

Ho Wang Alastair Ng  
MIS 3640  
Professor Li

## **Assignment #2- Elon Musk and Donald Trump**

### **Project Overview:**

In this project, I used the twitter API as my source of text-mining. I remember doing something similar in my business intelligence class where I had to use R to do a sentiment analysis on Hilary and Trump tweets. Similarly in this project, I want to see how the public viewed Trump's and Elon's tweets. I used the Twython module to gather my data. Then, I had to pickle the data as a means to store the data onto the computer. After pickling the tweets, I followed what Professor Li suggested to do for processing and analyzing the data.

### **Implementation:**

Twython uses a Twitter API to import 100 tweets given a specific query. The queries used were '@elonmusk' and '@realDonaldTrump'. I wanted to analyze the sentiments towards Elon Musk and Donald Trump, two of the most popular public figures on twitter. Once the tweets were gathered in the form of a list of string tweets, Pickle was used to export both public figures tweets data into two .pickle files. I first had to process and clean the data so I removed unnecessary white-spaces, punctuation, symbols and stop words. Furthermore, I only wanted to work with real dictionary words and not colloquial words so I removed any words that didn't appear in the Oxford dictionary, which is in the "word.txt" file.

I wanted to analyze the tweets in a few ways. First, I wanted to know how much difference the most common used words in the tweets pertaining to Elon and Trump will be different. I created a dictionary histogram of the words as they appeared in the list, similar to what I did for our wordplay exercise. The keys to this dictionary were the unique words appearing in the processed list of tweets. The values were the frequencies of the words. Then I created a function to display the top 20 most frequently occurring words in each of these dictionaries in descending order. This way, I could compare the most frequently used words by users tweeting about Elon and Trump. Next, I wanted to check the mean user sentiment in a tweet directed to Elon versus the average user sentiment in a tweet directed to Trump. In order to do that, I used the NLTK sentiment intensity analyzer module to obtain a summary of sentiment intensities for each tweet. After compiling these summaries to a dictionary each for Elon and Trump, I was able to compute the mean 'positivity' and 'negativity' intensities for tweets about Elon and tweets about Trump. I also made a graph to compare the sentiment intensities for Elon and Trump for every tweet, using the pyplot function of each of their sentiment summary list.

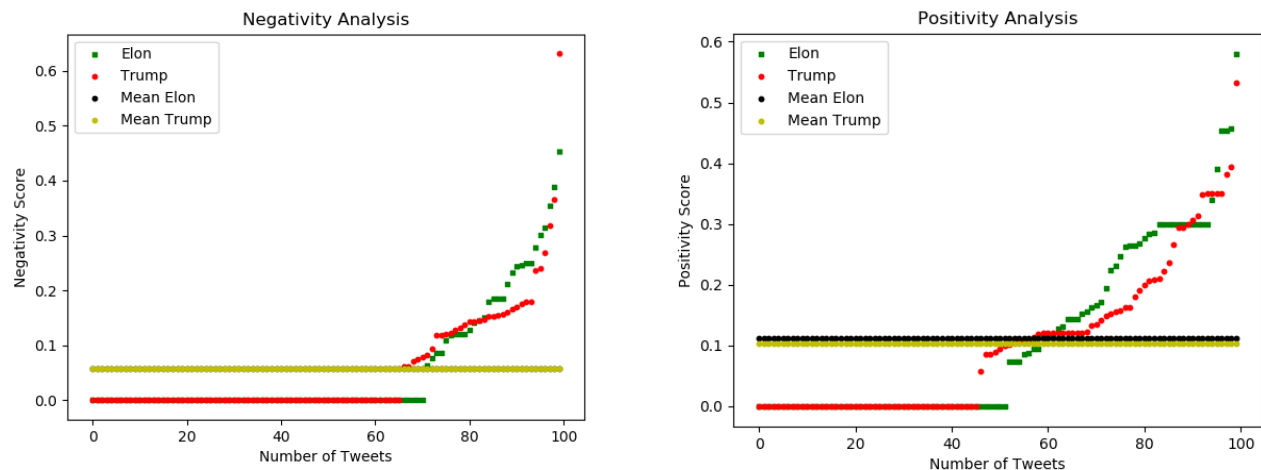
## Results:

The following are the results from the most common words:

```
Most common words in Elon tweets are:
[(22, 'tesla'), (16, 'time'), (12, 'pretty'), (11, 'win'), (11, 'secret'), (11, 'pin'), (11, 'much'), (11, 'least'), (11, 'infinite'), (11, 'giveaway'), (11, 'energy'), (11, 'close'), (11, 'adapters'), (7, 'well'), (7, 'one'), (7, 'car'), (6, 'like'), (5, 'would'), (4, 'week'), (4, 'thanks')]
Most common words in Trump tweets are:
[(20, 'election'), (19, 'protect'), (10, 'votes'), (9, 'corruption'), (8, 'trump'), (8, 'senate'), (8, 'new'), (8, 'must'), (8, 'match'), (8, 'call'), (7, 'signatures'), (7, 'know'), (7, 'electoral'), (7, 'county'), (7, 'conceded'), (6, 'want'), (6, 'confirmed'), (5, 'wants'), (5, 'president'), (5, 'people')]
```

It shows the most frequent phrases in each of the tweets about Elon and Trump. As expected though, amongst the Elon's most common words is Tesla. The most iconic public company for their groundbreaking innovations that Elon is the CEO of. For Trump, the most common word seen in Trump tweets is election. This is mostly due to the recent election that came up 2 weeks ago.

Next I used the `scat_plot` function to produce the following graphs:



This graph was created by using a loop to iterate through each tweet and appending each sentiment summary to a list. The items in this list were values in the dictionary, from which the 'pos' key contained the value for that tweet's positivity intensity. I used another loop to obtain a list of just the 'neg' values from the list of summaries. This is the data set for the scatter plot.

I computed the sentiments analysis using the `sentiments()` function in the `Sentiment_Analysis.py` file. The graphs were very interesting. I would expect that there should be a substantially higher negative sentiment for Trump, but comparatively, they are actually very similar according to the mean values. However, I do see that the mean positivity is higher for Elon meaning there is a more optimistic outlook in the public's eye keeping in mind that there are a larger number of '0'scores in Elon tweets, which drove the mean sentiment score down. Looking at the max sentiment result for both graphs, I can see that the public is more negative towards trump and more positive toward Elon.

**Reflection:**

I ran into a few bumps as a lot of modules I was unfamiliar with, which meant that I had to do a lot of google searching. It was straightforward to process the data as we have done this in class. However, gathering the data using os and pickle module was somewhat hard to figure out, as well as making a graph. The two main areas of improvement for the project are deriving better insights from the tweets by doing more with the sentiment analysis. In addition, I was limited to only scraping 100 tweets, which limited the accuracy of the insights.