**COSC 320 – 001**

***Analysis of Algorithms***

2020 Winter Term 2

**Project Topic Number: 2**

**Wall Street Bets Data Scraper**

**Group Lead: Tahmeed Hossain**

**Group Members: Barret Jackson, Omar Mourad, Kenji Niyokindi**

# Abstract

Our project will implement a data scraper that will traverse the blog post website Reddit and look through specific threads of information containing financial investment advice. Parsing this data, the program will find the stock tickers with the most value based on a weighted function that accounts for mentions, number of users who interacted with the post, and comments about that post. It is expected that this metric will provide insight to globally traded stocks regarding the volume of people speaking about individual stocks and could in the future help predict stock prices.

# Problem Description

The world of data science shows how to interpret data based on different inputs and try to generate a specific output. Dealing with data on Reddit threads there are millions of posts, comments, and mentions containing valuable information. Filtering this data to find trends that correlate to real-world outputs can be very valuable. Our goal is to take this information that provides financial advice and track how those stocks perform in relation to users being the indicators of stock price change. If we can identify a strong correlation between the increase in user activity with an increase in stock price, it will provide valuable information to future investors to see how interested people are in the stock and help them make the choice to invest or not. A detailed example can be found in the Dataset Collection section.

# Edge Cases

For our project, we will be comparing two sources of data with the goal of finding a correlation. Edge cases for this algorithm includes data that is missed or undetected by the Reddit scraper (empty input). This type of issue can be handled by running multiple tests and closely examining data for word structure variance to improve the algorithm’s detection of stock data. Other edge cases relate to the effective usability of our data. Since our main source of input is from a Reddit scraper, we are looking to sort and organize a collection of strings. The main set of strings we look for are stock ticker symbols which are an arrangement of characters (usually letters) that can range from 1 to 5 characters per symbol. Most current stocks are between 2-4 characters long, but we must also design our algorithm to accommodate the rare cases where a stock symbol has a length of size 1 or 5. We will also be utilizing a predefined list of stock symbols to validate data collected from Reddit posts. Finally, we must also consider when additional letters are added to a stock symbol as they may denote additional characteristics. This can be resolved by deconstructing the stock symbol into a substring where whitespace or anything that isn’t a letter denotes the end of the symbol. Collecting data for our second dataset will have similar edge cases.

# Expected Complexities

Complexities will mostly be found in the Reddit scraping and the creation of the two databases. The main challenge we will be facing will be to come up with an effective algorithm to perform Reddit threads scraping and data gathering from Yahoo Finance, with the smallest time-complexity possible. Reducing data redundancy is also a complexity we will be encountering. Data redundancy creates insertion anomalies, and unnecessary data increases the size of the database, which are problems we need to avoid. Data normalization will be performed to reduce redundancy and prevent any potential anomaly. Another challenge we are expecting to face is handling the real-time data we will be retrieving from Yahoo Finance. Data that is keen to be constantly changing will be hard to manipulate. Therefore, constructing a consistent database in order to find a correlation with a second database might be difficult.

# 

# 

# Dataset Collection

Collecting data will comprise a major part of this project, and will come from two different sources. We will be studying Reddit posts and stock information for the calendar year of 2020. Firstly, we will be using the Reddit API to search all posts in the subreddit r/wallstreetbets for the most popular daily stocks. Popularity will be determined by a weighted average of three factors: mentions of a given stock in original posts (OPs), number of comments on these OPs, and number of upvotes on the OPs. A weighted average will be used since we are guessing that there will be far more upvotes/comments on given posts than OPs themselves. The weights for the weighted average will be determined after data scraping has begun in order to ensure a fair measure is obtained. To determine which stock is the topic of an OP, we will use the Reddit convention of using a dollar sign before a stock symbol (e.g. $XYZ); our search algorithm will return the text from all OPs on the subreddit for each day, search for the dollar symbol, then use the following non-whitespace characters as the stock symbol (e.g. if the algorithm finds $XYZ, then XYZ is the stock symbol). The number of comments and upvotes for each symbol will then be aggregated in order to calculate the weighted average.

Next, once the top three stocks for a given day are determined, we will use the Yahoo Finance API to look up information on these stocks. Since we want to determine whether Reddit mentions have an effect on the price of the stock, we will look at the trend of the stock price over the week following the popular stock's mention. For example, if stock XYZ is very popular on Feb 1, then we will look at the trend in the price from Feb 1-7. We will use a naive algorithm for determining the trend, looking at the difference between the first day's price and both the maximum and minimum prices over the following week. Looking at our friend XYZ again, if it opened at $10 per share on Feb 1, dipped to $9 the following day, then shot up to $20 two days later, finally closing at $15 on Feb 7, then the downwards trend would be (9-10)/10 = -10%, and the upwards trend would be (20-10)/10 = +100%.

# **Program**ming Language

Python will be used as our main programming language.

# Task Separation and Responsibilities

Task Separation will be done through breakdown sessions where we breakdown the problem into steps. Then create tasks for individuals to complete and push/submit their work to the cloud server for all members on the team to review.

# Unexpected Cases/Difficulties

One type of unexpected difficulty we may encounter is the appearance of multiple stock symbols within the same post; where a post may disregard one and strictly promote another. This is unlikely to become an issue as our main goal is to look for a correlation between the existence of stock on Reddit and its effect on its future value.