CAPSTONE PROJECT DATA WRANGLING

ARUN S BHARADWAJ, SPRINGBOARD DATA SCIENCE INTENSIVE

DATA STRUCTURE

The company fundamental dataset was downloaded from the StockPup.com website. The data was downloaded individually for each company. The website has corporate fundamental data for 739 companies in the form of csv files. Each csv file has time series information, with quarter end as the first column. The quarter end column is followed by number of shares, assets, current assets, liabilities, current liabilities, shareholders equity, revenue, earnings, earnings per share, cash from operating activities, cash from investing activities, cash from financing activities, share price, P/E ratio and other information usually available in company financial statements.

For the second part of the capstone project, two language corpora are used. The first language corpora is used to train the classifier to identify the bullish investor. The second language corpora is used to train the classifier to identify the bearish investor.

DATA LOADING

All 739 csv files are loaded into Python Jupyter notebook using the glob.glob function. One of the problems with each csv file is that the name of the company is seen only in the filename. So, when using glob.glob, while all the data inside these csv files get imported, we will never be able to know the name of the actual company. To ensure that the name of the company is seen in each row of the csv file of the respective company, we use the assign function in conjunction with the pd.read_csv. Inside the assign function, the symbol attribute is made equal to os.path.basename of the respective file name. After all files are imported, the quarter end column is assigned a data type of datetime to ensure that future operations using timeseries data are performed easily. The assign function inputs companyname_quarterly_financial_data.csv into each row. Since we only want company name, the _quarterly_financial_data.csv is removed from each row using str.replace function.

For the second part of the project, the nltk library is used to load the language corpora. Further, the correct encoding is used to import the text files.

DATA WRANGLING

In the first part of the project, since we also want information about the industry the respective company is operating in, the secwiki_tickers file is imported and included in the combined dataset. By running the info function on the combined dataset, it is seen that many columns are of data type object. These columns are converted to numeric data type.