Final Project Proposal

Stock Market Tweet Analysis

1. **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.**

Kasam Dhakal (kdhakal2@illinois.edu)
Nisarg Mistry (nmistry2@illinois.edu)
Parth Shukla (pshukl21@illinois.edu)

Captain: Kasam Dhakal

2. **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?**

We are planning to create a program that analyzes tweets for sentiment analysis regarding stock market information. This program would allow us to see how users on Twitter are discussing a particular stock and would thus give us a general idea of what the market is feeling regarding a company. Our planned approach is to analyze a set of tweets corresponding to stock information and investigate the different words that are used in each. We can match the word choice to categorize a tweet as either positive or negative. This would give us an idea about how a stock would be either bullish or bearish and furthermore give us investing guidance. We can utilize a dataset composed of tweets in text format to analyze the contents to run a sentiment analysis and create an outcome. A very simple approach is to utilize a model we create using python and logistic regression to create a prediction about whether a tweet is positive or negative. Our expected outcome is to see that some stocks are positive, and others are negative in terms of user sentiment. We can evaluate our work first by testing and seeing simple phrases as being either positive or negative. Furthermore, we can test on individual tweets. Lastly, we can compare our results of whether a stock is positively trending with real research we conduct to see if our findings are like our program.

Tools, Systems, or Dataset

- http://www.tweepy.org/ - Python Library to access the Twitter API
- http://www.nltk.org/ - Natural Language Toolkit
- Twitter data from Kaggle: https://www.kaggle.com/datasets/utkarshxy/stock-markettweets-lexicon-data for Sentiment Analysis.

3. **Which programming language do you plan to use?**

We plan to use Python as the core programming language for this project.

4. **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**.

1. Week 9 - Project Proposal Submission for grading - (4-6 hours)
2. Week 10 – Getting familiarity with the tools (Tweetpy and other tools) - (4-6 hours)
3. Week 10 - Dataset Retrieval - (4-6 hours)
4. Week 10 - Dataset Cleaning - (4-6 hours)
5. Week 11 - Dataset Tokenization – (4-6 hours)
6. Week 11 - Create a Baseline(working) Model – (4-6 hours)
7. Week 12 - Model Evaluation – (10-12 hours)
8. Week 13 - Parameter tuning/improving over baseline – (10-12 hours)
9. Week 14 – Progress Report Submission for grading – (6-10 hours)
10. Week 15 - Final Model Evaluation – (4-6 hours)
11. Week 15 – Testing and Improvements – (10-20 hours)
12. Week 16 – Project Code Submission, Documentation and Presentation – (10-15 hours)