Xintong Song, Ivy Zhang

Professor Li

MIS 3640

26 March 2020

Assignment 2: Text Analysis Project

1. Project Overview

To explore different data analytic techniques and get ourselves familiar with multiple data sources, we imported data from Wikipedia and Twitter. From Wikipedia, we extracted information under two different but related keywords. This allowed us to generate summary statistics for two sets of data and compare them side by side due to their relativity. For Twitter, we looked up the keywords related to a current event and applied sentimental analysis to understand the society's (more specifically Twitter users') general reaction to the event.

2. Implementation

Wikipedia

There are several major components to the analytics of the Wikipedia text. First, in preparation for the code, we downloaded the mediawiki package, imported it along with other necessary packages. Secondly, we extracted the information from Wikipedia and processed it using a function called *process_file()*. This function takes the extracted, unformatted content from Wikipedia, which is a string, cleans it (leaving only the words), and provides a histogram (dictionary) which has each unique word as key and the number of occurance as value. Then, we filtered out commonly seen English words in the *stopword* text file so that all left-over words are related to the matter itself, and we created a list with reversed order of keys and values in the dictionary, which then helped us get the 10 most frequently-used words in the Wikipedia text. Having these major functionalities, we ran the program twice, first by using "patriarchy" as the keyword and second by using “feminism” as the keyword. By having the most common words in these two texts, we hoped to compare across them and understand quickly how these two concepts might differ. We also attempted to use sentimental analysis to understand whether the authors of these two texts hold any biases; however, the tests on both articles show a relatively neutral writing style. Therefore, we concluded that sentimental analysis is more suitable for analysis of statements and comments as on Twitter.

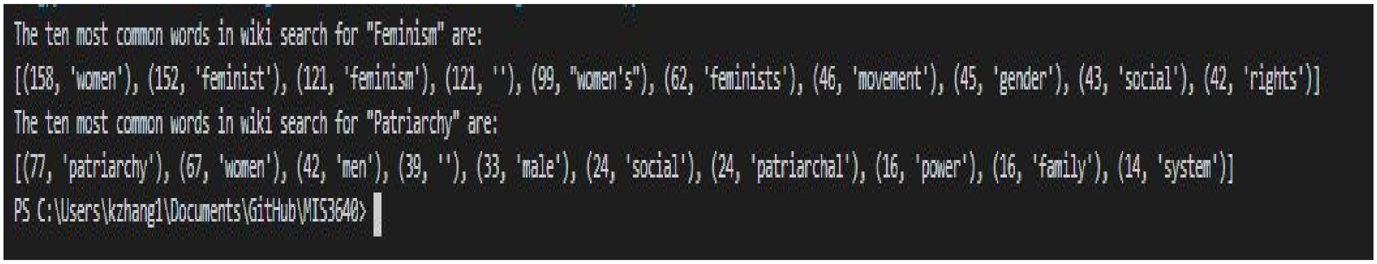
Twitter

First of all, we authorized a new application from Twitter and then retrieved data from Twitter’s API. We did some research and picked Tweepy Python library because of good reviews of it as an easy-to-use tool for accessing Twitter API. In addition, we imported *sys and re* modules for helping interact more functions and supporting regular expression through libraries, *csv* module for managing the csv file, *textblob* module for showing sentimental analysis results, and *matplotlib.pyplot* module for creating interactive visualization. Moreover, we used object-orientated programming approach, which is written in the form of the self-defined object and class. The class *SentimentAnalysis* is our blueprint for the object. Then we wrote a number of class methods by creating an object called *self*, which contains two empty lists to store data.

3. Results

Wikipedia

The output for the text analysis of Wikipedia search of “Patriarchy” and “Feminism” is as such:

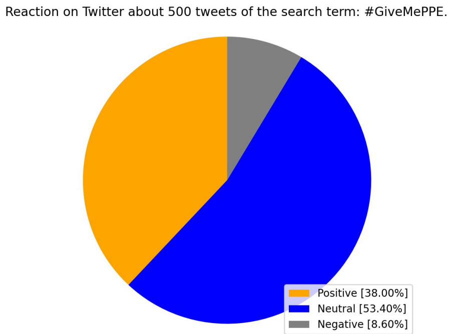


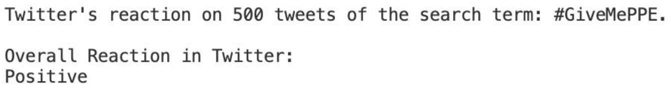
```
The ten most common words in wiki search for "Feminism" are:
[(158, 'women'), (152, 'feminist'), (121, 'feminism'), (121, ''), (99, "women's"), (62, 'feminists'), (46, 'movement'), (45, 'gender'), (43, 'social'), (42, 'rights')]
The ten most common words in wiki search for "Patriarchy" are:
[(77, 'patriarchy'), (67, 'women'), (42, 'men'), (39, ''), (33, 'male'), (24, 'social'), (24, 'patriarchal'), (16, 'power'), (16, 'family'), (14, 'system')]
PS C:\Users\kzhang1\Documents\GitHub\MIS3640>
```

As we can see, neglecting neutral, descriptive nouns like “men,” “women,” “gender,” and etc., we found the word “rights” specific to “Feminism.” In fact, by looking up the key “rights” in Patriarchy’s histogram, it is shown that the word has only been used once in Patriarchy, much fewer than 42 times in Feminism. Similarly, we can see “power” and “system” being used multiple times in the Patriarchy text. Interestingly, just by looking at these words, we can already sense the dynamics and general content of the text. Thus, simple text analysis can help us grasp the overall theme of a large chunk of text in seconds as demonstrated in this project.

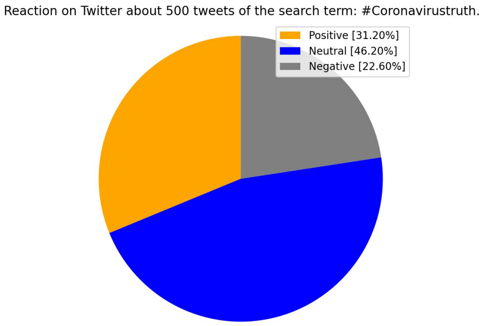
Twitter

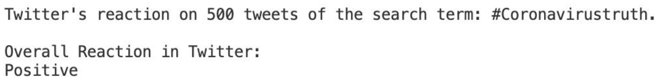
We asked our user to enter the search term or hashtag that they would like to know and enter the number of tweets wanted to be analyzed. Under current national coronavirus lockdown, we were curious about how people think about this unprecedented global pandemic. Therefore, we selected COVID-19 related words in trending topics on Twitter to run sentiment analysis in random 500 tweets and then plotted them in pie charts.



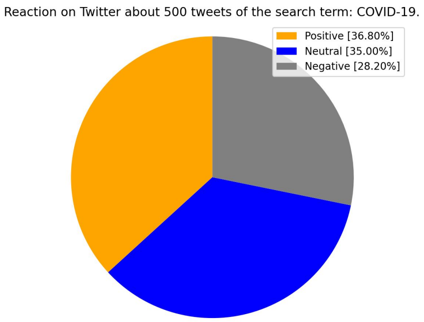


```
Twitter's reaction on 500 tweets of the search term: #GiveMePPE.

Overall Reaction in Twitter:
Positive
```

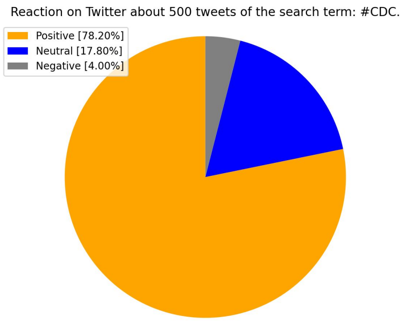




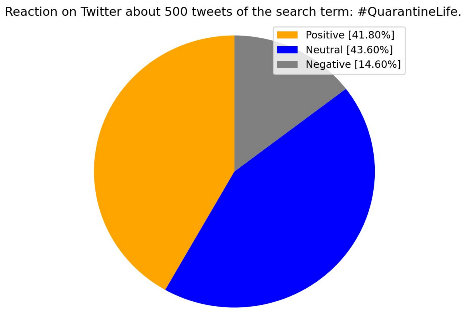
```
Twitter's reaction on 500 tweets of the search term: #Coronavirustruth.

Overall Reaction in Twitter:
Positive
```



```
Twitter's reaction on 500 tweets of the search term: COVID-19.

Overall Reaction in Twitter:
Positive
```



```
Twitter's reaction on 500 tweets of the search term: #CDC.

Overall Reaction in Twitter:
Positive
```



```
Twitter's reaction on 500 tweets of the search term: #QuarantineLife.

Overall Reaction in Twitter:
Positive
```

In conclusion, we were glad to see overall reaction for all five search terms above are positive. Compared to most traditional news media, such as newspapers and TV channels, who are telling people how bad current situations are, most tweets remain positive and neutral. We think this is a good sign because mental health is just as important as physical health around this period of stressful time.

4. Reflection

We both felt unsure and challenged at the beginning because we did not know what sources we should select and what kind of result we would generate. Since COVID-19 is in the air, we chose to avoid face-to-face communication. Then, we realized how hard it was to work on a group project without the benefit of looking at the same screen and interpreting meaning from body language, facial expression, and tone of voice. The miscommunication between us held us back from making progress and moving forward. Thus, we used a different approach by working on our own and then helping each other figure out what was going on. There are two things which we thought we did well on. Firstly, we ensured continuous communication throughout the project, which enabled us to pace ourselves and offer help when the other is struggling. Secondly, we frontloaded the task-splitting part and made key decisions early in the process together, so that the two parts of the project come together as a complete and complementary whole. This approach actually worked out greatly at the end because now both of us can look at each other's codes and we essentially learned about extracting information from both sources this way. Lastly, thank you Professor Li for giving us more time to work on this project.