BIG DATA PROJECT REPORT

SPARK STREAMING FOR MACHINE LEARNING

DATASET : Sentimental Analysis

TEAM DETAILS:

NAME	SRN
Nischitha P Chinnari	PES1UG19CS301
Nutheti Nikhila Priya	PES1UG19CS309
S S Priya	PES1UG19CS404
Kakumani Sai Deepika	PES2UG19CS903

About Dataset:

1. Two CSV files each for training (1520k records) and testing (80k records).
2. Sentiment is either 0 (negative) or 4 (positive).
3. Each record consists of two features,sentiment and text.

Task Specification:

1. Streaming the data
2. Processing the stream
3. Building the model
4. Testing the model
5. Clustering

Design Details:

1. Stream the dataset.
2. Receive the stream of input.
3. After streaming we applied preprocessing techniques on our dataset

4. After preprocessing we obtained a clean dataframe with required features.
5. Using sklearn we applied our first model which is Naive Bayes

Implementation Details:

- ➢ For each RDD, we first checked if it not None and then did streaming
- ➢ We converted it into a list and created a data frame

The steps used to complete preprocessing our data were:

1. Remove punctuation.
2. Remove special characters.
3. Make text lowercase.
4. Remove stopwords.

- ➢ The Naive Bayes classification technique is a simple and powerful classification task in machine learning.When used for textual data analysis the Naive Bayes classification yields good results.
- ➢ We found recall, precision ,F1 score
- ➢ Area under precision recall curve, Area under ROC

Reason behind design decisions:

We have used sklearn for the implementation of our models since it is versatile,free and easy to use. It helped us in model building and get meaning out of it.Scikit-learn is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression and clustering.

Takeaways:

This project has helped us learn about how applications in the real world modify their work on large data streams and how incremental processing can be leveraged to process and analyse streams over time