American President Speeches Text Mining Analysis

Bowen Liu, Emily Suan

Supervisor: Akram Mousa Almohalwas

University of California Los Angeles

Statistics 199 Directed Research

Abstract

Every president would have their inaugural speech after they won the election, to express their appreciation to the people who voted, address their focuses during the presidency, and unite people to fight for a better future. Therefore, these speech documents are very important resources to analyze the characters of the specific president as well as the party they belong to. This research concentrates on speech analysis using text mining, Latent dirichlet allocation (LDA), and sentimental analysis. First, president speeches will be grouped by the parties to perform text mining analysis and visualization with respect to unique words, bigram, trigram and correlations in order to summarize the general character of Democratic and Republican.

Second, four presidents will be selected from two parties at different periods as examples to perform LDA topic modeling and sentimental analysis, including text mining analysis as well, to study their personal characters during their presidency, and verify if their speaking styles could confirm the general character of the parties summarized in the first section.

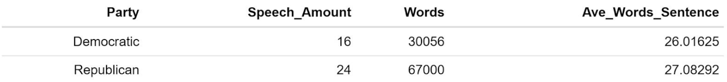
Ⅰ. Introduction

There have been 45 presidents since the United States was established since 1776. During this period, this country went through confederal to federal, civil war, world war, civil right movement and so many other historical events. With a two-party presidential election system, people usually regard Democratic as the representative of liberalism, focusing on human rights and social reform; on the other hand, Republican as the symbol of conservatism, emphasizing on restricting the government power, freedom and nationalism. Therefore, the goal of this research is verifying such a point using text mining.

The following report of this text mining research consists of two parts: party-wide speeches analysis and 4 specific presidents' speeches analysis. The first section will group the speeches by Democratic and Republican, then perform analysis including unique words, correlation plot, bigram and trigram, sentimental analysis and special words analysis; the second section will go over specific presidents from both parties, two before World War II and two after. The similar analysis will be performed again with LDA and common words as additional sections to study their concentrations of speeches. At the end of this research, a conclusion will be drawn from these analyses to summarize the general character of two parties.

Ⅱ. Democratic Vs Republican

A. Data Summary and Cleaning



Party	Speech_Amount	Words	Ave_Words_Sentence
Democratic	16	30056	26.01625
Republican	24	67000	27.08292

The presidential speech dataset comes from the “quanteda” package. Since this research focuses on the characters of both Democratic and Republican, it only deals with the presidents starting from Abraham Lincoln[1]. Therefore, there are a total of 16 speeches from Democratic with 30,056 words, and 24 speeches from Republican with 67,000 words.

This research uses the R package “tm” package to perform data cleaning. The steps are listed as below:

1. Convert the text to VCorpus
2. Remove whitespace
3. Convert the text to lowercase
4. Remove stopwords
5. Stem the text
6. Remove punctuation

These data cleaning steps are commonly used in the rest of this research, except the stop words which are customized words depending on the specific documents.

B. Unique Words Analysis

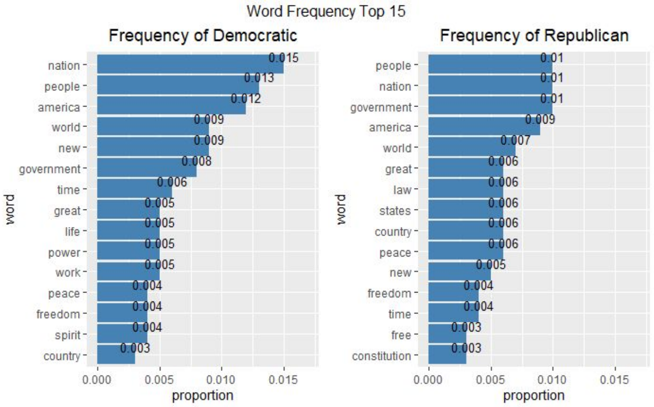


Figure 1. Top 15 Words of two parties

According to figure 1 that shows top 15 single words with highest proportion, even though both parties mentioned a lot about the “nation” and “people” , Democratic had more concentration on words like “new”, “work” and “life”, whereas Republicans focused more on words such as “government”, “country”, and “constitution”

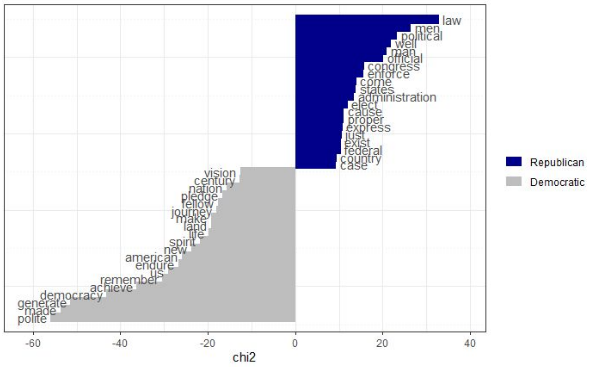


Figure 2: Keyness plot of two parties

Keyness[2] plot performs a function of feature words selection between two groups and plot the words that appear more frequently in one group than they do in another (measured by chi square). Unlike figure 1, keyness plot does not have words in common between two groups. According to figure 2, Democratic focused on words like “generate”, “democracy” and “new”, and Republican emphasized on words such as “law”, “congress“ and “political”.

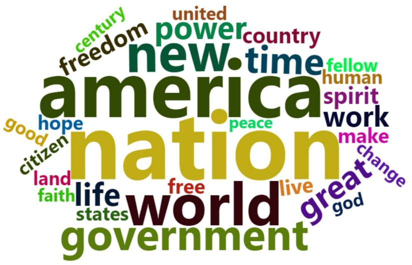


Figure 3: Democratic Word Cloud

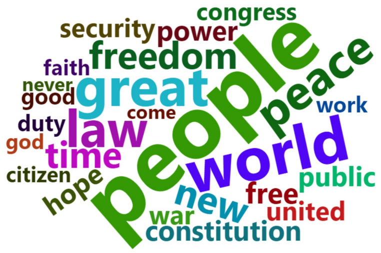


Figure 4: Republican Word Cloud

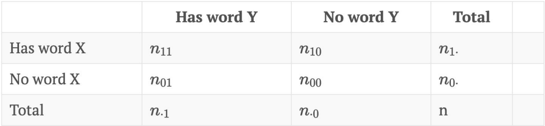
[1] Whig party had its last presidential candidate in 1856, then fade away after the civil war

[2] Keyness is a method using differential correlation to identify the words that have higher frequency than expected in target group, compared with the reference group

Except for the common words such as "america" and "nation", the world cloud of Democratic had larger spaces for words like "new" , "life" and "change", while Republican's word cloud concentrated on the words such as "security", "law" and "peace."

C. Bigram Analysis

Digging into the details about the bigram, which is measured in a binary form to show how often they appear together. A common measurement for such binary correlation is the phi coefficient, and it is calculated as the following:



	Has word Y	No word Y	Total	
Has word X	n_{11}	n_{10}	$n_{1\cdot}$	
No word X	n_{01}	n_{00}	$n_{0\cdot}$	
Total	$n_{\cdot 1}$	$n_{\cdot 0}$	n	

$$\phi = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{n_{1\cdot}n_{0\cdot}n_{\cdot 0}n_{\cdot 1}}}$$

Figure 5: Phi coefficient formula

According to figure 6, Democratic mainly had three clusters of words centered at "new" plus "work", "democracy" plus "freedom", and people.

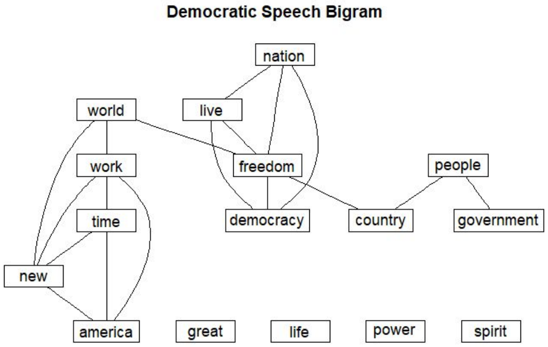


Figure 6: Bigram of top 0.5% Democratic speech with 0.5 threshold

Moreover, Republican have two main clusters centered at "power" plus "country", "nation" plus "world".

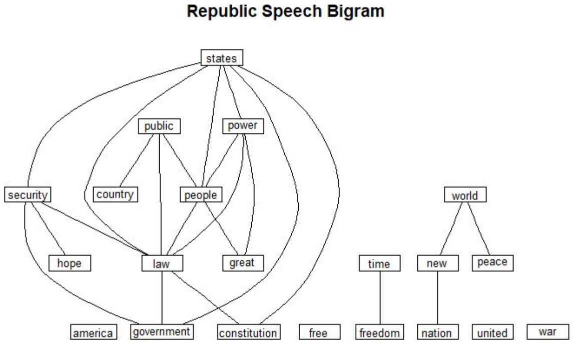


Figure 7: Bigram of top 0.4% Republican speech with 0.5 threshold

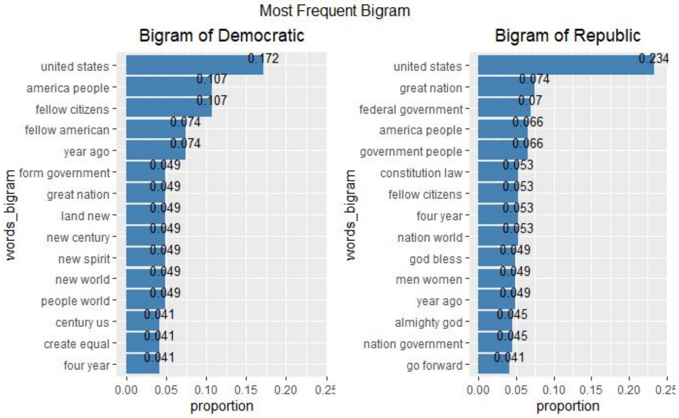


Figure 8: Top 15 Bigram in two parties

According to figure 8 that shows the top 15 adjacent bigrams with the highest proportion, Democratic delivered many words about changing the current situation such as "new spirit", "new world" and "create equal". This can also be reflected in figure 9, which is centered at "new" and "people" ; whereas Republican took more care of government, mentioning more about "federal government", "government people" and "constitution law". In figure 10, it is obvious that "government" and "world" are the center of the plot.

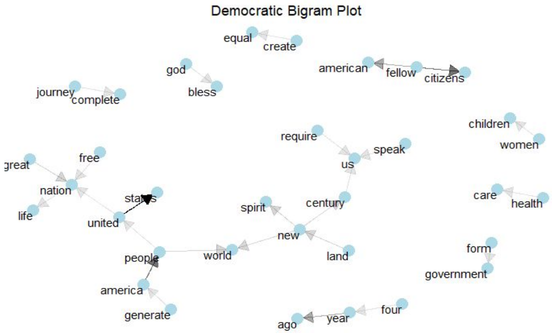


Figure 9: Bigram plot of Democratic

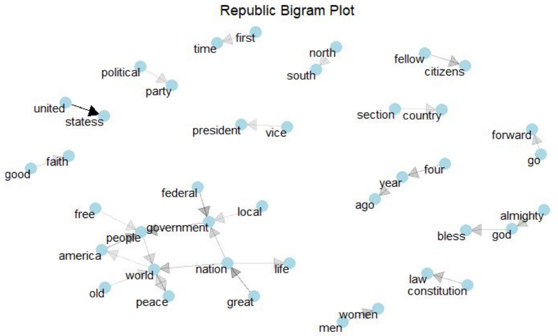


Figure 10: Bigram plot of Republican

D. Trigram Analysis

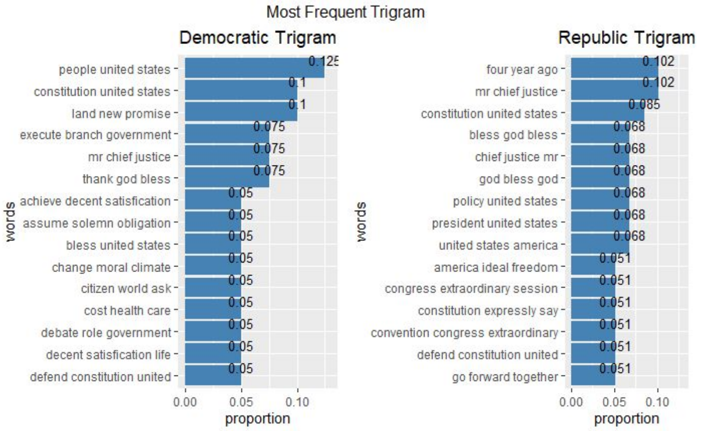


Figure 11: Top 15 Trigram of two parties

According to figure 11 that shows the top 15 adjacent trigrams with the highest proportion, the general characters of the two parties mentioned in the introduction becomes more obvious. Democratic had frequent trigrams such as "land new promise", "change moral climate", and "cost health care" , which represent social reforms. On the other hand, Republican's trigrams concentrated on the constitution and government, such as "constitution united states" and "congress extraordinary session".

At this point, the characters of two parties correspond to the point mentioned in the introduction: Democratic have more words related to social changes and human rights, in contrast, Republican pay more attention to government power and the national operation.

E. Emotion Distribution

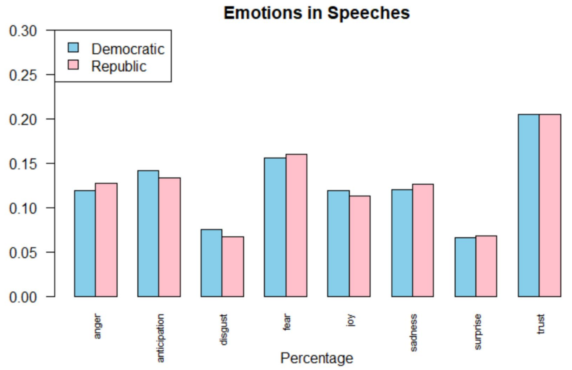


Figure 12: Emotion Comparison between two parties

According to figure 12 that shows the proportion of eight emotions among the party's speeches, there is nearly no difference between the general emotions of two parties. Both Democratic and Republican expressed "trust" mostly in their speech, but one interesting finding is the proportion of fear and anticipation in both parties, which are second and third highest among all emotions. In this case, "fear" should refer to the problems the president thought highly of, and "anticipation" should be the actions he was planning to take, which makes sense for the structure of an official speech: address the problems, express the solutions and encourage people to take actions.

F. Unique Words Distribution

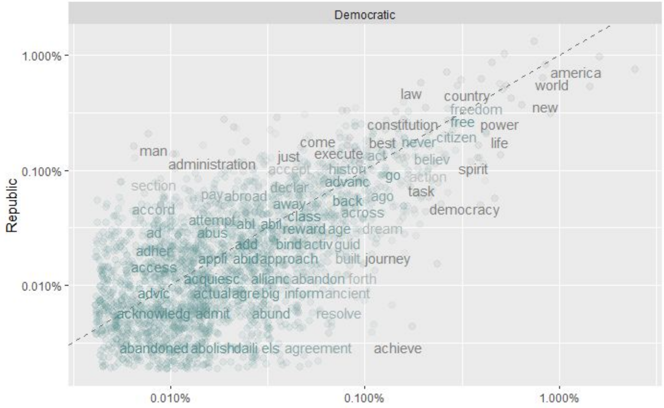


Figure 13: Unique words distribution among two parties' speeches

According to figure 13, the words around the dotted line are common words among the speeches of two parties, for example, "america", "world" and "freedom" appear frequently on both sides. Moreover, the words on the right lower corner, such as "spirit", "democracy" and "achieve", had a higher frequency in Democratic speeches, while the words on the left upper corner, such as "administration" and "constitution" appeared more in Republican's speeches.

G. Special Words Analysis

As the basic characters we summarized, Democratic speeches turn out to be more liberal, caring about social reform and human rights, while Republican is comparatively conservative, focusing on the political operation and society. Therefore, we plot the distribution of four specific words, representing human rights, social reform, politics and society, along the timeline and divided by parties to verify this idea.

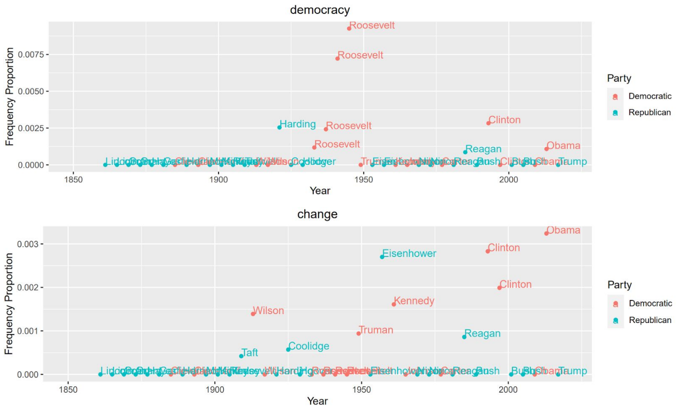


Figure 14: words represent human rights and social reform

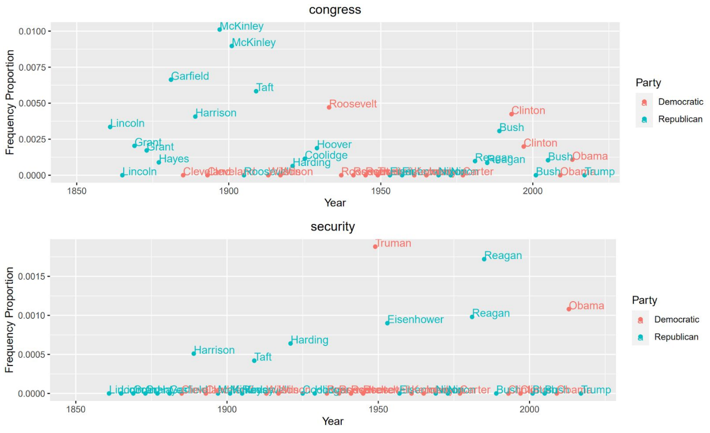


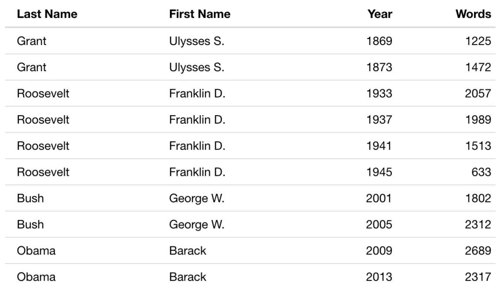
Figure 15: words represent politics and society

Through the figure 14 and 15, we can see that Democratic appear more in the area of democracy and change, and Republican show up more in the area of congress and security, which corresponds to their characters we summarized.

Ⅲ. Specific Presidents Analysis

After a general comparison between democratic and republican parties, analyses of Roosevelt(Democratic) Vs Grant(Republican) and Bush(Deomcratic) Vs Obama(Republican) will be discussed below. The reason why World War II was selected as a splitting point is because an International Monetary System based on the U.S. Dollar was established after World War II, indicating a new world-wide economy centered in the United States.

A. Data Summary



Last Name	First Name	Year	Words
Grant	Ulysses S.	1869	1225
Grant	Ulysses S.	1873	1472
Roosevelt	Franklin D.	1933	2057
Roosevelt	Franklin D.	1937	1989
Roosevelt	Franklin D.	1941	1513
Roosevelt	Franklin D.	1945	633
Bush	George W.	2001	1802
Bush	George W.	2005	2312
Obama	Barack	2009	2689
Obama	Barack	2013	2317

The table above shows the data background. Roosevelt had four speeches, but only the first two would be used here, and the average word per speech is 2023. Grant, Bush, and Obama all had two speeches, and the average word per speech is 1348, 2057, 2503 respectively. The speeches are getting longer and longer as time goes.

B. Unique Words Analysis

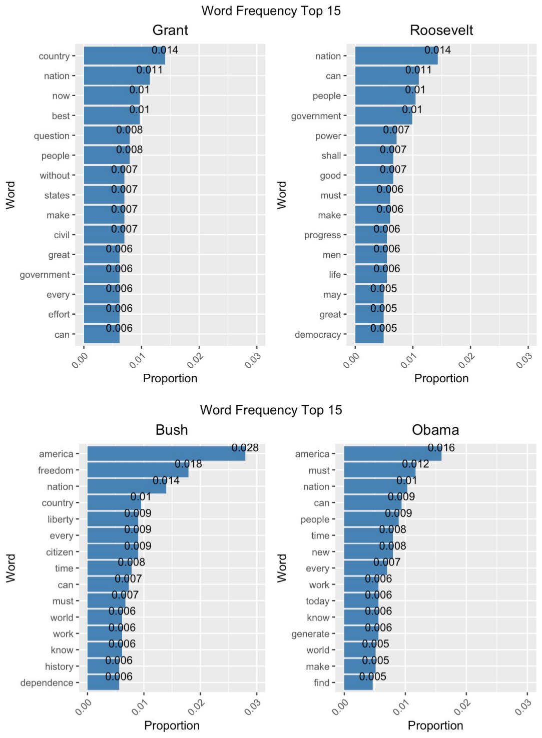


Figure 16. Top 15 Words of four presidents

According to the top 15 word frequency graphs, “nation” is the mostly mentioned word for all four presidents. “Country” is mentioned a lot for Grant and Bush, who represented Republican. Some distinguishing words were mentioned; for example, “civil” by Grant, “freedom” and “history” by Bush, indicating the conservatism of Republican and their advocacy of individual freedom. On the other hand, words such as “life,” “democracy,“ “spirit,” and “power” by Roosevelt, and “generate” and “work” by Obama indicate the liberalism of Democratic.

“Government” was mentioned frequently before World War II, while words such as “world”, “america”, and “work” were mentioned frequently after the war, because global relations and economics play a more important role for the presidents after World War II.

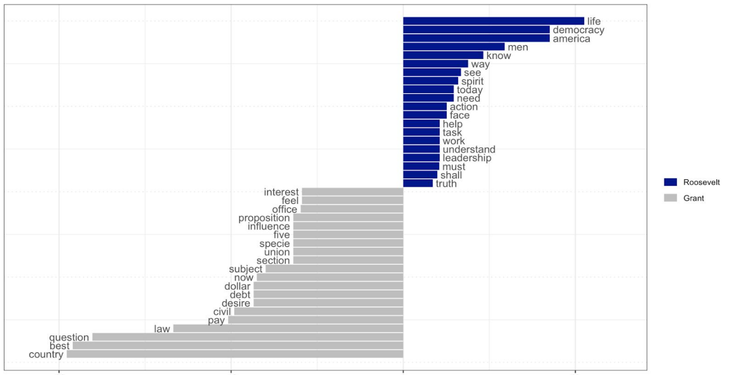


Figure 17. Keyness plot of Grant and Roosevelt

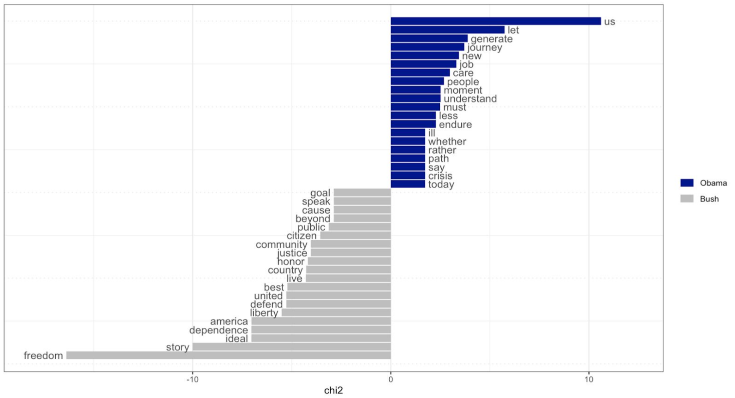


Figure 18. Keyness plot of Bush and Obama

Figure 17 and 18 show the featured words. For example, For figure 17, “civil” and “law” are Grant’s featured words, while “democracy” and “spirit” are Roosevelt’s featured words. For figure 18, “freedom” and “justice” are Bush’s featured words, while “job” and “care” are Obama’s featured words.

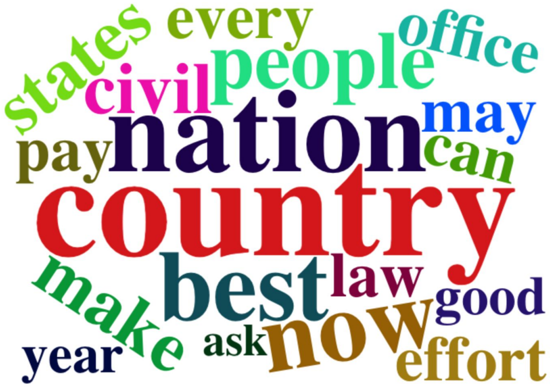


Figure 19: Grant Word Cloud

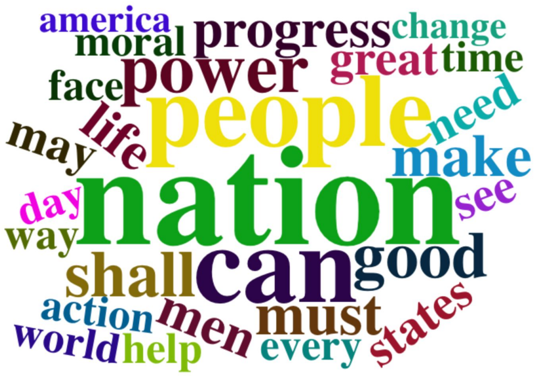


Figure 20: Roosevelt Word Cloud

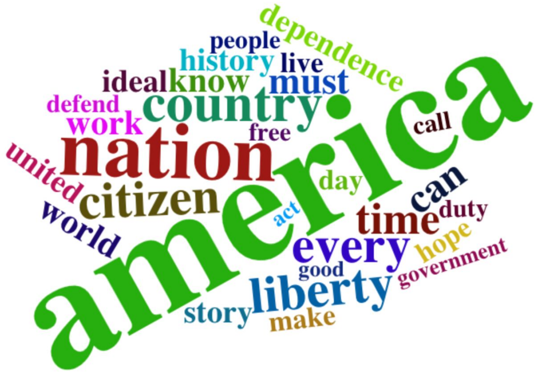


Figure 21: Bush Word Cloud

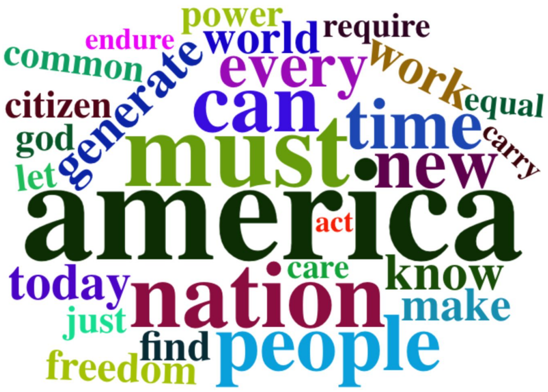


Figure 22: Obama Word Cloud

The word clouds above provide further detail of frequent words of presidents Grant, Roosevelt, Bush, and Obama respectively. Words such as “nation”, “america”, “people”, “make”, and “can” were mentioned a lot for all four presidents. From figure 19 and 21, “government,” “progress,” and “history” are frequent words for Republican presidents, while from figure 20 and 22, “work” and “new” are frequent words for Democratic presidents.

C. Bigram Analysis

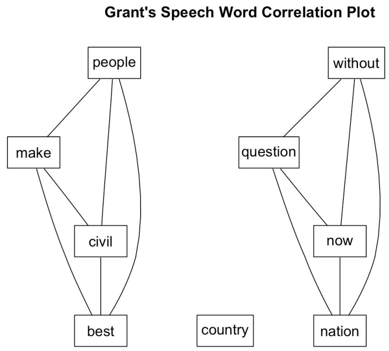

Grant's Speech Word Correlation Plot

Figure 23. Bigram of top 1% Grant’s speech with 0.5 threshold

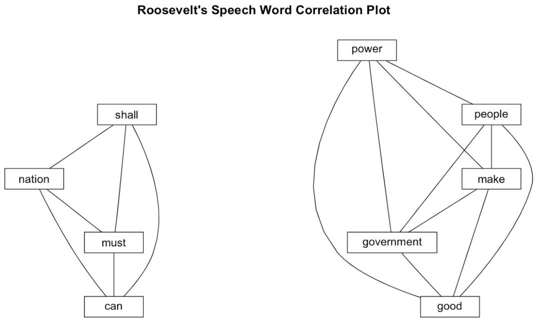

Roosevelt's Speech Word Correlation Plot

Figure 24. Bigram of top 1% Roosevelt’s speech with 0.5 threshold

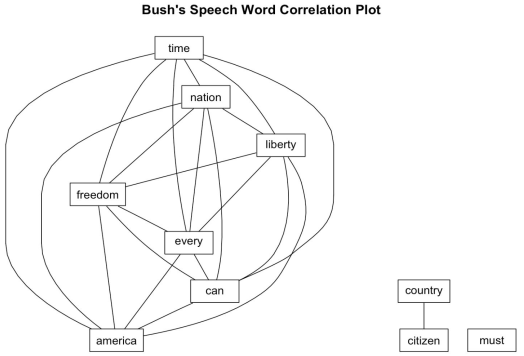

Bush's Speech Word Correlation Plot

Figure 25. Bigram of top 1% Bush’s speech with 0.5 threshold

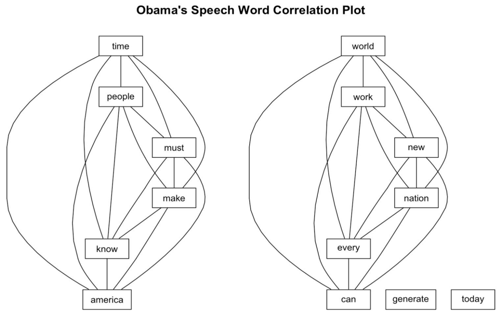

Obama's Speech Word Correlation Plot

Figure 26. Bigram of top 1% Obama’s speech with 0.5 threshold

From figure 23 to 26, three presidents except Roosevelt had two correlation clusters. One is about the people, and the other is about the nation. From plots of Grant and Bush, keywords are “civil” and “freedom.” From plots of Rosevelt and Obama, some keywords are “spirit,” ”democracy,” and “work.”

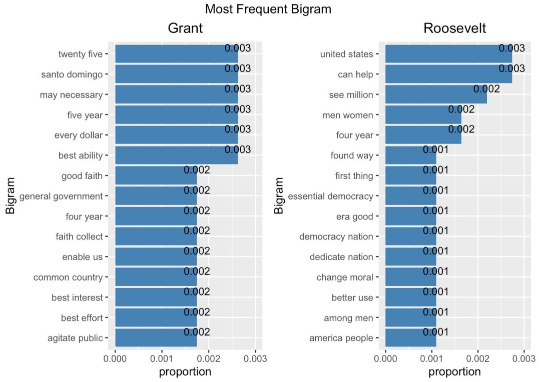


Figure 27: Top 15 Bigram of Grant and Roosevelt

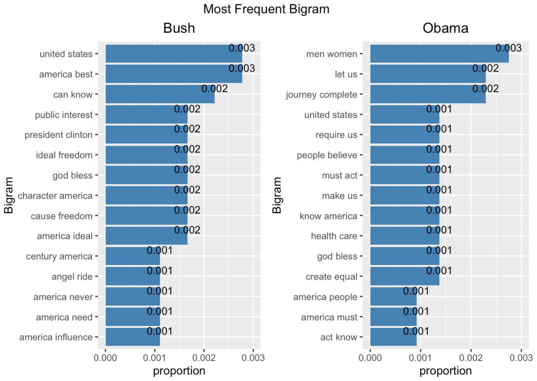


Figure 28: Top 15 Bigram of Bush and Obama

According to figure 27 and 28, the common adjacent words for all presidents are “united states”, “america people”, “men women”, and “four year.” Words such as “general government” and “ideal freedom” appear frequently in the speeches of Republican’s presidents. In contrast, words such as “spirit faith” and “create equal” show up more in the speeches of Democratic’s president.

Features of each president could also be seen. For instance, “santo domingo” shows the fact that the annexation of Santo Domingo was attempted by president Grant. “Health care” shows the policy that president Obama had greatly promoted during his presidency.

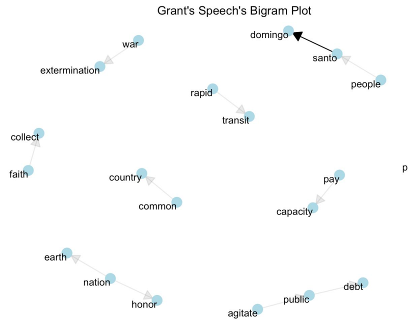


Figure 29: Bigram plot of Grant

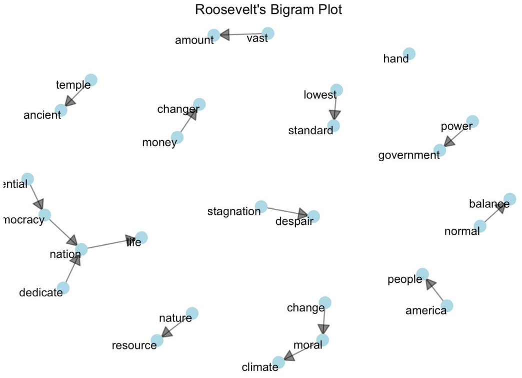


Figure 30: Bigram plot of Roosevelt

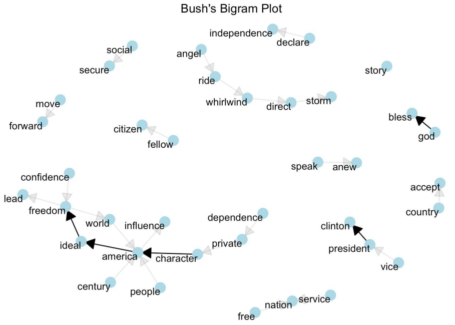


Figure 31: Bigram plot of Bush

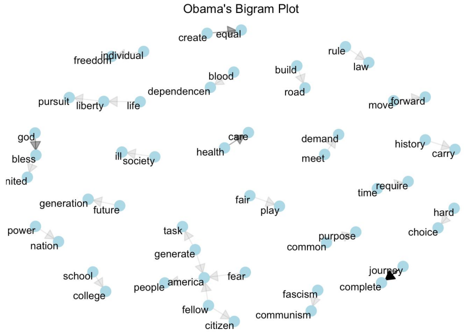


Figure 32: Bigram plot of Obama

Figure 29 to 32 provide further details of frequent adjacent words. Keywords of Grant and Bush, who represented Republican, are “war extermination,” “public debt,” and “social secure,” which are more conservative. While keywords of Roosevelt and Obama, who represented Democratic, are “change moral climate,” “act quickly," "individual freedom," and "liberty pursuit," which are more liberal.

D. Trigram Analysis

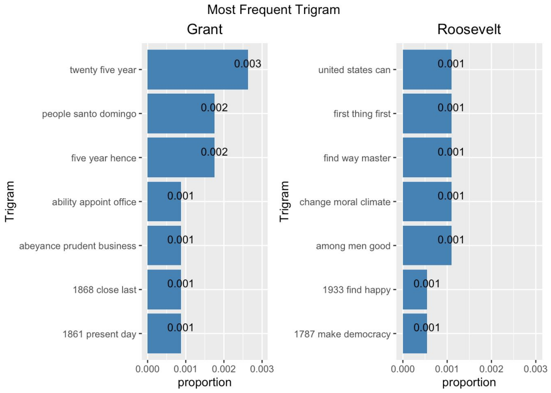


Figure 33: Top 15 Trigram of Grant and Roosevelt

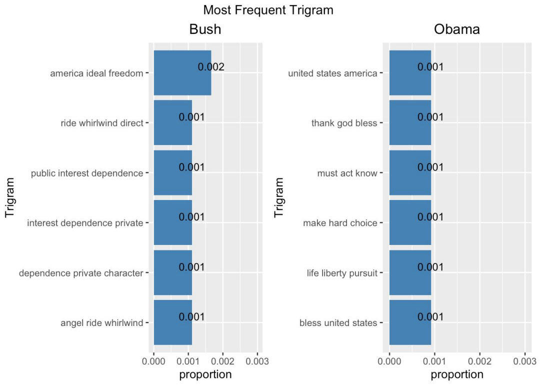


Figure 34: Top 15 Trigram of Bush and Obama

The distinguishing words of trigrams are "people santo domingo" and "ability appoint office" for Grant, and "public interest dependency" for Bush, which show their focus on national interests. On the other hand, "change moral climate" and "united states can" for Roosevelt and "must act know" and "life liberty pursuit" for Obama show their focus on social reform.

E. LDA

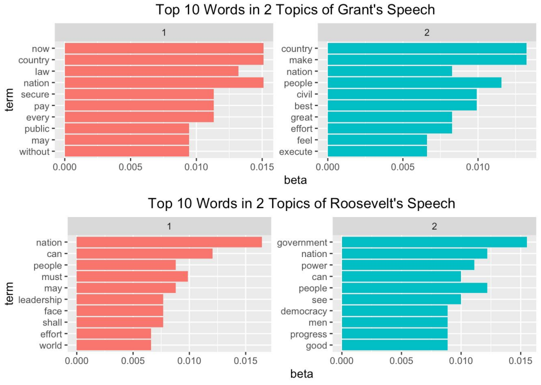


Figure 35: LDA of Grant and Roosevelt

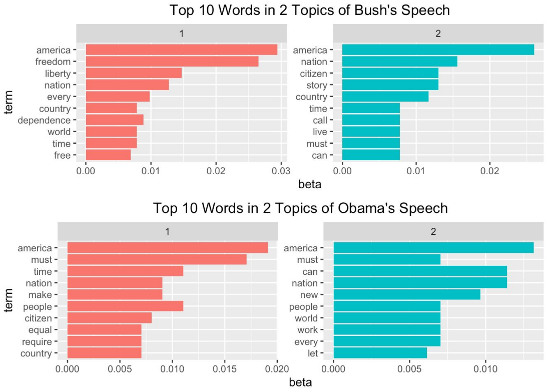


Figure 36: LDA of Bush and Obama

Latent dirichlet allocation (LDA) classifies words into different topics using beta. Beta represents topic-word density; with a high beta, topics are made up of most of the words in the corpus. According to figure 35 and 36, speeches are classified into two topics for all four presidents. Topic 1 is generally about the value of the nation. Topic 2 is generally about the future vision of the nation. Traits of two parties could also be seen here. For example, "secure" and "freedom" are shown in Republican's graphs; "life," "equal," and "democracy" are shown in Democratic's graphs.

F. Sentimental Analysis

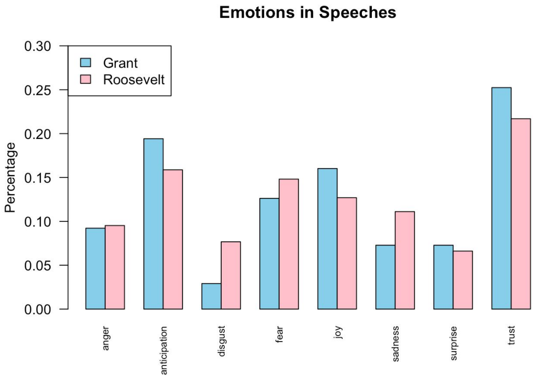


Figure 37: Emotion Comparison between Grant and Roosevelt

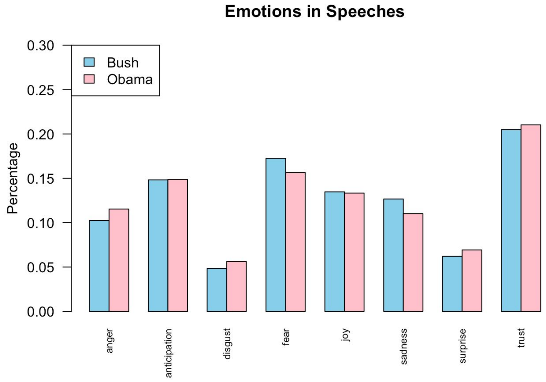


Figure 38: Emotion Comparison between Bush and Obama

According to figure 37 and 38, trust is the highest for all four presidents, thus it shows that inaugural addresses are important approaches for new presidents and governments to build trust with their people. In more details, Grant had a speech style that had relatively more anticipation and joy, while Roosevelt had a speech style that had relatively more sadness and fear. Bush had a speech style that had a slightly more fear and sadness, while Obama had a speech style that had slightly more anger and surprise.

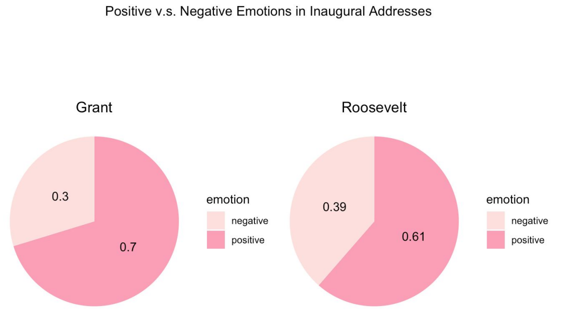


Figure 39: Emotion Comparison between Grant and Roosevelt

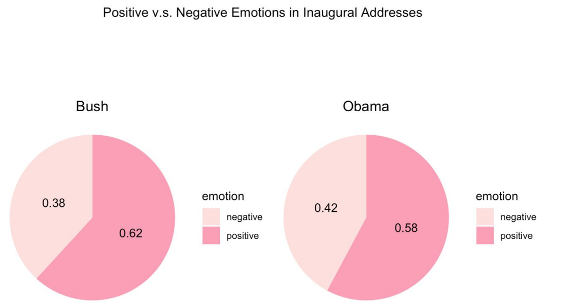


Figure 40: Emotion Comparison between Bush and Obama

Figure 39 and 40 show the percentage of positive and negative emotions in each president's speeches. All of them have relatively more positive speech styles for inaugural addresses. Grant had the most positive speech style among four presidents. The other three presidents had similar proportions of positive emotion. In general, Republican have a roughly more positive speech style than Democratic.

G. Unique Words Distribution

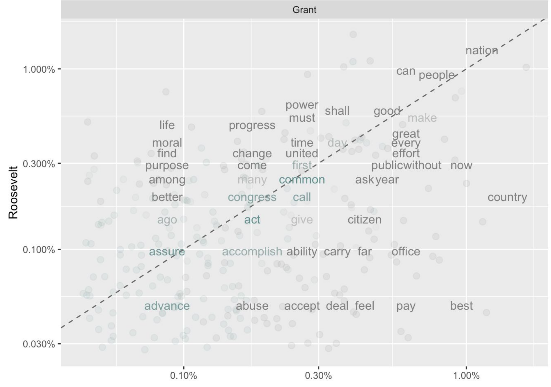


Figure 41: Unique words comparison between Grant and Roosevelt

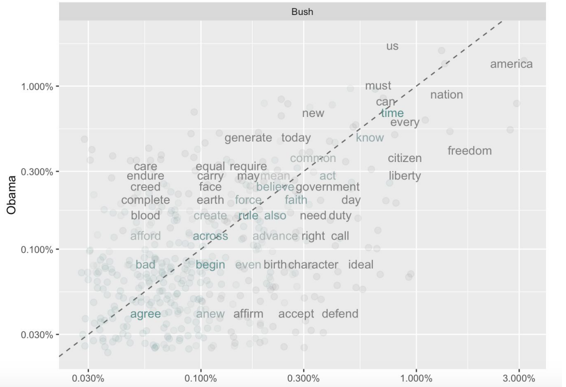


Figure 42: Unique words comparison between Bush and Obama

According to figure 41 and 42, the different words, shown in corners, are unique words that appear frequently in one president’s speeches but rarely in another. Clustered by the parties, words such as “defend,” “pay,” and “office” were expressed more for Republican’s speeches, while the words like “life,” “creed,” and “moral” showed up more in the speeches of Democratic. The results match the conservative style of Republican and libertal style Democratic.

H. Words in Common

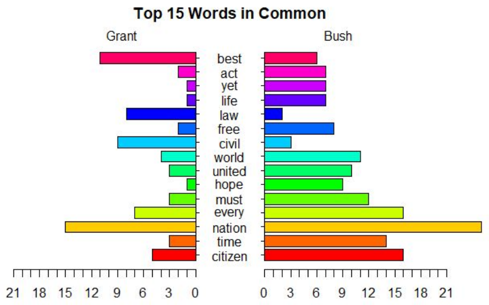


Figure 43: Top 15 Common words between Grant and Bush

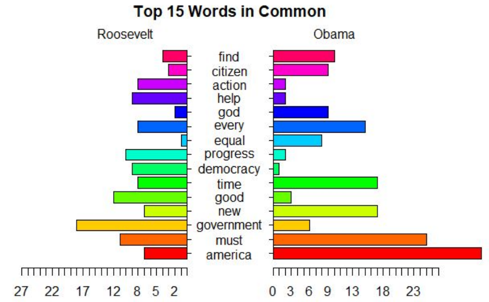


Figure 44: Top 15 Common words between Roosevelt and Obama

Figure 43 and 44 each shows the common words for presidents that belonged to the same party. Common words of Republican are, for example, “law” and “civil.” On the contrary, common words of Democratic are “equal” and “democracy.” The result corresponds to the traits of each party that Republican pursue conservatism and individual freedom, and Democratic emphasizes democracy and equality.

Ⅳ. Conclusion and Future Works

Conclusion:

According to the unique word, bigram, trigram analysis, LDA, and common word between two parties as well as their representative presidents, we can conclude that the point raised in the introduction is correct: the characters of Democratic are more liberal compared with Republican, whose character is more conservative. Such a general difference could be used as an important predictor for the modeling in the future.

Future Works:

As this research has reached this far, we summarized the general characters of two parties, and how to identify the special words, average sentence length and topic for specific presidents. Therefore, our future works are creating a dataset which contains paragraphs randomly selected from the presidents, then building models using the features mentioned above as predictors to identify the author of the speech.

Reference:

Ingo Feinerer, June 19, 2018, “Introduction to the tm Package Text Mining in R”, https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf

Kasper Welbersa, Wouter Van Atteveldtb, and Kenneth Benoit, 2017, “Text Analysis in R”, https://www.tandfonline.com/doi/full/10.1080/19312458.2017.1387238

Kenneth Benoit, Adam Obeng, and Stefan Müller, “Example: textual data visualization”, https://quanteda.io/articles/pkgdown/examples/plotting.html

Julia Silge and David Robinson, March 07 2020, “Text Mining with R”, https://www.tidytextmining.com/index.html

Marina Bondi and Mike Scott, 2010, “Keyness in Texts”, https://benjamins.com/catalog/scl.41

UC Berkeley Business Analytics R Programming Guideline, “Text Mining: Word Relationships”, https://uc-r.github.io/word_relationships?nsukey=3smQWQ9II2JxFQ%2FjooCOaee9oRWz5IF%2FKn%2BzfQOqlxDkBHTr8voSKwHnOvzhfSfMnfnLHuK6UehaJnQeP5K5%2BqmjUufvqDYCUrokJEOQ125qXKTnRZCRBivrm8RAWnOJezaT3QUDOqpKoV3lVzwBOi6IoHjRRwOe0lVrsZN03YX0JZyhvRMbhcihlcs7GvGl33r8WnIZxmeZeyPUjM9TAA%3D%3D