Names and NetIDs

Robert Marshall - rfm4
Teja Pitla - tpitla2
Stuart Jaffe - sijaffe2 (Captain)

Project

Detect sentiment and other useful metrics on popular/trending stocks using live text data from social media (Reddit and/or Twitter).

Progress made thus far

1. Research sources of data
    a. Twitter. Filed an application with Twitter to get access to their "Academic Research" tier instead of their basic "Essential" tier. The Academic Research tier will allow us to have a more extensive set of API endpoints, better searches, a higher API rate limit, and a higher monthly limit.
    b. Reddit. There is likely an avenue to access the content of Reddit posts through HTTP GET/POST requests that accessing static JSON data instead of having to deal with the Reddit API or use a headless browser like Selenium or pyppeteer to render dynamic JavaScript content.
    c. Financial data. There are plenty of websites that have historical time series financial data, including Yahoo! Finance. So we should be able to procure this data easily.

Remaining tasks

1. Implement Data Gathering
    a. As discussed above, this will be done through web scraping or API calls.
2. Topic and Sentiment Extraction
    a. We need to implement sentiment analysis by analyzing the text corpus to extract topics through an algorithm or set of algorithms/techniques, such as Latent Dirichlet Allocation or a more modern methodology using transformers, such as BERT combined with clustering and TF-IDF for interpretability.
    b. Decide on a sentiment analysis algorithm to implement, such as Naive Bayes, SVM, Logistic Regression and LSTM.
3. Comparison With Time Series
    a. We could try to implement contextual text mining through time series as discussed in Week 12 lectures.
4. Testing the Software
5. Formulate a demo/presentation

Any challenges/issues being faced

No issues thus far.