BIG DATA
SPARK STREAMING USING MACHINE LEARNING
REPORT

TeamID: BD2_217_243_394

The Topic Chosen is Email spam classification.

Design Details

The data consists of a set of emails and having columns feature0 (subject), feature1 (body of the email) and feature3 (class). The model predicts whether a particular email is spam or not using classifiers like Naive Bayes and Logistic regression.

Surface Level Implementation

The data was first streamed through a TCP socket, and recieved as a json string which was further converted into a spark dataframe.
Preprocessing was done by computing the length (number of words) of the body and then adding that column to the dataset. Average of the length grouped by class was found and sent for feature transformation.
Then, the data was divided into label and features by dimensionally reducing the body of the email .

The models we used were:

- Naive Bayes : Naive Bayes classification is a simple probability algorithm based on the fact, that all features of the model are independent. In the context of the spam filter, every word in the message is independent of all other words and we count them with the ignorance of the context.

- Logistic regression: Logistic regression is a supervised learning classification algoithm used to predict the probability of a target varibale. The nature of this variable is dichotomous, which means there would only be two possible classes (spam and ham for our dataset)

These were implemented by using spark mllib and importing the respective libraries and packages needed, and the accuracy of each model was determined.