Jerry Lee, Aryan Jain
Problem Solving and Software Design
Professor Li
10/30/2016

Analysis of Trump and Clinton’s Tweets

Project Overview:

For our project, Aryan and Jerry used Twitter as our data source. From Twitter, we wanted to compare tweets about Hillary and Trump to find out how the public viewed the candidates. We used ‘Twython’ to gather this data, which allowed us to use the Twitter API and collect 100 tweets from both presidential candidates. Once the 200 tweets were gathered, we used Professor Zhi Li’s method to pre-process the data. Summary statistics, word frequencies and sentiment analyses were done to analyze the differences in the tweets.

Implementation:

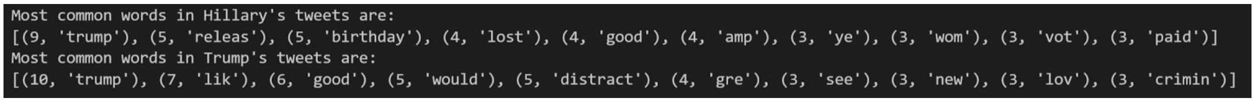
To implement the text analysis, tweets were gathered using Twython and were exported using Pickle. Twython uses a Twitter API to import 100 tweets given a specific query. The queries used were ‘@HillaryClinton’ and ‘@realDonaldTrump’ and a filter was added to remove retweets because a majority of the tweets resulting from these queries were just retweets from Clinton and Trump’s twitter profiles. What we wanted to analyze was the society’s sentiments towards Trump and Hillary. Once the tweets were gathered in the form of a list of string tweets, Pickle was used to export both candidates’ tweet data into two .pickle files.

Before the analyses starts, the data must be preprocessed to remove unnecessary white-spaces, punctuation, symbols and stop words. Furthermore, we only wanted to work with dictionary words and not slang for the purposes of this project so we removed any words that didn’t appear in the Oxford dictionary. Finally, it was important to extract the root of each word by lemmatizing it. This is so that in doing a word to word comparison, verb conjugations don’t make the words seem different. E.g., we wanted “watch” and “watched” to be treated as the same word. The result of this pre-processing was a list of stemmed words.

In order to analyze the data, we wanted to explore three possibilities. First, we wanted to test the hypothesis that the most common used words in the tweets pertaining to Trump and Hillary will be different but will have some common terms such as election. In order to do this, we created a dictionary histogram of the words as they appeared in the list. The keys to this dictionary were the unique words appearing in the processed list of tweets. The values were the frequencies of these words. Then we created a function to display the top 10 most frequently occurring words in each of these dictionaries. This way, we could compare the most frequently used words by users tweeting about Trump or Hillary. Second, we wanted to check the average user sentiment in a tweet directed to Trump vs. the average user sentiment in a tweet directed to Hillary. So we used the NLTK sentiment intensity analyzer function to obtain a summary of sentiment intensities for each tweet. After compiling these summaries to a list of dictionaries, we were able to compute the average ‘positivity’ and ‘negativity’ intensities for Trump tweets and Hillary tweets. Lastly, we wanted to compare the sentiment intensities for Trump and Hillary on a tweet by tweet basis, using an overlaid scatter plot of each of their sentiment summary lists.

Results:

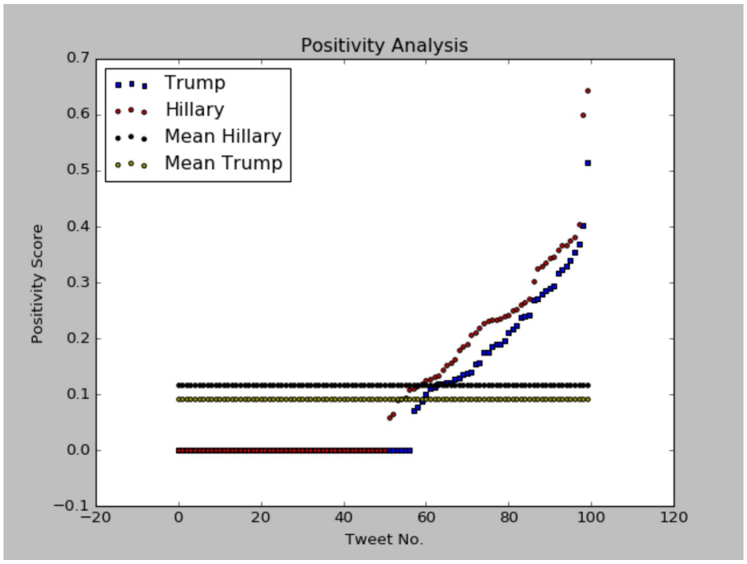
The following are the results from the Most_Common analysis. It shows the most recurring phrases in each of the presidential candidate’s tweets.

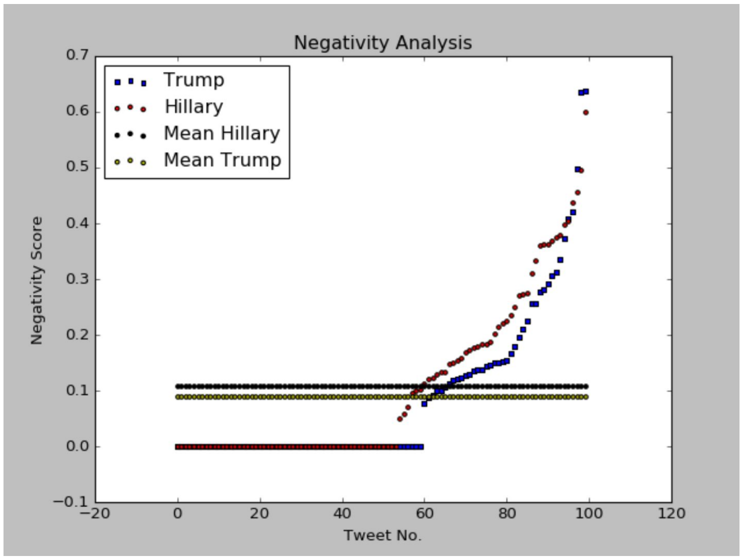


```
Most common words in Hillary's tweets are:
[(9, 'trump'), (5, 'releas'), (5, 'birthday'), (4, 'lost'), (4, 'good'), (4, 'amp'), (3, 'ye'), (3, 'wom'), (3, 'vot'), (3, 'paid')]
Most common words in Trump's tweets are:
[(10, 'trump'), (7, 'lik'), (6, 'good'), (5, 'would'), (5, 'distract'), (4, 'gre'), (3, 'see'), (3, 'new'), (3, 'lov'), (3, 'crimin')]
```

The biggest surprise for us here was that “Trump” was the most recurring word in tweets targeted to both Hillary and Trump. This could be because a lot of the tweets targeting Hillary on twitter were sourcing from Trump supporters trying to show Hillary down. In addition, many of the tweets from Hillary may be talking down Trump. As expected though, amongst the most common words in Hillary’s tweets was “woman”, seeing as one of her value proposition is that she will be the first woman president, which will bring a new perspective to the government. An interesting word to note in Trump’s tweets is “criminal”. This is probably because of all the hate Donald Trump has brewed by his constant politically incorrect behavior.

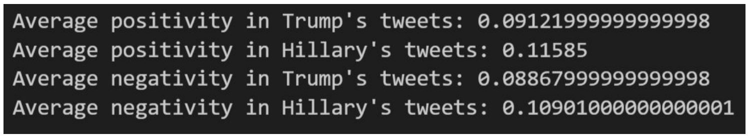
Next we used the scat_plot to produce the following graphs:





This graph was created by using a loop to iterate through each tweet, computing its sentiment summary, and appending each summary to a list of dictionaries. Each item in this list was a dictionary, from which the ‘pos’ key contained the value for that tweet’s positivity intensity. We used another loop to obtain a list of just the ‘pos’ values from the list of summaries. This gave us our data set for the scatter plot. The functions used to create this are present in Analysis.py and Visualization.py

A summary of the data displayed on the graph is also presented below:



```
Average positivity in Trump's tweets: 0.09121999999999998
Average positivity in Hillary's tweets: 0.11585
Average negativity in Trump's tweets: 0.08867999999999998
Average negativity in Hillary's tweets: 0.10901000000000001
```

This was created using the sentiments() function that we created in the Analysis.py file. This output was very interesting for us. Although, the graph makes it look like the positivity scores for Trump and Hillary are alike, the mean lines for both their scores shows a clear difference. Trump has more tweets than Hillary which had a positivity score of 0. This drove the mean positivity for Trump tweets down.

What's further interesting is that the average negativity is also higher in Hillary's tweets. However, we feel that this could potentially be because of the larger number of '0' scores in Trump's tweets, which once again drove the mean sentiment score down. But if you look closely, the maximum negativity score was derived from a Trump tweet, while the maximum positivity score was derived from a Hillary tweet. This could imply that voters feel stronger positivity towards Hillary and a stronger negativity towards Trump.

Reflection:

From a process perspective, most parts of this process went well. It was straightforward to gather the data, pre-process the data and analyze the data on a general level. The two main areas of improvement for our projects are deriving better insights from the tweets by doing more specific preprocessing and gathering more data. We could have better preprocessed the data by not removing Internet and general slang, acronyms and hashtags. In addition, we were limited by Twitter's API to scrape more data. We were limited to only scraping 100 tweets, which limited the types of insights we could have gotten.

In terms of dividing the work, we worked very well. We split the tasks evenly and had clear deliverable deadlines for each task. For example, one person would complete the tokenization and the other would do the analyses portion. In addition, for tasks that would break our code, we would sit down and discuss to figure out why the code was not functioning. Next time, it would have been better to have more sit down sessions to build upon each other's' logic for code.