PES UNIVERSITY - EC CAMPUS

BIG DATA PROJECT

Machine Learning with Spark Streaming

Team Member Names	SRN
Kale Pranav	PES2UG19CS174
Prachi Sengar	PES2UG19CS285
Raeesa Tanseen	PES2UG19CS310
Lavanya Yavagal	PES2UG19CS904

Dataset

Name: Spam Dataset

Description: The dataset contains two files, train.csv and test.csv. It is about spam messages. It has 3 attributes: Subject, Message, Spam/Ham.

All 3 attributes are of string data type. Based on the Subject and Message, the message has to be classified as Spam or Ham.

Design Details:

- ➔ Apache Spark is an open-source unified analytics engine for large-scale data processing that provides an interface for programming entire clusters with implicit data parallelism.
- ➔ PySpark is an interface for Apache Spark in Python. It allows us to write Spark applications using Python APIs. PySpark supports most of Spark's features such as Spark SQL, DataFrame, Streaming, MLlib (Machine Learning) and Spark Core.
- ➔ Spark SQL is a Spark module for structured data processing.
- ➔ Running on top of Spark, the streaming feature in Apache Spark enables powerful interactive and analytical applications across both streaming and historical data while being easy to use and fault tolerant.
- ➔ Built on top of Spark, MLlib is a scalable machine learning library that provides a uniform set of high-level APIs that help users create and tune practical machine learning pipelines.
- ➔ Spark Core is the underlying general execution engine for the Spark platform that all other functionality is built on top of. It provides an RDD (Resilient Distributed Dataset) and in-memory computing capabilities.
- ➔ Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python.

Surface Level Implementation Details:

- ➔ Streaming: We have used pyspark to stream the data onto a tcp server. Reading Dstream as RDD from the tcp socket, we convert the stream to a dataframe.
- ➔ Pre-processing: We applied Tokenizer, CountVectorizer, StringIndexer for class labels, Pipelining, Vector Assembler, IDF etc.
- ➔ Model Building for detecting spam mail: Implemented Naive Bayes classifier using MultinomialNB, SGD classifier using SGDClassifier, KNN using KNeighborsClassifier from Sklearn
- ➔ Model Testing: Testing the model using test dataset and finding confusion matrix, accuracy etc.
- ➔ Clustering: K means clustering using MiniBatchKMeans() from Sklearn
- ➔ Implementation uses dataframes and RDDs.

Reason behind Design Decisions:

- ➔ Spark Streaming allows us to use Machine Learning to the data streams for advanced data processing. It also provides a high-level abstraction that represents a continuous data stream.
- ➔ Spark MLlib is designed for simplicity, scalability, and easy integration with other tools. With the language compatibility, and speed of Spark, we can solve and iterate through data problems faster.
- ➔ Resilient Distributed Data set (RDD) is the basic component of Spark that helps in managing the distributed processing of data. Transformations and actions can be made on the data with fault tolerant RDDs. Each data set in RDD is firstly partitioned into logical portions and it can be computed on different nodes of the cluster parallely.
- ➔ Scikit-Learn is a higher-level library that includes implementations of several machine learning algorithms, so you can define a model object in a single line or a few lines of code, then use it to fit a set of points or predict a value.

Take away from the Project:

- ➔ With this project, we got a thorough understanding about how applications in the real world modify their algorithms to work on large data streams.
- ➔ We learnt how to handle this enormous data only in batches at any given point of time.
- ➔ We also learnt how incremental processing can be leveraged to process and analyze streams over time to achieve predictive modelling.
- ➔ We learnt how to analyze and process data streams for machine learning tasks using Spark Streaming and Spark MLLib to draw insights and deploy models for predictive tasks.