## DATA ANALYTICS PIPELINE USING APACHE SPARK

## Environment: Hadoop VM

#### By: Aditya Subramanian Muralidaran

### *Project Description:*

### To build a data pipeline that takes a document as an input and uses Apache Spark MapReduce and Machine Learning libraries to classify in one of the topics – Business, Sports, Politics and Technology. This pipeline is built using the following procedure:

### Data Collection: A large number of articles (1000 approx.) is collected from NY Times API using a python script on topics - Business, Sports, Politics and Technology.

### Feature Engineering: Apache Spark MapReduce framework is used in python programming language to get the top 40 words (representing each class) from each class using word count algorithm. Cumulatively these 120 words (approx.) will be used as feature in classification algorithms.

### The data gathered from NY Times is split into 80% training set and 20% testing set. Using the features extracted, a machine learning model is built (Naïve Bayes and Multi-Layer Perceptron) and the accuracy is determined.

### Once the model is built, a random document given to the model will be classified into one of the 4 classes - Business, Sports, Politics and Technology.

### Work Flow:

### 

### *Steps to calculate accuracy:*

For this project we have collected data from New York Times Articles.

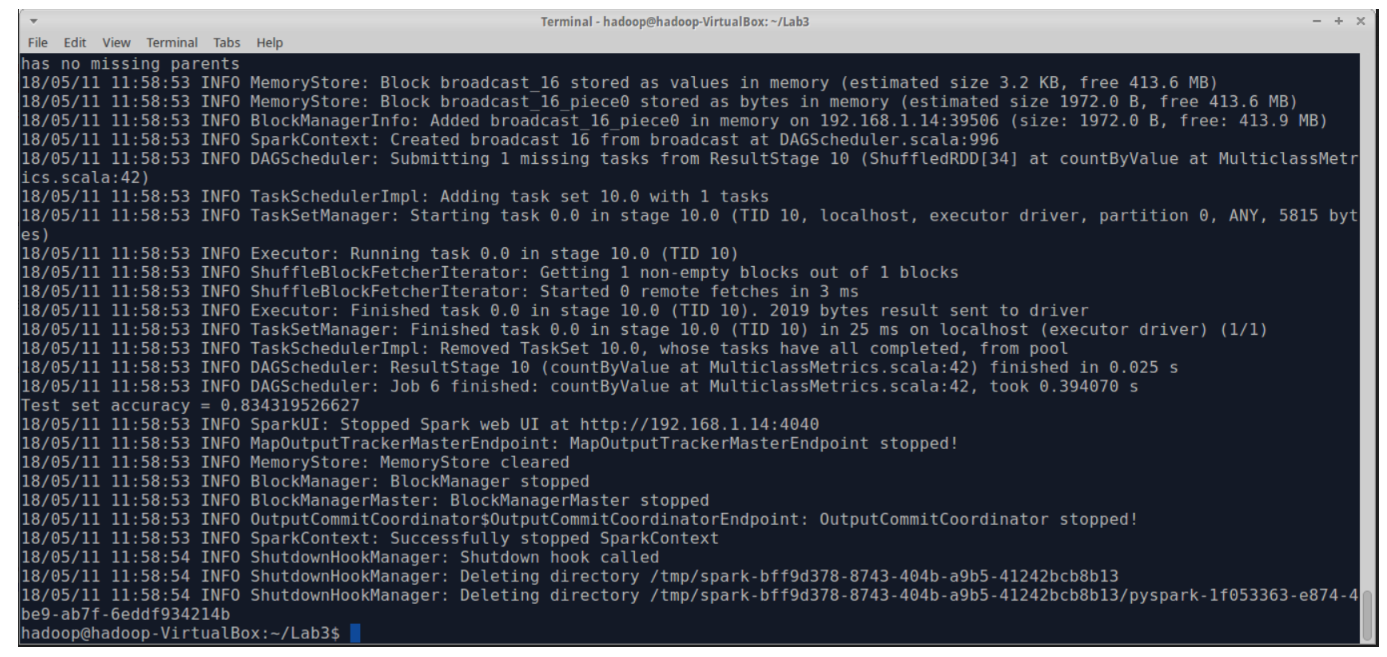
**Step - 1:**Run the following command :

*spark-submit WordCount\_MapReduce.py Business/ Politics/ Sports/ Tech/ classify\_input/*    
This generates the input file in the format needed (classify\_input) for calculating accuracy of classification based on naïve bayes and multilayer perceptron methods.   
The format of the output will be like: 0 5:2 6:1 12:7 16:1 18:1 24:4 32:2 43:1 55:2 57:2 60:5 74:1 77:3 92:1 104:3   
The General Format is given by: {class}{feature\_1}:{count}{feature\_1}:{count}....   
The above command will also create and save a file as ‘feature\_file’, that contains the list of words that are used as selected features.

**Step - 2:**Then run the following command for naïve bayes model:

*spark-submit classify\_ml\_naive\_bayes.py classify\_input/*

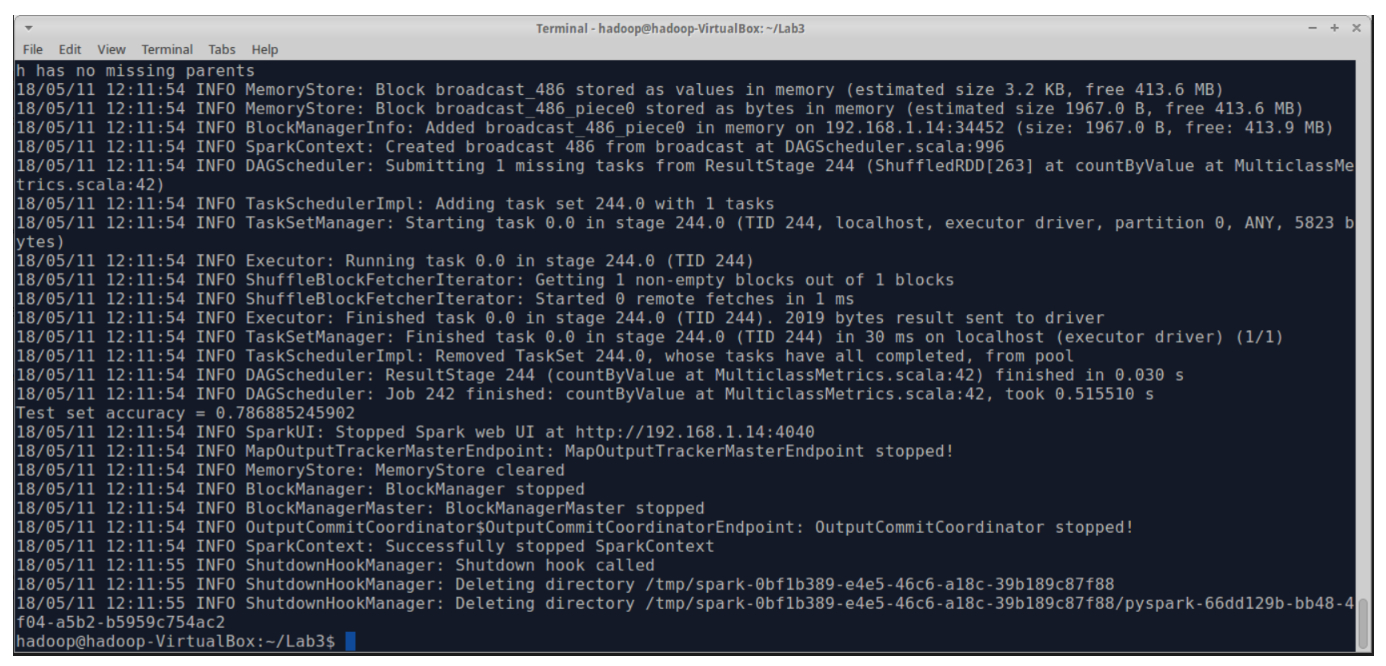
to get accuracy based on naive bayes method. From the navie bayes method we get an accuracy of 83.43%



**Step - 3:**Then run the following command for multi-layer perceptron model:

*spark-submit classify\_multilayer\_perceptron.py classify\_input/*

to get accuracy based on multi-layer perceptron method. Accuracy: 81.95%



***Classifying a Given Random File:***

Then follow the below steps to predict class of a given file:

* Run the command to format the given article:

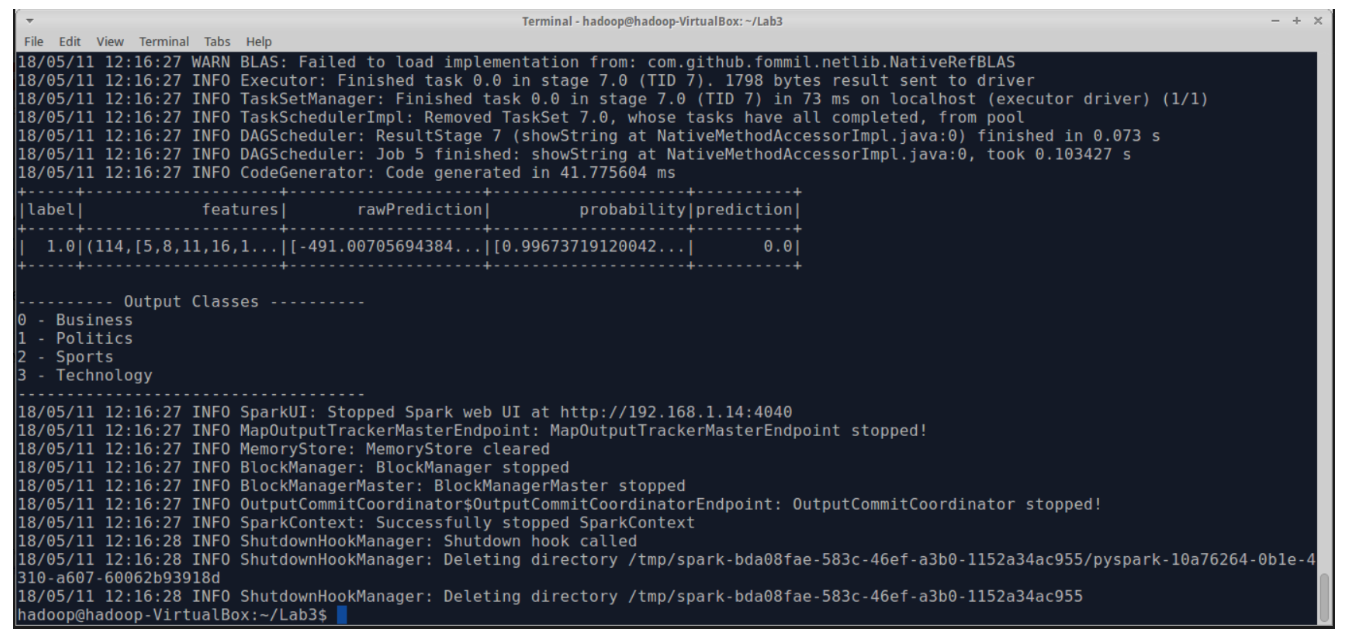
*spark-submit ClassifyFile\_MapReduce.py features\_file/ Article.txt fileclassify*

* The output from previous step ‘fileclassify’ is the input file in the needed format for classifying the article.
* Format of the output will be like : 1 1:0 2:0 3:0 4:0 5:0 6:1 7:0 8:0 9:1 10:0 11:0 12:3 13:0 14:0 15:0 16:0 17:1 18:0 19:1 20:0 21:0 22:0 23:0 24:1 25:3 26:0 27:0 28:0 29:1 30:0 31:0 32:2 33:1 34:1 35:0 36:1 37:0 38:0 39:0 40:0 41:0 42:0 43:22 44:0 45:4 46:0 47:0 48:0 49:0 50:1 51:0 52:0 53:0 54:4 55:1 56:0 57:0 58:5 59:0 60:1 61:0 62:0 63:1 64:0 65:0 66:0 67:0 68:5 69:1 70:0 71:0 72:2 73:0 74:0 75:1 76:0 77:0 78:0 79:0 80:1 81:0 82:0 83:0 84:0 85:7 86:4 87:5 88:0 89:0 90:0 91:0 92:0 93:0 94:1 95:19 96:0 97:0 98:3 99:0 100:0 101:0 102:1 103:0 104:0 105:0 106:3 107:0 108:0 109:3 110:0 111:0 112:0 113:0 114:0
* **Classification Based on Naïve Bayes Model:** Run the command:

*spark-submit FileClassification\_naive\_bayes.py classify\_input/ fileclassify/*

to classify the article based on naive bayes. The input for the naïve Bayes will be stored in the file ‘BusinessArticle.txt’

* We can see the classifiction output of the naïve bayes is in the screenshot below.



* **Classification Based on Multi-Layer Perceptron Model**: Run the command:

*spark-submit FileClassify\_multilayer\_perceptron.py classify\_input/ fileclassify/*

to classify the article based on multilayer perceptron. The input for the Multilayer Perceptron will be stored in the file ‘BusinessArticle.txt’

