BIG_DATA

Date: 06/12/2021

PROJECT REPORT

TEAM-MEMBERS

1. TEJAS KUMAR S PES2UG19CS428
2. MALAVIKA PES2UG19CS212
3. HIMANI OM PES2UG19CS149
4. SWETHA PRABHU PES2UG19CS422

Spark Streaming for Machine Learning-SSML
DataSet Used: Sentiment Analysis

DESIGN__

Importing libraries First, we import inbuilt libraries from spark such as spark context, streamingcontext, spark session, re (Reg_ex), etc. Using streaming data we fetch csv file required for working We then do Text Preprocessing for Tweet column on the data in order to remove irrelevant text like punctuations, white space, url removing techniques using lambda functions, tokenizer, stopwords remover, lemmatizer etc so that we get a cleaner and better data to work on. We do Feature Extraction for using HashingTF to get the final data used for model building. Model building (we use different types of models to increase the accuracy of the results such as Perceptron and PassiveAggresiveClassifier which has partial_fit method for incremental learning. We find accuracy, recall, precision etc for each model using confusion matrix implementation and perform classification for testing data.

SURFACE LEVEL IMPLEMENTATION__

Importing libraries: we import the required lib for processing of data.
Streaming: we stream data to get the tweets for processing using the stream.py file and get the required data.
Preprocessing: we clean the data to process and get the maximum accurate results after model building.
Model building: we choose some particular models which are most suitable for our data to get good accuracy

REASON BEHIND DESIGN__

Import lib: We have used pyspark library because it allows us to write Spark applications using python APIs.

Streaming: We have used data streams because it allows us to extract and process data in real-time.

Preprocessing: We have preprocessed the dataset to convert the raw data into a clean data set because data collected in raw format is not feasible for the analysis.
Model buildings: We have built a model to predict if a given Tweeet is negative or positive.

TAKE AWAY FROM PROJECT__

From this project we were able to classify if a given Tweet is negative or positive.after implementing variousmethods.