CSC 369 Winter 2021

Group Project - Individual Assessments (submit your own copy to Canvas)
Winter 2021

These will be kept confidential and are an opportunity to describe your experiences with our collaborative group project. This should be a first person narrative about your experience, not so much the specific results/details.

Project Title (your choice): An Analysis of WSB "meme" Stocks with Distributed Computing

Project Overview (3+ sentences describing your project):

- Use the Twitter, Reddit, and Yahoo Finance API to scrape data on these services for discussion on stocks (specifically volatile stocks 'meme' stocks such as GME and AMC). From this data, we will visualize the data and see what the trends are like and make comparisons between the different social media platforms. Spark will be used here to help us organize and process data much quicker through distributed work.

Project keywords (4-8 words): Web Scraping, Spark, WallStreetBets, GME, AMC, Sentiment

Describe your contribution (paragraph):

- My contribution to the project was organizing times to meet up as well as checking in with team members to see how progress is going. For the technical portion, I handled the Reddit data. I used the PSAW Reddit Api to get all posts for the month of December 2020, January, February, and March 2021 posts. I cleaned up the data and put it in a Spark DF to get the posts that mentioned AMC or GME. I then used Spark SQL to aggregate the data and visualize the data over the different months.

Best Outcome - What are you most pleased with from your project results?:

- I am most pleased that we were able to efficiently get so much data from Reddit and Twitter and process it. I had over 150MB of data and there were over 1,000,00 tweets that we had to process. I am very happy that we were able to take all this data, process it efficiently with Spark and create informative graphs.

Trickiest Part - What were the bottlenecks or sticking points and how did you resolve them (if you did)?

- The trickiest part was learning how to bring in the distributed component of our project. We originally wanted to distribute the API calls so that we did not have to wait a few hours to get the data. We realized that this was most likely not possible because of APIs

restricting the number of calls we can make at a time. We resolved this by pivoting and using what we learned on Lab 5 to use the power of Spark to help us process our data really efficiently.

Future Directions - if you had another year/$1,000,000 to work on this what would you do?

- If I had more time and money to work on this project, I would like to incorporate a real time analysis of the reddit posts, tweets, and all stock prices instead of just GME and AMC. I think this would be really cool because you would be able to see how social media relates to stocks in real time. I would also like to introduce a ML/NLP model to this data so that you would be able to see the sentiment of a stock and predict the behavior of a stock.

Please fill out the following table for your project group. In your estimation, please indicate what percentage of the project each member of your group was responsible for/contributed to/worked on, etc. It will be kept 100% confidential, and **in no direct way will this be used for grading purposes**. Below is a sample that you should modify.

Team Member	**Percentage**
Yourself	40%
Steven Taruc	40%
Callen Schwefler	20%