Doris: A tool for interactive exploration of historic corpora

Sreya Guha

Castilleja High School

Abstract. Insights into social phenomenon can be gleaned from trends and patterns in corpora of documents associated with that phenomenon. Recent years have witnessed the use of computational techniques, mostly based on keywords, to analyze large corpora for these purposes. In this paper, we extend these techniques to incorporate semantic features. We introduce Doris, an interactive exploration tool that combines semantic features with information retrieval techniques to enable exploration of document corpora corresponding to the social phenomenon. We discuss the semantic techniques and describe an implementation on a corpus of United States (US) presidential speeches. We illustrate, with examples, how the ability to combine syntactic and semantic features in a visualization helps researchers gain insights into the underlying phenomenon.

1 Introduction

One way of understanding social phenomenon or the behavior of groups is to analyze the discourse within the group. Such a discourse is often scattered across different documents in a large corpus. In any given document corpora, analysis of words, topics, and evolution of the discourse provide deeper insights into the social phenomenon. This is especially true for historical analysis of any social phenomena.

In recent years, computational tools have been developed to analyze large corpora of documents. A notable application of the use of computational techniques was the authentication of the authorship of the works attributed to Shakespeare. Computational techniques have enabled a deeper analysis of the playwright's work to uncover if he was the sole author [1]. Another area of development is the use of interactive user interfaces that enable a researcher to explore a corpus of documents. The work by Schmidt [10] has illustrated the power of interactive tools for deeper understanding of a corpus of speeches, such as those by United States (US) Presidents.

Most of the tools used currently for researching social phenomenon through associated text corpora perform a relatively shallow analysis. Much of this work is focused on the distributional properties of words and phrases in texts. However, many interesting questions cannot be expressed purely in terms of the words that appear in the documents.

Consider, for example, the corpus of US Presidential State of the Union speeches. The role and status of Native Americans have been an issue, to varying degrees, in this corpus. However, it is difficult to perform this analysis by looking at the distribution of words, since the terminology used to refer to them has evolved over time. Two hundred years ago, the group we now recognize as Native Americans were referred to as

'Indians', 'Red Indians' or by specific tribe names (Apache, Cherokee, etc.). Ideally, we would like a model that provides "Native Americans" as a first class model feature, allowing us to slice and dice by these features, just as we could using a word or phrase.

In this paper, we describe Doris[1], an interactive exploration tool that combines the generality of keyword based approaches with the deeper semantic understanding enabled by both by Semantic Web markup and statistical models. Our tool (available at http://pres-search.appspot.com) allows the user to search the corpus by a specific word, phrase or topic and see the distribution of mentions across Presidents. The user can also restrict the search by president or type of speech (such as State of the Union, Proclamation and Executive Action). The search results are accompanied by graphs plotting the distribution of documents satisfying the search criteria. These graphs enable researchers to gain insights into the evolution of the discourse, as captured by that search, over time. Our long-term goal is to enable interactive exploration of large document corpora at a semantic level. In this paper, we combine categorical filtering based on semantic categories together with classical keyword based search to create a tool, Doris, that can help explore a corpus of documents. We apply this tool to a corpus of over 12,000 documents of US Presidential statements, including the State of the Union speeches, Proclamations, Inaugural addresses and Executive Actions. We use this tool to explore this document space and illustrate the power of the tool with some conclusions that follow.

2 Related Work

Though the use of computational techniques for analyzing text corpora in the context of social science research is relatively new, there is a rich and growing body of work that uses various techniques drawn from the information retrieval community in order to better understand social and political phenomenon. The work by Shen, Aiden, Norvig, et al. [7], which performed a quantitative analysis of the unigrams and bigrams in millions of digitized books, though very simple in its analysis, was very influential.Ben Schmidt [10] uses the bookworm database to visualize the occurrence of certain words in the State of the Unions by American presidents. Given a particular State of the Union, the user can choose certain words and see a distribution across presidents.

Our interface is influenced by the work of Freeman and Gelernter [4], in which they introduced the idea of temporal presentation of a set of documents, which was adapted by Bergman, Beyth-Marom, et al. [2] to search interfaces.

3 Methodology

Our goal is to create a tool for interactive exploration of a corpus of documents that captures the discourse in some social phenomenon. With any large corpus, we need the ability to begin the exploration from different starting points. Keyword based search offers a good metaphor for this. We augment the 'raw' text of the documents with structured/semantic data, including metadata (author, date, etc.) and annotations that capture the higher level semantics of the topics discussed in these documents.

We apply our tool to a corpus of documents from US Presidency Project at the University of California, Santa Barbara [9], that includes all the US Presidential State of

[1] Named after Doris Goodwin, noted US Presidential historian.

the Union Speeches, Proclamations, Executive Actions, Proclamations, etc., giving us a total of 12,345 documents. We gather the following kinds of metadata: type of speech, author (president) and date for each. Since, the metadata is not directly embedded into these documents (like it often is in web pages), in addition to the text corpus, our processing pipeline accepts files containing annotations on these files expressed in RDF or RDFa. In addition to simple metadata about the texts, the kind of exploration we seek to enable benefits greatly from annotations that capture semantic aspects of documents. We now describe our work on each of two kinds of annotations.

3.1 Metadata

We extracted the metadata from the Presidency websites[9] by using a set of scrapers. Some of the vocabulary for expressing the metadata is already available in schemas such as those from Schema.org. Other aspects of the metadata, such as the kind of speech/document, are not part of any well-known vocabulary we know. In order to facilitate this, we have developed a number of vocabulary terms, which are in discussion, for inclusion into Schema.org. From existing Schema.org, we use the vocabulary items $datePublished$, $author$ and $title$. An important aspect of the documents in this corpus is the kind of document: State of Union, Proclamation, etc. Schema.org has a very general class called 'CreativeWork' and we can introduce subclasses under this to represent these kinds of documents. While it is easy to simply introduce four new types as subclasses of 'CreativeWork', it is clear that these are just four in a much large landscape of political documents. After a set of discussions involving the Schema.org community, we used the following vocabulary.

Political Discourse Vocabulary We propose a number of additions to the existing vocabulary at Schema.org. Schema.org already has the properties we require. We augment the existing vocabulary with the two classes: 1) subclasses of 'CreativeWork' (e.g., PublicSpeech, PressRelease, Proclamation, ExecutiveAction), and 2), subclasses of Speech (e.g., InauguralAddress, CommencementAddress, CampaignSpeech, StateOfUnionReport). Though some of these items, such as $StateOfUnionReport$, are specific to the political structure of the United States, most of the newly added terms apply not just to US politics, but more generally, to any political discourse.

3.2 Semantic Annotations

As discussed in the introduction, simple word level treatments are not capable of capturing trends that involve the significant, correlated variations in vocabulary (such as the language around partisan issues such as abortion and gun control), the evolution of vocabulary (such as the words used to refer to people of African origin) or the differences in granularity (such as Cherokee vs Native American Tribes). A number of techniques have been developed to extract more semantic abstractions, or topics, of documents. Since we would like to use different techniques, we accept any number of annotation files, each of which can be generated by different tools. Each annotation file typically contains a number of 'schema:about' statements about one or more of the documents, each associating a document with one of the 'topics' covered in the document. We now describe the mechanisms used for generating topics for the US Presidency corpus.

Since the only raw features that are available are typically words, clusters of words, either weighted or unweighted, are the most basic way of modeling topics. Each topic can be characterized by the mention of one or (preferably) more of a set of concepts, each of which in turn can be referred to by a set of alternative words phrases. Our goal is to bootstrap to a comprehensive set of words/phrases for each topic, with as little manual work as possible. We start with a very simple keyword cluster mechanism and use a variety of techniques to enhance these clusters.

In this work, we used a taxonomy of topics from earlier work [5].We began with a manually generated, small (4-8) keywords (positive and negative) for each topic. We extended each of these sets of keywords using Word co-occurrence based similarity, Word embeddings [8] and Topic Modeling [3].

We start with a small manually generated list of keywords for each of the topics. We augment these initial keywords with additional keywords by using the techniques listed above. We run the final set of keywords against the corpus to generate a list of topics that each document covers. This data is output into an annotations file, which together with a file containing the metadata (speakers, speech type, date) is consumed by the search engine run time.

3.3 Exploration interface

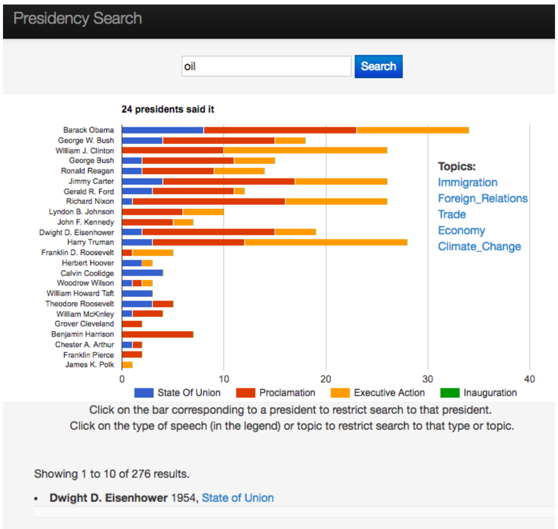
Doris is a hybrid between tools such as traditional search, which is aimed at enabling the user to find the most relevant result, and Google's Ngram viewer [6], which is primarily focused on distributions of terms. We extend both the charting and search capabilities of those systems with the addition of semantic categories and metadata. In this section, we describe some of the features of this tool and illustrate them with screen shots.

The user starts with a query, which could be a single word or a set of words. The results page, for a simple query 'oil', is shown in Fig 1. The bottom half of the page contains the first 10 results and is similar to a traditional search. The top of the page contains an interactive plot of the distribution of results. The results are collated by President, plotted on the Y-axis (or optionally X-axis), in temporal order. Each bar is split into sections for the different kinds of speeches. On the right/bottom, we show the top 5 topics that appear in these results.

Clicking on the legend on one kind of speech restricts the search to that kind of speech. Clicking on a topic restricts the search to that topic. Clicking on the bar corresponding to a president restricts attention to the speeches of that president. We can also aggregate a single president's speeches by year. The set of search results, which is in the bottom half of the page, is also kept updated through this exploration. In a different view of the interface (see figure 2), when the search is restricted to a topic node that has children, the bars for each president correspond to the subtopics of that topic node.

4 Analysis / Discussion

While it is possible to evaluate a tool, such as the one presented here, using traditional information retrieval metrics such as precision and recall, simply performing as well (or even slightly better) than current search tools would not justify such an effort. The



Click on the bar corresponding to a president to restrict search to that president.
Click on the type of speech (in the legend) or topic to restrict search to that type or topic.

Showing 1 to 10 of 276 results.

- **Dwight D. Eisenhower** 1954, State of Union

Fig. 1. Results page for the query 'oil'. The chart at the top gives the distribution of results, clustered by presidents, who are sorted in temporal order. The bar for each president is composed of segments corresponding to the type of document. On the right are the most frequently mentioned topics in the documents retrieved. The results can be restricted by clicking on the type of speech (in the labels), the president or the topic.

primary goal of this tool is to make it easier for students and researchers to gain insights from large document corpora. In that spirit, we discuss some patterns that are apparent through our tool that are not obvious in traditional systems. Given the huge increase in the number of Proclamations and Executive orders recently, in order to normalize for comparisons, unless otherwise mentioned, we restrict our attention here to State of the Union addresses.

Figure 1: We can see the frequency of the word 'oil' increasing over time, starting with Harry Truman. Furthermore, 'oil' appears mainly in Proclamations and Executive Actions and less in State of the Unions. Isolating oil and foreign relations together, in figure 5, shows peaks in Truman, Carter, and Obama.

Figure 2: While 'Economy' is currently one of the biggest issues and is discussed frequently in State of the Unions, it was not as prominent with early presidents. The 'Economy' started to gain prominence around the Great Depression, more specifically post Woodrow Wilson. We can see that in the early days of the United States, discussions of the 'Economy' was mainly focused on trade. Since Franklin D. Roosevelt's presidency, the focus of the economy has shifted to jobs and employment. Both the rising prominence of the 'Economy' and transition from 'Trade Relations' to 'Jobs and Employment' are seen in figure 2. These trends are not as evident in traditional search engine interfaces.

These examples illustrate the core strength of an interface that combines the ability to search across both words and topics, presenting them in an interactive graphical form. Semantic features (such as the topic 'Economy') are vital for performing this kind of analysis. This tool enables researchers to compare the evolution and relevance of certain topics such as how 'Economy' has evolved compared to 'Foreign Relations'.

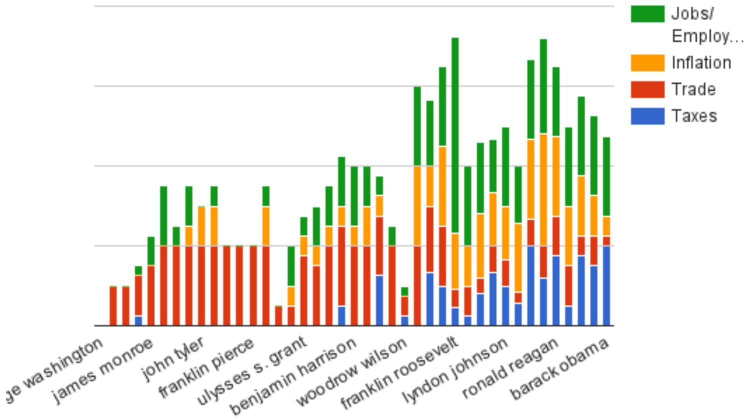


Fig. 2. Results for the topic 'Economy' in State of the Union addresses. Since the topic 'Economy' has subtopics, the bar for each president is broken down by subtopics.

Acknowledgements

I would like to thank Dr. Christy Story for her guidance and mentorship through this project. I would also like to thank Tasha Bergson-Michelson, Lillian Hoodes, Dave Lowell, Kyle Barriger and Ann Greyson.

References

1. Penn engineers network analysis uncovers new evidence of collaboration in shakespeare plays. *news.upenn.edu/news/penn-engineers-network-analysis-uncovers-new-evidence-collaboration-shakespeare-s-plays*.
2. O. Bergman, R. Beyth-Marom, R. Nachmias, N. Gradovitch, and S. Whittaker. Improved search engines and navigation preference in personal information management. *ACM Transactions on Information Systems (TOIS)*, 26(4):20, 2008.
3. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning*, 3(Jan):993–1022, 2003.
4. E. Freeman and D. Gelernter. Lifestreams: A storage model for personal data. *ACM SIGMOD Record*, 25(1):80–86, 1996.
5. M. Gupta and S. Guha. Topic based analysis of text corpora. *Computer Science & Information Technology*, page 33, 2016.
6. Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174. ACL, 2012.
7. J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.
8. J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.
9. G. Peters and J. T. Woolley. The american presidency project. *presidency.ucsb.edu/ws*, 2011.
10. B. Schmidt. State of the union in context. *benschmidt.org/poli/2015-SOTU*, 2016.