Midterm project Holly **104000036**

• Topic

The relationship between Trump's tweets, Newspaper context, and stock price movement in a one-year period

(For the midterm project, I narrowed the scope to the relationship between Trump's tweets and stock price movement in a one-year period)

• Introduction

Because of the rapid development of the internet, our everyday life bombarded with lots of information every second. We can know what happened right at the moment by simply taking out our cellphone and going to some social media platform. Our behaviors are increasingly influenced by social media, including the behavior in the stock market. I'm interested in the extent social media affect behaviors of investors, and twitter is a great example of social media. As a result, in the midterm and final project, I would like to find the relationship between Twitter tweets context and stock price movement. By the way, I choose Trump's tweets to do the research because I think Trump is a common factor in the international stock market recently, and it will be good to see the influence of some international affairs such as US& China trade war or North Korea& US summit.

• Datasets

A. Kaggle

- Donald trump tweets 09/17-09/18
  https://www.kaggle.com/jersey33/donald-trump-tweets-09170918

B. Yahoo Finance historical price 09/05/17 - 09/10/18

- Dow Jones (^DJI)
- S&P 500 (^GSPC)
- NASDAQ (^IXIC)
- Russell 2000 (^RUT)
- Xiaomi Corporation (1810.HK)
- Taiwan Semiconductor Manufacturing Company Limited (2330.TW)
- Samsung Electronics Co., Ltd. (005930.KS)
- Sony Corporation (6758.T)
- Apple Inc. (AAPL)
- Amazon.com, Inc. (AMZN)
- Alibaba Group Holding Limited (BABA)
- Facebook, Inc. (FB)
- Alphabet Inc. (GOOG)

• Until midterm project, what I have done?

I. Try to extract News (ex. BBC, New York Times, CNN, The Washington Post, Bloomberg, and Trump tweets data using Tweeter Developer API) - failed (Used Kaggle data instead)

II. Try to extract historical prices on yahoo finance - finished

III. Make a word cloud for Trump's tweets context between 09/17 - 09/18 - finished

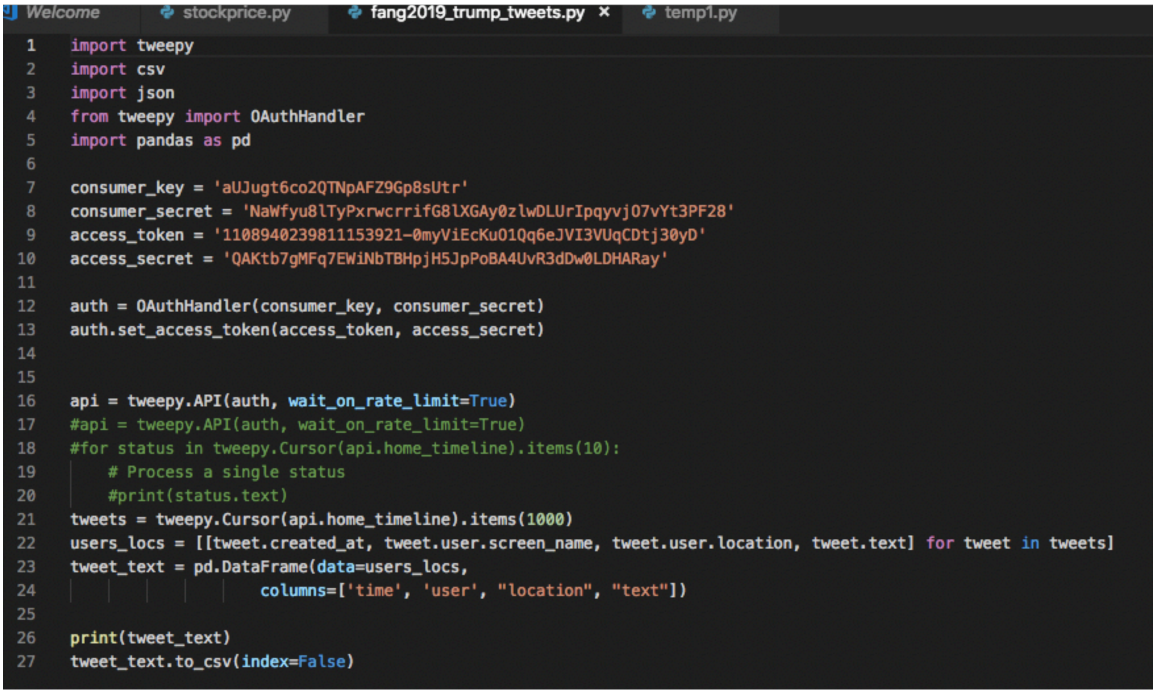
IV. Pre-process data so that I can put tweets and stock prices in the same time period, comparing them in 256 open days of the stock price - finished

V. Plot the daily closed price and daily tweets on a graph - finished

• Detail about what I did in each work

I. Try to extract News (ex. BBC, New York Times, CNN, The Washington Post, Bloomberg, and Trump tweets data using Tweeter Developer API) - failed (Used Kaggle data instead)

I registered a Twitter developer account and used python to extract data from the Twitter timeline. Because I follow all data providers, including BBC, New York Times, CNN, The Washington Post, Bloomberg, and Trump, I extract data from my timeline. However, Since Twitter limits each developer a maximum request time and request number, I couldn't download all data I need in a limit time. In order to finish the midterm project on time, I found some data on Kaggle that contains Trump's tweets in a one-year period, between 09/17 - 09/18. I thought the data is appropriate to fulfill our research purpose, so I decided to use it.



```python
import tweepy
import csv
import json
from tweepy import OAuthHandler
import pandas as pd

consumer_key = 'aUJugt6co2QTNpAFZ9Gp8sUtr'
consumer_secret = 'NaWfyu8lTyPxrwcrrifG8lXGAy0zlwDLUrIpqyvj07vYt3PF28'
access_token = '1108940239811153921-0myViEcKu01Qq6eJVI3VUqCDtj30yD'
access_secret = 'QAKtb7gMFq7EWiNbTBHpjH5JpPoBA4UvR3dDw0LDHARay'

auth = OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_secret)


api = tweepy.API(auth, wait_on_rate_limit=True)
#api = tweepy.API(auth, wait_on_rate_limit=True)
#for status in tweepy.Cursor(api.home_timeline).items(10):
    # Process a single status
    #print(status.text)
tweets = tweepy.Cursor(api.home_timeline).items(1000)
users_locs = [[tweet.created_at, tweet.user.screen_name, tweet.user.location, tweet.text] for tweet in tweets]
tweet_text = pd.DataFrame(data=users_locs,
                    columns=['time', 'user', "location", "text"])

print(tweet_text)
tweet_text.to_csv(index=False)
```

II. Try to extract historical prices on yahoo finance - finished

I downloaded the historical price of some US index, US tech companies, Chinese companies, Taiwanese companies, Japanese companies, and Korean companies. I chose these companies because I thought they would be great to research US & China trade war and the relationship between Trump and North Korea.

III. Make a word cloud for Trump's tweets context between 09/17 - 09/18 - finished

- Because the dataset is a JSON file at first, I convert it to a CSV file. Later, I found that the data is really in disorder. I extract "date" and "full text" for each tweet.
- Then, I tokenized full text and removed some punctuation and strange words and URL.
- In the next step, I define some common words and then remove stop words. Actually, I tried to remove stop words together with punctuation, in the previous step, but it didn't work. However, I removed stop words in this step, outside that for-loop, it worked. I still don't know why.
- After all the pre-processing, I put all the text together and made a word cloud.

- Finally, I count the most common words.

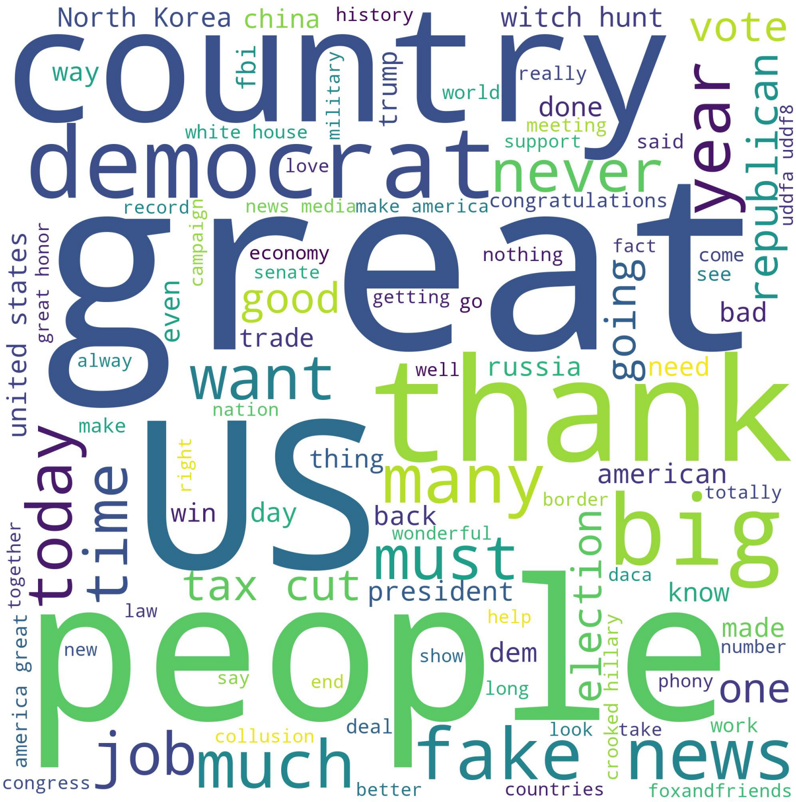
IV. Pre-process data so that I can put tweets and stock prices in the same time period, comparing them in 256 open day of stock price - finished

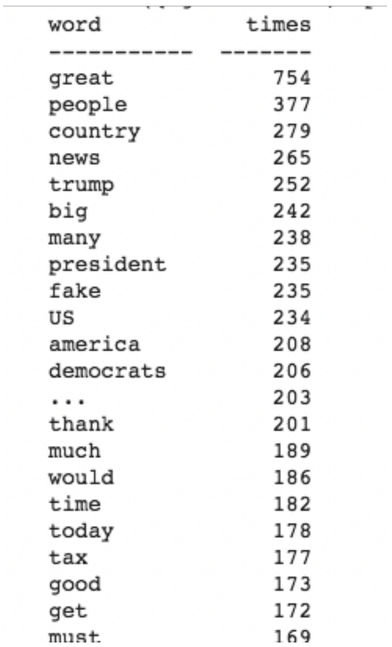
- I took “North Korea” for example and counted its frequency in each tweet.
- Then, I counted how many tweets each day.
- Next, I counted how many “North Korea” appeared in each day.
- Following, I think this is the most troublesome step, to find what date Trump didn’t tweet, what date he tweeted but the stock market didn’t open. I added Sat. Sun. to Mon. tweets and also added the tweets on market-closed day to the next day’s tweets. I did this because I used daily closed price for stock price data and because I think tweets before stock price closed will affect the closed price.
- Finally, I have 256 data for both stock price and the time “North Korea” appeared in tweets.

V. Plot the daily closed price and daily tweets on a graph - finished

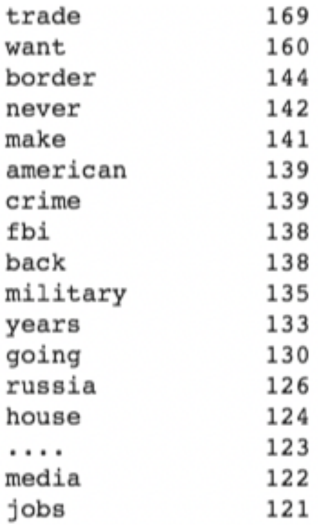
I took Google stock price for example. Using matplotlib.pyplot library, I plotted the time “North Korea” appeared in tweets in each day and closed price for each day on the graph.

• Result of my work

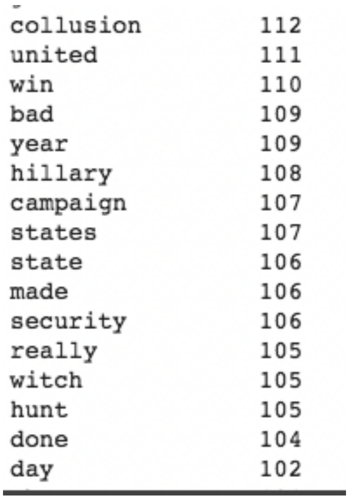




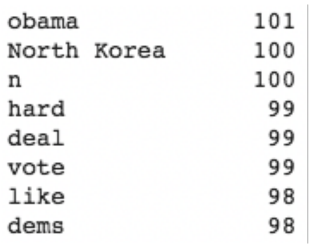
```
word              times
-----------     -------
great               754
people              377
country             279
news                265
trump               252
big                 242
many                238
president           235
fake                235
US                  234
america             208
democrats           206
...                 203
thank               201
much                189
would               186
time                182
today               178
tax                 177
good                173
get                 172
must                169
```



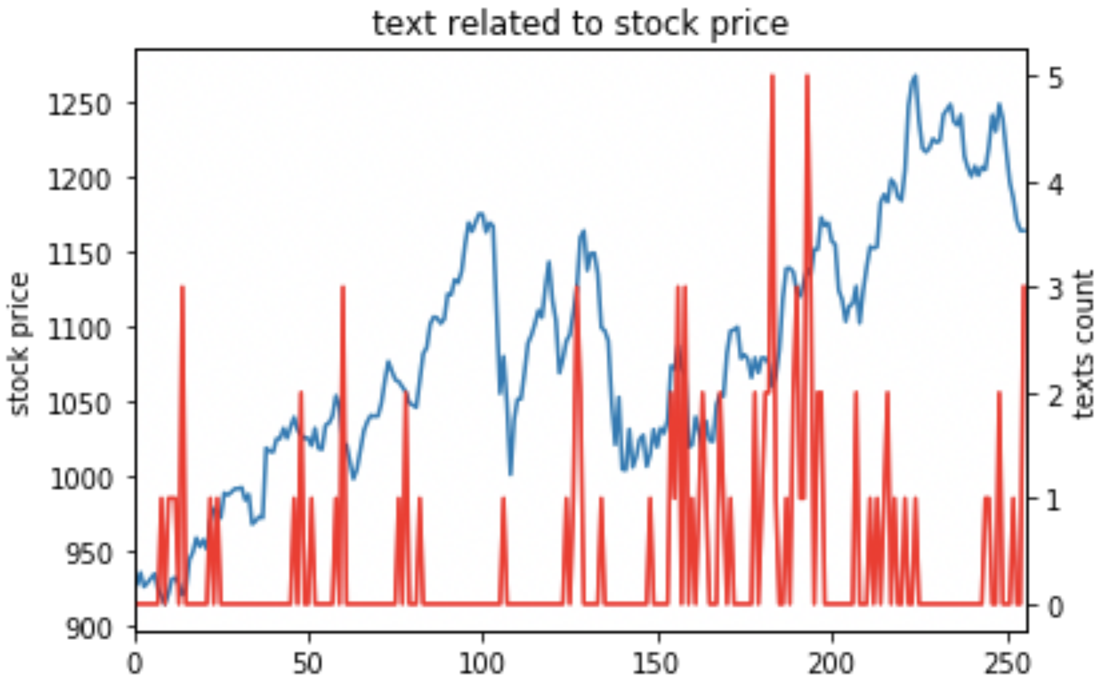
```
trade                169
want                 160
border               144
never                142
make                 141
american             139
crime                139
fbi                  138
back                 138
military             135
years                133
going                130
russia               126
house                124
....                 123
media                122
jobs                 121
```



```
collusion          112
united             111
win                110
bad                109
year               109
hillary            108
campaign           107
states             107
state              106
made               106
security           106
really             105
witch              105
hunt               105
done               104
day                102
```



```
obama              101
North Korea        100
n                  100
hard                99
deal                99
vote                99
like                98
dems                98
```



- **What I plan to do in the final project?**

In this midterm project, I did lots of data pre-processing work. I didn't use any model to find the relationship between stock price and text, nor do I take into account any random factors in the process. I plan to keep finding the relationship for the final project. In an attempt to widen the dimensions of our research, I would like to add more data from different sources in the same time period, including news data that I couldn’t get before. After I have those data, I am going to find an appropriate model to diagnose the relationship between tweets, news words, and the stock market. Perhaps I will use the method from panel data research.