetry written mostly by women from 1815 to 1835.

3.3 Limitations of Topic Modeling

Using topic modeling to perform text analysis is not without its limitations, and any researcher should proceed with caution before embarking on a topic modeling project. Two limitations with using topic modeling to effectively identify and analyze latent topics within a collection of documents are the inputs required by the researcher into the model and the time, effort, knowledge, and expertise required to analyze the results.

3.3.1 Inputs Make a Difference…

Several researchers have noted the critical effect that the researcher-provided inputs into the topic modeling tool can have on the results produced by the model. Researchers must provide the scope of the collection i.e. determine which documents should form the collection, the number of topics the topic modeling tool should return, and the stopwords in the text. Variations in all of these inputs can yield different word clusters, and it is often necessary to adjust the number of desired topics from the corpus to extract coherent word clusters that provide a basis for further analysis [5, 6, 14].

3.3.2 …as does the Interpretation of the Results

Other researchers have pointed to the need to devote considerable time and energy to analyzing, interpreting, and understanding the results [1]. Quinn et al. [10] acknowledge in their study on the Congressional Record that while topic modeling offers lower costs in terms of time and effort spent on each text, the approach nonetheless requires a "moderate" level of person-hours spent interpreting the results.

A word cluster resulting from a topic model also may fail to represent a coherent theme or discourse. Schmidt [12] emphasizes that understanding the results of a topic model requires a thorough understanding of the words that build each cluster. A complete topic might consist of several hundred words, and the top few may only give a limited perspective into the topic that it represents. While examination of a word cluster may lead to profound new insights, a topic model could also produce misleading or ambiguous results.

4. APPLYING LESSONS LEARNED

The above literature review yields several important insights—as well as words of warning—in before embarking on a topic modeling project. On practical matters, the literature review gives some insight into the topic modeling tools and the preparations that must be performed on the documents before the topic modeling can commence. Text must be tokenized and the stop words removed before analyzing the documents.

The literature on topic modeling also highlights the importance of adjusting the inputs into the model, such as the number of topics, and examining each set of results to determine how effectively they represent the corpus' latent themes. The review also highlights the importance of examining the results closely; for example, looking carefully at all of the words that make up a topic instead of only the initial few, to uncover unique topics uncovered by the model and examining them further to determine in what documents they are prevalent and what insights they yield into the collection.

The literature also highlights the importance of comparing unusual or noteworthy topics together over time to determine how they correlate. In short, the model will not spit out new insights wrapped up with a bow. The researcher must uncover them through extensive analysis and interpretation of the topic modeling results.

Figure 2. Topic modeling results from 622 US presidential speeches using *gensim*

5. TOPIC MODELING PRESIDENTIAL SPEECHES

Topic modeling 622 presidential speeches located on the UVa Miller Center web site corroborated many of the advantages and limitations of this technique that are documented in the literature on topic modeling, in particular the importance of effectively choosing and preparing the documents, devoting time and effort to interpreting the results, and the potential for new and unexpected insights to emerge. The topic modeling analysis was performed using the *gensim* toolkit, which returned 25 topics out to 20 words for each topic (see *Figure 2*).

5.1 Identifying Key Themes across History

Many of the topics produced by the topic modeling are consistent with what one would expect from the presidential speeches. Such results demonstrate the power and utility of this form of computational text analysis to identify underlying themes over time that influenced the preparation of the 622 speeches analyzed in this project.

These results demonstrate the capability of probabilistic topic modeling to accurately identify and represent the underlying semantic content within a corpus; however, many of the results also reflect what one might expect from the history of presidential speeches and probably could have been determined from a close read of the speeches themselves.

5.1.1 Comparing Presidents Bush and Obama

A close look at the distribution of topics present in speeches from President George W. Bush and President Barrack Obama is consistent with the type of subjects that one might expect from these two leaders (see *Figure 3*.). For example, Topic #16, which includes words such as "Iraq, nation, terrorists, Iraqi, freedom, forces, and democracy," and which I have labeled "War on Terror," has a strong and consistent presence in many of President Bush's speeches.

As one familiar with US politics might expect, "War on Terror" becomes much less prominent in speeches from President Obama. However, Topic #20, which includes words such as "work, tax, health, care, jobs, children, jobs, budget, and economy," and which I have labeled "People and Society," increases significantly in Obama's speeches.

Two prominent exceptions in Bush's and Obama's speeches to the above pattern also make sense when taking into account the particular nature and content of each of these exceptions. Bush's speech on June 7, 2001 does not contain any portion of Topic #16 "War on Terror," and instead features Topic #20 "People and Society" and another topic

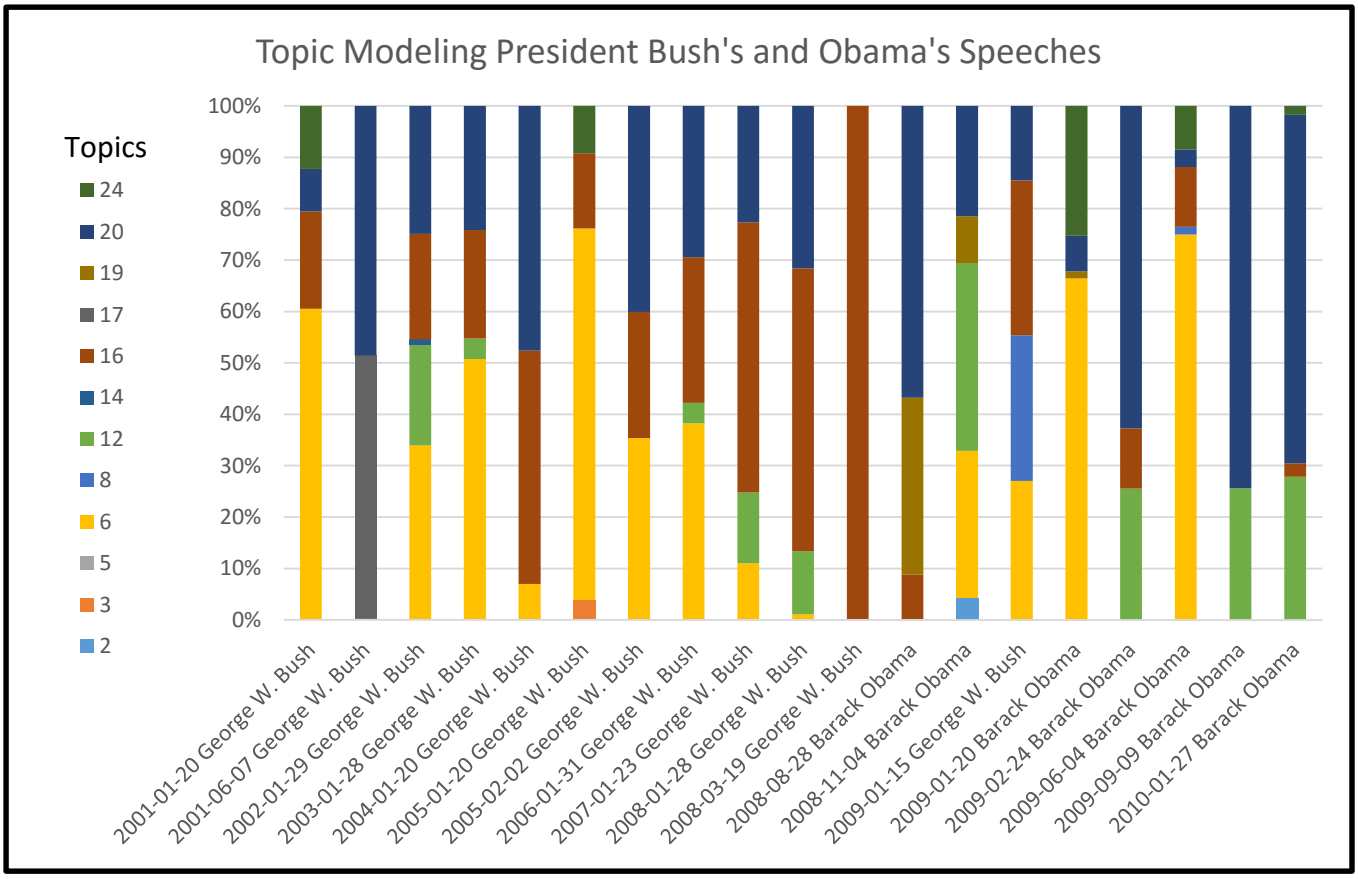

Figure 3. A comparison of the topic distribution within speeches by President Bush and President Obama

(Topic #17) that contains words associated with domestic political and economic issues. This speech, which was entitled "Remarks on Signing the Economic Growth and Tax Relief Reconciliation Act," discusses US domestic economic issues, indicating that these topics make sense according to the subject of this particular President Bush speech.

Similarly, President Obama's speech on June 4, 2009, contains relatively little of Topic #20 and instead features Topic #6 (a topic that I have labeled "20th Century US Foreign Policy") and to a much lesser extent Topic #16 "War on Terror." The distribution of topics in this speech also makes sense because this particular President Obama speech was given at Cairo University and focused on US foreign policy.

5.1.2 Tracing a topic over time

An examination of a topic's presence within this corpus over time also yields interesting results that validate the utility of this technique. Topic #24, which I have labeled "Energy and Oil," consists of words that suggest it will be found in speeches that address US energy needs. If the topic modeling is accurate, a look at this topic over time should reflect a greater presence in speeches delivered in the 20th century, and in particular during the 1970s and 1980's during US oil shortages. *Figure 4* indeed shows this trend.

5.2 Finding New and Unexpected Insights.

Topic modeling the presidential speeches also reinforced the ability of topic modeling to yield new insights into the corpus that might otherwise not be apparent. A close analysis of several interesting topics, whose underlying theme is not readily clear, and their distribution within the presidential speeches provides some astonishing results. These results further suggest the capability of topic modeling to identify similarities in the semantic content, as well as the tone and rhetoric, within a corpus. Several topics produced by the probabilistic topic modeling, such as Topic #10, Topic #10, and Topic #13, show this capability.

These findings reinforce some of the lessons espoused by Rhody [11] and Underwood [14] on the ability of topic modeling to yield insights based on figurative as opposed to direct subject-oriented language. This analysis also highlights the importance identified by Schmidt [12] of examining all of the words that comprise a topic and not just the first 10 or 20; indeed, these additional words may have a great impact on generating the final topics and provide insight into why there are similarities between different documents that otherwise may not be apparent in the first 10 or 20 words in the topic.

5.2.1 Topic # 10

Topic #10 consists of words such as "statute, law, purpose, men, company, company, tobacco, antitrust, and corporations." These words defy an easy label and represent the kind of topic that bears further analysis. The topic is present in many speeches, mostly in small percentages, but there are three speeches that contain this topic exclusively, and a fourth that is represented by 86 percent of this topic. An examination of these speeches shows some interesting results.

Two of the documents (it turns out that they were not actually speeches) that consist exclusively of this topic are George Washington's pardon to those accused of treason during the Whiskey Rebellion, and John Adams pardon five years later to those engaged in the Fries Rebellion. These results suggest that the topic may be influenced by legalistic language as well as concepts such as treason, rebellion, insurrection, and pardons. It also suggest that this topic is influenced more by the kind of words or language germane to written documents as opposed to oral speeches.

A third document that exclusively contains this topic is President Abraham Lincoln's public letter to James Conkling in which President Lincoln discusses and defends the Emancipation Proclamation. A fourth document, 86 percent of which consists of Topic #10, is Thomas Jefferson's written instructions to Lewis and Clark ahead of their expedition.

These results—all of which were originally written documents and not oral speeches— suggest that the topic is heavily influenced by written language associated with the late 18th and early 19 centuries, and probably as well by the semantics associated with freedom, pardons, treason, rebellion, and negotiation.

5.2.2 Topic # 2

Topic #2 also defies easy categorization based on the top 20 words in the cluster, but an examination of four speeches that contain a high proportion of this topic provides interesting insights. The topic consists of words such as "people, children, time, good, church, faith, freedom, and place."

Two speeches exclusively contain this topic. The first is a speech by President Reagan delivered on May 5, 1985 in Germany in which President Reagan remembers and discusses the horrors of the Holocaust. The second is a campaign speech by President Clinton on November 3, 1996 in a church in which he discusses, among other things, the conflict in Bosnia, the Holy Land, Africa, and the US Oklahoma bombing. Another speech by President Clinton, which consists of 89 percent of this topic, is his remarks at the US national cemetery on June 6, 1994, the anniversary of the D-Day invasion. Finally, John F. Kennedy's "Ich bin ein Berliner" speech on June 26 1963 consists of 64 percent of this topic.

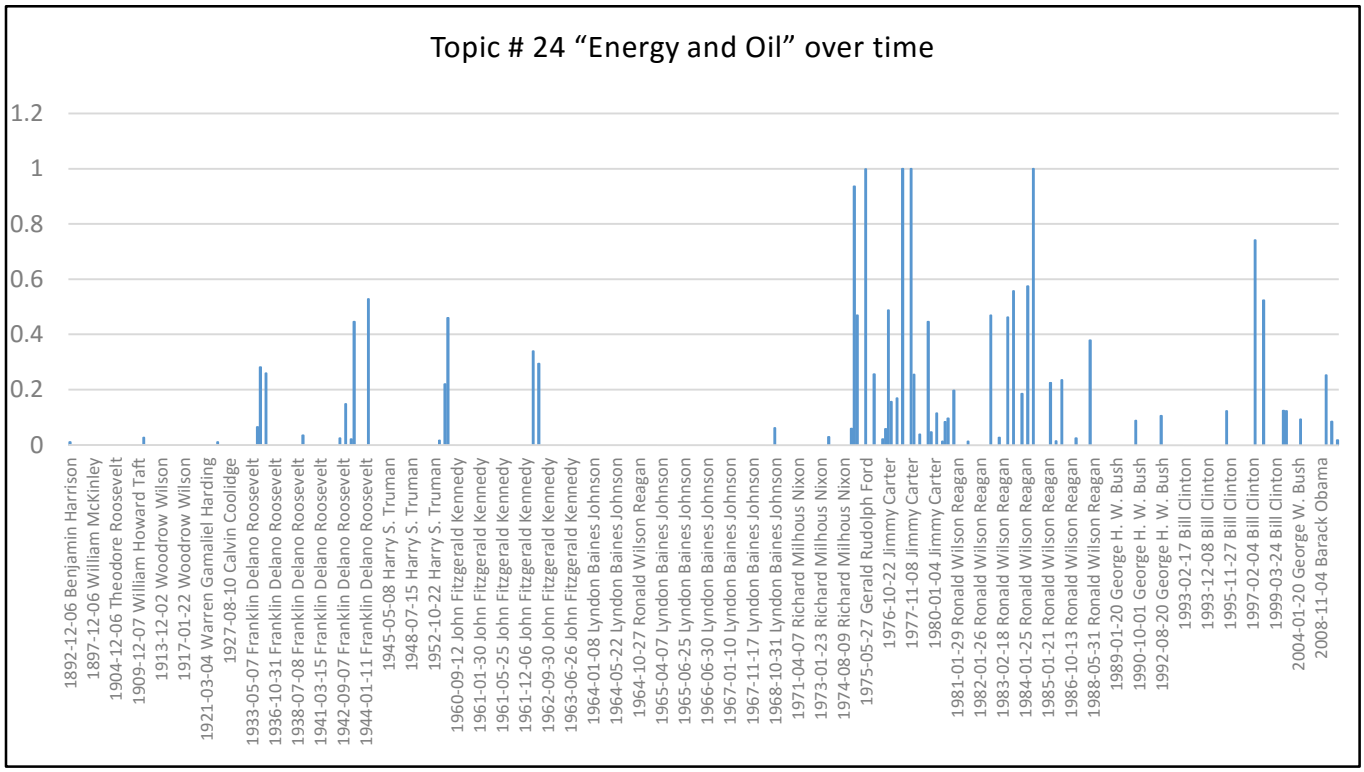

Figure 4. Tracing the appearance of Topic #24 starting with its emergence in 1892.

Examining these speeches in light of this topic's prevalence within them suggests that the topic is capturing content related to historical conflict and tragedy, particularly in Europe during World War II, the importance of not repeating the horrors of the past, and of joining and working together for the future. It is very likely that many of the words within this cluster aka topic beyond the first 20 heavily influence the presence of this topic within the speeches.

5.2.3 Topic #13

Another very interesting topic is Topic #13, which consists of words such as "beloved, rescue, people, Cherokees, good, men, Iran, give, nation, operation, made, states, united, agent, lands, indian, man, advice, nations, great." The topic is found in only six speeches, and only two of those speeches contain the topic to a significant degree.

President George Washington's speech on August 29, 1796 to the Cherokee Nation exclusively contains this topic. President Carter's statement on the failed Iran Rescue Mission on April 25, 1980 consists of 86 percent of this topic.

These two speeches are very different in terms of their direct subject matter; one is about US relations with Native Americans and the other about a failed US combat mission in Iran. The top 20 words in the topic also are apparent in these speeches, with "Cherokees" and "lands" mixed in with "Iran" and "operation."

A close examination of these two speeches suggests, however, that they have the same pensive, reflective, gentle, and explanatory tone, which perhaps yielded the kind of figurative language that ultimately joined these two speeches together within the same topic. As with Topic #2, the words within this cluster beyond the first 20 probably would provide greater understanding of the influence of this topic on these speeches.

Tips for Successful Topic Modeling

- Carefully prepare the corpus by tokenizing the text and removing stopwords. The first few iterations of the model may identify additional stopwords for removal that appear in the word clusters.
- Run several iterations of the model, adjusting parameters such as the number of topics each time. Fewer topics will produce broader themes, and more topics produce narrower themes.
- Be sure adjust the tool e.g. MALLET or *genism* to produce topics out to at least 50 words; the extra words may yield important insights into the nature of the topics when interpreting the results.
- Examine the results closely to look for themes that are easy to identify and those whose meaning is more ambiguous. Often the latter will identify unusual similarities in the documents where these topics are prevalent.
- Use graphics and other methods of visualization to represent the results over time or compare the distribution of topics in a set of documents.

5.3 Preparing for and Performing the Analysis

5.3.1 Preparing the Speeches

The speeches resided on the UVA Miller Center's web site, and the first step in the project was to grab and convert them into a usable form using "web scraping" techniques. To perform the web scraping I used an online and publicly available tool called *import.io*. After training the tool, I was able to use it to download the name of the president, date, title, and content of the speech into separate columns in a spreadsheet. I then used an Excel VBA macro to convert the text of the speeches into 622 separate text files and name them according to the president and date of the speech e.g. 1789-04-30 George Washington.txt.

5.3.2 Performing the Topic Modeling

I used the *gensim* toolkit to perform the topic modeling, which first required writing a program in the Python programming language to prepare the text documents i.e. clean and tokenize the texts and remove stopwords. I also wrote the script so *gensim* would generate and print 25 topics out to 20 words, as well as the distribution of the topics in each speech.

5.3.3 Analyzing the data

To assist with analyzing the data, I wrote a separate Python program to help convert the distribution of topics in each speech (the text output from the Python program above) into a spreadsheet, which I then was able to use to manipulate and visualize the results.

6. CONCLUSION

Probabilistic topic modeling is an effective technique to identify the prevailing themes or subjects that influenced the preparation of a large corpus of written material, based on the results of this project analyzing 622 US presidential speeches and a review the literature on topic modeling.

Topic modeling nonetheless requires a high level of understanding of the corpus and the ability and willingness to spend considerable effort analyzing the results. Making the most out of the results of the topic modeling often requires examining unusual word clusters, i.e. "topics," that lack an obvious or consistent theme. Examining such topics against the documents in which they are prevalent will often yield the most profound insights into the corpus and identify similarities in the documents' content that otherwise may not be apparent.

Topics whose words have a more easily identifiable or consistent theme may not necessarily yield new insight into the corpus, but they can be useful for validating the success

of the topic modeling. Such topics can also yield new insights by examining their prevalence in the corpus over time or in proportion to the use of other topics. Using graphics or other visualization techniques to represent the topics and their distribution in documents is an effective, and arguably necessary, way to identify these insights.

Running several iterations of the topic modeling tool and adjusting parameters such as the stopwords and number of topics to be produced is critical to producing usable results. A successful topic modeling project will also depend on the choice and preparation of the corpus itself. Effectively preparing the corpus i.e. tokenizing the documents and removing all stopwords will have a notable impact on the results.

7. REFERENCES:

[1] David M. Blei. 2012. Topic Modeling and Digital Humanities. *Journal of Digital Humanities* 2, 1 (Winter 2012), 8-11.

[2] Cameron Blevins, 2010. Topic Modeling Martha Ballard's Diary. (April, 2010). Retrieved from http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary/

[3] Megan R. Brett. 2012. Topic Modeling: A Basic Introduction. *Journal of Digital Humanities* 2, 1 (Winter 2012), 12-16.

[4] Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. *Poetics* 41, 6 (December 2013), 570-606.

[5] Michael S. Evans. 2014. A Computational Approach to Qualitative Analysis in Large Textual Datasets. *PLoS ONE* 9, 2 (February 2014), 1-10.

[6] Andrew Goldstone and Ted Underwood. 2012. What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship. *Journal of Digital Humanities* 2, 1 (Winter 2012), 39-48.

[7] Shawn Graham and Ian Milligan. 2012. Review of MALLET, produced by Andrew Kachites McCallum. *Journal of Digital Humanities* 2, 1 (Winter 2012), 73-76.

[8] Elijah Meeks and Scott Weingart. 2012. The Digital Humanities Contribution to Topic Modeling. *Journal of Digital Humanities* 2, 1 (Winter 2012), 2-6.

[9] John W. Mohr and Petko Bogdanov. 2013. Introduction—Topic Models: What they are and why they matter. *Poetics* 41, 6 (December 2013), 545-569.

[10] Kevin M. Quinn, Burt L. Monroe, Michael Colaresi, Michael H. Crespin, Dragomir R. Radev. 2010. How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science* 54, 1 (January 2010), 209-228.

[11] Lisa M. Rhody. 2012. Topic Modeling and Figurative Language. *Journal of Digital Humanities* 2, 1 (Winter 2012), 19-35.

[12] Benjamin M. Schmidt. 2012. Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities* 2, 1 (Winter 2012), 49-65.

[13] Stephan Sommer, Andreas Schieber, Kai Heinrich, Andreas Hilbert. 2012. What is the Conversation About? A Topic Model Based Approach for Analyzing Customer Sentiments in Twitter. *International Journal of Intelligence Information Technologies* 8, 1 (January-March 2012), 10-25.

[14] Ted Underwood. 2012. Topic modeling made just simple enough. (April, 2012). Retrieved June 29, 2014 from http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/