Project Deliverable 2

Group Members: Conor Walsh, Yash Bengali, Patrick Kuzdzal, Carlos Lopez

---

1. Collect and pre-process a secondary batch of data

Last week, we found a kaggle dataset which we had used for the preliminary analysis, and have since begun working on a twitter scraper to collect a more recent dataset. The issue we ran into however, is that twitter's official API limits us to tweets from the past 7 days. We looked into possible workarounds, in the form of various python modules that parse tweets using HTML, but found that twitter has fixed such functionality earlier this year. Thus, we are still limited to this short timeframe for collecting tweets. Nonetheless, we have run our scraper for two consecutive weeks, and have gathered upwards of 60,000 relevant tweets. In addition to analysing Apple, we also collected the price of ethereum for the last 5 years.

2. Refine the preliminary analysis of the data performed in PD1

We had to perform preliminary analysis on the data we collected recently using the twitter api. The dates were given to us in the following format (date, time) in one cell, and we could not use that because the time is different on every tweet. We first split that column into 2 columns by splitting on a space. We then ran the SentimentIntensityAnalyzer from the NLTK toolkit and then performed a groupby on the dates. We then computed the mean on each group from the groupby (shown in the table below).

3. Answer another key question

2021-10-24	0.12249221719967056
2021-10-25	0.12084845886499654
2021-10-26	0.12251482273575175
2021-10-27	0.08574175865076618
2021-10-28	0.05431917319108719
2021-10-29	0.09663675589173433
2021-10-30	0.10916118716692737
2021-10-31	0.07210526405355783
2021-11-01	0.11107361504671975
2021-11-02	0.136572077185017
2021-11-03	0.11744762068788885
2021-11-04	0.1295232129906945
2021-11-05	0.17669836441621298
2021-11-06	0.10633574005046967
2021-11-07	0.17161863304504243
2021-11-08	0.10586804580511221

Sentiment values of tweets between October 24, 2021 and November 8, 2021

We looked at our dataset of the past 2 weeks of apple stocks. In the actual stock price, we noticed a dip in the stock price from October 28th to November 1st where the price went down from 152 to 148. Although all the sentiment values in our table are positive, we can see a correlation where the sentiments from October 28th to November 1st are comparatively smaller. On the other days we can generally see much higher sentiment values, going all the way up to 0.17. In the future, we want to try normalizing this data between -1 and 1 to see if we can try and match these sentiments up to the apple stock price better. This will make it so a positive sentiment would predict an increase in stock.

4. Refine scope, list of limitations with data, and potential risks of achieving project goal
Our project scope is staying the same to analyze the correlation between tweets and apple stock. However, we are potentially looking into seeing if we can expand into also seeing if there is a correlation between cryptocurrency and tweets. This would be fascinating because cryptocurrencies fluctuate much more than the apple stock so it would be interesting to see how that affects the analysis.

In terms of limitations with data, as explained above, there exists a bound on the timeframe from which you can scrape tweets. We attempted numerous proposed workarounds, however to no avail, due to updates twitter rolled out earlier this year. Therefore, we are ultimately limited to tweets from the past 7 days, and are collecting a csv of tweets, currently sized at around 60,000.

We do not see any potential risks of achieving our project goal, as we are just looking to determine whether a correlation exists between tweets about a particular company, and their stock price.