BIG DATA

Spark Streaming for Machine Learning

Team : BD1_430_435_450

Sanjana – PES1UG19CS430

Sathwik - PES1UG19CS435

Sharan - PES1UG19CS450

Dataset – Email Spam Analysis

Design Details :

We first streamed the dataset from spark streaming. The streamed dataset is then preprocessed which gives us the cleaned dataset.

For our first model we chose Multinomial Naïve Bayes

For our second model we chose Perceptron

For out third model we chose Bernoulli Naïve Bayes

Implementation :

We stream by checking if each rdd is empty or not.

Then we cleaned the data from stop words using the ENGLISH_STOP_WORDS

We also calculated the Accuracy, recall and precision for each of the models, from using the training dataset.

Reason behind the design decisions :

We chose the models by hit and try method to find out which gives the best scores for our outcome.

Takeaways :

This project helped us explore all the various functions of sklearn, pyspark. Also opened our minds to see how realtime data is processed for the maximum learning and helped us understanding big data better.

ThankYou