Case Study : Spam Detection Domain: Telecom

A telecom software provider is building an application to monitor different telecom components in the production environment. For monitoring purpose, the application relies on log files by parsing the log files and looking for potential warning or exceptions in the logs and reporting them.

The POC we had been working on, for SPAM Detection on the data of telecom operator forum, has been accepted and the stakeholders has asked us to work on the real-time example for predicting SPAM messages.

Tasks:

This POC will focus on saved machine learning model for spam prediction with streaming data to do real-time prediction. Now with model and data pipeline ready, you are required to predict the spam message on the steaming data.

1. Modify the model application to train the model and persist it.

2. Create a new spark streaming application to predict the spam messages

3. Application will connect to the flume to retrieve the data

4. Load the model

5. Predict the SPAM messages and print the SPAM in the logs

6. Test the application by sending dummy data rows from the consumer

Hint:

Below is the example for Spam Messages:

• HAM: What you doing? How are you?

• SPAM: Sunshine Quiz! Win a super Sony DVD recorder if you can name the capital of Australia? Text MQUIZ to 82277

Solution:

- Jupyter Notebook has been attached with code , comments and outputs
- Sample Input/output of the Streaming request is shown below
- Flume Configuration file has also been attached

Input:

curl -X POST -H 'Content-Type: application/json; charset=UTF-8' -d '[{"headers" : {"a":"b"},"body": "{\"spam\" : \"spam\" , \"message\" :\" Hello how are you \" }"}] ' http://localhost:9000

Output:

```
|label|prediction|
+-----+----------+
+-----+----------+

-------------------------------------------
Time: 2020-07-21 04:21:15
-------------------------------------------
1

+-----+----------+
|label|prediction|
+-----+----------+
|  1.0|       0.0|
+-----+----------+

-------------------------------------------
Time: 2020-07-21 04:21:30
```

Command to run Flume :

./flume-ng agent --conf conf --conf-file /mnt/home/edureka_960126/flume_spark.conf --name httpagent -Dflume.root.logger=INFO,console

Architecture of the solution:

HTTPSOURCE → FLUME AGENT → SPARK STREAMING PULLING DATA & PREDICTING DATA BASED ON THE SAVED MODEL

Notes:

Note that the Jupyter Notebook has the code continued over the Module 8 Spam Detection ML Model building application , so last portion is the streaming portion.

Note Inputs are sent in the Json Format otherwise in the streaming portion DStreams(sequence of RDDs formed by micro batches collected by the Spark Streaming) can be mapped to convert message stream into list by splitting the message on tab and then putting them in Dataframe with same schema