What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.

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

What is your free topic? Please give a detailed description. What is the task? Why is it important or interesting? What is your planned approach? What tools, systems or datasets are involved? What is the expected outcome? How are you going to evaluate your work?

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

We intend to develop a system that can interface with and retrieve raw text data from popular social media sites such as Twitter or Reddit. The data is to be related to stocks and investing, including related markets such as equity options. We will then perform natural language processing on this data (tokenization, lemmatization, named entity recognition, sentiment analysis, etc.) as needed to detect current sentiment on popular stocks and other investment topics. This information may provide utility to drive investment decisions.

We aim to make use of official APIs for Twitter or Reddit to obtain raw text data but may have to explore other methods if the official APIs do not expose sufficient data. These methods could include utilizing a web scraper, through a number of different Python packages such as scrapy, Beautiful Soup, Puppeteer, and Selenium. Additionally, we will have to evaluate and select the best NLP library for our task. This will require some research and exploration.

The outcome could be presented in a number of different ways, including a GUI or from an API call, depending on the data being displayed and the preferences of the team members as the project progresses.

We can evaluate our results by comparing the output to other publicly available indicators of market sentiment, as well as to changes in the stock's price, the stock's daily trading volume, and other indicators in the equity options markets.

Which programming language do you plan to use?

Python

Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.

N=3, 60 hours.

There are a substantial number of unknowns, and we will need to potentially evaluate many different approaches for acquiring our data and then processing it to get our desired outcome. It may be that the official APIs for the target social media platforms do not expose the desired data in a convenient manner or the data may require a significant amount of wrangling before analysis. If we end up utilizing web scraping, figuring out the details such as XML or CSS parsing, whether a headless browser is required for dynamically loaded data, and not getting blocked by our target websites will all take time to get up and running before even doing the NLP analysis. We will also need to evaluate popular NLP libraries such as SpaCy, CoreNLP, Stanza, etc. and select the ideal tool. Finally we will need to develop a reasonable way to display our output data.

At a high level we may broadly split the time allocation into 3 major tasks:

Text/data retrieval – 25 hours

Data processing/NLP – 25 hours

Data presentation – 10 hours