Meiling (Olivia) Qi, Myat Shwe Yee Maung

Text Mining Assignment

Professor Li

October 30, 2019

Assignment #2 Report

1. Project Overview

For this project, we chose Twitter as the data source and analyzed tweets that contain #andrewyang. The overall goal was to see what words people use when they talk about the presidential candidate, Andrew Yang, and how they feel about him. We also hoped to learn web scraping, data cleaning, text and sentiment analysis through the utilization of various packages such as tweepy and nltk.

2. Implementation

2.1 Web Scraping

The first step was to determine which data sources we wanted to learn to scrape from and analyze. Initially, we chose IMDB, but due to problems that we were unable to troubleshoot, we settled on using Twitter as a data source. In the *web_scraping.py* file, we used the tweepy package to scrape 1000 tweets containing #andrewyang since October 1, 2019 to only get the text of the tweets (excluding username and retweets). The retrieved tweets were stored into a list for easier processing and cleaning. The scrubbing step was crucial as we got rid of any special characters and URL included in the tweets. We then saved all the cleaned tweets into a text file, on which we conducted further text analysis and sentiment analysis.

2.2 Text Analysis

In the *analyze_text.py* file, we implemented a text analysis consisting of word frequency, unique words, average words, and total words used. During the first iteration, we noticed the most common words were filler words such as ‘the’, ‘is’, ‘to’, ‘at’ as well as words related to Andrew Yang. From further research, we were able to remove these stopwords using the nltk package. We also created a list containing words such as ‘Andrew Yang’, ‘YangGang’ , ‘Yang2020’ to make sure the text analyzed exclude the words in the list, yielding results with more relevant insights.

2.3 Sentiment Analysis

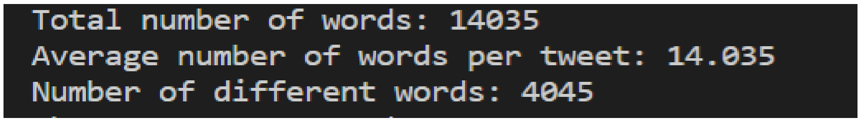
We were first deciding between conducting a Sentiment Analysis or Markov Text Synthesis. As the overall project goal is to see how people feel about the presidential candidate Andrew Yang, we eventually pursued Sentiment Analysis as it better fits our needs.

Since we previously stored all the tweets into a separate text file, we converted it to a list with each tweet being an item for further analysis. In the code construction process, we took advantage of the built-in Sentiment Intensity Analyzer function within the nltk package. A score is calculated for each tweet using the for loop, and the specific compound score generated is used to determine the positive, neutral or negative sentiment. We then added up each of these tweets and divide them by the total of 1000 tweets to get the percentage of each type of sentiment. To better communicate our findings, we created a pie chart to visualize people’s attitudes.

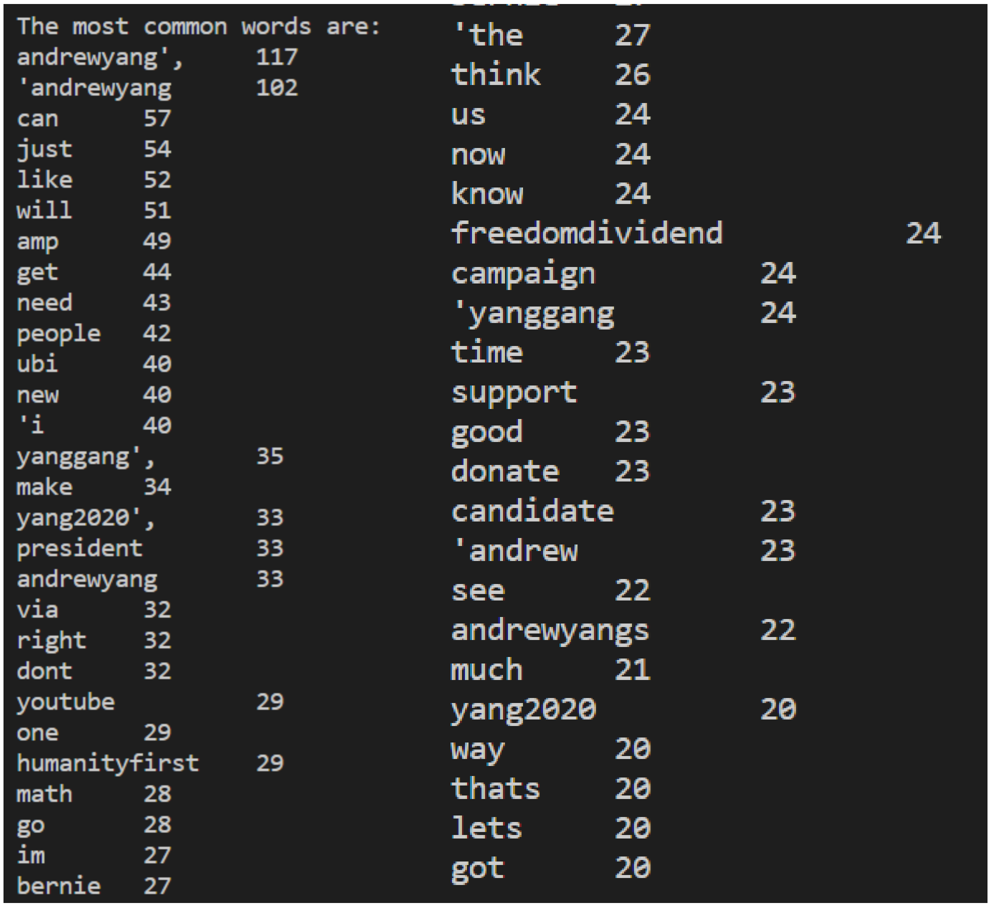
3. Results

3.1 Text Analysis Insight

Looking at the results of the text analysis, the average words per tweet is 14.035, the total number of words is 14,035 and there are 4,045 different words. In addition, the top three most common words are “andrewyang’”, “‘andrewyang”, and “can” with respective counts of 117, 102, and 57. The first two words appeared because we were not expecting “ ‘ ” as part of the string. In addition, the word “amp” occurred 51 times because *&amp;* is an HTML markup for “&”. Due to the scrubbing of special characters “&” and “;”, only “amp” remained. However, other frequent words such as “humanityfirst”, “freedomdividend”, “support”, “math”, “good”, and “donate” gave us useful insights on what people are associating with Andrew Yang’s campaign. We could infer that there are positive sentiments, but to be sure, we also ran a sentiment analysis.



```
Total number of words: 14035
Average number of words per tweet: 14.035
Number of different words: 4045
```

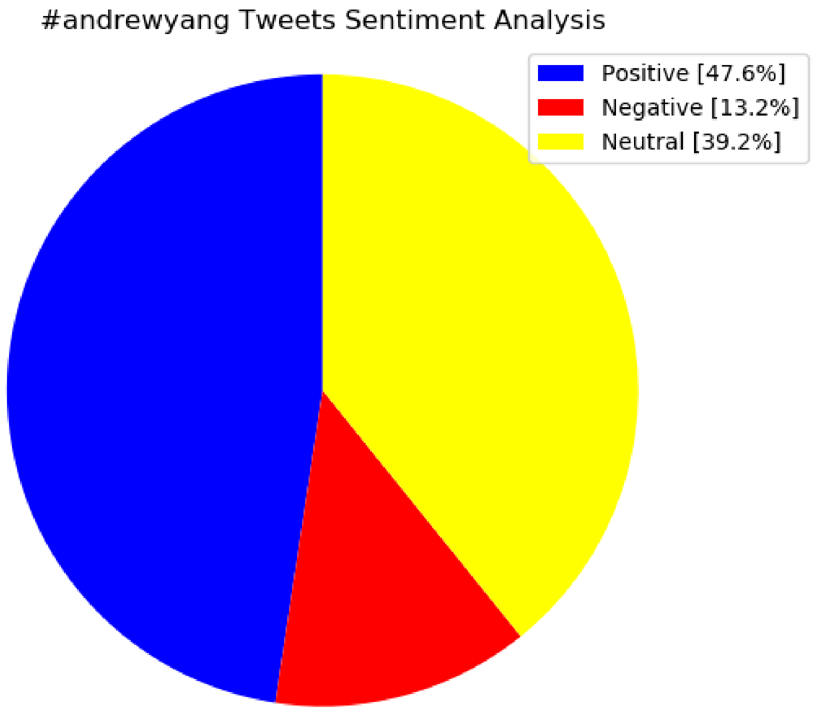


```
The most common words are:
andrewyang',      117
'andrewyang       102
can       57
just      54
like      52
will      51
amp       49
get       44
need      43
people    42
ubi       40
new       40
'i        40
yanggang',        35
make      34
yang2020',        33
president         33
andrewyang        33
via       32
right     32
dont      32
youtube           29
one       29
humanityfirst     29
math      28
go        28
im        27
bernie    27
```

```
'the      27
think     26
us        24
now       24
know      24
freedomdividend           24
campaign          24
'yanggang         24
time      23
support           23
good      23
donate    23
candidate         23
'andrew           23
see       22
andrewyangs       22
much      21
yang2020          20
way       20
thats     20
lets      20
got       20
```

3.2 Sentiment Analysis Insight

From conducting sentiment analysis on the 1000 tweets, we found that 476 of them have positive sentiment while 132 tweets displayed negative sentiment, the rest are neutral. This means that between Oct.1st and when we finalized our code (Oct.30th), 47.6% of Twitter users who used #andrewyang showed positive attitude towards the candidate, while respectively 13.2% and 39.2% of them conveyed negative and neutral sentiment. We believe our analysis could be helpful for presidential candidates such as Andrew Yang to determine how the public feels about them. If a larger sample size is adopted, the result would be even more helpful for further reference.



4. Reflection

We believe the overall project went well because we effectively distributed the tasks and communicated learning outcomes with each other. Myat was responsible for web scraping and text analysis while Olivia was responsible for conducting sentiment analysis. When we encountered difficulties, we would ask one another, consult the professor and conduct more research to understand how to construct the code. Throughout the process, we gained a better understanding of how to scrape data from Twitter using the tweepy package, how to process natural language using the nltk package, and visualize data for better interpretation.

With that being said, there are also several areas for improvement. First of all, we could clean the text more thoroughly and get rid of all the symbols so that we could present better text analysis. This is currently beyond our capability but we will continue to improve our skills for the next project. Secondly, we could create more visualizations such as plotting the word frequency. We believe such a graph could help the viewers better understand the results of our analysis.