

Chih-Hsien Lin

DSCI 510

Dr. Yigal Arens

23 April 2021

Maintainability and Extensibility

While structures for my sources are unlikely to change in the short-term, numerous changes need to be made to the program in order to adapt to these changes. For web scraping parts of my program, I would have to learn the HTML structure of the new format and perhaps locate classes of tags in order to extract information. For instance, if Reddit decides to revamp its website with React.js, I would probably have to use the “.text” attribute instead of the “.contents” attribute to access React text (like I did for the Yahoo Finance scrape). If the structure of AlphaVantage API changes, I would simply have to look over the documentation to figure out how to extract the information I need.

The program is designed such that for any reasonable changes made to my sources can be dealt with by changing the “web scraping” and “API calling” portions of code. The data processing portion shouldn’t change much, if at all.

A potential additional data that I may want to add is the dates of the Yahoo Finance headlines. Since those dates are in the form of “x hours/days ago”, I would have to write code that converts those types of strings into datetime format and then compute “x hours/days ago” from the current time/day in order to obtain the date posted. This data will be added to the Yahoo Finance output dataframe.

Additional functionality that I may add into the code is the option to filter Reddit's result by weekly. This means that I should access the stock data of the past week instead of the past month from the AlphaVantage API. This would mean that the Reddit scrape portion of my code would have to be converted into a function that has inputs such as "monthly" and "weekly". Moreover, I would have to add some "if" statements into the Reddit scraping portion as well as the AlphaVantage API portion, which are relatively easy changes.