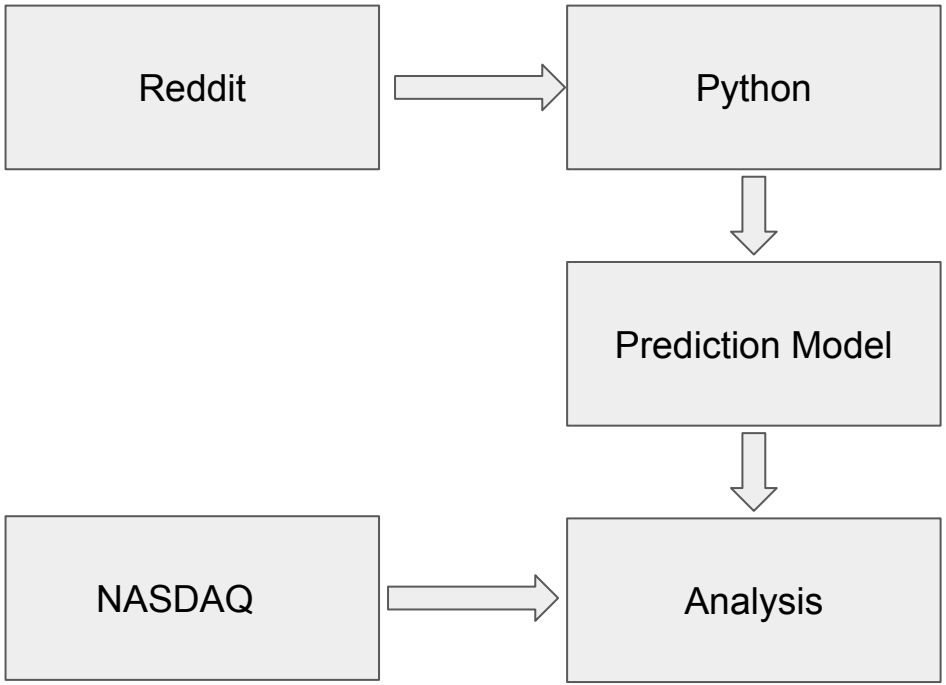
Scraping Reddit Stocks

W. Nhan, I. McCormick, B. Charlton, G. Freifeld

Project Description

- Find correlation between sentiment towards tickers in r/maxjustrisk and actual stock performance
  - To do this we will use sentiment analysis using the NLTK Python module and the Python Reddit API Wrapper (PRAW)
- We can fetch actual stock performance data from NASDAQ

Solution Diagram



Deliverables

Reddit scraping script

- Takes a Reddit thread as input, collects comments from the thread, and analyses them for sentiment.

Predictor script

- Takes a date range as input and makes a prediction for stocks based on the data collected by the other script.

Data Collection

- Scrape data from a year’s worth of daily discussion threads from r/maxjustrisk
- Use sentiment from these threads to predict stocks in the near future
- Compare these predictions with actual stock data

Tools

- Python
  - PRAW
  - NLTK (VADER)
- Reddit account

Timeline

Week 5: Reddit scraper and stock data collector working
Week 6: Predictor working
Week 7: Start writing
Week 8:
Week 9: Finish
Week 10:

Challenges

- No training data, we have to work from lexicon based sentiment analysis
- Cleaning data
  - Ignoring sarcastic comments
  - Redditors are aware of bots scraping and intentionally obfuscate at times
- May be difficult to determine what ticker the sentiment is on

Expansions

1. Automate data collection in specified subreddits.
2. Combine data from multiple subreddits.
3. Make predictions for the future, then wait and check how well the prediction did.