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.

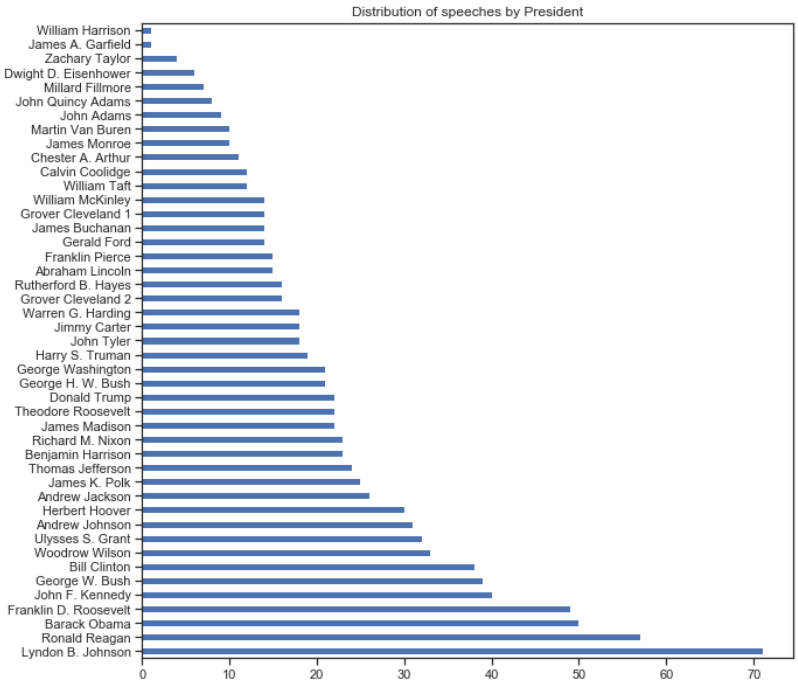

Figure 1 Distribution of speeches by President

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.

term_id	term_rank	term_str	pos_max	tfidf_sum
38177	9	we	PRP	0.624529
39105	36	you	PRP	0.527677
25482	10	our	PRP$	0.39673
0	46	,	,	0.384525
31841	95	shall	MD	0.327172
37713	315	vietnam	NNP	0.32373
35988	171	treaty	NN	0.322539
38442	1553	whereas	NNP	0.310942
6640	205	bill	NN	0.307515
38839	65	world	NN	0.298911
4522	88	america	NNP	0.29632
9851	154	constitution	NNP	0.289202
37313	69	upon	IN	0.276452
17892	186	her	PRP$	0.264482
39120	94	your	PRP$	0.261413
26288	82	peace	NN	0.260694
4541	136	americans	NNPS	0.258377

Figure 2. Vocab table with stop-words

term_id	term_rank	term_str	pos_max	tfidf_sum
0	46	,	,	0.384525
31841	95	shall	MD	0.327172
37713	315	vietnam	NNP	0.32373
35988	171	treaty	NN	0.322539
38442	1553	whereas	NNP	0.310942
6640	205	bill	NN	0.307515
38839	65	world	NN	0.298911
4522	88	america	NNP	0.29632
9851	154	constitution	NNP	0.289202
37313	69	upon	IN	0.276452
26288	82	peace	NN	0.260694
4541	136	americans	NNPS	0.258377
38003	78	war	NN	0.253523
9678	60	congress	NNP	0.25212
35538	137	today	NN	0.250775
35625	246	tonight	NN	0.242611
21334	173	laws	NNS	0.238489

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.

pres_id	speech_id	term_id	n	c	tf	tfidf	term_str	pos_max
17	747	16146	5	1	0.0183824	0.038393	freedmen	NNS
21	668	11207	20	1	0.0165017	0.035422	deck	NN
22	656	33386	10	1	0.0185185	0.0347727	statue	NN
18	695	30857	8	1	0.0196078	0.0335866	salary	NN
36	321	12772	5	1	0.0203252	0.0335213	dominican	NNP
22	655	25039	5	1	0.0138889	0.033189	officeholders	NNS
36	285	37713	14	1	0.0333333	0.0320762	vietnam	NNP
45	2	32676	6	1	0.0115163	0.0309862	soleimani	NNP
41	166	25947	20	1	0.0259067	0.0301967	panama	NNP
23	632	6738	4	1	0.0132013	0.0292214	blaine	NNP
22	657	26414	17	1	0.0208333	0.0287268	pension	NN
31	484	35633	41	1	0.0189815	0.0277166	tons	NNS
43	73	5747	17	1	0.0162524	0.0274769	auto	NN
17	747	29334	5	1	0.0183824	0.0273257	refugees	NNS
36	325	20969	7	1	0.00910273	0.0272324	klan	NNP
1	966	35966	4	1	0.0108108	0.0271843	treasons	JJ
39	227	29909	16	1	0.0196319	0.0270701	rescue	NN
32	438	20475	10	1	0.019305	0.0270387	japanese	JJ

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.

```
President PC0+ klan selma beirut hitler laos luther nancy nazi flight afghan
President PC0- treaty constitution shall whereas upon territory laws congress mexico persons
President PC1+ soviet iraq world americans tonight thank thats weve energy vietnam
President PC1- whereas treaty territory texas constitution mexico persons laws hereby duties
```

Figure 6. PC0 and PC1 loadings of presidents

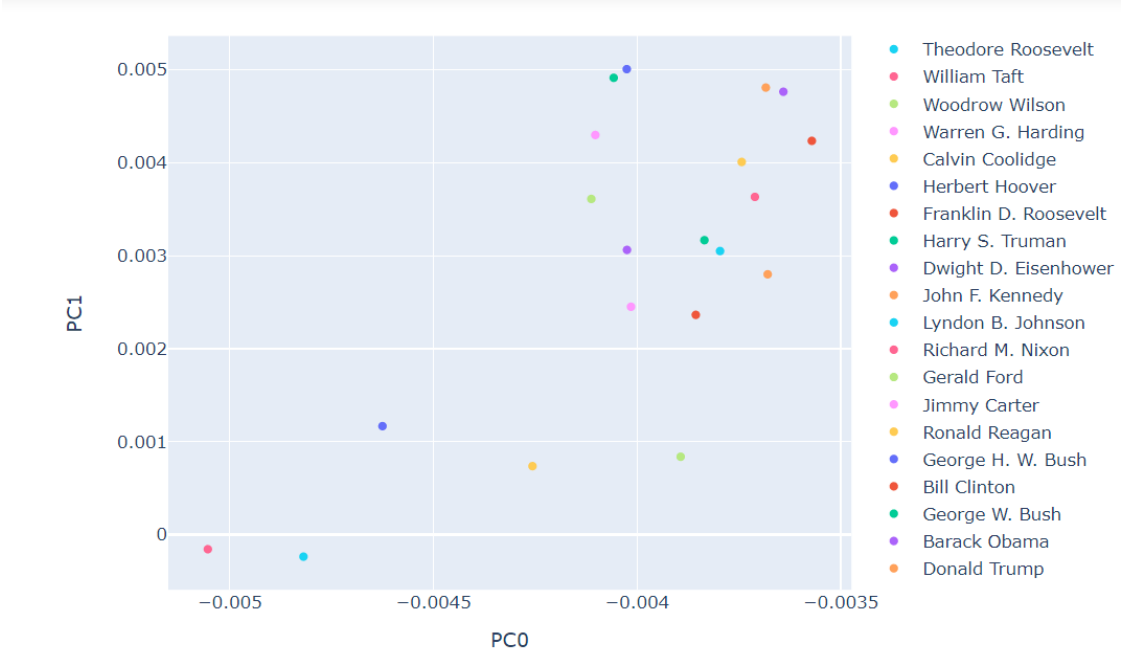

Figure 7. Presidents mapped onto PC0 and PC1

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.

pres_id	Barack Obama	Donald Trump	Theodore Roosevelt	George Washington	topterms
topic_id					
9	0.516864	0.678677	0.0810348	0.00643015	people world years country time freedom nation government peace life
2	0.439541	0.276557	0.00519675	0.00115506	people year tax years jobs time economy health care country
4	0.0114943	0.0178099	0.0889233	0.0115458	people war men time country forces peace world day action
0	0.00444043	0.00996549	0.16754	0.0697442	law men time people number service court business act man
8	0.0159729	0.00926052	0.0324452	0.00323097	world peace people nations war nation freedom years men time
5	0.00200915	0.003623	0.271451	0.00115509	country business people year tariff labor legislation work trade conditions
6	0.00018161	0.00199611	0.0270043	0.195704	people government country power time interests duty service bank duties
7	0.00146463	0.00135957	0.101259	0.324678	treaty country citizens year time war subject relations nations commerce
3	0.000181601	0.000477437	0.065198	0.00897499	treaty article silver currency gold notes banks coin law circulation
1	0.00785057	0.000273983	0.159948	0.377381	power law laws people act authority right country government time

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).

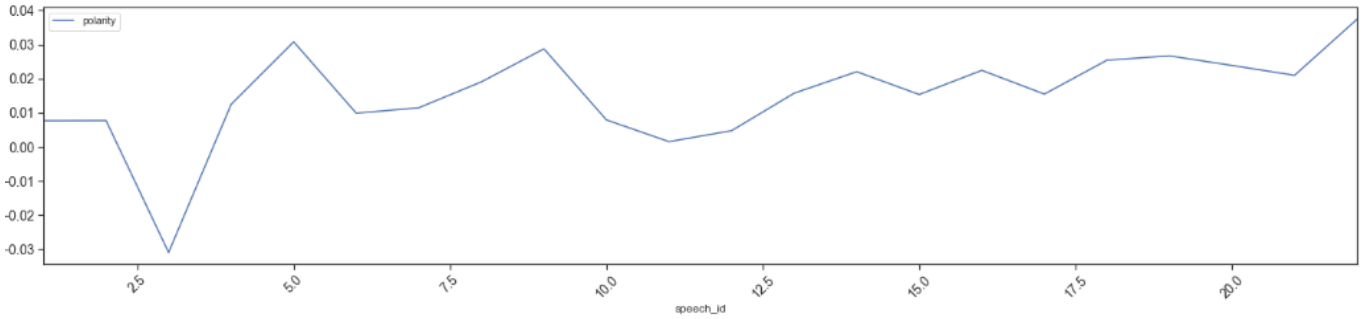

Figure 9. Sentiment analysis of Donald Trump-polarity

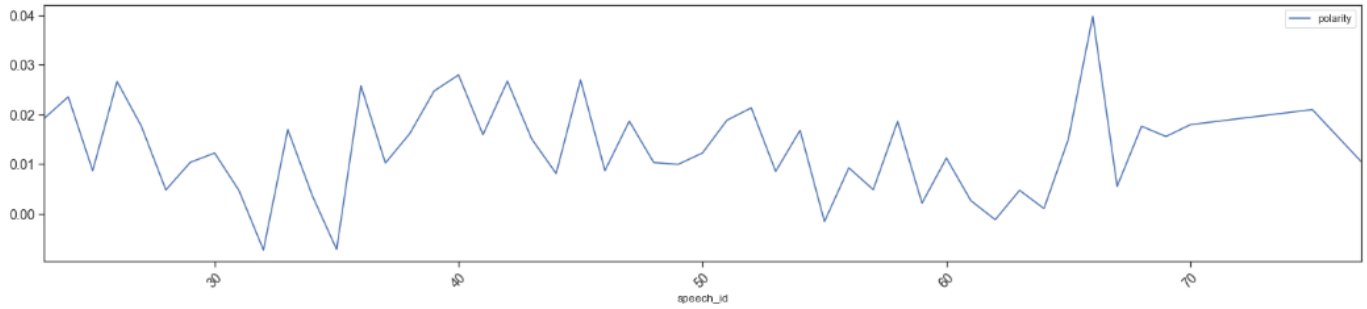

Figure 10. Sentiment analysis of Barack Obama-polarity

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