Milestone Report 1

Customer/Client: Corporate Public Relations department

Problem Statement:Gauge how change in company executive leadership is received by the general public and investors.

1. Data gathering
    1.1. Data sources: Data sources are Twitter full archive search that allows access to tweets from up to 2006 and News API with access to 1 month previous news. Both allow only 100 results at a time.

    Due to restrictions on how many results can be brought back each time and how many times a search can be performed for free access, I was only able to get 100 results per data source per day. Because of this, in some days when there is a large number of tweets or news generated, the time range where the data gathered can be less than a day - in some cases, only 20 minutes. Conversely, if there is less activity in the day, there will be less than 100 tweets or news for that day.

    Data is accessed via Web APIs provided by Twitter and Newsapi.org. Once the data is pulled from APIs, they are stored as txt files.

        1.1.1. Twitter - Full archive search https://developer.twitter.com/en/docs/tweets/search/overview/enterprise
        1.1.2. News API: https://newsapi.org/

    1.2. Data storage and data cleanup:
        1.2.1. Data downloaded and saved as txt files in local machine is stored into MongoDB.
        The reason for this is for me to review the data and add flag indicating whether the tweet/news is relevant.
        Many of the tweets are not related to the company being searched or even if they are, they only have the URL to the web article which is not very useful. It is not in the scope of this project to pull in and parse the web articles.
        Also for News articles from NewsAPI.org, they only display up to

1024 characters. Content related to the company being searched may be in the article but not included in the first 1024 characters, in which case, it will be considered not relevant.
I also manually added topics for each of the relevant tweets/articles to understand what are the common topics.

1.2.2. Notebooks:
Get_Twitter_Feed_Full_Archive_WC
Get_NewsFeed_WC

2. Data wrangling
After the data cleanup by removing non-relevant tweets and articles, the concern is that not enough tweets/articles are useful. This may have to be handled via resampling.
Notebook: Wrangle_data_in_Pymongo

2.1. Read data from MongoDB and add to dataframe
Data stored in MongoDB is read in and added to dataframe. Only data necessary for analysis is added to dataframe.
Although topic information is within each tweet or news article row, topic can be in 1 to many relationship with a tweet or news so topic is stored in separate dataframe and will be joined with tweet or news dataframe as needed.

2.2. Remove duplicates from News.
News data has duplicates and have to be removed to avoid double-counting. Duplicates can be identified by same url and timestamp.

3. Exploratory data analysis
This is exploratory analysis based on relevancy and topic I manually input. After sentiment analysis using Vader and TextBlob packages, additional exploratory analysis will be performed.

Notebook: Explore_data_in_Pymongo_Twitter and Explore_data_in_Pymongo_News
Total Number of Tweets:
- WF Tweets: 12120
- SAP Tweets: 13740

Total Number of News Articles:
- WF Articles: 3000
- SAP Articles: 3984

3.1. **Wells Fargo**

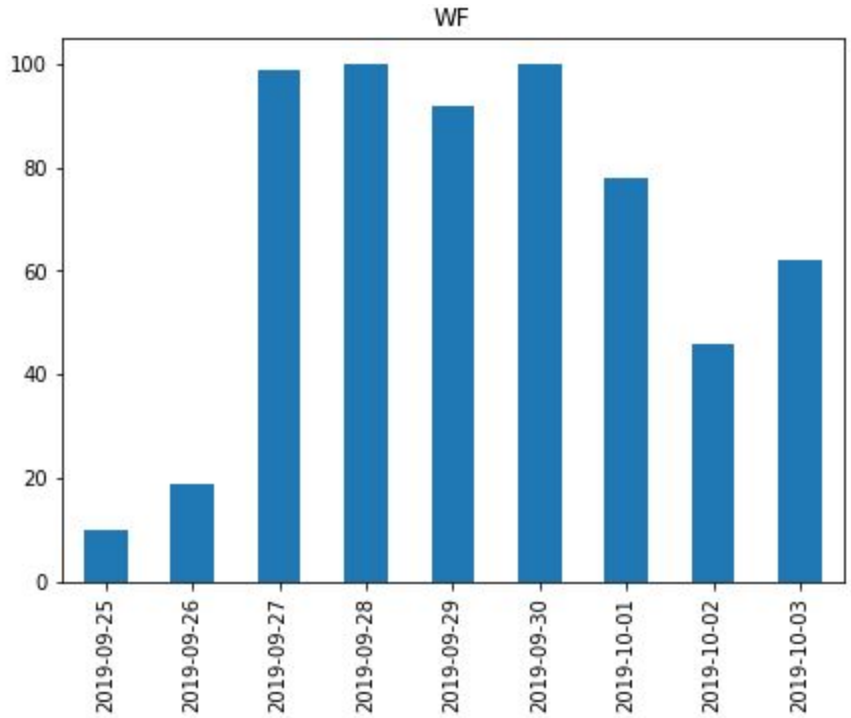
Wells Fargo announced the CEO on 9/27.

Below shows the number of tweets for each day. You can see that tweet activity increased starting 10/27. Similarly, there were some differences in number of not-relevant, there is sharp increase in relevant tweets starting from 9/27 and the number of relevant tweets don't go below the number of not-relevant tweets until 10/2.

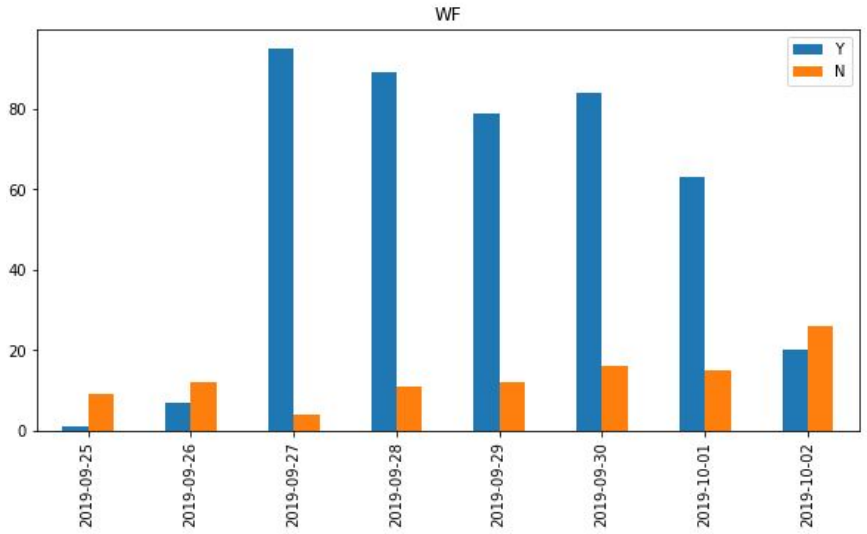
Number of News articles shot up on 10/27, went down during the weekend but the following week, there were not as many articles as 10/27 even though still slightly higher than before the announcement. In same vein, number of relevant articles was much higher than not-relevant ones on the day of the announcement but was quickly replaced by not-relevant articles the next day.

This shows that announcement of new CEO did not go unnoticed and public is interested for at least 5 days after the announcement. However, in terms of news articles, reporters tend to focus on new news the day of and move onto other news quickly.

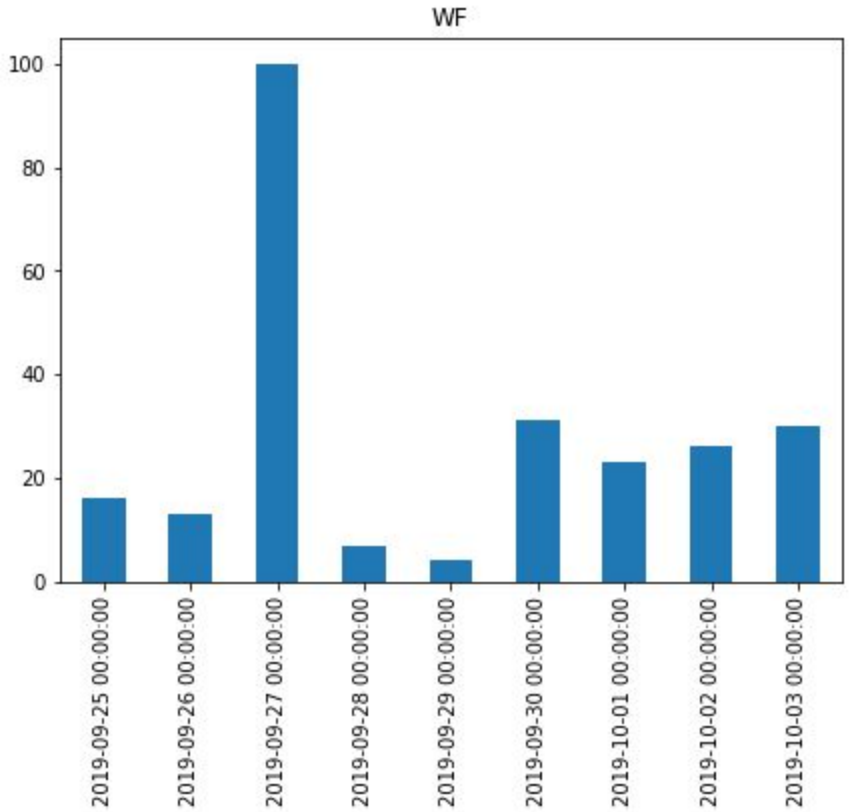
The top topics are Dimon, Diversity, New York, Pay, Stock and Tough Job. This tells me the story that people focused on Charles Scharf having been Dimon's former mentee, yet another white male CEO based in New York. Stock went up modestly, people acknowledge that this will be a tough job given the circumstances and the reason why Scharf accepted the position is because of significant pay bump as an enticement.



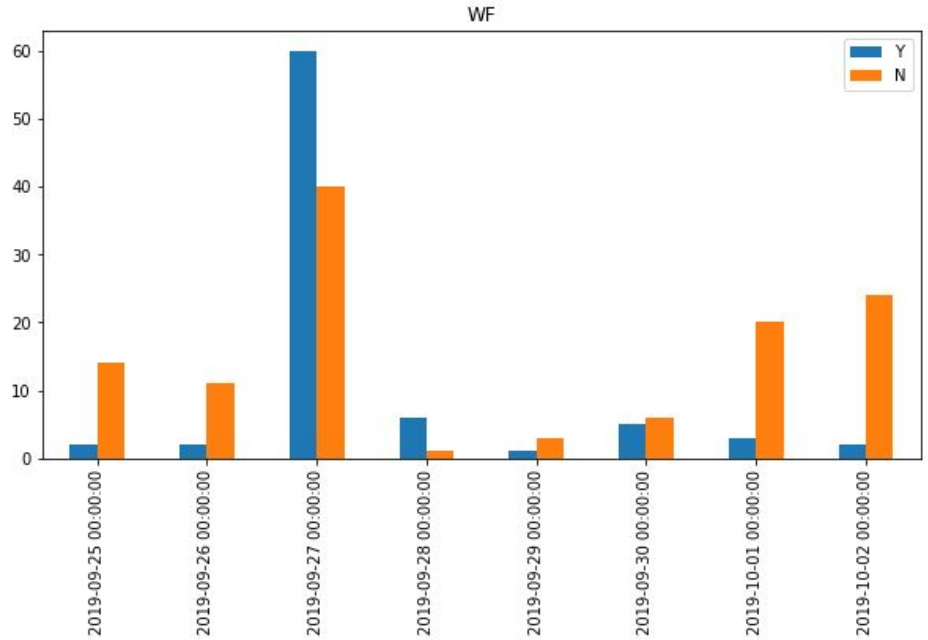
# of Tweets per day.



# of tweets per day - divided by relevant (Y) and not relevant(N)



# of News articles per day

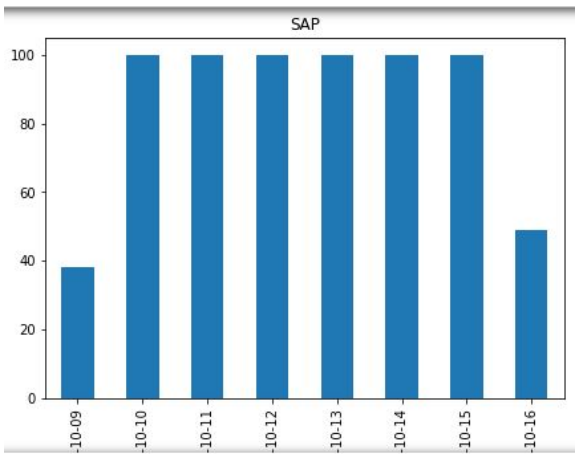
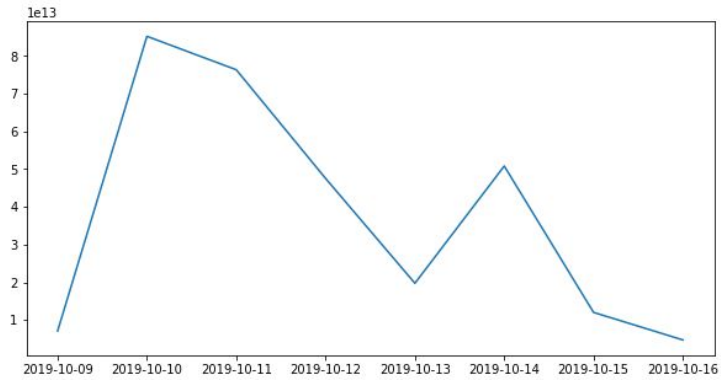


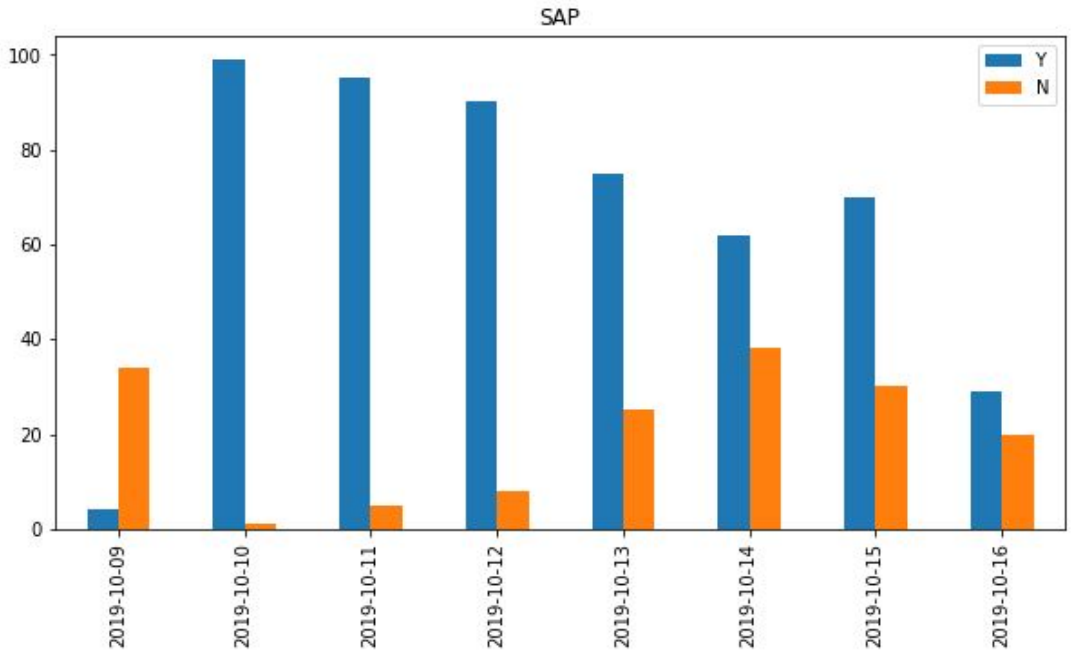
# of articles per day - divided by relevant (Y) and not relevant(N)

3.1. SAP

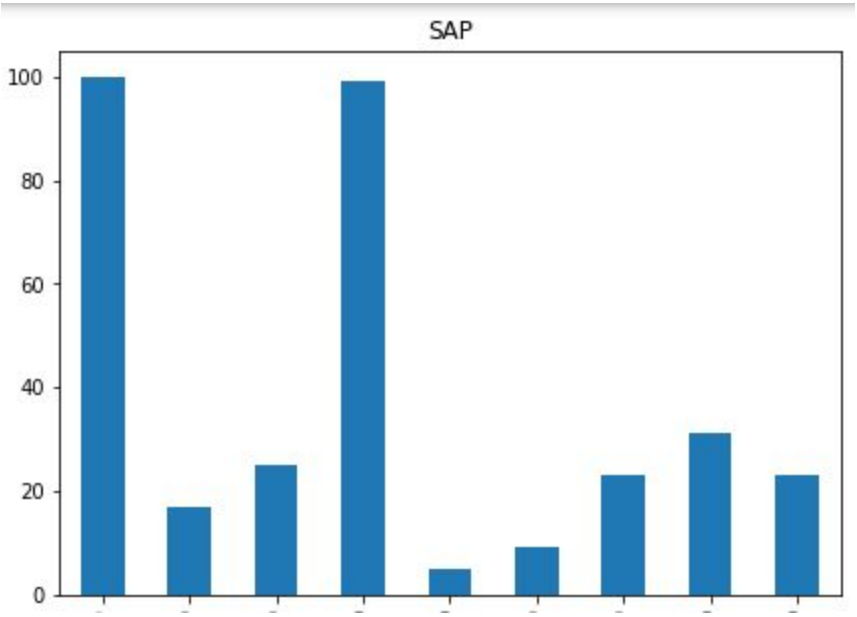
SAP announced the change in leadership on 10/11 in Europe which was 10/10 in U.S.
Tweets started increasing on 10/10 and due to the large amount of activity, there were more than 100 tweets for the day. For the days where there were more than 100 tweets, 2nd chart shows the range of time collected for the day (inverted. The amount of time 100 tweets were generated was less on 10/10, the day of the announcement, and started falling in the next few days and down close to the amount of activity before the announcement on the 6th day after the announcement.
Similar to Wells Fargo, the number of relevant tweets compared to not-relevant tweets sharply grow the day of the announcement and slowly dies off towards the 6th day of the announcement.
In terms of news articles, again similar to Wells Fargo, on the day of the announcement, there was a sharp increase in number of articles but quickly died down the day after.

Top topics for SAP are Diversity, New CEOs/New Era and Previous CEO. People focused on previous CEO's achievement and new co-CEOs. One of the co-CEO, Jennifer Morgan is the first woman to lead DAX index company.

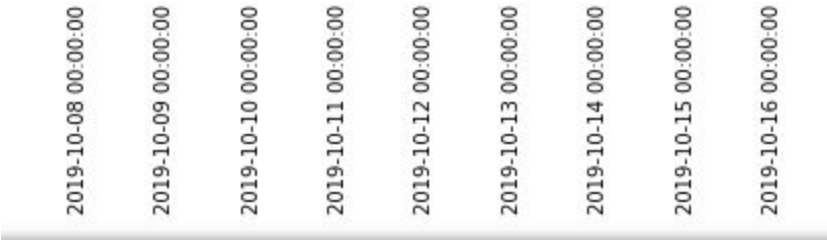
	
# of tweets per day	1 day - range of time 100 tweets were collected for the day. (inverted)



# of tweets per day - divided by relevant (Y) and not relevant(N)



,



# of news articles per day.

4. Notebook location: https://github.com/yulmee/springboard/tree/master/CapStone2
5. Presentation deck: https://docs.google.com/presentation/d/1vSeZPmi88EC8Akm_QU7KXGT_a0qJlfgWB9MDs2PzrUs/edit?usp=sharing