This project is a basic twitter analysis of certain key words that would help trigger if it may be a immediate response to the current pandemic covid-19. This would be accomplished by streaming via our local machine using our twitter development accounts that we created earlier in the semester. The data would be retrieved using tweepy and placed in Spark RDDs that would then transfer the spark.SQL data. Lastly our data would be placed in a visual graph to represent the current hastags that are used most frequently. These hastags are updated every 10 seconds to keep a user up to date on if there was a sudden increase of the filtered words.

Errors:

While the code for my project does work it only accomplished about 65-70 percent of what I sought to do. I can filter the key words but I could not filter the state or distinguish retweets or original tweets. While the usefulness for the filtration would be to the user discretion this caused massive errors on the server side. Other problems that occurred was that since covid is ambiguous and world wide there are also tweets that are coming from other languages.

Final thoughts/ Resolution:

I accomplished a useful tool to capture tweets but may have looked to capture to many key words that makes this project more difficult than it should have been. Also I couldn’t fix the data so that it could be represented using Spark.SQL. The main restraint for this project is time, I believe if I would attempt this project again I would only use one word to filter so that I can use the twitter data more effectively to curate a synopsis on the problem.

To set up the serverless website I used the tutorial on this website:

<https://douglasduhaime.com/posts/s3-lambda-auth.html>

Source code for starting the streaming using tweepy was obtained from here:

<https://www.bmc.com/blogs/track-tweets-location/>

Link to project code :

<https://github.com/Rclark2019/Final.git>