Twitter Data Analysis
Task 1: Cindy Ho
Task 2: Gurvinder Mann
Task 3: Daniel Martinez Garzon

Project Introduction

Our project revolved around building a learning model for Twitter analysis. In our tasks, we cleaned a given dataset, selecting the relevant attributes. We assigned the relevant attributes to each tweet based on the most frequent list. From there, we were able to assign a topic for each tweet using a relationship between the tweet and features. Applying this model to then predict one topic for each tweet. For this project, we used Scala to access Spark SQL for the data preparation and use the Spark MLlib to run the machine learning algorithm.

Task 1: Data Preparation 1

In this section, I was asked to clean the given dataset by saving only the necessary attributes in a new JSON file, and with the clean dataset, find the top 20 hashtags that are frequently used and keep it as an array of keywords.
The top 20 keywords found in the 10k dataset are:

ALDUBxEBLoveis	no309	FurkanPalalı	LalOn	chien
job	Hiring	sbhawks	Top3Apps	perdu
trouvé	CareerArc	trumprussia	trndnl	Job
Jobs	hiring	ShowtimeLetsCelebr8	impeachtrumppence	Music

Task 2: Data Preparation 2

In Task 2, we are required to add a new column named topic to the Twitter dataset obtained in Task 1, which will indicate the most frequent hashtag for each tweet. If a tweet contains more than one top hashtag, any of them can be used. We are given a sample file to work with which is the Task 1 output.

We load the file as a JSON file using sparkSession.read.format("json"), and then create a temporary view of the data frame to enable us to use SQL to manipulate it.

Next, we define an array of the top 20 hashtags, which we will use to filter out hashtags not in the top 20.

We then filter out hashtags that are not in the top 20 using array_contains and exists functions in the SQL query. We add a temporary column, temp1, and another column, row1, to enable us to keep the first element of the intersection of the relevant hashtags array with the top 20 hashtags using the array_intersect function.

Finally, we select the required columns id, text, topic, user_description, retweet_count, reply_count, and quoted_status_id and save the resulting data frame as a JSON file to be used in Task 3. The total number of records in the tweets_topic dataset for the 10k dataset is 269.

Task 3: Topic Prediction

For this task, I was asked to build a machine learning model that assigns a topic for each tweet based on the classified tweets. The machine learning model had to create a relationship between features and the topics. The machine learning model had to use a tokenizer, hashing, string indexer, and linear regression in order to try and correctly predict the corresponding topic given the tweet text.

We tested the input using a training-test split to train on one set and test on the other using a multiclass evaluator. We computed the precision and accuracy on the 10k dataset.