TOPIC BASED ANALYSIS OF TEXT CORPORA

Madhumita Gupta[1] and Sreya Guha[2]

[1]Palo Alto, USA
madhumita@gmail.com
[2]Palo Alto, USA
sreyaguha@gmail.com

ABSTRACT

We present a framework that combines machine learnt classifiers and taxonomies of topics to enable a more conceptual analysis of a corpus than can be accomplished using Vector Space Models and Latent Dirichlet Allocation based topic models which represent documents purely in terms of words. Given a corpus and a taxonomy of topics, we learn a classifier per topic and annotate each document with the topics covered by it. The distribution of topics in the corpus can then be visualized as a function of the attributes of the documents. We apply this framework to the US State of the Union and presidential election speeches to observe how topics such as jobs and employment have evolved from being relatively unimportant to being the most discussed topic. We show that our framework is better than Vector Space Models and an Latent Dirichlet Allocation based topic model for performing certain kinds of analysis.

KEYWORDS

Text Analysis, Machine Learning, Classification

1. INTRODUCTION

In recent years, researchers in fields as diverse as biology, law and the social sciences have started using computational models to analyze corpora of scientific papers, judgments, speeches, etc. These models have enabled researchers to discern patterns and gain insights into their respective fields. For example, such models have been used to corroborate the authorship [6] of several of the Federalist Papers, a collection of 85 essays that promoted the ratification of the constitution of the United States, written by Alexander Hamilton, James Madison, and John Jay, under the pseudonym Publius. The computational analysis of these essays has led researchers to dispute the authorship of 11 of them.

Computational models of text corpora aim to find patterns that capture underlying phenomena in the domain discussed by the documents. A prominent feature of the approaches that are used today is that they are bottom up. As pioneered by Salton et. al. [15] in the Information Retrieval community, the representations, learnt concepts, etc. are all built up purely from the words in the document. As these approaches are driven only by the words in the document, they can be applied to a new corpus without any manual pre-processing { a huge advantage. This generality,

Jan Zizka et al. (Eds) : ACSTY, NATP - 2016
pp. 33– 47, 2016. © CS & IT-CSCP 2016 DOI : 10.5121/csit.2016.61403

however, can also be a shortcoming. 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 (a State of the Union speech is an annual speech given by the President of the United States). In the early years of the US, Native American relations were talked about more frequently than they are today. If we wanted to understand how this is reflected in these speeches, or when this change started, we would need a model that provides "Native Americans" as a model feature, which a pure word based model would not.

Every field has concepts/topics that are central to the discourse in that field and many important questions are in the vocabulary of these concepts/topics (henceforth, referred to as topic). The above question, for example, needs to reference the concept/topic "Native American". While the words in the document do capture the topic(s) discussed, it is difficult to express certain questions only in terms of the words in the document. This is especially the case when:

- Different combinations of words can be used to refer to the same topic. E.g., Sometimes, Native American relations are discussed within the context of certain tribes, requiring the model to be able to recognize the topic from the tribe names.

- Terminology evolves over time. Over the years, Native Americans have been referred to by other names, such as "Indians", a term which is now used to refer to a different community.

- Different participants use different words to talk about the same topic. For example, for the topic "Abortion", different parties use different terms. Democratic Party speakers tend to use the phrases such as "access to contraceptives," and, "womans right to choose," whereas Republican Party speakers tend to use "rights of the unborn," and, "right to life," and so on.

In this paper, we propose a framework for addressing the mismatch between the queries researchers want to ask and the vocabulary of the modeling tools. We are given a corpus of documents, each with a set of attributes, and a taxonomy of topics. Assuming that a single paragraph in a document is restricted to one topic, we build a set of classifiers, one for each topic such that each paragraph is labeled as belonging to one or zero topics. We use our framework to analyze the distribution of topics discussed as a function of the attributes of the documents. We apply our framework to two corpora:

1. The State of the Union Speeches, from 1790 to 2016 (226 speeches),
2. Speeches from the primary and general elections in the presidential elections from the years 1996 to 2016 (17,718 speeches)

Each speech has the date of the speech, a title and the speaker. These two corpora are part of a larger corpus of 19,572 political speeches obtained from [20], [16] and [4]. We created a simple taxonomy of 26 topics corresponding to the most popular issues in political discourse, such as the economy, human rights, etc. (see figure 1 for the full taxonomy). Given the nature of political speeches and the significant time frame over which these speeches were made, there is a wide variation in the words used to discuss a given issue, both across speakers and over time. Though the list of topics in our taxonomy is by no means exhaustive, we believe that it is adequate in size and variety to demonstrate the benefits of our approach.

2. METHODOLOGY

Our goal is to develop a framework with which we can better understand the distribution of a set topics across the documents in a given corpus, as a function of the attributes of the documents. We develop the framework and apply it to the speeches in two corpora mentioned earlier. We analyze the distribution of these topics across these speeches as a function of the speaker, speaker affiliation, year, etc. We do this using not only our proposed framework but also using two popular techniques for quantitative analysis of text corpora. We now discuss these two techniques before describing the details of our framework.

2.1. Existing Models

The models described below are currently the two most widely used models:

Bag of Words Model: Each document is modeled as bag of words, where a "Term Frequency Inverse Document Frequency" (TFIDF) score is computed for each word document pair. This measures the number of times the word occurs in the document, normalized by the frequency of occurrence of the word in the corpus as a whole. The most significant terms in a document can be said to capture the main points in the document. The most widely used application of this model is document search (e.g. Web search).

Latent Semantic Models: Starting with the work by [2], a variety of "latent semantic" models have tried to create more abstract, implicit representations of meaning, which capture the fact that different combinations of words can be used to express the same concept. Latent Semantic Indexing has been used in a variety of applications such as patent discovery, document classification and determining authorship of documents. However, the resulting implicit representations are hard to interpret and words with multiple, evolving meanings cannot be disambiguated easily. Recently, there has been work in machine learning, under the term "Topic Modeling" that computes a generative model for a corpus of documents. Documents are assumed to be generated, according to some distribution (typically, a Dirichlet distribution) from a set of latent topics. The goal is to generate these topics, which should provide insight into what the corpus is about. Though latent semantic approaches such as topic models do try to go beyond words, since their representation of semantics is "latent", it is hard, if not impossible to express questions about a particular topic that is of interest to us.

2.2. Classification

Each speech was divided into several "documents", i.e. paragraphs. Short paragraphs with fewer than 50 words were merged to create bigger paragraphs. A single paragraph is henceforth the equivalent of a document for our classification purposes. Each paragraph may discuss one of our given topics. Note that there will be many paragraphs that don't discuss any of our given topics. One of our key technical problems is to identify when a paragraph is discussing one of the given topics. We do this by using machine learning to build automated classifiers for each topic.

In this work, we build classifiers only for the leaf nodes in our taxonomy. The interior nodes are assumed to be unions of their child nodes. The framework we develop can easily be extended to taxonomies that don't make this assumption. We build our classifiers against the combined

corpus, i.e., there is a single classifier for a topic such as 'Immigration', not one for each of our corpora.

2.2.1. Limitations of Pure Keyword Matching

One very simple way of matching paragraphs with topics is by keyword matching. For example, a topic such as "Jobs/Employment" could be associated with a keyword 'jobs' and every paragraph that contains the word 'jobs' (or a stemmed version of 'jobs') can be associated with this topic. Unfortunately, this method has significant limitations. The same phrase can be used in different contexts with different meanings. For example, the phrase "right to choose" is often used to discuss "Abortion". However this phrase is also used in many political speeches on democracy, to discuss the right of the Vietnamese or the Iranians to choose their own government. Similarly, using jobs as an indicator of a document that matches the topic of Jobs/Employment leads to large numbers of spurious matches (let them do their job, the book of Job, and so on).

Table 1: Comparison of terms used to discuss Nuclear Weapons in 1950's vs in 2010's

Nuclear Weapons in 1950's	atom, missile, communist, soviet, bomb
Nuclear Weapons in 2010's	korea, israel, sanction, iran, deal

Vocabulary changes over time, and the manner in which words are used changes over time. For example, the term "Drone" in today's context typically refers to remote piloted aerial vehicles. However, it also appears in speeches by George Washington to refer to something quite different. A striking example of the changing vocabulary is shown in table 1, which illustrates the change in some of the terms associated with Nuclear Weapons.

Support Vector Machines [9] and other machine learned classifiers, on the other hand, don't rely on simple keyword matches. By using more complex functions that combine partial support for a given topic from different words in the document, they get around some of the limitations of simple keyword matches.

2.2.2. Creating Training Data

In order to learn a classifier, we need a set of labeled examples. We use the following procedure to create a training set.

1. For each topic, we manually specify a set of phrases that, with high probability, identify paragraphs about that topic (Table 2).

2. For each phrase, we extract the paragraphs in the corpus containing that phrase, giving us a set of paragraphs for each topic. For some topics, for example, Drugs, we had as few as 300 matches in all. For others, such as Jobs/Employment, we had 13,000 matches.

3. We manually check a small sample of the paragraphs to check whether the paragraph corresponds to the topic. The correct ones go into the positive training set and the (small number of) wrong ones go into the negative training set.

4. We extract top K ($\approx$ 10) words/phrases in these positive training set that have a high TFIDF. These are added to the list of phrases for the issue.

5. We extract additional matches using these new phrases and repeat the labelling into positive and negative training sets.

6. The negative training set for each issue is augmented with samples from the positive training set of other issues.

For each training set, we took positive and negative examples in the ratio of 1:4, i.e., about 250-500 positive examples, and 1000-2000 negative examples.

2.2.3. Pre-processing Data

Each paragraph is pre-processed using the statistical package R [14] as follows.

1. The text is lower-cased, punctuation and whitespace removed.

2. All numbers and stop words are removed (words such as a, is, after, before, etc. that are very frequent, but do not provide any information).

3. The words are stemmed (reduced to their base form).

4. Finally, each paragraph is broken into a set of "features" - the single word phrases (also known as unigrams) it is composed of.

These paragraphs and features are then composed into a "Document Term Matrix" (dtm): a matrix whose rows are the paragraphs and columns are the union of all features across all paragraphs. The values are the TFIDF scores of the features:

Table 2: Sample of phrases used to create training set.

Topic	**Keywords / Phrases**
Jobs/Employment	middle class, GDP, unemployment, recession, job creation, jobless, great depression, economic recovery, minimum wage
Immigration	Anchor babies, illegal immigration, H1B, deportation, border guard
Trade	tariff, laissez faire, nafta, free trade, tpp, import duty
Taxes	income tax, death tax, redistribution, 1 percent, trickle down, tax cut
Healthcare	medicare, medicaid, obamacare, medical insurance, public option
Social Security	retirement age, safety net, social security fraud, government handout
Nuclear Weapons	thermonuclear, ballistic missiles, arms limitation, mutually assured destruction, north korea, kim jong un, nuclear disarmament, plutonium
Climate Change	rising temperatures, global warming, fracking, carbon, Kyoto, Paris Climate Agreement, Fossil Fuel, sea level rise, glacier, greenhouse
Race Relations	separate but equal, racism, Brown v. Board of Education, Martin Luther King, voting rights act, desegregation, negro, emancipation, slavery
Drugs	just say no, war on drugs, heroin, cocaine, drug rehabilitation, opioid, overdose
Terror	Al Qaeda, Taliban, bin laden, ISIS, drones, 911, twin towers, world trade center, benghazi, Tora Bora, Hizballah
Inflation	bimetallism, gold standard, de ation
Native American	indian, apache, cherokee, western settlements, indian affairs, cheyennes

$$\text{dtm}[i,j] = (\text{the number of times feature } j \text{ occurs in paragraph } i) \times \log_2\left(\frac{\text{total number of paragraphs}}{1 + \text{total number paragraphs in which feature } j \text{ occurs}}\right)$$

2.2.4. Tuning the pre-processing

There are many parameters that can be tuned during the pre-processing, each of which can affect the classification. Some of these include:

- **Stop Words:** TFIDF based scoring alone is not enough to eliminate the effect of very frequent terms. The programming package we used (tm for NLP) removes a standard set of stop words such as 'the', 'and', etc. However, there are many other words such as "will", "campaign", "political", which occur very frequently in this corpus, which do not provide any significant information. The TF-IDF measure for scoring features is intended to reduce the importance of exactly such words. However, they are so frequent within the documents, that merely counting the documents that they occur in does not have a significant enough impact on their score, and they are not dropped from the feature matrix, and sometimes end up as or more important than high information words. As an example, in a set of 7980 documents used to train the concept of Abortion, "will" occurred 4551 times in 2454 documents. "Partial birth" on the other hand occurred 18 times in 17 documents. The resulting IDF makes "partial birth" 9 times more significant as compared to "will", but the high TF of "will" counteracts IDF, and "partial birth" and "will" end up with similar TFIDF scores.

- **Stemming Variants** of a word (legislation, legislate, etc.) should usually be treated similarly for purposes of classification. We therefore 'stem' each word occurrence to a root and treat them all similarly. However, occasionally, stemming can change its meaning. For example, for the topic "Jobs/Employment", for the training set, we looked for documents containing "recession". The stemmed form of "recession" is "recess", and we had many spurious matches such as "Congress is in recess". However when we trained without stemming, the results were slightly worse - e.g. for Safety the F-measure went from 92.85% to 91.11%, for Social and Health from 76.45% to 74.33% while for Human Rights it went from 85.78% to 87.41%. So, despite losses such as "recession", we continued using stemming.

- **Unigrams, bigrams, etc.** In addition to unigrams, we experimented with bigrams and trigrams (two word and three word phrases). Example bigrams and trigrams are "Star Wars" and "National Reconstruction Act". If we use only bigrams and trigrams, our specificity increases to a point where we are left with too few matches, and if we use both unigrams and bigrams, the number of unigrams swamp out the bigrams. So, we used only unigrams.

- **Imbalanced Data** Our training data set had fewer positive examples for each topic as compared to negative examples. As a result, the initial classifiers labeled all documents as negative (not matching the topic). We solved this problem by assigning weights to the positive and negative examples, in inverse proportion to the number of examples of that type.

2.2.5. Learning Algorithms

With the Document Term Matrix as input, we used three different algorithms to learn classifiers for each topic: logistic regression [10], Classification and Regression Trees (CART)[1], and Support Vector Machines (SVM) [9]. For each algorithm, we measured its precision on a hold back from our training set. Of the three, Support Vector Machines with a radial basis kernel [18] were the most accurate.

3. EXPERIMENTAL RESULTS

We had a total of 19572 political speeches, made by 245 speakers. Our taxonomy (Fig. 1) has 26 topics with an average of 11 phrases per topic for creating the initial training set. We used the tm (Natural Language Processing, [5]), XML ([11]) and plyr ([19]) packages of the R [14] programming language to preprocess the data and then to convert it into the document term matrix. For classification, we used the randomForest ([12]) and rpart ([17]) packages for building the CART classifiers, and the e1071 package ([3]) for building Support Vector Machine (SVM) classifiers with linear and radial basis kernels. We obtained the best results with SVM with a Radial Basis kernel.

As with all classifiers, we have to trade off precision (the fraction of sentences classified as belonging to an issue that are indeed of that issue) with recall (the fraction of all sentences of an issue that are identified as being of that issue). We used the F-measure, the harmonic mean of precision and recall to estimate the performance. The precision and recall numbers used in the tuned classifiers, for a sample of topics, are given in table.

Topic	Precision	Recall	F-measure
Safety	93.42%	92.28%	92.85%
Social & Health	81.87%	71.71%	76.45%
Human Rights	83.07%	88.68 %	85.78%
Economy	64.0%	66.36%	65.16%
Defense & Foreign Relations	74.72%	72.15%	73.41%

To compare our framework with word based techniques, we also analysed our corpora both by modeling each document as a Bag of Words and by using Topic Modeling to identify a set of topics (that the corpora are 'about'.) The words with the highest TFIDF from the speeches of a selected set of speakers in the two corpora are given in tables 3 and 4.

We used two different Topic Modeling tools ([13] and [8]) to generate topics for our combined corpus (remember that build our classifiers against the combined corpus, so that there is a single classifier for a topic such as 'Immigration' that works for all speeches). The results generated by [13] were better and are given in table 5. As we can see, the automatically generated topics are not easily understandable. In particular if we are trying to answer questions such as "how did discussion about `Native Americans' in the State of the Union speeches evolve", since none of the topics corresponds to `Native Americans', it is difficult to frame the question in terms of the automatically generated topics of table 5.

Table 3: Words with highest TFIDF used by Presidents in State of Union Speeches

George Washington	indian, treaty, militia, session, tribes, rendered, hostile, cherokee, frontier, tranquility, deliberations, orins, insurrection, expedient, attaching, observed
Andrew Jackson	treaty, treasury, indian, session, deposits, france, exercise, minister, vessels, branches, deemed, payment, portion, intercourse, rendered, ports, possessions
Abraham Lincoln	slavery, territorial, nebraska, missouri, compromise, emancipation, negroes, clay, repealing, prohibition, framed, douglas, ordinance, rebellion, sacr
Theodore Roosevelt	panama, interstate, island, navy, forests, canal, wageworker, republic, isthmus, railroad, exercise, treaty, philippine, territorial, supervised, tariff
Franklin D. Roosevelt	nazis, germany, axis, planes, pacific, farmers, japan, sea, material, british, hitlerism, italy, britain, island, agriculture, recovery, continent
Ronald Reagan	soviet, in ation, missile, lebanon, strategic, nicaragua, laughter, gorbachev, grenada, revolution, space, treaty, sdi, israel, rgeneva, sandinistas, totalitarian
Bill Clinton	medicare, police, bridges, bosnia, guns, somalia, russia, bipartisan, somalis, brady, laughter, kosovo, covenants, nato, cold, doctorate, black, teaches

Table 4: Words with highest TFIDF used in Presidential primary speeches

Ted Cruz	pastor, churches, houston, baptist, trump, cochair, rubio, activist, religious, christian, rep, coalition, polk, tea, texas, islamic, marriage, ministries
Bernie Sanders	billionaires, vermont, climate, burlington, saturday, superpac, sunday, isis, inequalities, donald, friday, turnout, weaver, mondays, minimum
Rick Santorum	verona, obamacare, romneycare, marriage, rpa, gingrich, abortions, hogan, mandates, healthcare, bailouts, newt, radical, contrast, repealing, tea
Hillary Clinton	Activist, mortgages, nevada, foreclosure, coverage, manchester, rural, nurse, guns, des moines, lgbt, latino, healthcare, hispanic, longterm, treatment, green
Donald Trump	illegal, grafton, georgia, merrimack, hillsborough, cheshire, israel, falwell, tremendous, lewandowski, cochair, belknap, palin, patrol, ballot
John McCain	Palin, acorn, usn, biden, admiration, anncr, arlington, ayers, coal, surge, drilling, afghanistan, blogs, usaf, withdrawal, iraqi, abc, mortgages
Mitt Romney	medicare, michigan, Ryan, obamacare, illegal, huckabees, recovery, journal, bain, editorial, marriage, fox, giuliani, newt, biden, veto, ret, solyndra

4. DISCUSSION

One of the primary motivations in creating this framework is to enable us to analyze how the discourse (in a given corpus) has evolved with respect to a given set of topics. This kind of analysis often facilitates better understanding of the underlying phenomena.

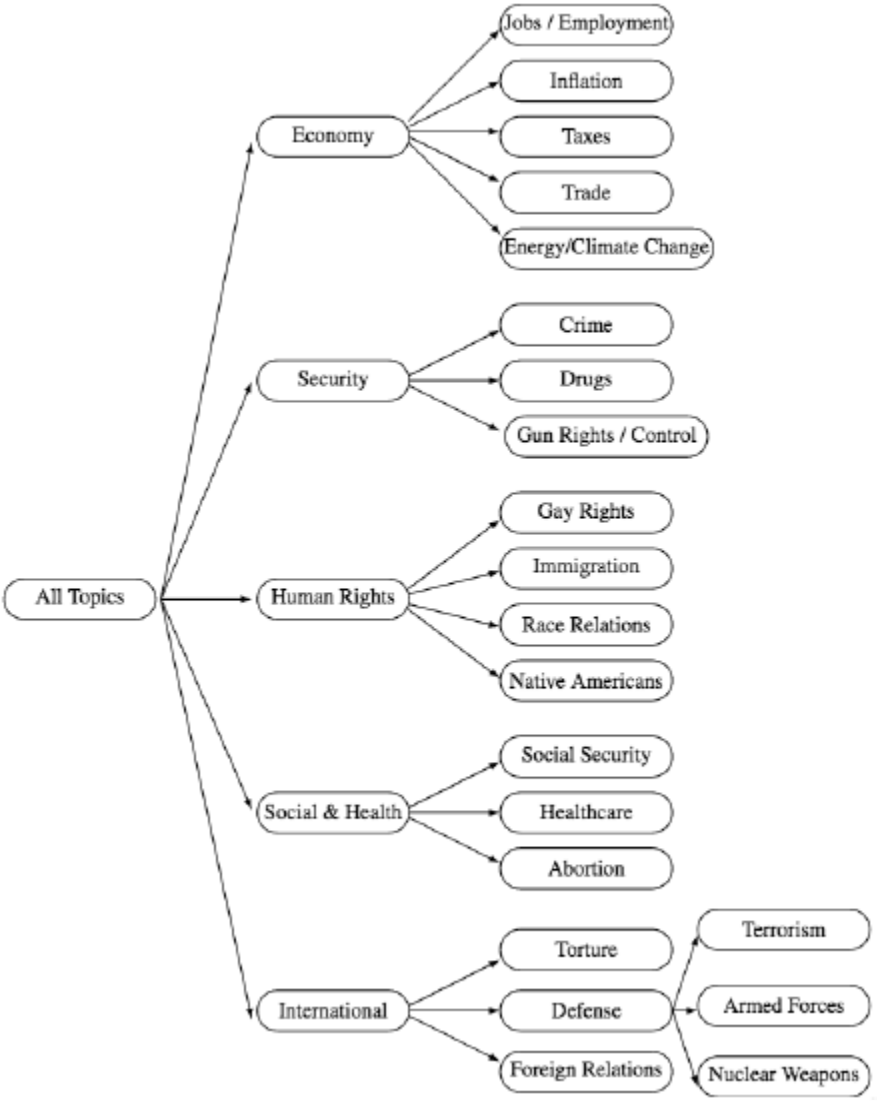

Figure 1: Topic Hierarchy

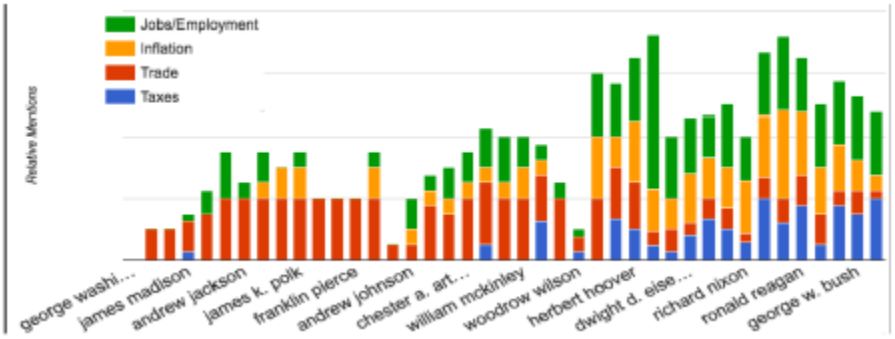

Fig 2 : Relative mentions of the 'Economy' topic in State of Union Speeches

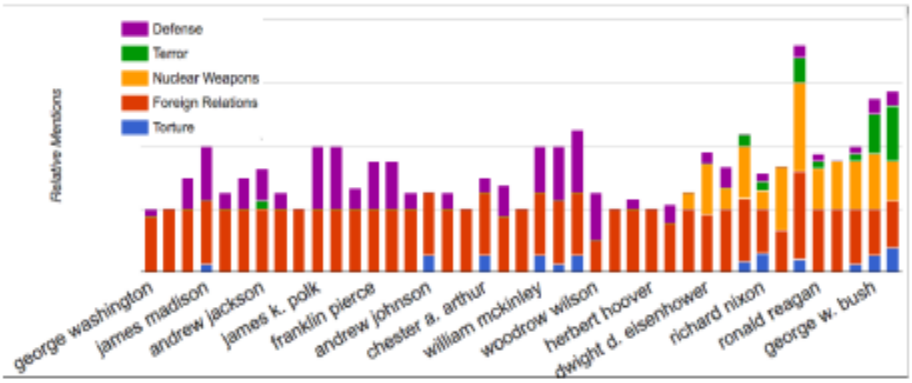

Fig 3: Relative mentions of the 'International' topic in State of Union Speeches

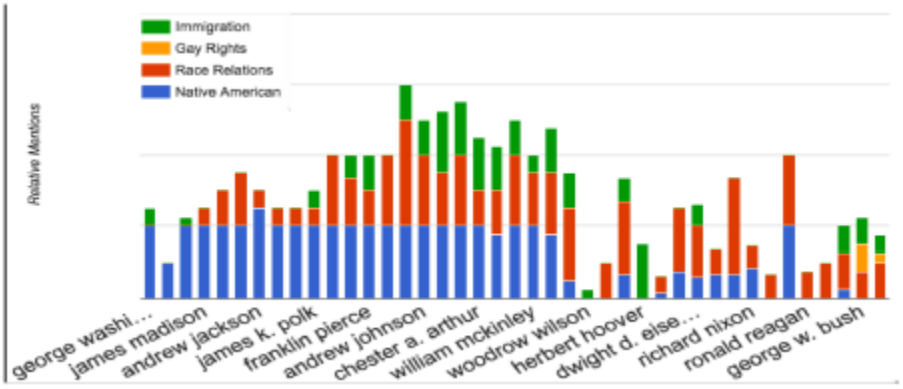

Fig 4: Relative mentions of the 'Human Rights' topic in State of Union Speeches

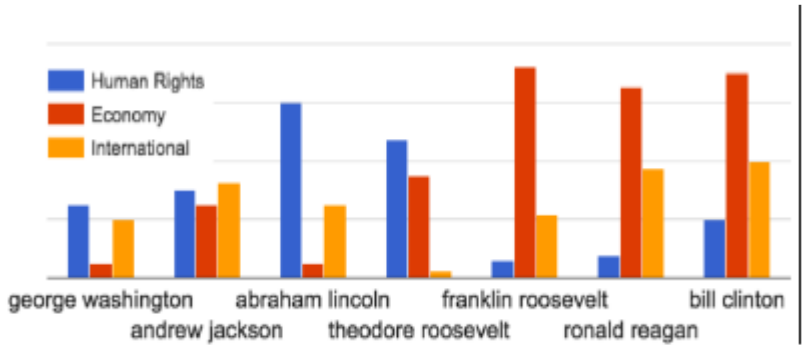

Fig 5: Change in relative coverage of topics over time in State of Union Speeches

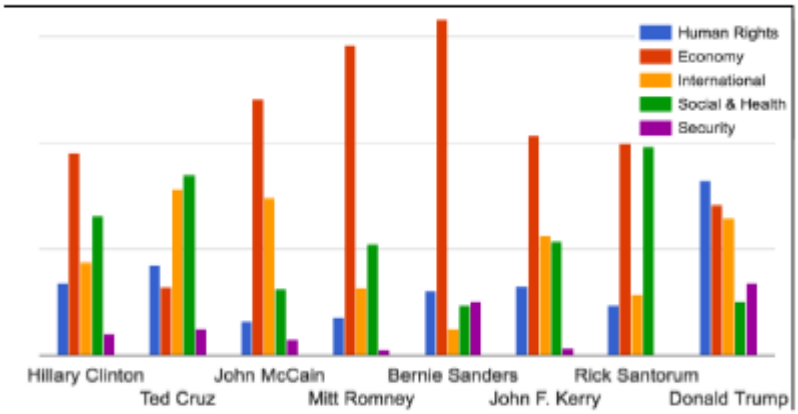

Fig 6: Relative coverage of topics in primary election speeches

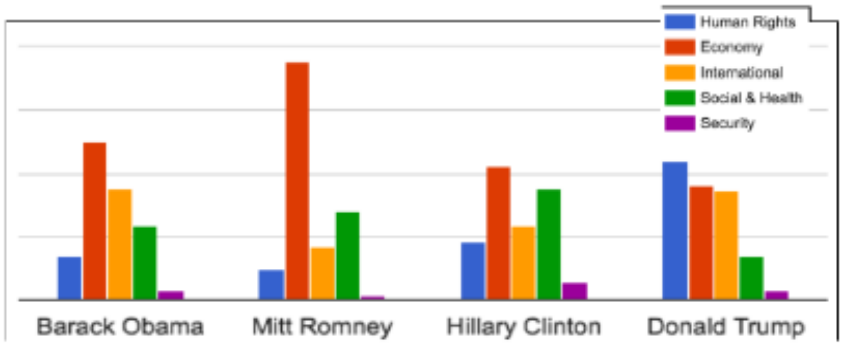

Fig 7: Relative coverage of topics in general election speeches

Table 5: Autogenerated topics using LDA

Topic 1	government, american, country, america, united, promise, times, states, new
Topic 2	santorum, state, rick, government, vote, united, romney, president, life
Topic 3	tax, plans, jobs, romney, percent, federal, energy, taxes, American
Topic 4	government, states, united, congress, year, 000, law, people, public
Topic 5	santorum, great, wall, shall, country, present, people, government, time
Topic 6	santorum, congress, public, life, campaign, war, rick, romney, said
Topic 7	public, government, congress, country, law, 000, people, shall, American
Topic 8	perry, romney, rick, gov, president, state, santorum, said, America
Topic 9	world, people, america, life, government, great, country, peace, time
Topic 10	search, romney, public, health, year, massachusetts, care, choose, month

The output of the analysis is a database of paragraphs with annotations for topic, speaker, date, context (i.e., primary election, general election or State of Union). We use the graphing functions provided by Google Sheets [7] to help us detect patterns in the data. In this section, we present some insights/observations as illustrated by the relevant graphs. We also use the bag of words model to generate a list of the words with the highest TFIDF scores and compare the two approaches from the perspective of their ability to enable us to make these observations.

4.3. Evolution of Topics

Figures 2, 3 and 4 plot the number of mentions of the topics `Economy', `International' and `Human Rights', in the state of union address by each president. Since the number of state of union speeches given by a particular president varies from 1 to 13, we normalize the number of mentions by the number of speeches given by that president. Looking at these graphs, we make the following observations

- Economy is the biggest issue nowadays, but this was not always the case. Economy as the theme of the presidency seems to started at around the time of the Great Depression. Further looking at figure 2 we see the mixture of topics related to the Economy has changed. In the earlier days of the Union, the Economy related discussions were more around the topic of `Trade'. The topic of 'Jobs / Employment' has become an increasingly important part of the discussion in the last 80 years. Taxes start coming up repeatedly only around 1900, when income tax was first introduced.

- In figure 3, we see that Foreign Relations and Defense are a constant theme, but we see the recent emergence of subtopics such as Terror.

- Looking at figure 4, we see a recent fall in discussions about Native Americans. The issue of Gay Rights is relatively recent and has made its appearance in State of Union speeches only in the last few presidencies. It is interesting to see that race relations have always been a topic in these speeches. Indeed, comparing figures 3 and 4, we see that Foreign Relations and Race Relations are the only two topics that have figured prominently in over 90% of the speeches.

- We see that different topics have very different temporal behaviors. Some (such as "Race Relations") have maintained a relatively steady presence through the history of these speeches. Others, such as "Native American" have remained steady for a period of time and then declined. Some topics such as the "Jobs/Employment", which are now central, took a long time to develop. Some other issues, such as "Nuclear Weapons" sprung into prominence relatively fast.

4.4. Relative Coverage of Topics

The relative coverage of topics gives us insight into the changing priorities of a community.

Graph 5 shows the relative coverage of topics across seven presidents drawn from different time periods. As we can see, the biggest issues in the early years of the country were "Human Rights" (actually, "Native American" issues) and "Foreign Relations", and now it is "Economy". Discussions about the topic "Foreign relations" have stayed relatively the same throughout time; however, coverage of the topic "Jobs/Employment", and of the topic "Economy" overall have significantly increased since the Great Depression. Comparing graph 5 to the TFIDF table 3 for the same presidents, we see that the analysis produced by our framework is substantially better for answering these kinds of questions.

Graph 6 shows the relative coverage of topics for a number of recent presidential candidates. We see that "Economy" is the most important topic for most, but not all of them. The graph clearly

shows which issues were most important to each candidate. Graph 7 shows the relative coverage of topics during the general election. Comparing graphs 6 and 7, we can see that the coverage of topics during general elections is more even than during the primaries.

4.5. Word Based Analysis

As can be seen from table 3, the words used during different time periods have evolved significantly. It is interesting to see how well the words with the highest TFIDF capture the major issues of each presidency. For example, the most significant words for Lincoln include 'slavery', 'rebellion' and 'emancipation' and the most significant words for Franklin Roosevelt include 'nazis', 'japan' and 'recovery'. The significant words from the primary speeches on the other hand do not exhibit such clarity. Despite the words in table 3 accurately reflecting the major topics of each presidency, we can see that would be difficult to draw conclusions such as those enabled by figures 2, 3, 4 from these word lists.

5. CONCLUSIONS AND FUTURE WORK

We introduced a framework for analysing the documents in a corpus, from the perspective of a given taxonomy of topics. The framework facilitates annotation each document with the topics covered and subsequent analysis of the distribution of topics as a function of the attributes of the documents. We implemented this framework and applied it to to two corpora of political speeches. We showed how the framework enables us to draw insights into into the relative composition of speeches and to the evolution of topics.

There are many different areas of future work. Looking at the topic hierarchy (fig. 1), we can see that it is hard to organize topics into a clean tree. Some topics, like "Climate Change", legitimately belong in multiple higher level categories | "Economy", "International", etc. We would like to extend our approach to taxonomies that are allow for multiple parents for each topic. We have assumed that the topic hierarchy is relatively static. However, just as the terminology used to discuss a topic evolves, the topic hierarchy itself evolves. E.g., In the early years, "Native American" issues were probably more part of "Defense" than "Human Rights". Over time, they have become more part of "Human Rights". Handling evolving taxonomies is another direction for future work.

The biggest limitation of our approach is the cost of building classifiers. A taxonomy with thousands of topics could be very expensive to build classifiers for. As we can see from table 3, the high TFIDF words do sometimes capture the main points of a document/ speaker. Though words by themselves are inadequate for capturing more abstract concepts like `Race Relations' they are good at capturing more specific topics like `emanicipation proclamation'. One line of future work involves combining the two approaches, wherein for the more general topics, we use machine learnt classifiers, but as the topics get more specific, a combination of topic modeling and word vector based approaches can be used to fill out the taxonomy.

REFERENCES

[1] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. Classification and regression trees. CRC press, 1984.

[2] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. Journal of the American society for information science, 41(6):391, 1990.

[3] E. Dimitriadou, K. Hornik, F. Leisch, D. Meyer, A. Weingessel, and M. F. Leisch. Package e1071. R Software package, avaliable at http://cran. rproject. org/web/packages/e1071/index. html, 2009.

[4] M. E. Eidenmuller. American Rhetoric: The Power of Oratory in the United States. http://www.americanrhetoric.com, 2001-2016. [Online; accessed 7-July-2016].

[5] I. Feinerer and K. Hornik. Text mining package, 2015.

[6] G. Fung. The disputed federalist papers: Svm feature selection via concave minimization. In Proceedings of the 2003 Conference on Diversity in Computing, pages 42{46. ACM, 2003.

[7] Google.com. Google Sheets : Create and edit spreadsheets online. https://www.google.com/sheets, 2016.

[8] B. Grun and K. Hornik. topicmodels: An R package for fitting topic models. Journal of Statistical Software, 40(13):1{30, 2011.

[9] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf. Support vector machines. IEEE Intelligent Systems and their Applications, 13(4):18{28, 1998.

[10] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. Applied logistic regression, volume 398. John Wiley & Sons, 2013.

[11] D. T. Lang. Xml: Tools for parsing and generating xml within r and s-plus. R package version, pages 3{9, 2012.

[12] A. Liaw and M. Wiener. Classification and regression by randomforest. R News, 2(3):18{22, 2002.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R.Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825{2830, 2011.

[14] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2016.

[15] G. Salton and M. J. McGill. Introduction to modern information retrieval. McGraw-Hill, Inc., 1986.

[16] The Miller Center. American President: A Reference Resource. http://millercenter.org/president, 2016. [Online; accessed 7-July-2016].

[17] T. M. Therneau, B. Atkinson, and B. Ripley. rpart: Recursive partitioning. r packageversion 3.1{42. Computer software program retrieved from http://CRAN. R-project.org/package= rpart, 2010.

[18] J.-P. Vert, K. Tsuda, and B. Sch olkopf. A primer on kernel methods. Kernel Methods in Computational Biology, pages 35{70, 2004.

[19] H. Wickham. The split-apply-combine strategy for data analysis. Journal of Statistical Software, 40(1):1{29, 2011.

[20] J. Woolley and G. Peters. The American Presidency Project. http://www.presidency.ucsb.edu/index.php, 2009-2016. [Online; accessed 7-July-2016].