Vicky McDermott
Text Mining Project
February 20, 2017

Project Overview *[Maximum 100 words]*

I used the 200 tweets published by the accounts of Donald Trump and Hillary Clinton leading up to the November 8th presidential election. I compiled a list of the most commonly used words by each candidate to compare which words were more effective in convincing the electorate to vote. I also used the sentiment analyzer to determine which candidate used more positive, negative, and neutral language in their tweets. I hoped to investigate and learn if certain Twitter strategies might have played a role in Trump's win in the presidential election this past November.

Implementation *[~2-3 paragraphs]*

I created two main python files, one which handles gathering the text from Twitter and another which handles analyzing the gathered text. I have a function which retrieves the tweets from a certain twitter screen name and then dumps them in a pickle file to access later. I have two other functions, one of which handles gathering the Trump tweets either from the pickle file or, if that does not exist, from the Twitter API, and another which does the same thing for the Clinton Tweets.

Initially the tweets are stored in a list. I then converted the list to a string so I could get rid of all characters which are not letters such as punctuation and numbers. I converted it back to a list and created a function which counted the number of times each word appeared and put them in a dictionary. I converted the dictionary to a list of tuples and sorted the words so that the words which showed up the most frequently were shown last. I then printed the words and their frequencies. In addition, I imported the Vader Sentiment Analyzer and printed the polarity scores of the entire list of Tweets from each candidate.

At first, I was writing out all the code to gather the Clinton and Trump Tweets from the API separately. I noticed that I was copying much of the same code and only changing a few key variables. To make my code more robust, I decided to compile much of it into a function to retrieve Tweets from the API for any screen name. I also decided to use tuples to compile the list of words with frequencies so that the words would not be separated from their frequency count when I sorted the list.

Results *[~2-3 paragraphs + figures/examples]*

When I ran my program I excluded words that I did not think were meaningful and printed only words that appeared more than 10 times in the respective candidate's Tweets. My results are case sensitive and are grouped only when the exact same word appears. I get the following results from a single run of my program:

Hillary's Words:

(10, 'Beyonce')

(10, 'Clinton')

(10, 'election')

(10, 'live')

(10, 'today')

(11, 'just')

(11, 'she')

(12, 'Donald')

(12, 'Tomorrow')

(13, 'not')

(13, 'president')

(13, 'who')

(14, 'all')

(14, 'country')

(14, 'your')

(15, 'go')

(15, 'want')

(16, 'POTUS')

(16, 'believe')

(16, 'get')

(16, 'up')

(16, 'what')

(17, 'make')

(18, 'us')

(20, 'America')

(20, 'can')

(21, 'Trump')

(24, 'are')

(27, 'We')

(39, 'I')

(56, 'vote')

(59, 'we')

(62, 'you')

(65, 'Hillary')

{'neu': 0.825, 'neg': 0.035, 'pos': 0.14, 'compound': 0.9999}

Trump's Words:

(10, 'Colorado')

(10, 'Crooked')

(10, 'Michigan')

(10, 'as')

(10, 'can')

(10, 'from')

(11, 'AMERICA')

(11, 'GREAT')

(11, 'MAKE')

(11, 'Obamacare')

(11, 'now')

(13, 'rally')

(13, 'watch')

(14, 'ICYMI')

(14, 'MAGA')

(15, 'tomorrow')

(15, 'your')

(16, 'Florida')

(16, 'VOTE')

(16, 'here')

(16, 'we')

(18, 'Get')

(21, 'Join')

(21, 'We')

(21, 'are')

(23, 'Hillary')

(24, 'me')

(27, 'Clinton')

(29, 'DrainTheSwamp')

(34, 'Thank')

(40, 'I')

(59, 'you')

{'neu': 0.708, 'neg': 0.075, 'pos': 0.217, 'compound': 1.0}

One interesting thing to note is that in the Tweets leading up to the election, Trump used the word "you" more than any other word while Hillary actually used her own name more than any other word in her Tweets. It is also interesting to note that Hillary used

much more neutral language than Trump. It appears that 82.5% of the language she used was neutral while only 70.8% of the language Trump used was neutral. Trump used much more emotional language, beating Hillary out in both his use of positive and negative language.

Reflection *[~1 paragraph]*

I could have done more unit testing along the way and better documented my code as I went instead of having to go back through and document everything after finishing. I think my project was fairly well scoped, but although I can speculate greatly about my results, I am not sure I can come to any definitive conclusions based on my findings. I will definitely make a more thorough plan for the functions and things I want to organize before beginning in the future. I wish I knew that when you write everything from scratch it can get messy and confusing if you don't plan it out, document everything, and make unit tests.