About our project:

Investing can be daunting and for beginners especially it can be hard to know where to start. Our goal was to create a platform that gives entry level investors the ability to do their due diligence without having to do hours of research and analysis of the stock market. With our website you can get a general sense of how people are viewing the most desirable stocks with an easy to understand graph showing the sentiment analysis that, while not perfect, is a good indication of how the stock will perform in the near future. We also have a dataset that will continue to scrape the internet to provide close to real time sentiment towards these stocks. If we had more time we would have implemented a way to corroborate our data based on the actual changes in the stock prices of the companies.

Built with

We mostly used python and it’s libraries to acquire and process our data. In the early stages of the project we used various news and social media apps in combination with basic web scraping to get the stocks that we should track and data that reflects these stocks. We used the TextBlob sentiment analysis in order to analyze the reddit posts focusing on finding relevant information by searching for specific keywords. Ideally we would have liked to use a Naive Bayes model which probably would have given better results. We felt we did not have sufficient data to train the model in order for it to be accurate so we opted for a more general model.

Our Challenges

The main challenges that we faced were mostly with our data pipeline. To build an accurate model we needed lots of data.The other issue with our dataset is that Reddit is not the best place to get our data because it is often long and deviates off topic. We first attempted to use the Twitter API to get real time tweets however ran into trouble since we did not have the elevated access API. Unfortunately a lot of the API’s we would try to use would often have request limits and as a result we weren’t always able to get data consistently from one source which greatly reduced the size of our training sets. By the end we were able to refine our process and ended up with a pretty large dataset that continuously collects and new data. that boosted the overall accuracy of our model. This was good for our model but also created a rather tedious data wrangling process as some of the text we gathered contained some difficult characters like emojis which can’t be understood by the model and greatly reduced the yield of our data collection. In the end however, datasets that are clean and also useful to the problems you’re trying to solve may not always be available so tidying is a necessary skill and can be rewarding in the end.

What We Learned

Future

Since our project was very data driven we would have liked to accumulate a lot more data and then implement a better model that could understand the data better. We would have also increased the avenues in which we would receive data using stuff like News Headlines and Twitter posts to get more accurate data and just making for overall a better dataset. I think a Naive Bayes Dataset would be ideal for this and with the help of some known sentiments it would have been very powerful. Also there are probably extra ways we could have preprocessed the data and made the overall website more sleek and interactive. Although I think we did a good job considering our time constraint and needing to learn a lot of things on the fly.