CS410 Project Proposal

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.

   Group Name: HODL
   Name: Yi Hao Tan
   NetID: yihaoht2@iliionis.edu

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?

   My free topic is to perform sentiment analysis on the Bitcoin cryptocurrency using Reddit subreddit as the data source. With more focus and interest in the cryptocurrency market, there are increasingly more investors and more cryptocurrency being invented. The cryptocurrency market is extremely volatile, and with the uprising of Reddit users in 2021, it has shown that social media has become a powerful tool that affects financial markets. With this analysis, we could scale it to cover more coins. It will be interesting to explore the relation of social media and its users on cryptocurrency.

   My task is to scrape posts / comments relating to Bitcoin from Reddit, perform sentiment analysis, and compute a daily Bitcoin sentiment score between 2021-10-01 and 2021-10-31.

   My planned approach is to scrape posts / comments from the popular Bitcoin subreddits such as r/CrypoCurrency or r/Bitcoin between 2021-10-01 and 2021-10-31 using PRAW. The next step is to utilize VADER, an open-source sentiment tool for social media data, to perform sentiment analysis on each post / comment. This would produce a sentiment score per post / comment. In general, a higher sentiment score means that investors are bullish about Bitcoin while a lower score means the investors are bearish about Bitcoin.

   To compute a daily Bitcoin sentiment score, the sentiment score will be normalized per day. In the event where there is a huge discrepancy between the expected outcome and actual outcome, I will add new positive / negative words to the VADER sentiment model. Some new words may include HODL, Diamond Hand etc. Afterwhich, I will use Plotly to visualize the time series graph of Bitcoin sentiment score.

   The main tools that I will be using are PRAW or PushShiftAPI for scraping Reddit data, VADER for sentiment analysis and Plotly for visualizations. The main dataset will be

Reddit posts / comments retrieved in the time period between 2021-10-01 and 2021-10-31.

My expected outcome is to display Bitcoin's sentiment score daily in a time series graph between 2021-10-01 and 2021-10-31. Based on the daily sentiment score, I will be able to categorize Bitcoin sentiment into three distinct categories such as Bearish, Neutral and Bullish.

I will evaluate my work by comparing my results to the Bull and Bear Index, to check if my model and analysis is in line. If successful, there should not be any major differences in the scores or categorization.

3. What programming language do you plan to use?

   Python

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.

As I will be doing the project on my own, I estimate the work hours will be broken down as follow:

Task	Hours planned
Research and getting familiar with tools	2
Establish a Reddit instance via API	2
Identify key subreddits and obtain posts / comments on Reddit	5
Preprocess comments	2
Apply a sentiment analyzer model (VADER)	5
Obtain sentiment results and tweak model	2
Create visualizations	2
Demo and project summary	5