TEXT ANALYTICS ON PRESIDENTIAL SPEECHES

A PREPRINT

keyu chen
School of Data Science
University of Virginia
Charlottesville, VA
km5ar@virginia.edu

May 12, 2021

1 Introduction

After the collapse of the Soviet Union, the United States of America is the one and the only existing superpower in the world, its influences, both soft and hard power, has a huge impact on the average people's lives around the world. Therefore, the president of the United States (the commander in chief) is the most powerful man on earth, which makes his words extremely important. How he or she think, talk and react to different issues in the speech could be a very important key information to understand the president and what he will do next step.

In this paper, I will apply the exploratory text analytic methods on the speeches of US presidents then take a closer look of 2016 presidential campaign speeches made by Trump and Hillary Clinton. The 2016 presidential election, although to be very controversial, but I think will be very interesting to see what's the differences between Trump and Hillary Clinton based on their speeches during the presidential campaign.

2 Corpus of Presidential Speeches

- Provider: The Grammar Lab
- 3.5 million words
- speeches made by all previous presidents of US
- plus speeches delivered at campaign events by Hillary Clinton and Donald Trump, "beginning with their acceptance speeches at their respective party conventions and continuing up to the election. The corpus contains approximately from 114, 000 words from Clinton and 440,000 words from Trump." according to the Grammar Lab.

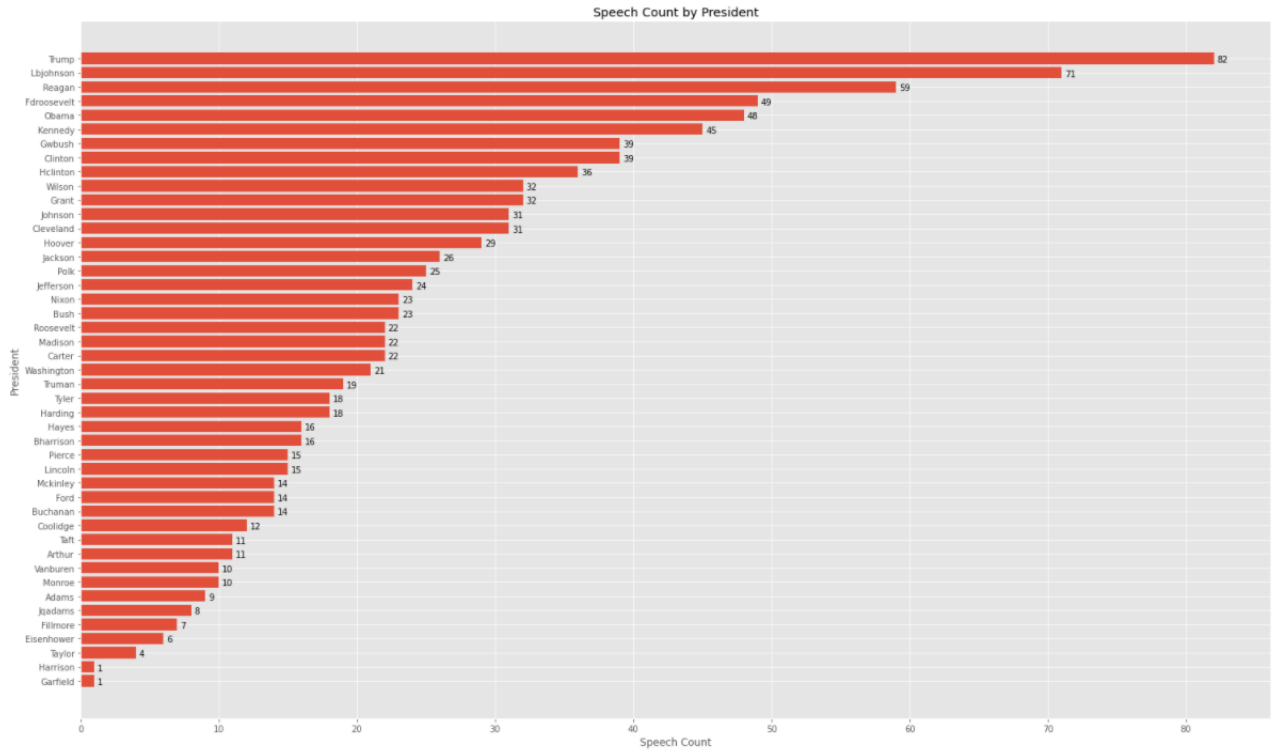


Figure 1: Speech Count by President

3 Number of Speeches by presidents

- Trump: 82
- Hillary Clinton: 36
- please see Figure 1 for all information regarding previous presidents

4 Hierarchical clustering by president

In dendrogram Figure 2, D starts for ”Democratic Party", R stands for "Republican Party", the length of the branches shows the similarity of the grouped presidents. According to Professor Raf’s lecture, "Hierarchical clustering is more informative than flat clustering." we can see, it clearly distinguished president from a different era, for example:

Recent 20 years:

- obama
- clinton
- bush
- carter

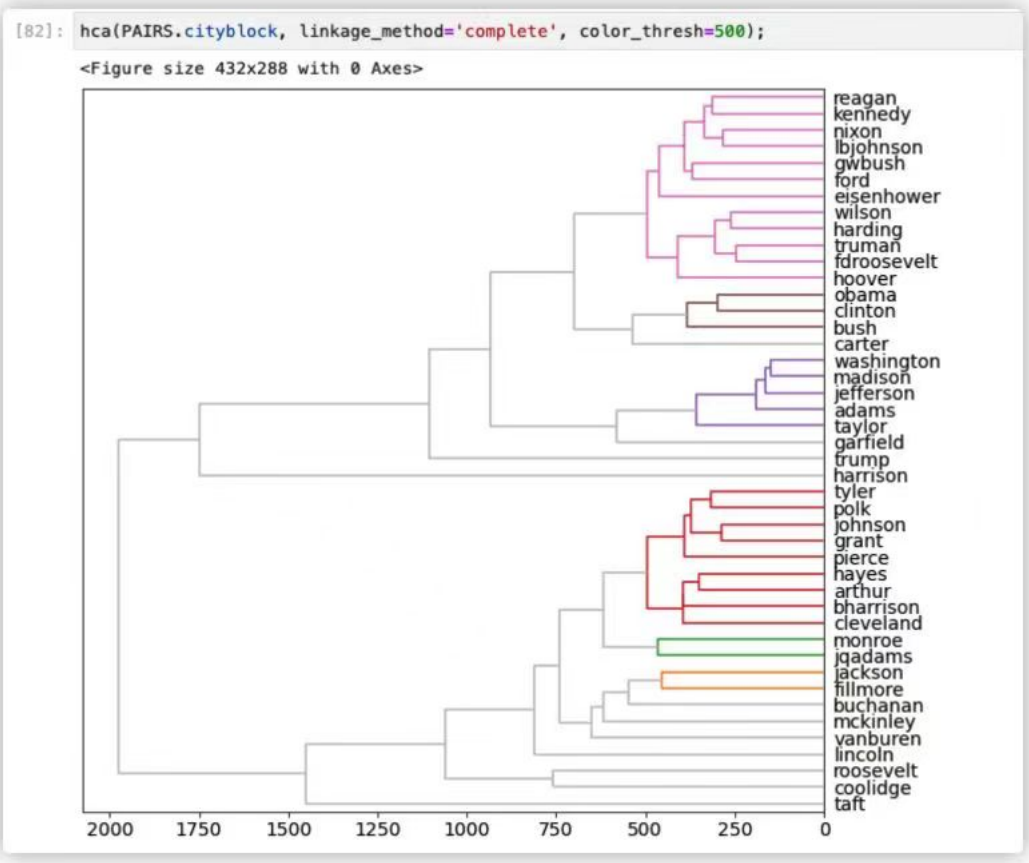


Figure 2: Hierarchical clustering by president

Funding fathers:

- washington
- madison
- adamas
- jefferson

President from 1840s to 1890s:

- James K polk, 1845 - 1849
- John Tyler, 1841 to 1845
- Ulysses S. Grant, 1869 to 1877
- Andrew Johnson, 1865 to 1869
- Franklin Pierce, 1853 to 1857
- Rutherford B. Hayes, 1877 to 1881

- Chester A. Arthur, 1881 to 1885
- William Henry Harrison, 31 days in 1841
- Grover Cleveland, 1893 to 1897

1900s to 1950s:

- Franklin D Roosevelt 1933 until his death in 1945
- Calvin Coolidge, 1923 to 1929
- William Howard Taft, 1909–1913

5 PCA

In our homework, we were able to use PCA to distinguish genres and authors. So I think we can probably use PCA to discovering some insight from Trump and Hillary's speeches since both of them are also so different, in addition to that, the level of vocabulary they use is also different, Trump tend to like to use simple words while Hillary Clinton, have a more solid training as an elite politician. I should able to use PCA on them as well.

In the Figure 3. We can see that Trump and Hillary have some overlap but not much on Scatter plot. Trump has significant differences compared to all other presidents. Hillary has half of it close to trump while another half close to other presidents.

In Figures 3 and 4. We can see Trump and Hillary had some overlap in both Box plots scatter plot, while clearly different, that's maybe due to the educational background and the language they have been using in the speech. Trump's word choice and the way he talks to his voter is just completely different than any other president and traditional politicians in the US history. Trump's vocabulary is more toward working-class people, while Hillary Clinton's supporters are mainly college students[**2**].

Another observation we can see is that traditional politician elites such as Hillary, Obama, and Bush(junior and senior), reagan has a similar pattern which makes sense, they all coming from an elite politician training, went to similar school and speak a similar language at the same time, they all belong to the same ear(recent 40 years).

By looks at the graph in detail, we can see that it follows a chronological order from Washington at the right upper corner, then gradually moved to Reagan at the underneath center then moved to Trump at the left upper corner. The order of the president is surprisingly good.

The presidents after the 1980s are all located on the left side of the graph. Why? based on my own understanding of US history. I think the words and the way they communicate to the mass is different compare to decades ago(the 1950s to 1980s). The most important thing happened after 1980s, is the period where globalization became the central policy of the United States, the democratic party has increasingly become a right-wing party(it used to the party in the middle, but after the disappearance of the US communist party, the current democratic party has been shifting to the right from the middle). The foreign policy of the United States changed dramatically as well compare to the 1930s. It used to be

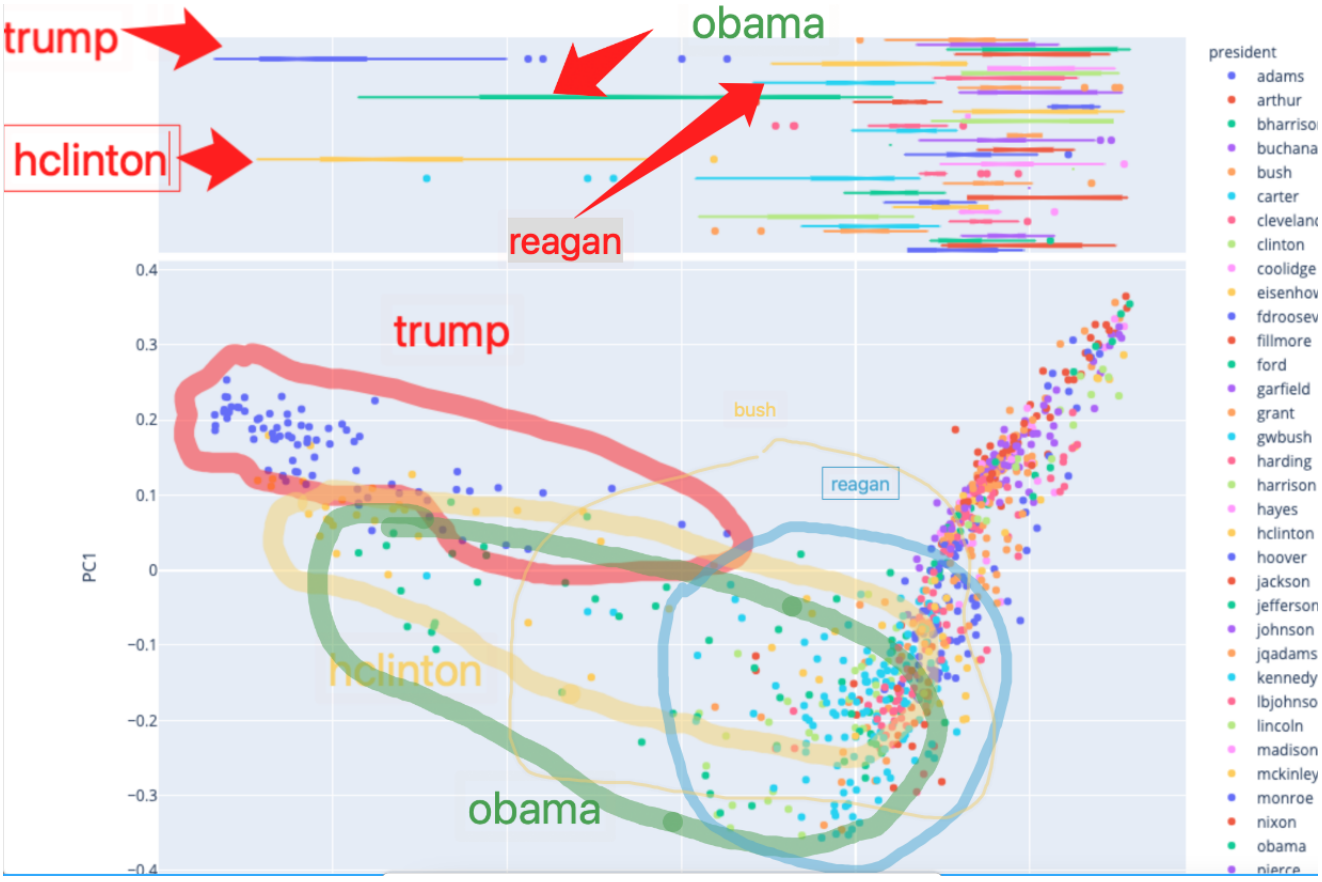
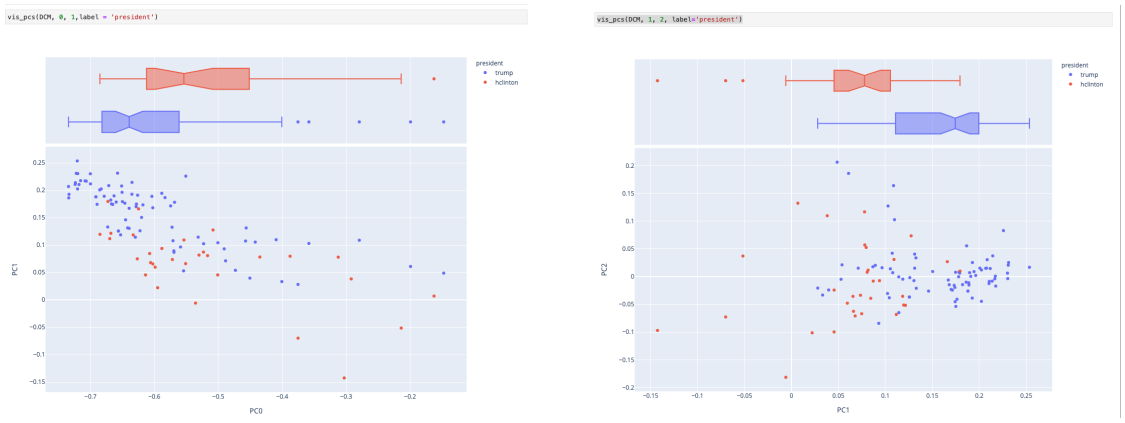


Figure 3: PCA for all presidents

characterized by isolationism, while after the 1980s, there is a rise of neo-liberalism in the US, the Democratic elite adopted the language of the left while embraced a right-wing economic doctrine start from President Ronald Reagan in 1981, and a increasing military overreach which shifted from isolationism to interventionism. I'm just trying to connect the important historical event to trying to explain the data, further study needed to ensure if this make sense.

6 Word Embeddings

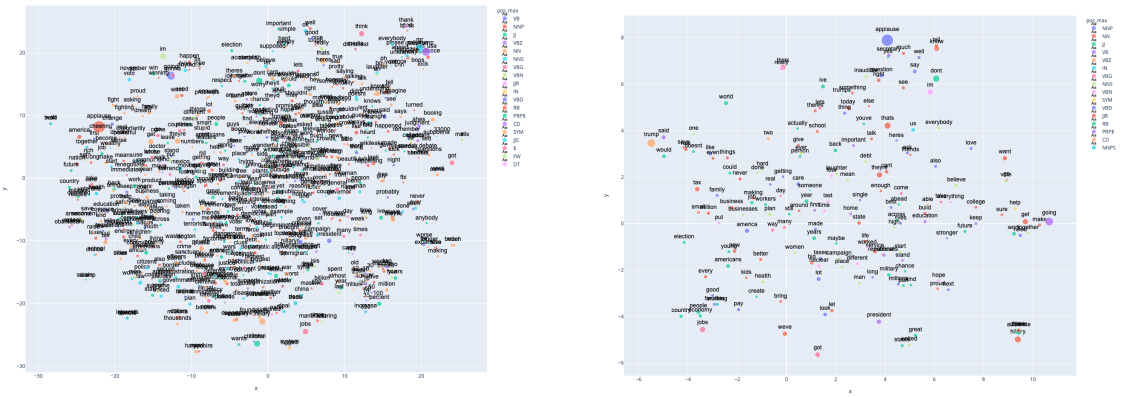
I also decided to perform a word embedding analysis, in this case, I utilized word2vec through Gensim Library. I first made two scatter plots by using Trump and Hillary Clinton. We can see that in both scatter plots. the word "applause" is significant (size is larger than any other word). I have no idea what's that means initially, after search inside the text files, I realized in this Corpus, every time if there is a large number of people "applause" during the speech, the text file will have "[applause]" in it. So through the both scatter plot we can see that both "applause" sizes are significantly large, we can interpret it as both candidates are extremely popular when they gave a speech to their supporters(left-right corner for Trump, top center for Hillary). Both names appear in each other's scatter plot with a large size, shows both talks about each other a lot in their speech, which also makes sense since they attack each other a lot.



(a) Trump vs Hillary Clinton, PC0 vs PC1

(b) Trump vs Hillary Clinton, PC1 vs PC2

Figure 4: Top 50 words, Trump vs Hclinton



(a) Trump

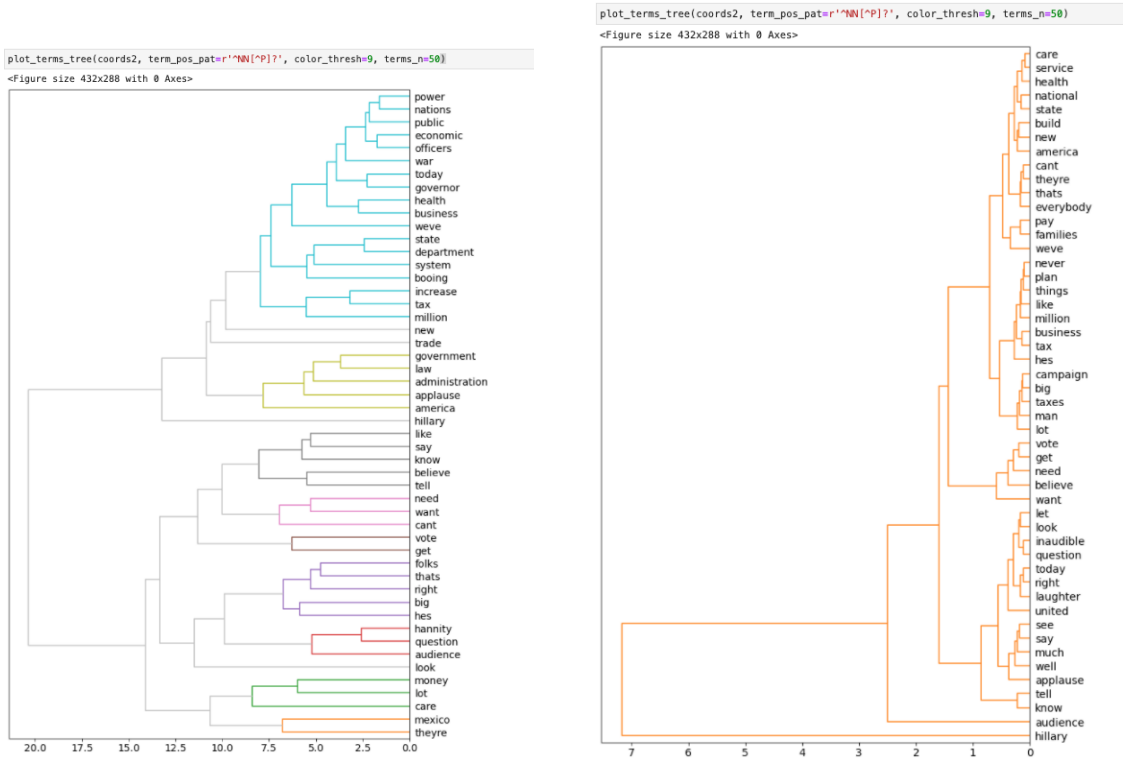
(b) Hclinton

Figure 5: Trump vs Hclinton

In order to analyze it more in detail, I chose Hierarchical clustering and display the top 50 significant words. We can see both Trump and Hillary focus on business and tax, the difference is Trump also talk about "mexico" "state department" "war" and "power" while Hillary talks about "health care", "plan", and "united". It fit my impression of them as well, Since we know Trump attack "mexico" a lot in his speech, moreover he also talks about the "state department" as a "deep state" in many interviews/speeches. While Hillary is just like a traditional democratic elite politician, shows a more elite style than Trump. Hillary talk about "right", "vote", "taxes", "campaign", "business", "women", "families", "pay", "state", she is more focus on the female voter than trump, and more towards those college-educated background voters.What's surprises me is that both Russia and China did not make it to the top 50. Then I increase from top 50 to top 100, now we can see China appears, but still not seeing Russia. I think the overall top 100 is a better choice for this particular situation because it covered the holistic picture of what's going on in their speeches.

We can see that the top 50 words used by Trump and Hillary Clinton.

Trump word:



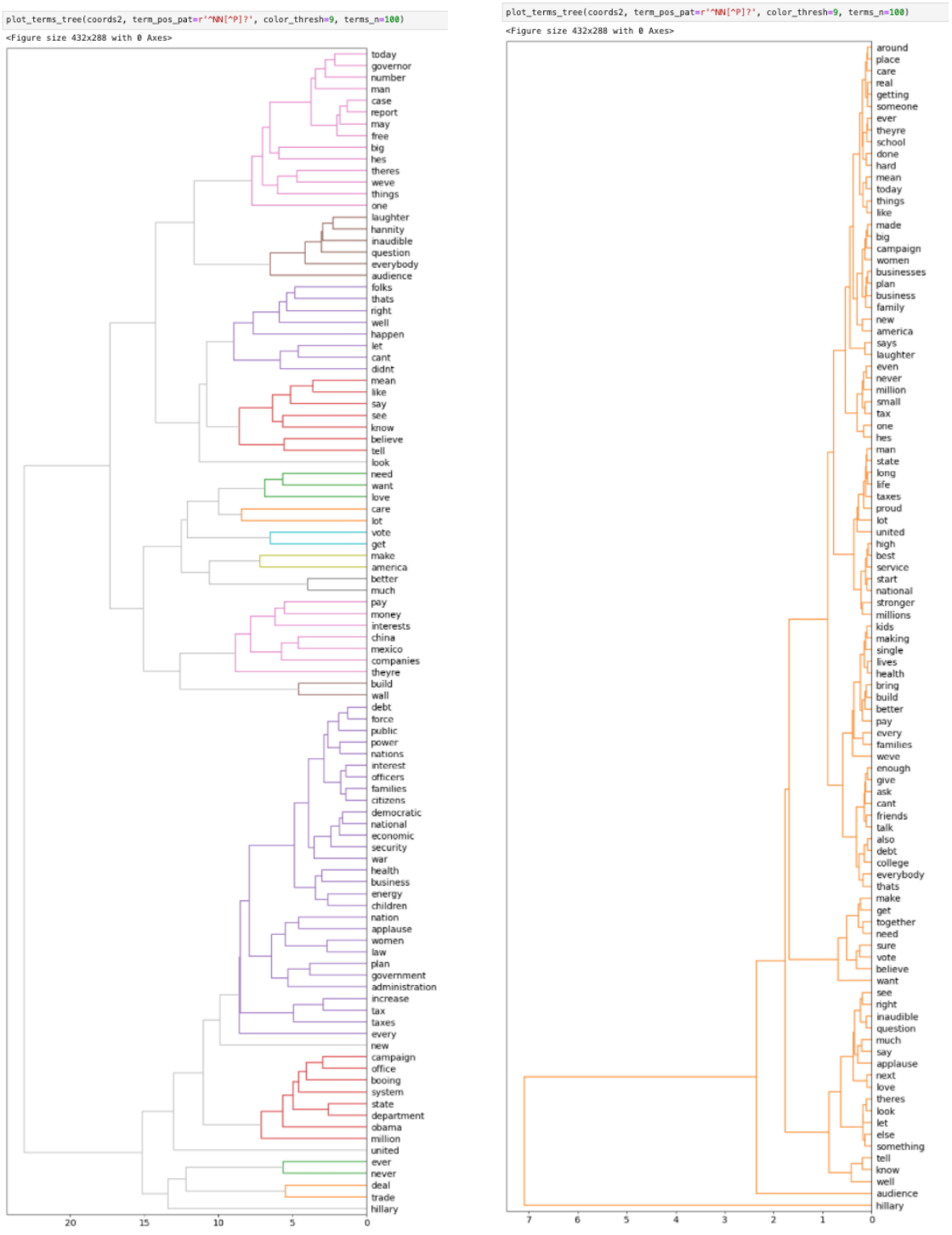
(a) Trump (b) Hclinton

Figure 6: Top 50 words(Noun,Proper noun), Trump vs Hclinton

- mexico
- money
- tax
- million
- state department
- business
- war
- power
- nation
- lot

hillary clinton:

- health care
- America
- build



(a) Trump (b) Hclinton

Figure 7: Top 100 words(Noun,Proper noun), Trump vs Hclinton

- plan
- business
- tax
- vote
- united
- million

7 Sentiment Analysis

Since we don't have a Corpus of how and what Trump and Hillary speak on everyday life.

Their speeches became the the best source we can analyze the emotion of them. By analyzing it by per speech through the date of their speech, we can see how their emotion changed over the presidential campaign through 2016 July to mid-Nov.

In Figure 8 and 9, we plug-ed in 4 emotions, trust, fear, joy, and polarity. the emotional fluctuation between Trump and Hillary are very similar, the peak and the fall often appear in the same period, since all those speeches are given during the presidential period, all those fall and peak between trump and Hillary are due to they are addressing the hottest topic which related to their campaign. The closer to the late stage of the campaign, the appears more peak on trust, joy, and fear.

The difference between Trump and Hillary in this chart, Trump's emotional fluctuation is more unstable compared to Hillary's. Hillary's "polarity" has been very stable from 2016.07 to 2016.11. During the speeches in 2016, Hillary shows a classical elite politician while Trump is more like a TV show host.

What happened in 2016/11? To the end of the election circle, especially we see there is a huge drop in Trust Fear and Joy on Hillary's sentiment at Nov 9th, while Trump has an increase of trust and fear at Nov 9th, the projected winner of the election, becoming president-elect. Then later Nov 11th, According to New York Time, "Donald Trump Is Elected President in Stunning Repudiation of the Establishment[3]."

- List of important dates which correlated to our sentiment analysis(Timeline of the 2016 United States presidential election. (n.d.). Retrieved May 10, 2021):
- July 21 – Donald Trump formally accepts the Republican nomination. - Increase in trust and joy,
- July 28 – Hillary Clinton accepts the nomination from the Democratic Party, becoming the first female presidential nominee of a major party in U.S. history. - Small increase in trust and joy(in july 29),
- September 26 – First presidential general election debate between the two major candidates was held at Hofstra University in Hempstead, New York. Hillary Clinton ends up taking the majority support after the debate. - a increase in trust and joy in sep 29 to sep 30

Figure 8: Sentiment analysis on Trump, by speech. The last day there is a hike, it is because He got elected as president elect

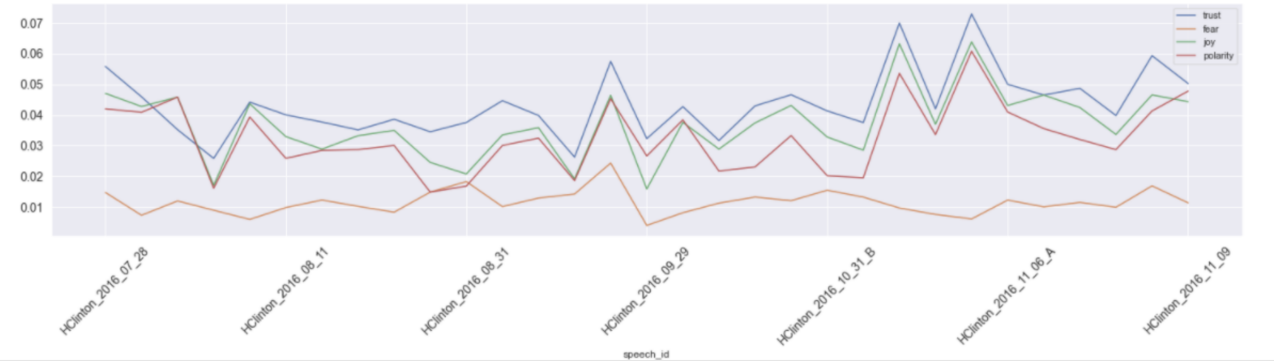


Figure 9: Sentiment analysis on Hillary Clinton, by speech. The last day there is a huge drop because she lost the election

- WikiLeaks begins publishing thousands of emails from the personal Gmail account of Clinton campaign manager John Podesta, revealing excerpts from Clinton's paid speeches to Wall Street. - Hillary doesn't have a speech from Oct 3 to Oct 24th because in October 7
- Second presidential general election debate was held at Washington University in St. Louis in St. Louis, Missouri.[163] Hillary Clinton ends up narrowly winning over Donald Trump. - October 9
- October 19 – The third and final presidential debate between the two major candidates was held at the University of Nevada, Las Vegas in Paradise, Nevada[163] Hillary Clinton ends up winning with a very close margin over Donald Trump. - Hillary had a increase of turst and joy in Oct 24th speech
- November 6 – James Comey tells Congress there is no evidence in the recently discovered emails that Clinton should face charges over handling of classified information. - decrease of trust and joy in Trump speech and increased trust and joy in Hillary.
- Nov 9, Donald Trump is the projected winner of the election, becoming president-elect. - increased in trust and joy in Trump and decrease in trust and joy in Hillary.

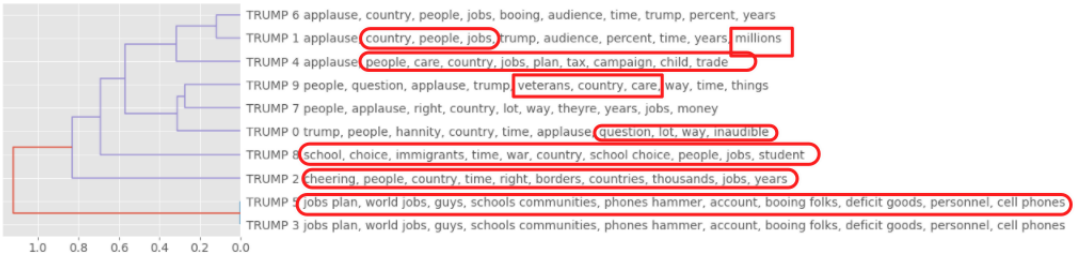


Figure 10: Clutser Topics, Trump by speech, he focus on topics more related to immigrants, war, borders and trade

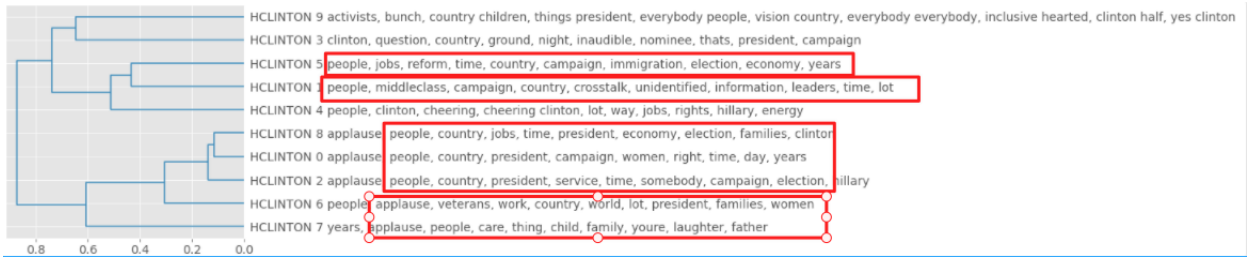


Figure 11: Clutser Topics, Hillary Clinton by speech, she focus more on topics related to reform, economy, family/women

8 LDA Topic Modeling

To see the topics of Trump and Hillary, I cluster the topic model (Figure 10 and 11), we can see that, both trump and Hillary like to talk about jobs, people, country, tax. The difference is Trump seems to be more appealing to working-class males and conservatives with topics related to "veterans", "war", "world jobs", "folks", "deficit goods". While Hillary's topics are more appearing to middle class and upper-middle class-related things such as "right", "women right", "reform", "immigration"

9 Conclusion

Through applying exploratory text analytic techniques to the Speeches of Trump and Hillary, I was able to see the cultural and linguistic behind all the speeches, and what's their main topics in their speeches, which help us how their speech can influence and attract the attention of the average American people. In addition to that, we are also able to connect it with the important event happened in the US history (need further study to confirm the connection). For further study, I think it will be very cool to compare all the presidents after WW II, and maybe to think of a way to compare all US presidents in a time series so we can see what's the cultural shift of the US president over the whole history of US.

References

[1] Timeline of the 2016 United States presidential election. (n.d.). Retrieved May 10, 2021

[2] An examination of the 2016 electorate, based on validated voters. (2020, September 22). Retrieved May 12, 2021, from https://www.pewresearch.org/politics/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/

[3] Flegenheimer, M., amp; Barbaro, M. (2016, November 09). Donald Trump is elected president in Stunning repudiation of the establishment. Retrieved May 12, 2021, from https://www.nytimes.com/2016/11/09/us/politics/hillary-clinton-donald-trump-president.html