Jeffrey Kao
IST652

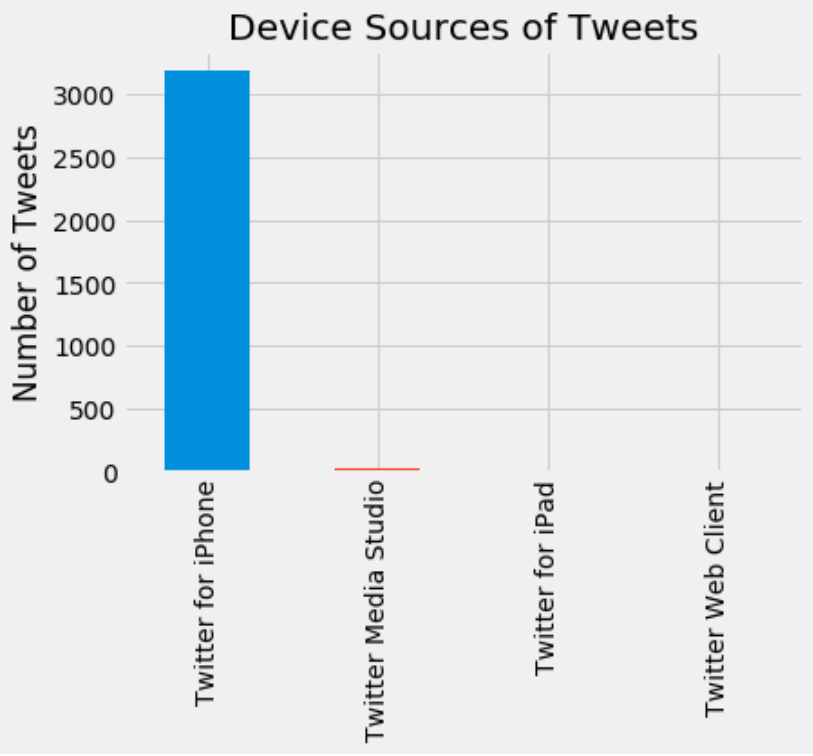
Final Project - Trump's Tweets

Data and its sources:
For the final project I decided to use the Twitter API with the tweepy package to obtain data of Donald's Trumps most recent tweets. Using the Twitter API I was able to to obtain 3232 tweets in a json file which I saved for analysis. Using a series of functions I was able to optimize the program so I wouldn't have to keep reloading tweets if I ran the program more than once because obtaining all the tweets for each API call took a couple minutes and I also did not want to go over the rate limit from the Twitter API.

I only wanted to take a look at a couple data fields so I took the fields that I wanted to look at and created a dataframe. The dataframe had the index, retweet_count, text, source, and time of each tweet. I cleaned up the columns but removing HTML from the source column. I also converted the time column to EST because that where Trump mostly resides so it was best to put convert it to that time zone. I also added an hour column to the dataframe so I could make a distribution plot. Also I had to remove punctuation for my sentiment analysis.

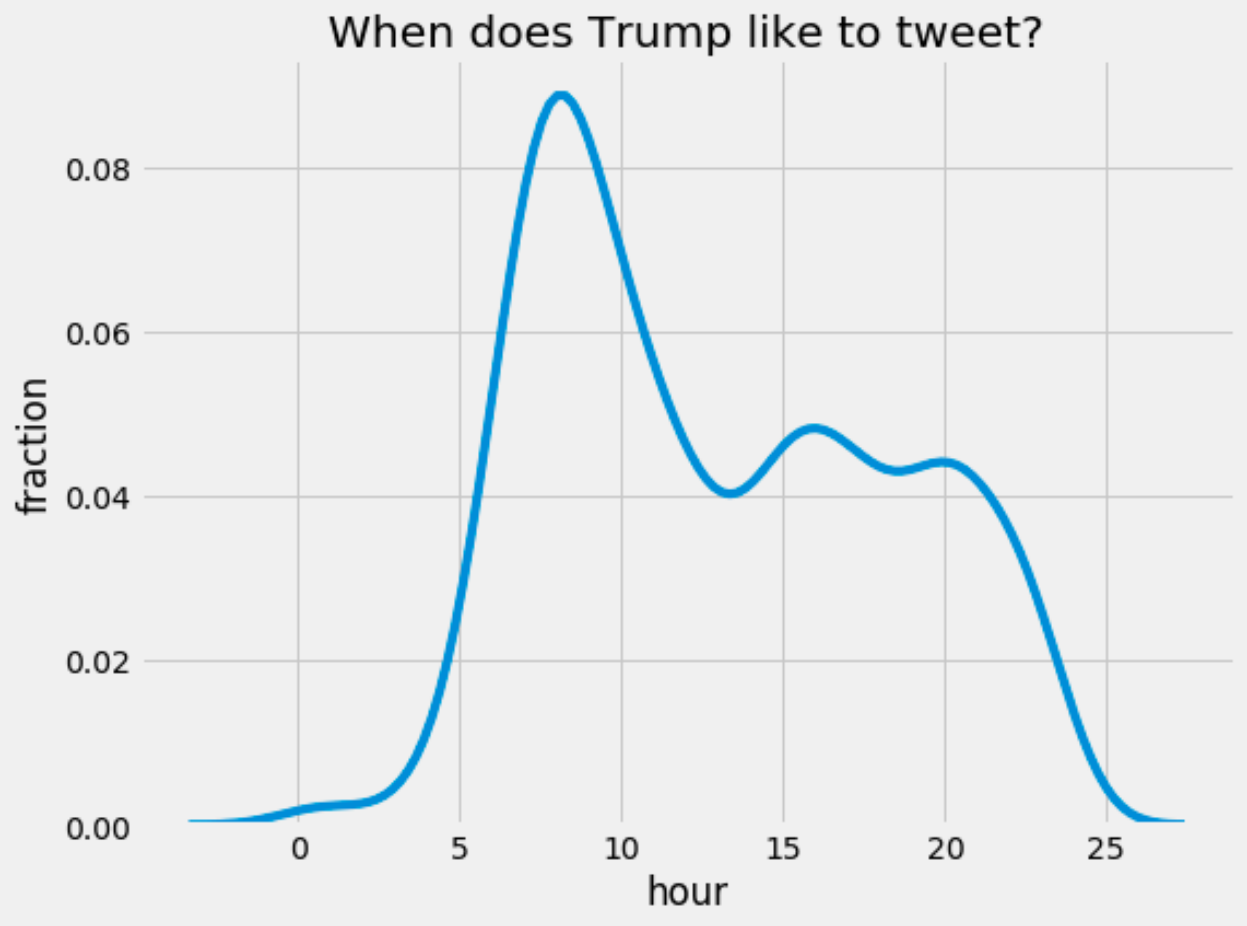
Question 1: What was the source of Trumps tweets?
I wanted to take a look what source Trumps tweets came from. I obtained the data from the dataframe that I had made.

```
Twitter for iPhone      3194
Twitter Media Studio      30
Twitter for iPad           7
Twitter Web Client         1
```



As we see here Trump mostly tweets from his iPhone, so he is just as addicted to his smartphone as the rest of us. Thirty of his tweets came from Media Studio which use involves some sort of image or video. And only 7 game from his iPad and 1 single tweet came from the browser. I plotted a bar chart of the distribution of sources for the tweets.

Question 2: When does Trump like to tweet?
For this question first converted the time column to EST because that where Trump mostly stays. I then created another column for hour in float so I could make a distribution plot that is below.



It looks like the majority of Trump's tweets are in the morning from around 8-10am. This goes hand in hand with the leaked schedule for President Trump. On his schedule it is called "Executive Time."

Question 3: What is the sentiment of Trump's tweets?
I will use the VADER (Valence Aware Dictionary and sEntiment Reasoner) lexicon to analyze the sentiment of Trump's tweets. VADER is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media which is great for our usage. The VADER lexicon gives the sentiment of individual words.

I created a dataframe for the VADER lexicon and then I did a join on a new table that contained the text of the tweets. I added a polarity column to the table. The polarity column should contain the sum of the sentiment polarity of each word in the text of the tweet.

Here are the most negative and positive tweets:

```
Most negative tweets:

   the trump portrait of an unsustainable border crisis is dead on. “in the last two years, ice officers made 266,000 arrests of aliens with criminal records, including those charged or convicted of 100,000 assaults, 30,000 sex crimes &amp;amp; 4000 violent killings.” america’s southern....

   it is outrageous that poisonous synthetic heroin fentanyl comes pouring into the u.s. postal system from china. we can, and must, end this now! the senate should pass the stop act — and firmly stop this poison from killing our children and destroying our country. no more delay!

   the rigged russian witch hunt goes on and on as the “originators and founders” of this scam continue to be fired and demoted for their corrupt and illegal activity. all credibility is gone from this terrible hoax, and much more will be lost as it proceeds. no collusion!

   ...this evil anti-semitic attack is an assault on humanity. it will take all of us working together to extract the poison of anti-semitism from our world. we must unite to conquer hate.

   at the request of many, i will be reviewing the case of a “u.s. military hero,” major matt golsteyn, who is charged with murder. he could face the death penalty from our own government after he admitted to killing a terrorist bomb maker while overseas. @petehegseth @foxnews

Most positive tweets:
```

congratulations to a truly great football team, the clemson tigers, on an incredible win last night against a powerful alabama team. a big win also for the great state of south carolina. look forward to seeing the team, and their brilliant coach, for the second time at the w.h.

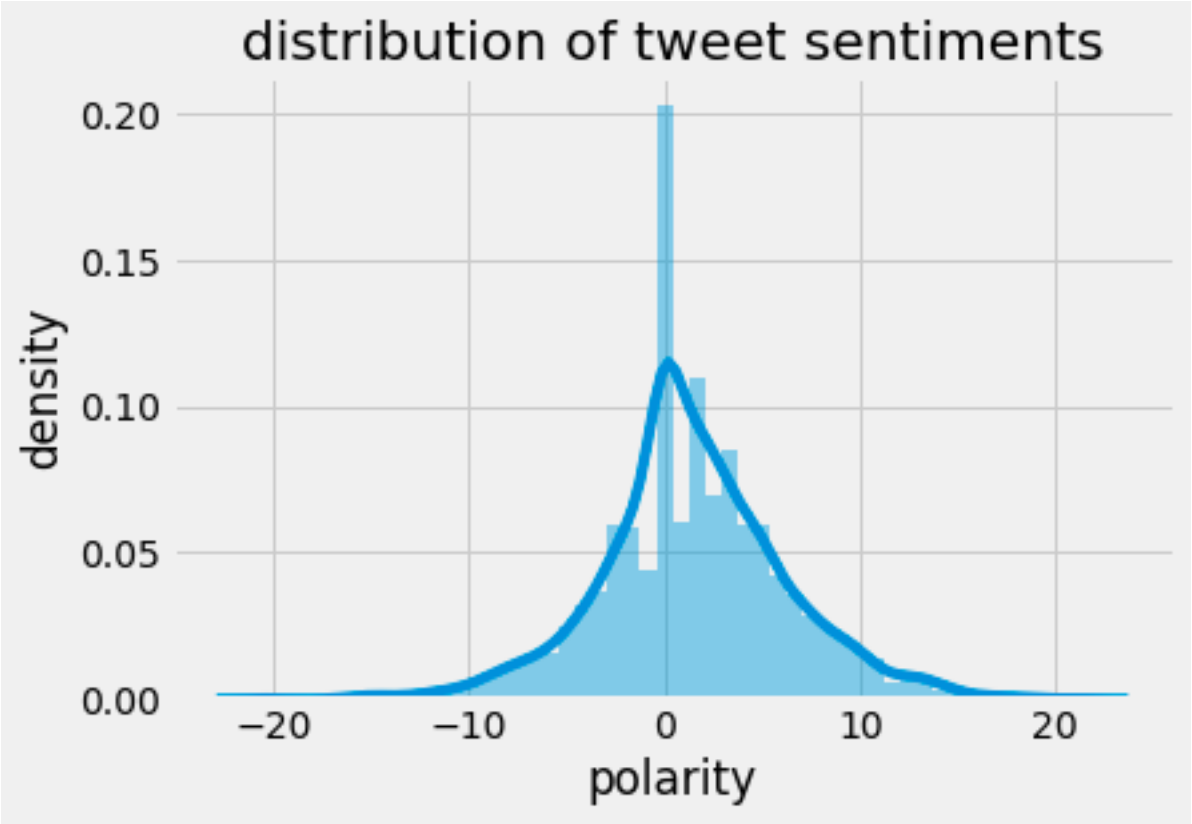
my supporters are the smartest, strongest, most hard working and most loyal that we have seen in our countries history. it is a beautiful thing to watch as we win elections and gather support from all over the country. as we get stronger, so does our country. best numbers ever!

thank you to all of my great supporters, really big progress being made. other countries wanting to fix crazy trade deals. economy is roaring. supreme court pick getting great reviews. new poll says trump, at over 90%, is the most popular republican in history of the party. wow!

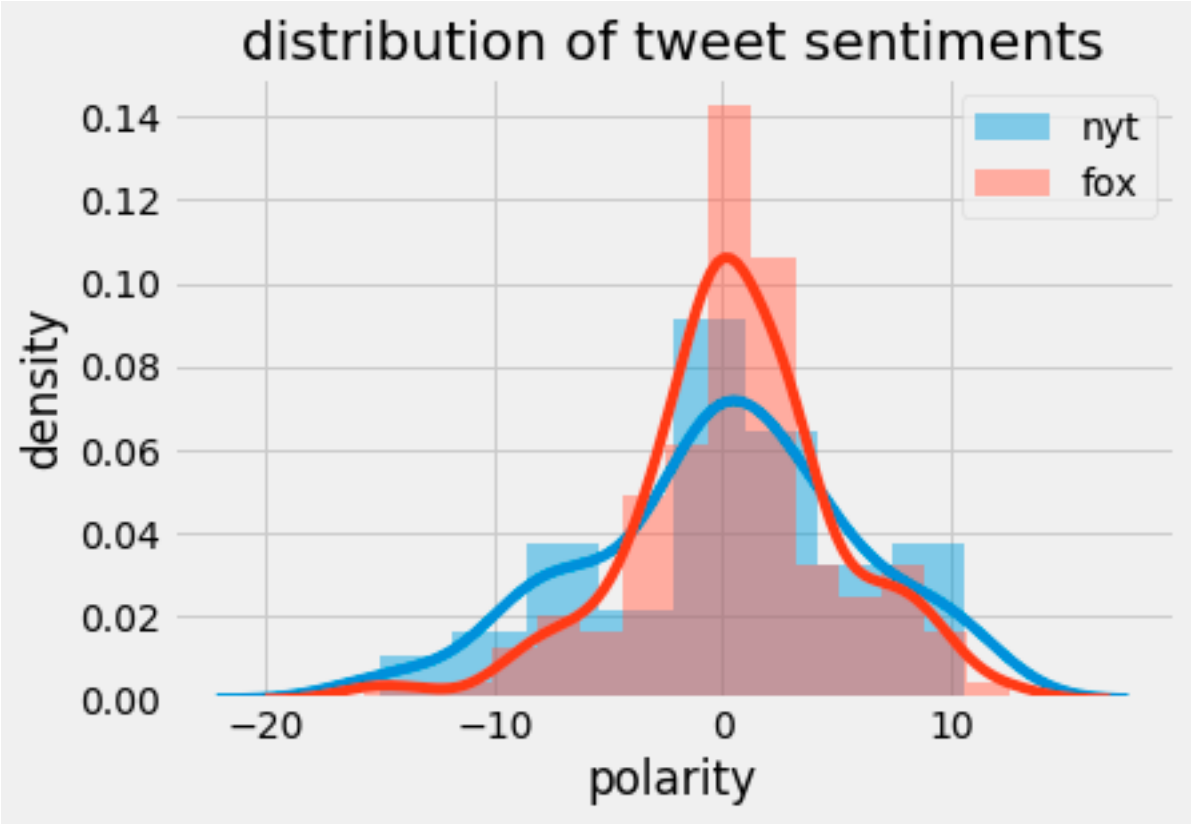
thank you, @wvgovernor jim justice, for that warm introduction. tonight, it was my great honor to attend the “greenbrier classic – salute to service dinner” in west virginia! god bless our veterans. god bless america - and happy independence day to all! https://t.co/v35qvcn8m6

president george h.w. bush led a long, successful and beautiful life. whenever i was with him i saw his absolute joy for life and true pride in his family. his accomplishments were great from beginning to end. he was a truly wonderful man and will be missed by all!

Fairly normal distribution of sentiment for all the tweets as well.

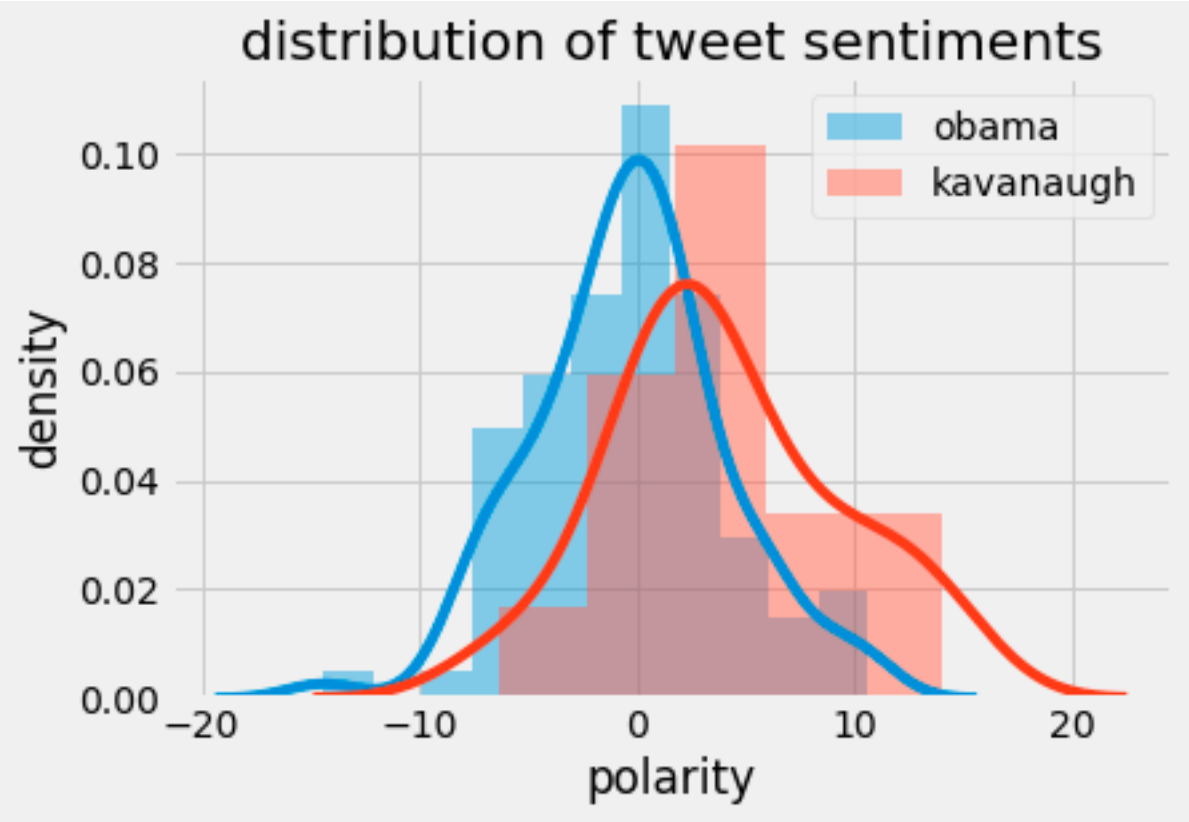


Also i wanted to look at the sentiment of tweets containing “fox” and “nyt.” I isolated the tweets that contained these words and looked at the polarity of these tweets.



Looks like both “nyt” and “fox” have a fairly normal distribution. But Trump seems have more positivity about “fox.” This makes sense because Fox is known to be fairly conservative and Pro-Republican.

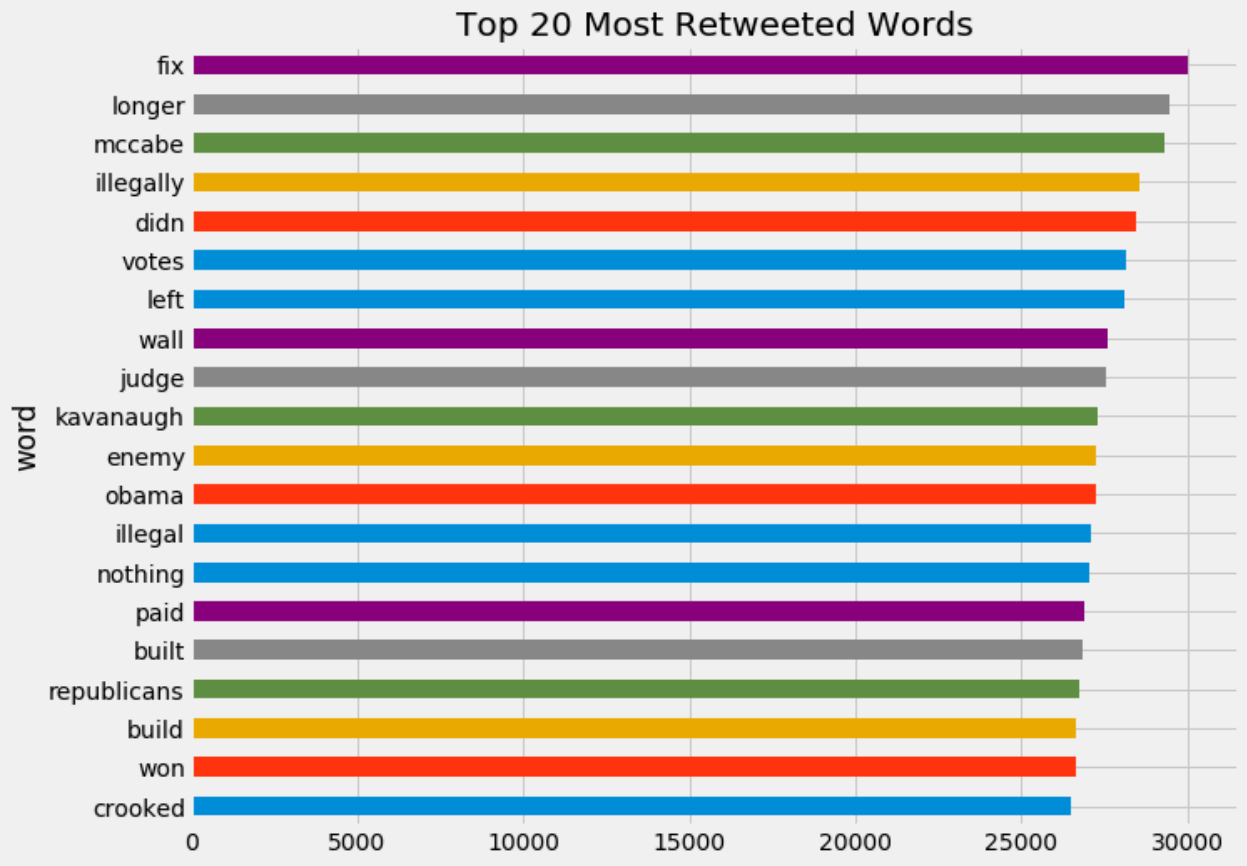
I also looked at the sentiment of tweets containing “obama” and “kavanaugh.”



Clearly here there is a skew of positivity towards “kavanaugh” and negativity towards “obama.” Trump was trying very hard on Twitter to defend his Supreme Court nominee!

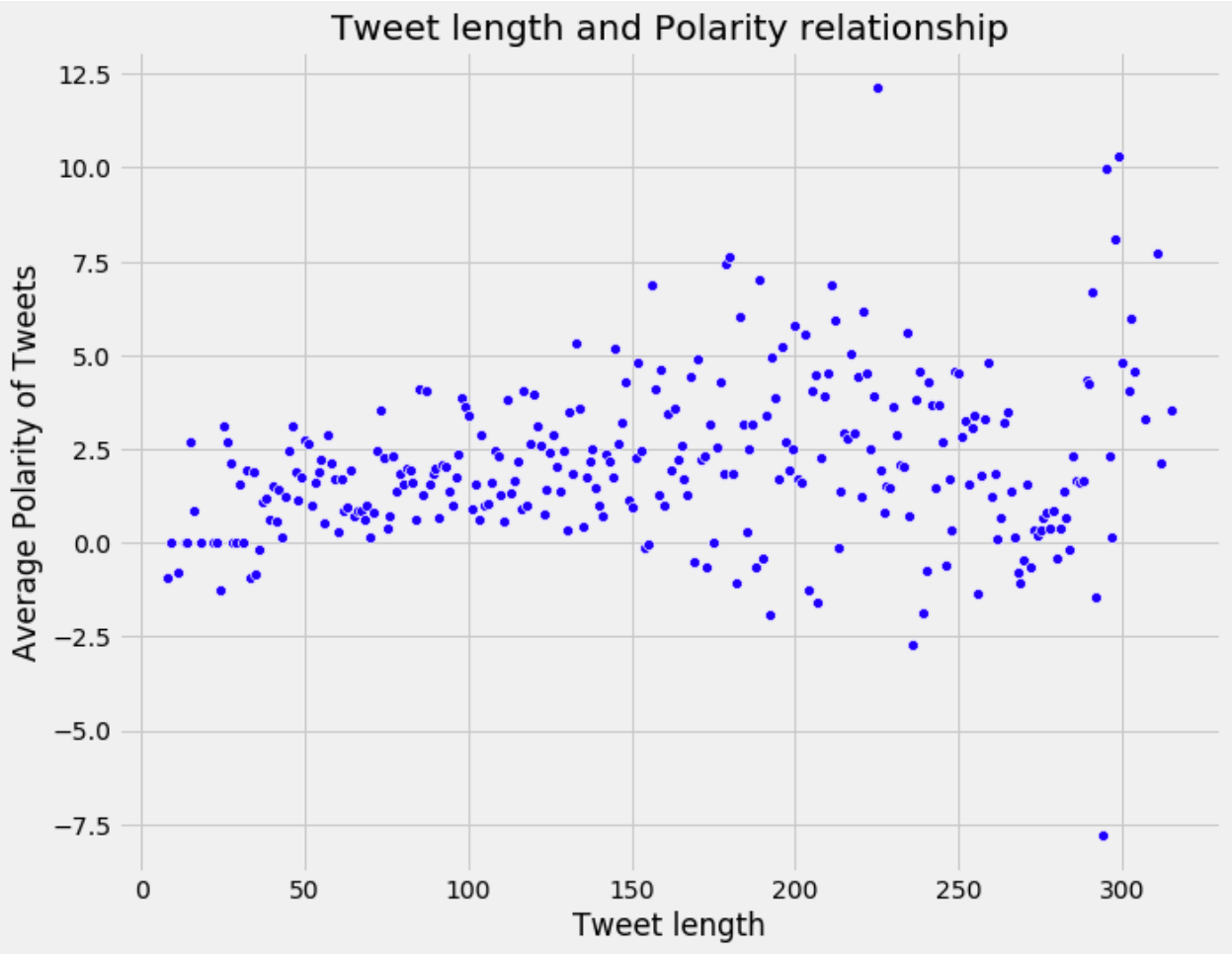
Question 4: What are the most retweeted words from Trump's tweets?

I wanted to see what were the most retweeted words, but I only included words that appeared in at least 25 tweets. First created a table that only had the words that were appeared 25 times or more, then sorted the values to show the top 20 retweeted words. I then created a bar plot to show the results.



Question 5: What is the relationship between polarity and tweet length?

I wanted to compare the tweet length with the polarity of the tweet so I created a new dataframe with new column tweet_len and compared it with polarity column. I created a scatterplot of the results.



As the length of tweets increased, in general there is a positive correlation between the tweets length and average polarity of the tweets. The longer the tweets, the stronger the polarity with some outliers. Also we can see that from tweets length 0-150, the average polarity of Tweets is stabilized in the line that is slightly above 0. More length leads to more variation of the polarity change as it is more spread out as well. When the length of the tweets is longer than 300, it's very hard to predict the polarity of the tweets.

Program Description:

- Used the Twitter API to collect Donald Trump's most recent tweets.
- Prevented rerunning collecting tweets after json file was created.
- Clean the data I wanted place into a dataframe.
- Process data to analyze and graph source of tweets.
- Process data to analyze and graph when Trump tweets.
- Used VADER for sentiment analysis. Process and clean lexicon and twitter data.
- Print most positive and negative tweets. Overall sentiment of tweets graph was produced.
- Compare tweet sentiment of certain keywords: "nyt", "fox", "obama", "kavanaugh"
- Psrocess and clean data to see the most retweeted words and graph.
- Process and clean data to compare tweet length and sentiment with graph.

Conclusions:

My analysis of Trump's twitter data came up with a couple conclusions from the questions that I posed:

- The majority of Trump's tweets come from his iPhone.
- Trump normally tweets around 8-10am during his "Executive Time."
- Overall Trump's tweets have a normally distributed sentiment.
- Trump has a more positive sentiment towards Fox News versus New York Times.
- Trump had a very positive sentiment towards his Supreme Court nominee, Kavanaugh.
- There is a slight positive correlation between tweet length and polarity.

Output can be seen on this report and the ipynb file.