Gavin Wiehl

Gtw4vx

DS5001

**An exploratory text analysis of Presidential speeches**

Introduction

For this project, I have chosen to analyze the text of president speeches, from George Washington to Donald Trump. The dataset comes from Kaggle user Kristof Boghe [1], who in turn got the data from University of Virginia’s Miller Center. Along with the President and their speeches, the dataset includes metadata such as their years in office, party, date of the speech, title of the speech, and basic summary information about the speech.

The goal of this project is to gain some understanding of how a certain president’s speech relates to other presidents, as well as to the world around them at their place and time. Are there some presidents who sound similar in their speeches? Are there certain topics that come up most frequently? Which word choices stick out among the speech corpus? These are among the questions I would be interested in answering.

Outline of the data

The dataset includes 994 speeches from all 45 presidents. Since some of these speeches are debate transcripts, which will be difficult to parse through since different people will be speaking throughout the speech, I remove all the debates from the dataset. This leaves 981 speeches to analyze. Figure 1 shows the distribution of speeches for each president present in the dataset.

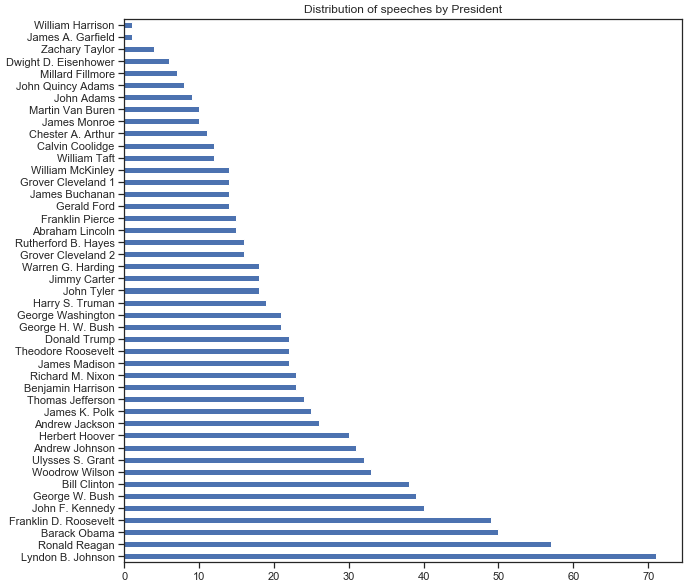


Figure 1 Distribution of speeches by President

Lyndon B. Johnson, Ronald Reagan, and Barack Obama have the largest amount of speeches in the dataset, while William Harrison, James A. Garfield, and Dwight Eisenhower have the least. Since the top 6 presidents by speech count are post World War Two, we may see more influence of words or topics that are attributed to a WW2/Post-WW2 world later.

After tokenization, stemming, adding stop-words, part-of-speech tagging, etc., next is compute the term-frequency inverse-document-frequency (TDIDF) measures for the vocabulary present in the text. Figure 2 is the vocabulary table with stop-words included, figure 3 is the table with stop-words removed.

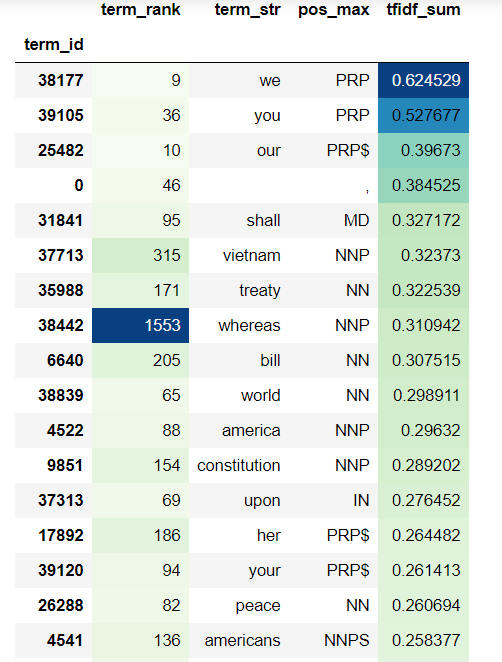
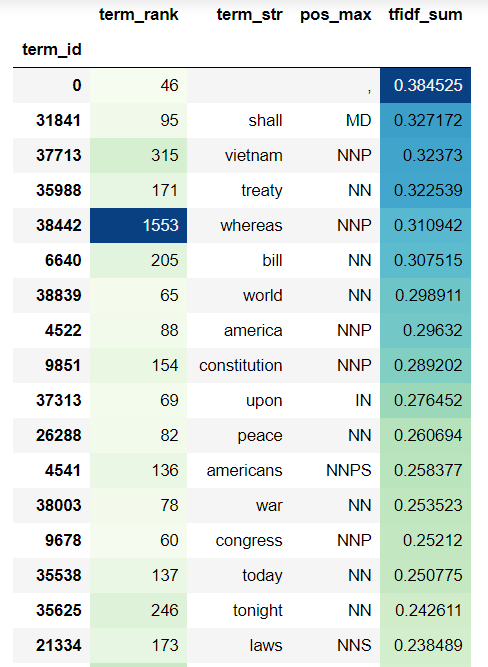
 

Figure 2. Vocab table with stop-words Figure 3. Vocab table without stop-words

Looking at the table with stop-words removed, we see high TFIDF rankings for “shall”, “vietnam”, “treaty”, and “whereas”. For two of the words, “shall” and “whereas”, those suggest older speeches, since those are antiquated terms. For “vietnam” and “treaty”, these are both terms related to foreign affairs (with “vietnam” obviously referencing the Vietnam war). Figure 4 shows the terms bagged by speeches. The terms “freedman”, “deck”, “statue”, “salary”, and “dominican” rank high in TFIDF. A lot of these terms resemble important historical events/contexts, such as “klan” and “japanese” which obviously relate to the Jim Crow south and World War Two.

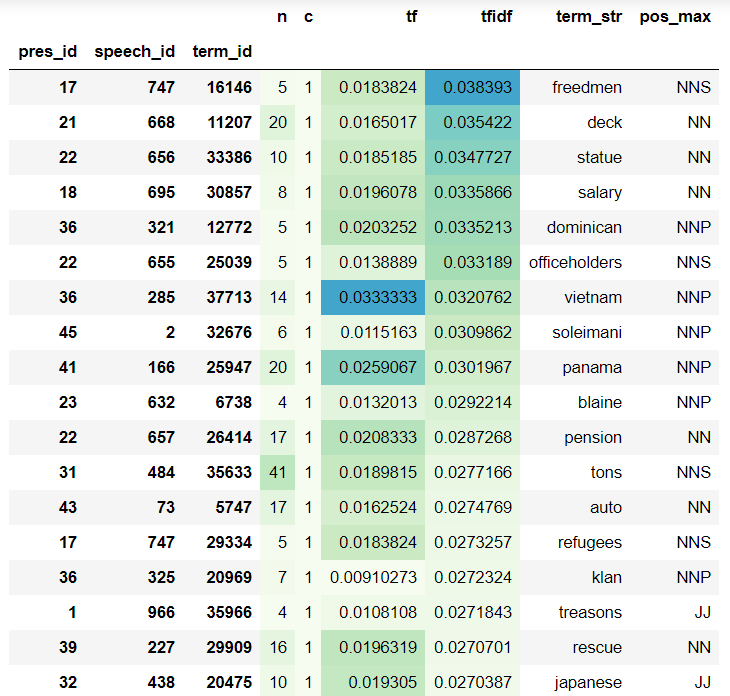


Figure 4. Terms bagged by speeches

Clustering speeches

Hierarchical clustering can be used to get a sense of similarity between different texts based on the TFIDF tables created. The cityblock, euclidean, and cosine similarity measures are used to judge the closeness of the speeches. Figure 5 is a sample of the cosine similarity tree, which I will only be presenting, since the tree, with 981 speeches, is very big.



Figure 5. Hierarchical Clustering Tree, cosine similarity

Often the groups will cluster around the president, but it is interesting to have clusters of different presidents. In black you can see a cluster of different presidents speaking about US infrastructure.

Principal Component Analysis

Principle component analysis seeks to find the axis on to which the data has the greatest variance, then project the data onto that basis. Here we use the top 4000 terms to compute the covariance matrix. When grouped by presidents, the first to principle component loadings are presented in Figure 6. The presidents (from Theodore Roosevelt to Donald Trump) mapped onto the principle component space appears in Figure 7.

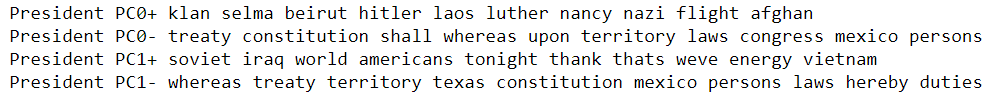


Figure 6. PC0 and PC1 loadings of presidents

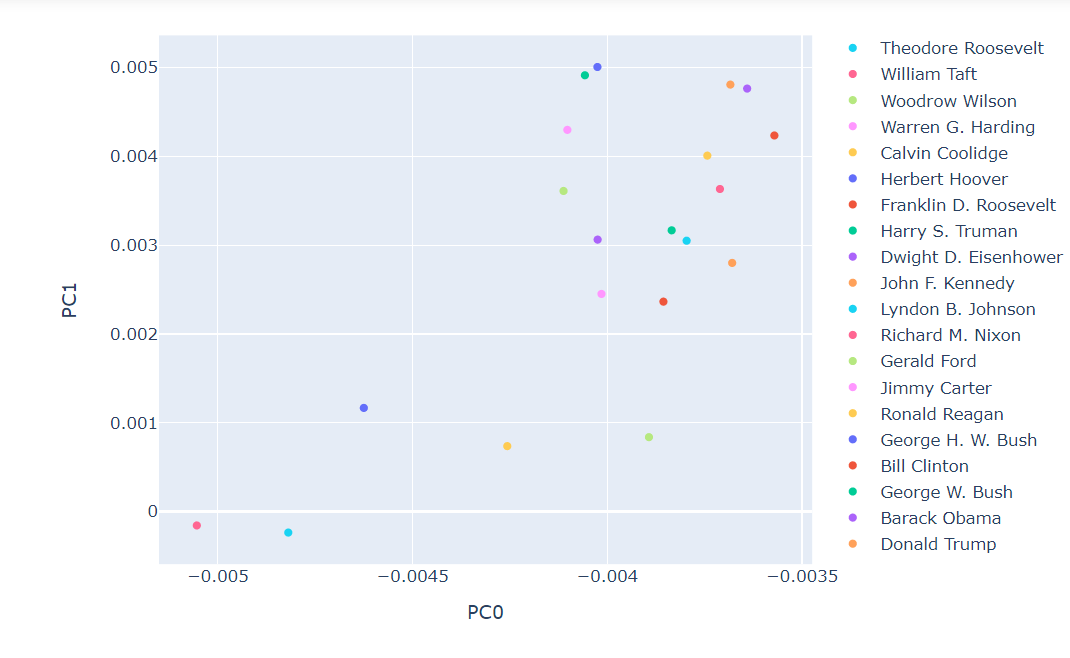


Figure 7. Presidents mapped onto PC0 and PC1

Here we can see William Taft and Theodore Roosevelt in the bottom left corner, low on both PC0 and PC1. Given the loadings, this is in the direction of dealing with treaties, Mexico, territories, Texas, etc. A topic more prominent during the turn of the century. In the top right-hand corner sit Barack Obama, Donald Trump, Bill Clinton, etc. There is clearly a time component here, the more recent presidents trend toward the top right and vice versa. These presidents deal with topics of the WW2/post-WW2 era, with terms like “Hitler”, “Selma”, “Beruit”, “Laos”, etc.

Topic modeling with LDA

Latent Dirichlet Allocation (LDA) is a generative statistical model that allows us to group certain terms in a text into a set of topics which can then be analyzed. The text will be grouped by presidents, in order to see which presidents spoke on similar topics. Figure 8 shows the topics discussed by Barack Obama, Donald Trump, Theodore Roosevelt, and George Washington.

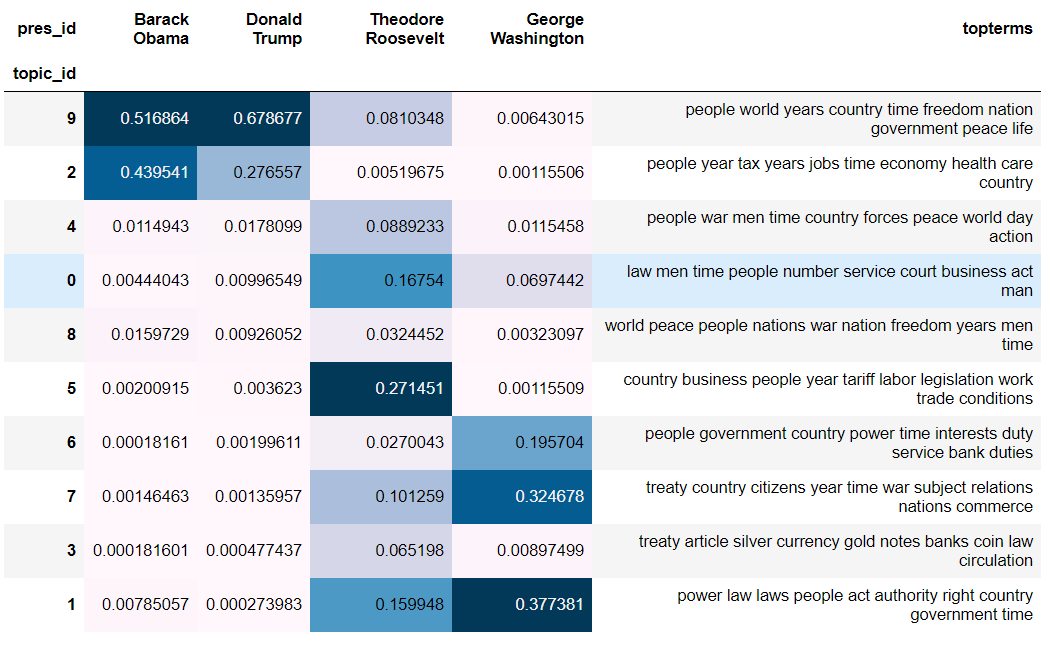


Figure 8. Topic model with Obama, Trump, Teddy Roosevelt, Washington

Both Obama and Trump speeches talk about topic\_id 9, which looks like it is discussing the general health of the nation/world, general aspirations like “peace” and “freedom” etc. Teddy Roosevelt is fond of talking about trade and business conditions. Washington talks about authority and government, which might speak to the fact he was the first president, who had to lay the groundwork for the nation.

Sentiment analysis

Sentiment analysis aims to capture the emotions underlying a text. I will be using a lexicon-based approach which associates certain words with certain sentiments (positive/negative). This will be mapped onto the president’s speeches to get a grasp of how the sentiment of president speeches change throughout his term. Figure 9 and figure 10 show the sentiment of Trump speeches and Obama speeches, respectively, by polarity (positive/negative).

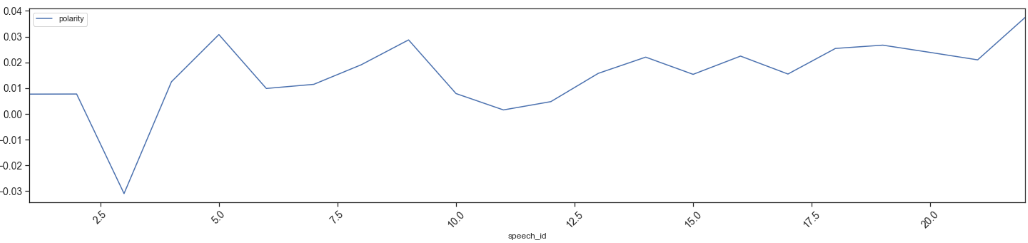


Figure 9. Sentiment analysis of Donald Trump-polarity

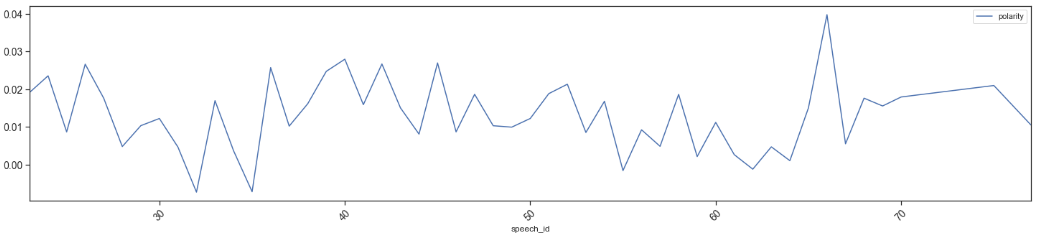


Figure 10. Sentiment analysis of Barack Obama-polarity

Trump’s speeches are more consistently positive in polarity (outside of speech\_id 3, which is really about the killing of Abu Bakr al-Baghdadi, the founder and leader of ISIS.) Obama’s speeches are less consistent, with a lot more spikes, though still generally positive.

Conclusion

Throughout the analysis one can see a few themes. One is that there is an obvious time difference in presidential speeches, which is obvious in that presidents will not be discussing the same issues throughout history, nor will they be using the same words. There does seem to be a similarity (per the PCA’s and topic model) between pre-WW2 and WW2/post-WW2 presidential speeches. This suggests a noticeable difference in the role of the president and the role of the country after WW2. Clustering does a good job of not just capturing the different presidential vocabularies but also certain topics as well. Further analysis can be done by perhaps separating by parties, segmenting by eras, or adding in debate and interview transcripts.

References

[1] <https://www.kaggle.com/kboghe/presidentialspeeches?select=2presidential_speeches_with_metadata.csv>