Mason Bouffard
MIS3640-01: Assignment 2 Write up
Professor Li
10/21/20

Assignment 2 Write Up

Project Overview

For my text-mining project I pulled two articles from Wikipedia. These articles were the Donald Trump article (https://en.wikipedia.org/wiki/Donald_Trump) and the Joe Biden article (https://en.wikipedia.org/wiki/Joe_Biden#Climate_change). I used general analysis techniques in the form of word frequencies and summary statistics to give a quick comparison of the two articles. I also used NLTK to perform a sentiment analysis of the intro paragraphs of each article. I hoped to compare the wording of each article because with the upcoming election I figured there would be some linguistic differences in each of the articles.

Implementation

In terms of system architecture I first focused on loading the full articles off of wikipedia and into text files onto python. This way I could manipulate the document to create word frequency dictionaries in order to analyze the similarities and differences of the linguistics in both articles. As a result, I was able to compare the most common words in each article as well as the most common exclusive words to each article. This took form in multiple functions worked together to provide this information. I split up each analysis into different files, each pertaining to the specific article it was analyzing.

I then wrote two different files to analyze the sentiment of each of the intro paragraphs of the articles using the Natural Language Toolkit. I chose to use the intro paragraphs because Wikipedia usually does a pretty good job of covering the majority of the article in broad terms in the intro, so I figured the sentiment of the intro would be relatively representative of the whole article. I used the NLTK platform to first prep the introductions to be analyzed, and then

performed a sentiment analysis on each of the sentences of the intro. I also used the NLTK module to make a frequency distribution plot of the most common words in the intros. I decided that the overall goal of this portion of my program was to look at how negative each intro was to compare each article and look for bias towards one person or the other. To do this I also designed my program to make a dictionary of the more negative sentences so I could compare them.

When it came to deciding between different design alternatives, I think the biggest thing I struggled with is what exactly I wanted to perform sentiment analysis on and how it would affect the output. At first I was going to write a program that looked at the linguistics of the entire article and decided if it, as a whole, had a negative, positive, or neutral sentiment. I decided that this was too broad and didn't really show evidence of why the article had the sentiment that it did, so instead I decided to look at the introductions and break them up into sentences. This way I could look at the sentiment of each sentence and analyze what made the sentences negative and how that lends to the sentiment of the complete article.

Results

Overall, with the presidential election coming up, my goal was to look at how Wikipedia portrayed each candidate. My assumption was that Donald Trump would have a much more negative article just because of the nature of who he is and the fact that when it comes to politics Wikipedia has been known to lean a little more to the left than to the right. In terms of my text analysis and looking at summary statistics, I found that just looking at the words used in each article showed how the Trump article was more negative than the Biden article. When generating a list of the 20 most common words in the articles, I didn't find too many differences only because they were pretty general words. However, after I found the top 15 exclusive words to each article I found what I was looking for. If you look at the 15 most common words that are in

the Trump article and not the Biden one (See Exhibit 1), and compare those to the most common words in the Biden article that aren't in the Trumpa article (See Exhibit 2) it is pretty apparent that the linguistics in Trump's article focus on a lot more criticisms. The top exclusive word for the Trump article is Pandemic, and the fourth is Covid, so it is clear that there are criticisms on his mishandlings of the virus. Also in Trump's top 15 are words like Racist, Obstruction, Mueller, and Iran which all signal criticisms of Trump about the Mueller Report, him being a racist, and his handling of the Iran situation. If you then look at the exclusive words to Biden, there are no negative words at all except maybe the name of his son because I would assume there is criticism surrounding that. So just by looking at the most common words in each article gave me a baseline in showing that Trump's article is much more negative than Biden's.

Next, I performed sentiment analysis on each of the introductions to the articles and I found similar results. First, I generated a frequency plot of the 20 most popular words in each intro, these can be found in Exhibit 3 and Exhibit 4. After doing so it was clear that I would have to dive in deeper to the intros to get any good results, because as seen in the graphs the most common words don't really tell anything about the paragraphs. I then performed a sentiment analysis on all the sentences of each intro to look at how the NLTK program rated each sentence. Since the introductions were relatively long, and I wanted to look at which one had a more negative sentiment, I then generated only the sentences that had a negative rating above .2 so I could look at the most negative sentences. This gave me the output shown in Exhibit 5 and Exhibit 6. When looking at these sentences it is clear that the Trump intro is much more critical than the Biden one. The biden negative sentences aren't even negative towards Biden, they just have negative words in them which probably is what makes the program mark them as "negative". But each of the Trump sentences are extremely critical of how he is labeled and the

decisions that he has made. This also went to prove that there is definitely a more negative feeling to the Trump Wikipedia article in comparison to the Biden one.

Reflection

I think overall the project ended up going better than I thought it was going to. I was pretty overwhelmed to begin with and didn't really know where to go after getting the articles. That being said, I think that there are definitely a lot of flaws in what I did that could be improved. If I had more time to work on this project, I would definitely have focused more on removing pointless words. Each program I used to clean up the words worked to an extent, but my results would always end up having words in them that I didn't think were worth looking at. Also, I wish I could've spent more time wrapping my head around the NLTK module because I think it could have given me a lot more insight in the text than I used it for. It took me long enough to figure out what I did, but reading the documentation website showed me that it could do a lot of things I didn't have the time to wrap my head around. I would say that for the time I had and all the other work I had going on, I did fine. But if I had more time and less other work I would have been able to write a much more quality program. Going forward, if I ever plan to write a detailed program I would give myself more than enough time because problems always come up.

Exhibits

Exhibit 1: Top Exclusive Words to Trump Article

Word:	Frequency:
pandemic	29
19	25
–	23
covid	23
statements	20
business	19
organization	17
fbi	16
order	15
iran	15
mueller	14
coronavirus	14
businesses	14
racist	11
obstruction	11

Exhibit 2: Top Exclusive Words to Biden Article

Word:	Frequency:
delaware	40
busing	18
amendment	14
wilmington	12
hearings	12
catholic	11
head	10
beau	10
r	9
joseph	8
ticket	7
syracuse	7
sponsored	7
scranton	7
mccain	7

Exhibit 3: Frequency Distribution of Biden Intro 20 Most Common Words

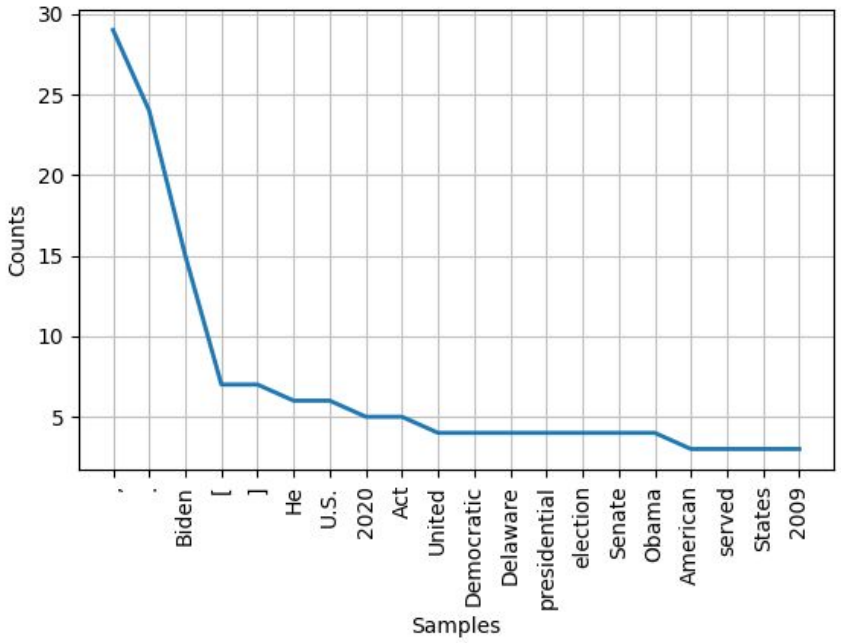

Exhibit 4: Frequency Distribution of Trump Intro 20 Most Common Words

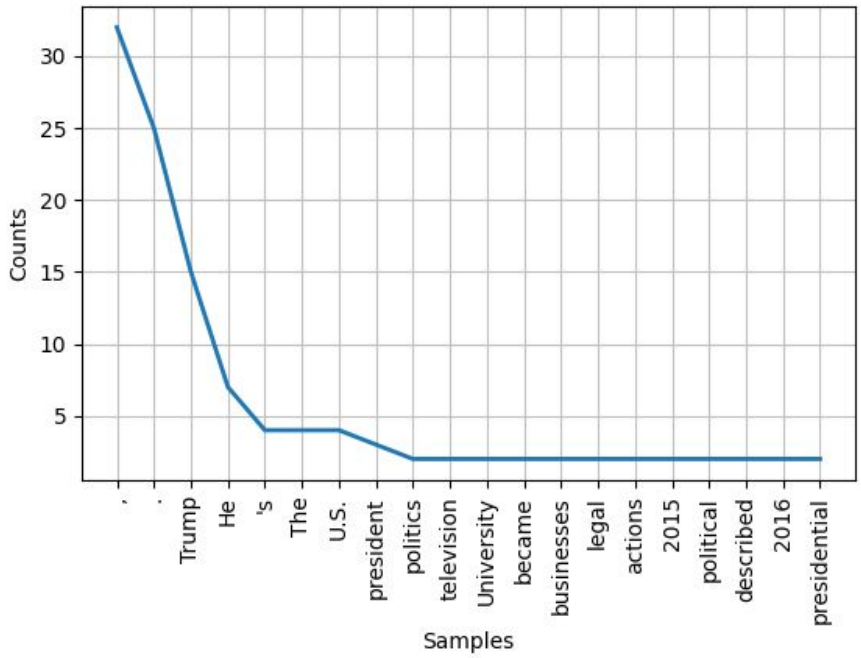

Exhibit 5: Trump Intro Sentences with Neg Sentiment Value over 0.2

{'**His election and policies have sparked numerous protests**.': {'neg': 0.213, 'neu': 0.787, 'pos': 0.0, 'compound': -0.2263}, **'Many of his comments and actions have been characterized as racially charged or racist.**': {'neg': 0.326, 'neu': 0.674, 'pos': 0.0, 'compound': -0.7003}, '**Trump reacted slowly to the COVID-19 pandemic; he minimized the threat, ignored or contradicted many recommendations from health officials, and promoted false information about unproven treatments and the availability of testing.'**: {'neg': 0.212, 'neu': 0.714, 'pos': 0.074, 'compound': -0.6369}}

Exhibit 6: Biden Intro Sentences with Neg Sentiment Value over 0.2

{**'He opposed the Gulf War in 1991 but supported the expansion of the NATO alliance into Eastern Europe and its intervention in the Yugoslav Wars of the 1990s.'**: {'neg': 0.208, 'neu': 0.708, 'pos': 0.084, 'compound': -0.6597}, **'Biden led the efforts to pass the Violent Crime Control and Law Enforcement Act and the Violence Against Women Act, and oversaw six U.S. Supreme Court confirmation hearings, including the contentious hearings for Robert Bork and Clarence Thomas.'**: {'neg': 0.272, 'neu': 0.656, 'pos': 0.072, 'compound': -0.8779}, **'Following the Sandy Hook Elementary School shooting, Biden led the Gun Violence Task Force, created to address the causes of gun violence in the United States.'**: {'neg': 0.344, 'neu': 0.529, 'pos': 0.127, 'compound': -0.8481}}