Topic: Crypto Sentiment Analysis

Team: HODL

NetID: yihaoht2@illinois.edu

Progress Report

Below is the progress made in the project:

- [Done] Researched and identified subreddit with the highest traffic for analysis
  I used the subreddit r/Bitcoin, as it has the most number of members as compared to the other subreddit. c/Bitcoin currently has 1.1 million members.

- [Done] Establish Reddit instance via PRAW API
  Created a Reddit app to obtain client id and client secret.

- [Done] Retrieved submission IDs in subreddit r/Bitcoin
  I had to use 2 different APIs to obtain submissions and comments by date. Pushshift for filtering submissions by date, and PRAW for obtaining comments. This took around **5** hours.

  For the first step, I pulled a maximum of 1000 submission IDs per day between 2021-10-01 and 2021-10-31

  The dataset contains ~3096 submissions. Each submission includes the ID and title.

- [Done] Retrieved all comments of relevant posts into a dataset
  For every submission ID in the dataset, I pulled the top 10 related comments as well.

  The dataset contains ~12172 comments. Each comment includes the ID and body.

- [In progress] Preprocess comments
  Have written the submissions / comments per day into each file with the following format **bitcoin_subreddit_YYYY_MM_DD.txt**

  The next step is to remove stop words and emojis in the text file

Remaining Tasks

- Apply sentiment analysis model (VADER)
- Obtain results and tweak model
- Visualize results with plotly
- Project demo and documentation

Challenges

- Had to use 2 different APIs in order to obtain post via dates
- Takes time to retrieve all the posts and comments