## BIG DATA PROJECT CS644

## GROUP MEMBERS:

## 1.Nomrata Deb

## 2.Jedidiah John Nagaranjan

## 3.Manali Khokhal

## 4.Oun Muhammad

**Classification through Apache Spark streaming using Twitter data**

## Objective

The objective of this assignment is to learn streaming of data from social media platform like Twitter and applying machine learning techniques like classification algorithms and pattern recognition using big data analytical tools like Apache Spark.

## Task Description

The task of this assignment was to perform sentiment analysis on stream of tweets using Apache spark and storing the results in CSV format. We are using python 2.7 for this assignment and Spark 2.3.1. We wrote python code to stream real time Twitter using Apache Spark. We then trained the model with labeled training data using MLlib (Apache Spark Machine Learning Library). Using the trained model we carried out sentiment analysis on the batches of streamed data to classify them into positive, negative or neutral. To avoid issues with our cloud instance we terminated the streaming process after analyzing 2000 tweets using our trained model.

## Sentiment Analysis Algorithm

For this Assignment we have used RandomForestClassifier because from our analysis it gave us the maximum accuracy. We trained the model with 500 tweets and the accuracy is 57%. Despite our various attempts, we were not able to increase the accuracy of the model more than 57%. Since it is a classification problem, the model which was trained had 3 labels, making it a 3-class classification problem.

## Labeling Training Data

We already received labelled data to train our model. The data is in a CSV format with mostly all the emojis removed. The file has 6 fields and 3 labels. the polarity of the tweet (0 = negative, 2 = neutral, 4 = positive).

## Output

The result after the completion of sentiment analysis is stored in a csv file containing tweets and the result. The result is in given in range 0 to 2 with positive, neutral and negative respectively.

Graphical user interface, text, application

Description automatically generated