INTRODUCTION TO BIG DATA PROJECT CS644

GROUP MEMBERS:

1. Pradyumna Jangiti Ramchander – pj27

2. Hardik Rajendra Pakhale – hp495

3. Rakesh Reddy Nareddy – rn68

4. Illavenil Prabhakaran – ip223

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 description of this assignment was to perform sentiment analysis of tweets using Apache Spark and to store the data collected in CSV format. The following two software's were used for this project:

1. Python
2. Spark

We used python to write the code for it to stream real-time Twitter with the help of Spark. Furthermore, we trained the model with the data using MLlib. We carried out sentiment analysis on the batches of the previously streamed data to classify them into positive, neutral, or negative using the trained model. Tweets were auto terminated after analyzing 2000 tweets using the trained model.

Sentiment Analysis Algorithm:

We have used RandomForestClassifier in this assignment as from our analysis it gave us the most accurate result. The model was trained with 500 tweets and the accuracy is 57%. After multiple attempts, we were not able to increase the accuracy of the model. Since it is an issue of classification, the model, which was trained had 3 labels, making it a 3-class classification problem.

Labeling Training Data:

We have 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 output/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.

A	B	C
TEXT	PREDICTION	
Has anyone seen web3? I can't find it.	1	
RT @This might be the vivo V23 Pro globally— the vivo S12 Pro which is I	1	
Tesla opened another 250kW Supercherger in SLC. Can't say it's unneede	1	
nyc https://twitter.com/BrandonJHavard/status/1473156214267826183/	1	
Buying crypto has often felt exclusionary. In order to democratize who ca	1	
RT @Please donate directly to NGO. I hope! The gov will use the money t	1	
RT @I know a few of these kinda people , don't worry they are usually th	1	
don't lie to yourself and think you're better than the person whose sins a	1	
iOS 15.1 and iOS 15.1.1 are not signed by Apple anymore. This means yo	1	
RT @LardDevis #crypto is probably the only market where someone with	1	
Old fashion. #teampixel https://twitter.com/JeremiahBonds/status/1473	1	
Top 0.01% Bitcoin owners own 27% of Bitcoin wealth For comparison, th	1	
RT @Twitter FINALLY https://twitter.com/Davideml/149288	1	
Airplane wifi is god's way of showing the youngs what dial-up was like.	0	
CANT GET OVER THIS https://twitter.com/DurvidImel/status/147304739	0	
#Bitcoin en vue jour tente le break du biseau et du canal descendant Rocl	1	
RT @LinusTechTips thanks bing, i didn't want the answer to my question	0	
Happy Monday, especially to all Android users out there! Today we're rel	1	
low key, Dyson makes one of the best looking smart home control apps o	1	
RT @DavidRuddock it's a big, colorful, but extremely readable interface.	0	
Apple Releases Safari Technology Preview https://macrumors.com/2021	1	
**reminder to get up, walk around, check your posture, & do some stretc	1	