# Analyzing Presidential Debates using NLP

**Introduction**:

During each Presidential election in the US, three Presidential debates are held in which the main candidates of the Republican & Democratic parties engage in debates. The topics discussed are often controversial, so we thought that we might stumble upon some interesting insights.

**Data Set:**

We analyzed the debates for 5 Presidential Elections, starting from 2000 up to 2016, so we have data for 15 debates. These debate transcripts were available on the American Presidency Project hosted at the University of California - Santa Barbara. Data pre-processing and cleaning took some time because usually we have a moderator presiding over the debate & 2 candidates' debate. Sometimes we might have 2 moderators or even members from the audience asking the candidates some questions. We had to go through the debate transcripts individually and then combine them at a candidate level as well as at a party level.

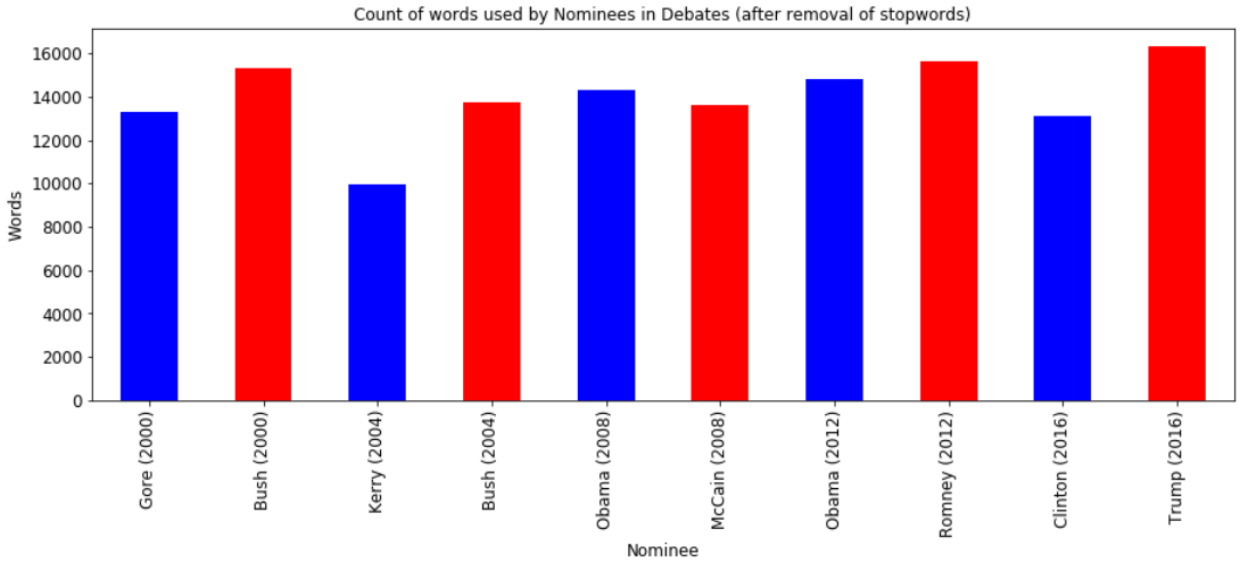
**Data Cleaning:**

The format of each debate transcript was different than the other. Some files had the candidate's title before their first name (e.g 'President', 'Senator', etc). Looking at the inconsistencies, we decided to convert all that in a proper format so that it is easier to extract specific text based on first names. Based on this pre-processing it was easier for us to compile each candidates' transcript for each year. This helped us to get a word count for each candidate for each year after removing appropriate stop words and unnecessary punctuations.

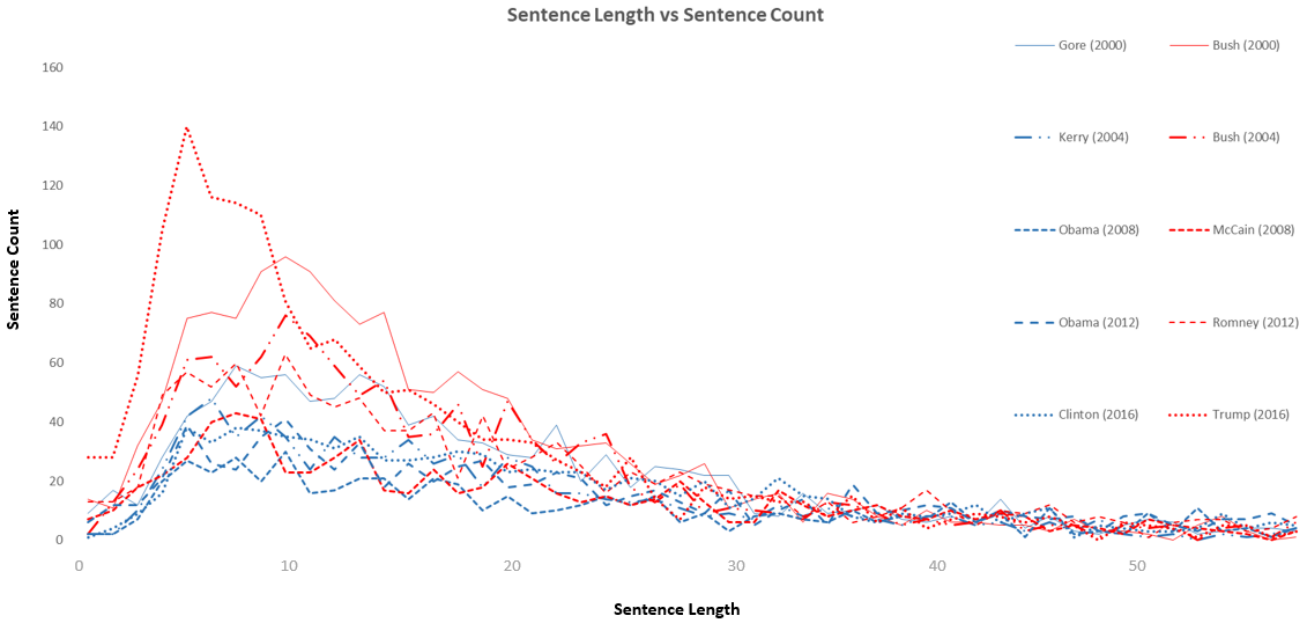
**Analysis**:

***Sentence, Word Length:***

On plotting the word count for each of the candidates, we saw an interesting statistic - In 4 out of 5 elections, the candidate who usually talks more wins. Bush, both in 2000 & 2004, Obama in 2008 & Trump in 2016 won the election. Romney, in 2012, though spoke more, lost.




Instead of just looking at word counts, we added another dimension to this analysis. For each candidate, we counted the number of words in each sentence and tried to create a graph that compares Sentence Length against Sentence Count. We were running short on time so we did this bit on Excel. On comparing with the previous graph too, we can see that Republicans talk more than Democrats. Also, they use shorter sentences more. The spike on the left, is for Trump. In the 3 debates, Trump spoke 140 sentences each with a length of 5 words.




***Pronoun Usage:***

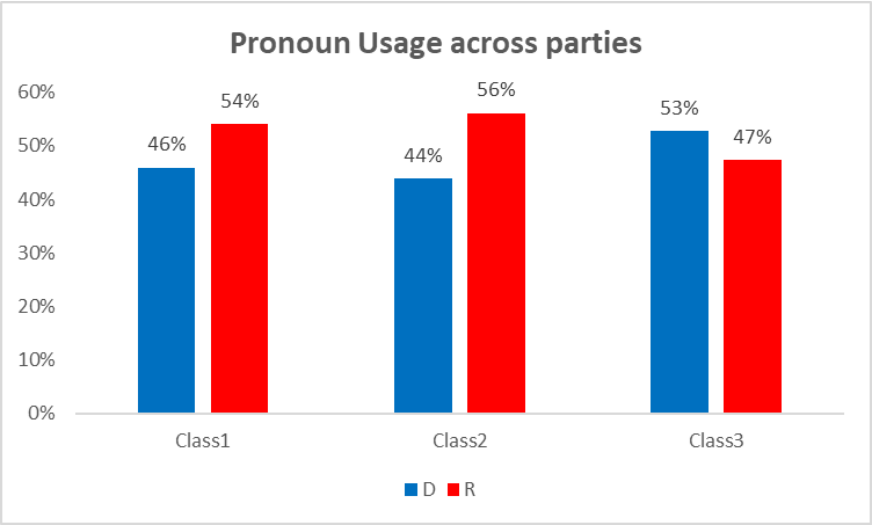
We also tried to do a bit of analysis on the usage of pronouns in the debates. We grouped the pronouns into 3 buckets. The first group had pronouns referring to oneself. The second one had pronouns referring to either an individual or a group of individuals & the 3rd was referring to everyone together as a group.

| Group | Pronoun |
|---|---|
| Class1 | I |
| Class1 | I've |
| Class1 | Me |

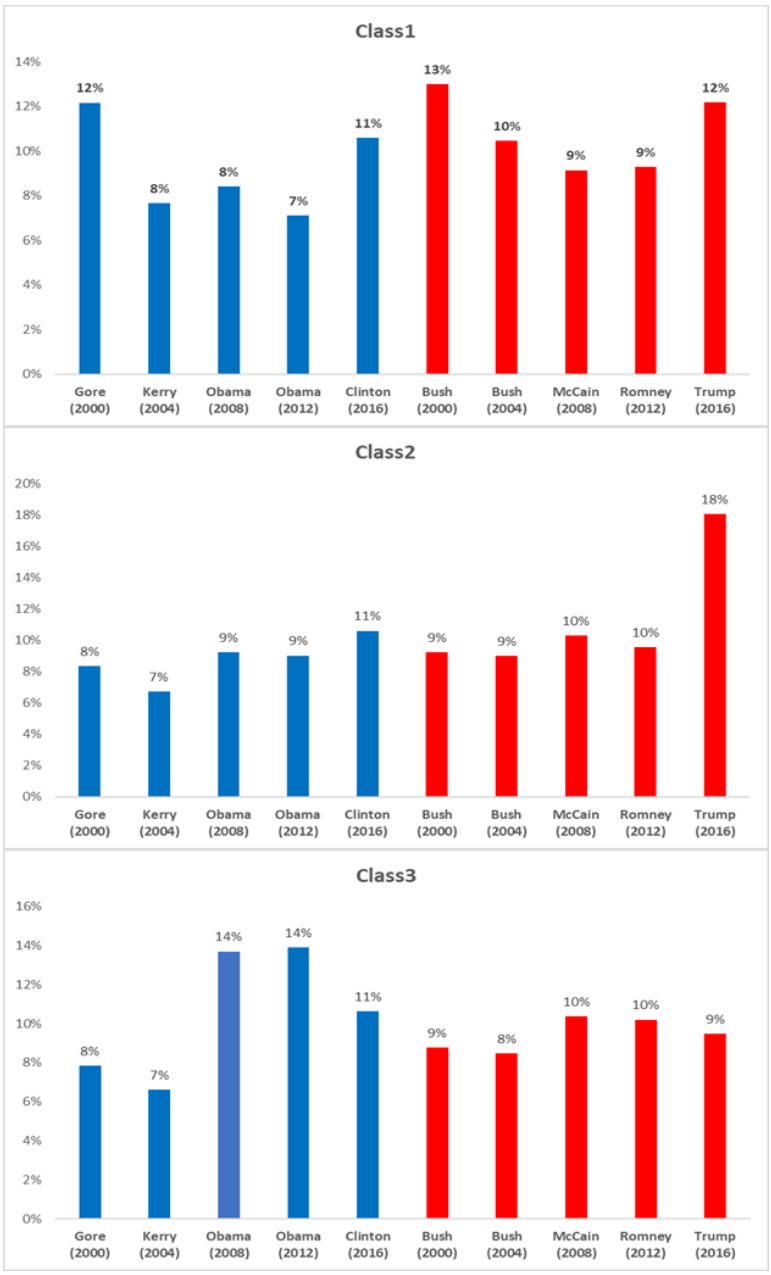
| Group | Pronoun |
|---|---|
| Class2 | He |
| Class2 | Her |
| Class2 | Him |
| Class2 | His |
| Class2 | She |
| Class2 | Their |
| Class2 | Them |
| Class2 | They |
| Class2 | They've |
| Class2 | You |
| Class2 | Your |

| Group | Pronoun |
|---|---|
| Class3 | Our |
| Class3 | Us |
| Class3 | We |
| Class3 | We've |

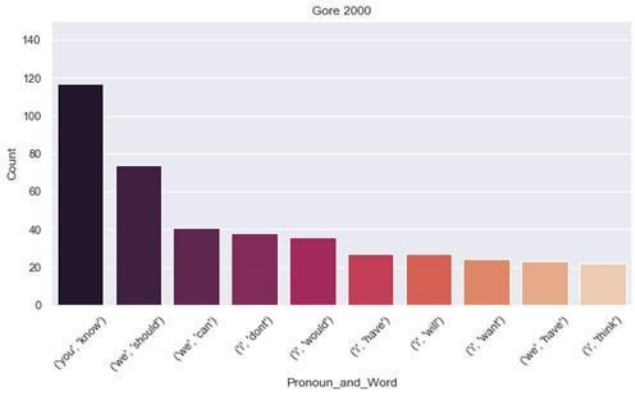
We saw that Republicans use more pronouns referring to themselves or other people or individuals. Democrats use words from group 3 more than Republicans.

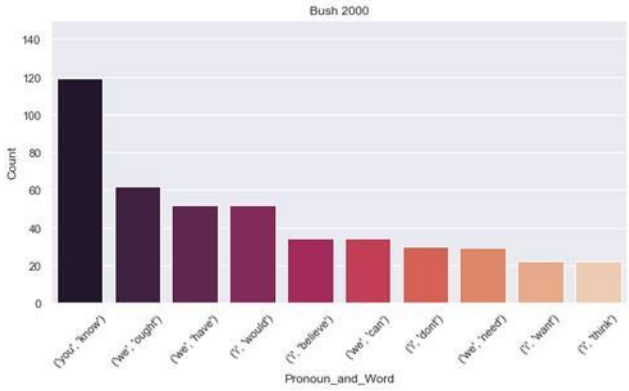


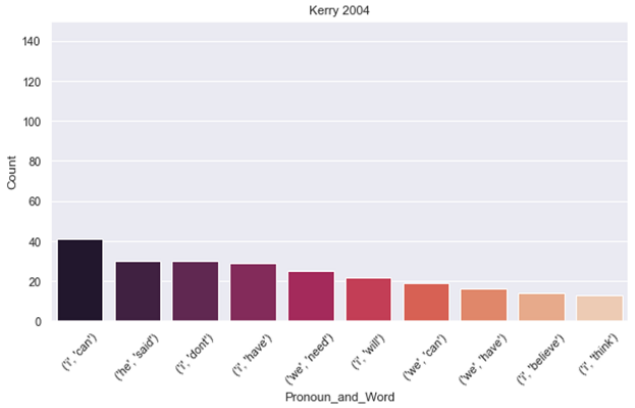

In the first chart we can see that Bush in 2000 & Trump in 2016 used pronouns from Class 1 the most. Trump used it most in Class 2. Obama in 2008 & 2012 used pronouns from Class 3 the most.

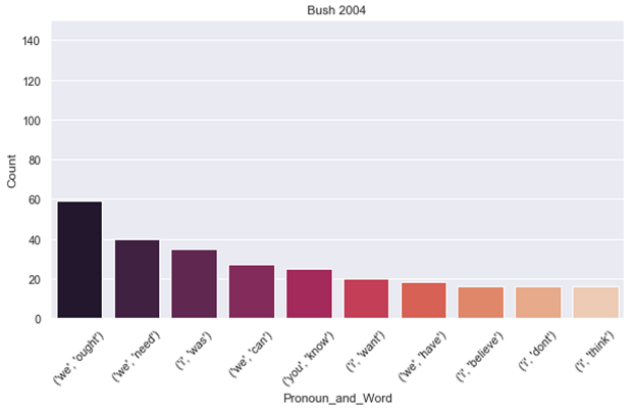


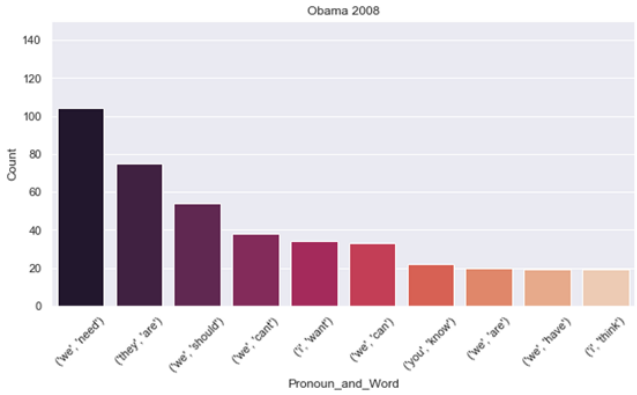

In order to have a better idea about the usage of pronouns, we created a function that finds the frequency of the top n bigrams and displays the results in a plot.

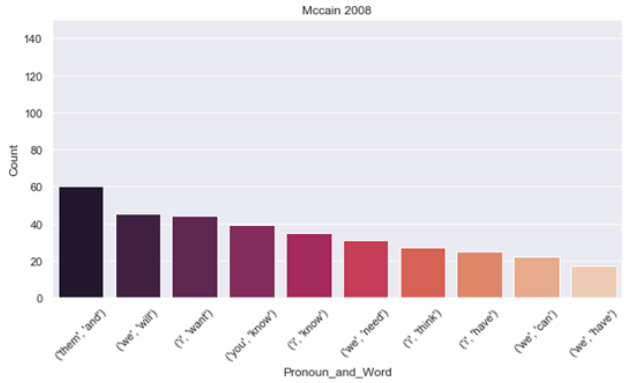

Gore 2000
Count
Pronoun_and_Word
('you', 'know')
('we', 'should')
('we', 'can')
('i', 'dont')
('i', 'would')
('i', 'have')
('i', 'will')
('i', 'want')
('we', 'have')
('i', 'think')

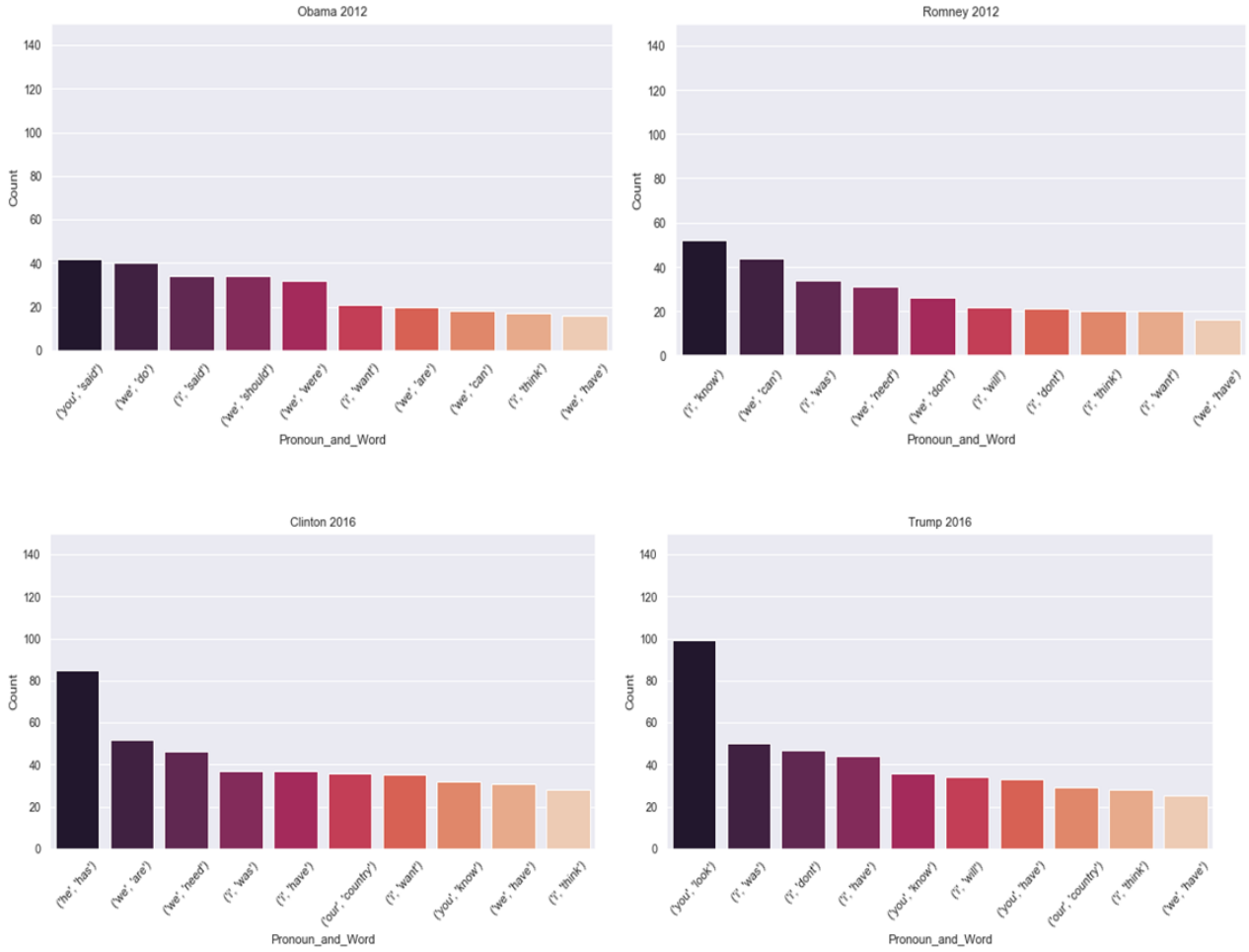

Bush 2000
Count
Pronoun_and_Word
('you', 'know')
('we', 'ought')
('we', 'have')
('i', 'would')
('i', 'believe')
('we', 'can')
('i', 'dont')
('we', 'need')
('i', 'want')
('i', 'think')


Kerry 2004
Count
Pronoun_and_Word
('i', 'can')
('he', 'said')
('i', 'dont')
('i', 'have')
('we', 'need')
('i', 'will')
('we', 'can')
('we', 'have')
('i', 'believe')
('i', 'think')


Bush 2004
Count
Pronoun_and_Word
('we', 'ought')
('we', 'need')
('i', 'was')
('we', 'can')
('you', 'know')
('i', 'want')
('we', 'have')
('i', 'believe')
('i', 'dont')
('i', 'think')

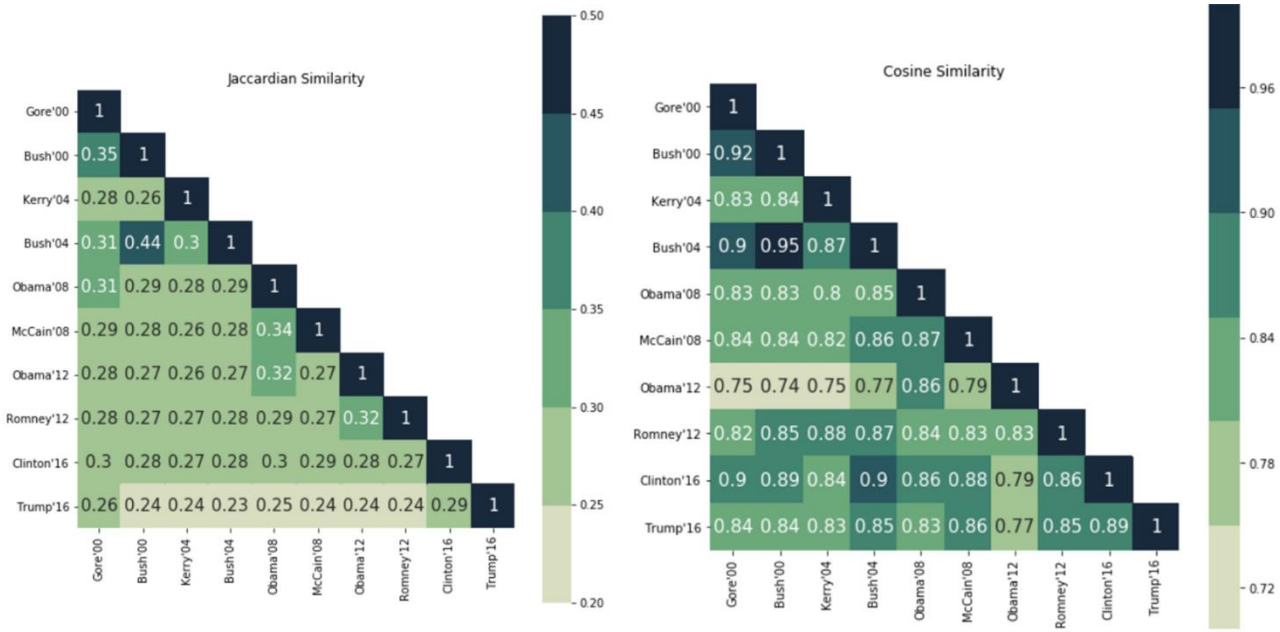

Obama 2008
Count
Pronoun_and_Word
('we', 'need')
('they', 'are')
('we', 'should')
('we', 'cant')
('i', 'want')
('we', 'can')
('you', 'know')
('we', 'are')
('we', 'have')
('i', 'think')


Mccain 2008
Count
Pronoun_and_Word
('them', 'and')
('we', 'will')
('i', 'want')
('you', 'know')
('i', 'know')
('we', 'need')
('i', 'think')
('i', 'have')
('we', 'can')
('we', 'have')




## ***Text Similarity:***

In order to check the similarity or dissimilarity of dialogues between each of the candidates, we calculated the Similarity Indexes. As we can see in both the indexes, Bush in 2000 & in 2004 had the highest index (i.e the context of his dialogues was largely similar (both during 2000 & 2004)). Also, Bush and Gore in 2000 discussed mostly the same topics & that explains the 2nd highest index across each of the indexes.




***Latent Dirichlet Allocation:***

To make the analysis more interesting, we implemented a Latent Dirichlet Allocation in order to identify the main topics for each candidate. We used the gensim package to preprocess and train the model. First, we slightly reprocessed the data before we proceed to extract the bags of words. Second, we used the corpora dictionary from gensim to create the bags of words. Finally, we trained our models to identify eight topics using LDA multicore for ten epochs on the exclusive bags of words.

Most of the topics identified from the model results were expected. The topics focused on many themes such as health (health insurance), economy (wage, growth, jobs, recession), security, immigration (border), international relations. We did not find any difference between the party affiliation because we merged all the debates into one document. Hence, one might consider to run the LDA model on the answers provided by each candidate after the question asked by the moderator. We also found that it was easier to depict the topics for most of the candidates but not

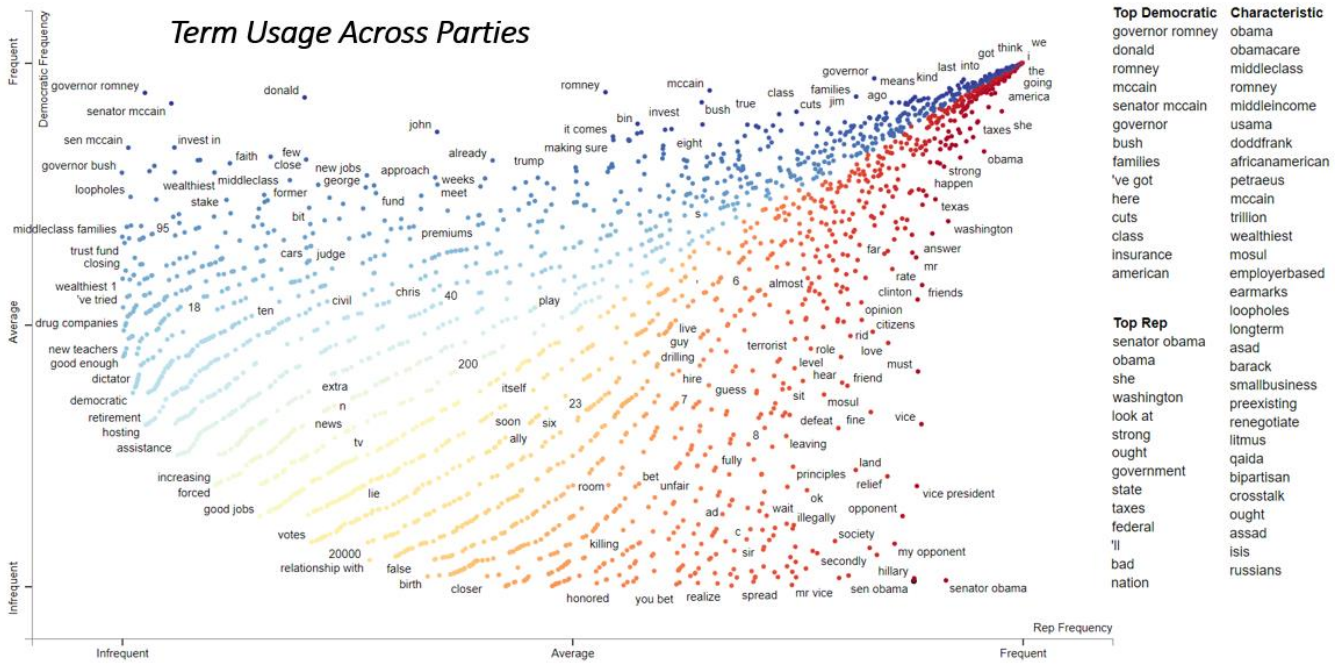
all. That might be an indicator of the level of coherence of the candidates agenda during the campaign.

```
'0.034*"health" + 0.020*"spend" + 0.017*"invest" + 0.017*"insur" + 0.016*"program" + 0.015*"budget" + 0.012*"feder" + 0.012*"peopl" + 0.012*"polici" + 0.012*"propos"'
```

Topic - Healthcare ?

***Scattertext:***

We used this package for finding distinguishing terms in our dataset, and presenting them in an interactive scatter plot with non-overlapping term labels. On one axis, we have the frequency of words being used by the Republicans & on the other, the Democrats. On the top right part of the chart, we see the words that are used in high frequency by both the parties.




***Tools:***

Major Python packages - re, NLTK, Gensim, Scattertext

***Work Distribution***:

*Data Collection & Preprocessing* :

Purvi Thakor, Akshay Kamath

*Data Analysis:*

- Word Count - Akshay Kamath
- Text Similarity, TF-IDF - Purvi Thakor
- Word Usage, Pronoun Usage, Scattertext - Akshay Kamath
- LDA - Junior Ovince

***Future Work***:

We can work on adding another segment to compare complexity of speeches. We can also have another bit where we can have a text similarity done on the questions asked by the moderators/audience & the answers given by the candidates.