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:

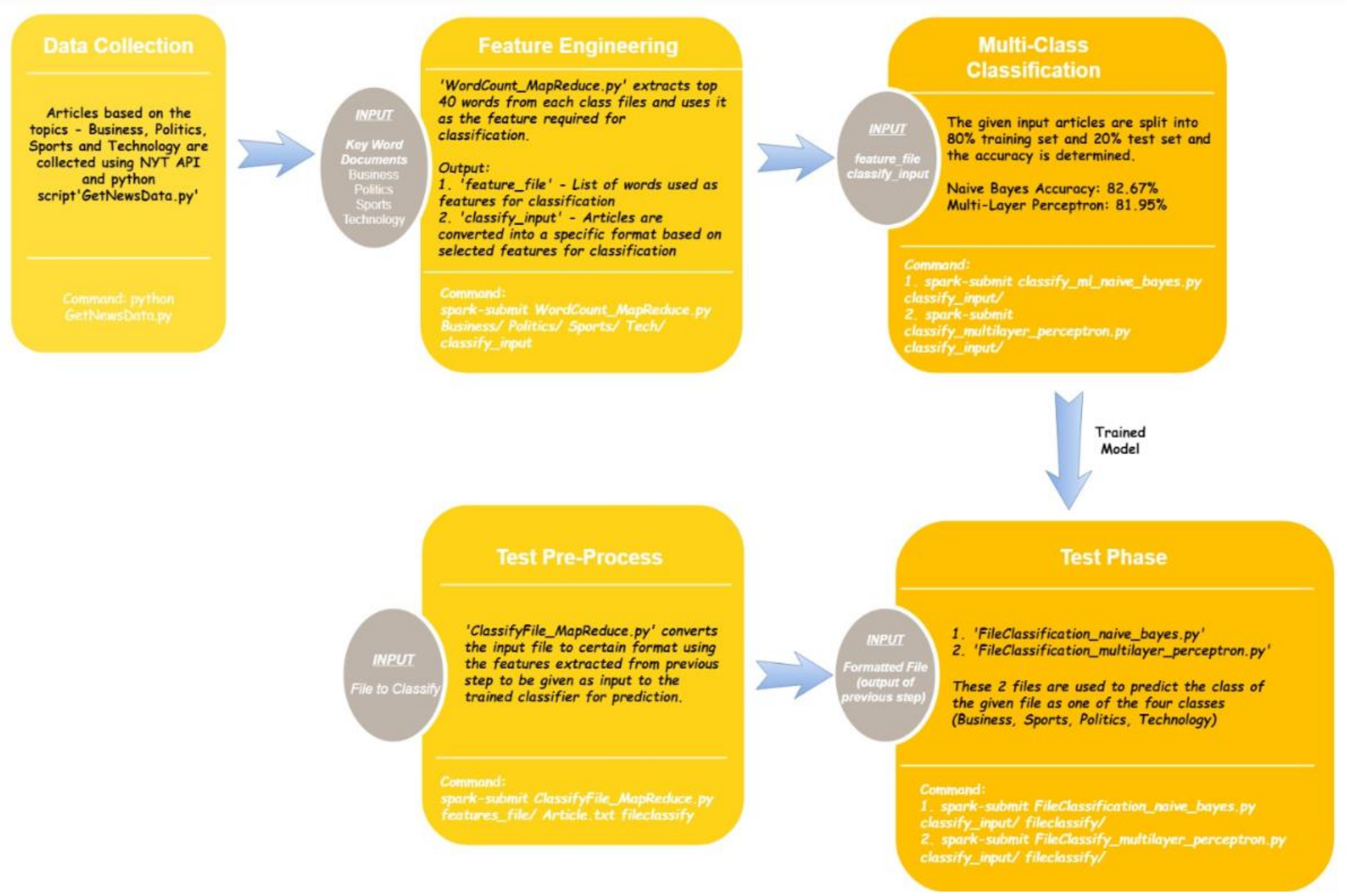
- Data Collection: A large number of articles (1000 approx.) is collected from NY Times API using a python script on topics - Business, Sports, Politics and Technology.

- 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:



Data Collection

Articles based on the topics - Business, Politics, Sports and Technology are collected using NYT API and python script'GetNewsData.py'

```
Command: python GetNewsData.py
```

Feature Engineering

'WordCount_MapReduce.py' extracts top 40 words from each class files and uses it as the feature required for classification.

Output:
1. 'feature_file' - List of words used as features for classification
2. 'classify_input' - Articles are converted into a specific format based on selected features for classification

```
Command:
spark-submit WordCount_MapReduce.py Business/ Politics/ Sports/ Tech/ classify_input
```

Multi-Class Classification

The given input articles are split into 80% training set and 20% test set and the accuracy is determined.

Naive Bayes Accuracy: 82.67%
Multi-Layer Perceptron: 81.95%

```
Command:
1. spark-submit classify_ml_naive_bayes.py classify_input/
2. spark-submit classify_multilayer_perceptron.py classify_input/
```

Test Pre-Process

'ClassifyFile_MapReduce.py' converts the input file to certain format using the features extracted from previous step to be given as input to the trained classifier for prediction.

```
Command:
spark-submit ClassifyFile_MapReduce.py features_file/ Article.txt fileclassify
```

Test Phase

1. 'FileClassification_naive_bayes.py'
2. 'FileClassification_multilayer_perceptron.py'

These 2 files are used to predict the class of the given file as one of the four classes (Business, Sports, Politics, Technology)

```
Command:
1. spark-submit FileClassification_naive_bayes.py classify_input/ fileclassify/
2. spark-submit FileClassify_multilayer_perceptron.py classify_input/ fileclassify/
```

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

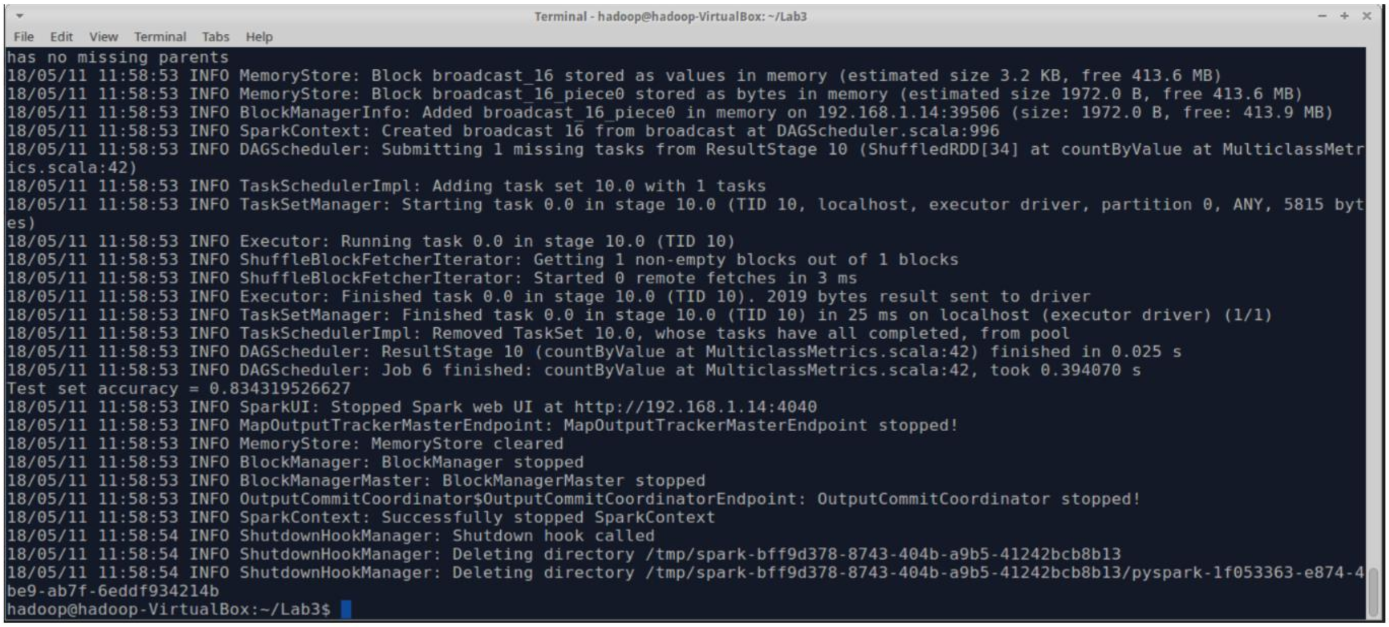
The General Format is given by: {class}{feature_1}:{count}{feature_1}:{count}....

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%



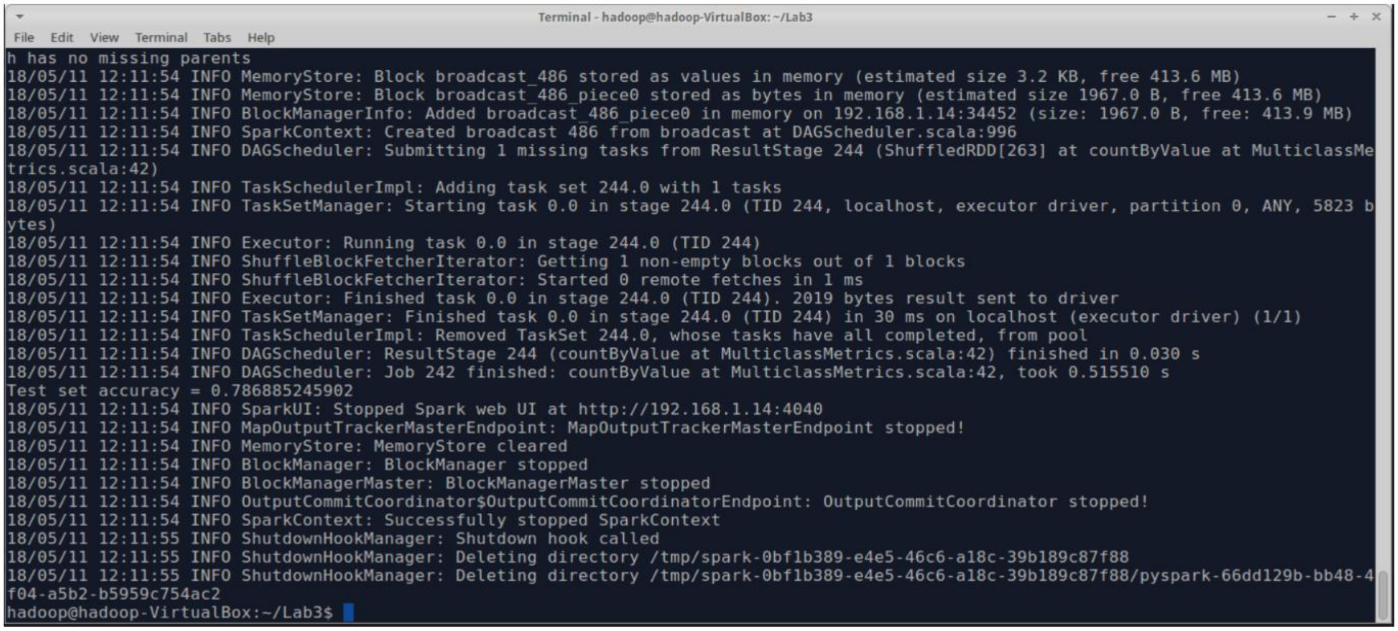
Terminal - hadoop@hadoop-VirtualBox: ~/Lab3
File Edit View Terminal Tabs Help

```
has no missing parents
18/05/11 11:58:53 INFO MemoryStore: Block broadcast_16 stored as values in memory (estimated size 3.2 KB, free 413.6 MB)
18/05/11 11:58:53 INFO MemoryStore: Block broadcast_16_piece0 stored as bytes in memory (estimated size 1972.0 B, free 413.6 MB)
18/05/11 11:58:53 INFO BlockManagerInfo: Added broadcast_16_piece0 in memory on 192.168.1.14:39506 (size: 1972.0 B, free: 413.9 MB)
18/05/11 11:58:53 INFO SparkContext: Created broadcast 16 from broadcast at DAGScheduler.scala:996
18/05/11 11:58:53 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 10 (ShuffledRDD[34] at countByValue at MulticlassMetr
ics.scala:42)
18/05/11 11:58:53 INFO TaskSchedulerImpl: Adding task set 10.0 with 1 tasks
18/05/11 11:58:53 INFO TaskSetManager: Starting task 0.0 in stage 10.0 (TID 10, localhost, executor driver, partition 0, ANY, 5815 byt
es)
18/05/11 11:58:53 INFO Executor: Running task 0.0 in stage 10.0 (TID 10)
18/05/11 11:58:53 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/11 11:58:53 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 3 ms
18/05/11 11:58:53 INFO Executor: Finished task 0.0 in stage 10.0 (TID 10). 2019 bytes result sent to driver
18/05/11 11:58:53 INFO TaskSetManager: Finished task 0.0 in stage 10.0 (TID 10) in 25 ms on localhost (executor driver) (1/1)
18/05/11 11:58:53 INFO TaskSchedulerImpl: Removed TaskSet 10.0, whose tasks have all completed, from pool
18/05/11 11:58:53 INFO DAGScheduler: ResultStage 10 (countByValue at MulticlassMetrics.scala:42) finished in 0.025 s
18/05/11 11:58:53 INFO DAGScheduler: Job 6 finished: countByValue at MulticlassMetrics.scala:42, took 0.394070 s
Test set accuracy = 0.834319526627
18/05/11 11:58:53 INFO SparkUI: Stopped Spark web UI at http://192.168.1.14:4040
18/05/11 11:58:53 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/11 11:58:53 INFO MemoryStore: MemoryStore cleared
18/05/11 11:58:53 INFO BlockManager: BlockManager stopped
18/05/11 11:58:53 INFO BlockManagerMaster: BlockManagerMaster stopped
18/05/11 11:58:53 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/05/11 11:58:53 INFO SparkContext: Successfully stopped SparkContext
18/05/11 11:58:54 INFO ShutdownHookManager: Shutdown hook called
18/05/11 11:58:54 INFO ShutdownHookManager: Deleting directory /tmp/spark-bff9d378-8743-404b-a9b5-41242bcb8b13
18/05/11 11:58:54 INFO ShutdownHookManager: Deleting directory /tmp/spark-bff9d378-8743-404b-a9b5-41242bcb8b13/pyspark-1f053363-e874-4
be9-ab7f-6eddf934214b
hadoop@hadoop-VirtualBox:~/Lab3$ 
```

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%



Terminal - hadoop@hadoop-VirtualBox: ~/Lab3
File Edit View Terminal Tabs Help

```
h has no missing parents
18/05/11 12:11:54 INFO MemoryStore: Block broadcast_486 stored as values in memory (estimated size 3.2 KB, free 413.6 MB)
18/05/11 12:11:54 INFO MemoryStore: Block broadcast_486_piece0 stored as bytes in memory (estimated size 1967.0 B, free 413.6 MB)
18/05/11 12:11:54 INFO BlockManagerInfo: Added broadcast_486_piece0 in memory on 192.168.1.14:34452 (size: 1967.0 B, free: 413.9 MB)
18/05/11 12:11:54 INFO SparkContext: Created broadcast 486 from broadcast at DAGScheduler.scala:996
18/05/11 12:11:54 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 244 (ShuffledRDD[263] at countByValue at MulticlassMe
trics.scala:42)
18/05/11 12:11:54 INFO TaskSchedulerImpl: Adding task set 244.0 with 1 tasks
18/05/11 12:11:54 INFO TaskSetManager: Starting task 0.0 in stage 244.0 (TID 244, localhost, executor driver, partition 0, ANY, 5823 b
ytes)
18/05/11 12:11:54 INFO Executor: Running task 0.0 in stage 244.0 (TID 244)
18/05/11 12:11:54 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/11 12:11:54 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/05/11 12:11:54 INFO Executor: Finished task 0.0 in stage 244.0 (TID 244). 2019 bytes result sent to driver
18/05/11 12:11:54 INFO TaskSetManager: Finished task 0.0 in stage 244.0 (TID 244) in 30 ms on localhost (executor driver) (1/1)
18/05/11 12:11:54 INFO TaskSchedulerImpl: Removed TaskSet 244.0, whose tasks have all completed, from pool
18/05/11 12:11:54 INFO DAGScheduler: ResultStage 244 (countByValue at MulticlassMetrics.scala:42) finished in 0.030 s
18/05/11 12:11:54 INFO DAGScheduler: Job 242 finished: countByValue at MulticlassMetrics.scala:42, took 0.515510 s
Test set accuracy = 0.786885245902
18/05/11 12:11:54 INFO SparkUI: Stopped Spark web UI at http://192.168.1.14:4040
18/05/11 12:11:54 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/11 12:11:54 INFO MemoryStore: MemoryStore cleared
18/05/11 12:11:54 INFO BlockManager: BlockManager stopped
18/05/11 12:11:54 INFO BlockManagerMaster: BlockManagerMaster stopped
18/05/11 12:11:54 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/05/11 12:11:54 INFO SparkContext: Successfully stopped SparkContext
18/05/11 12:11:55 INFO ShutdownHookManager: Shutdown hook called
18/05/11 12:11:55 INFO ShutdownHookManager: Deleting directory /tmp/spark-0bf1b389-e4e5-46c6-a18c-39b189c87f88
18/05/11 12:11:55 INFO ShutdownHookManager: Deleting directory /tmp/spark-0bf1b389-e4e5-46c6-a18c-39b189c87f88/pyspark-66dd129b-bb48-4
f04-a5b2-b5959c754ac2
hadoop@hadoop-VirtualBox:~/Lab3$
```

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:

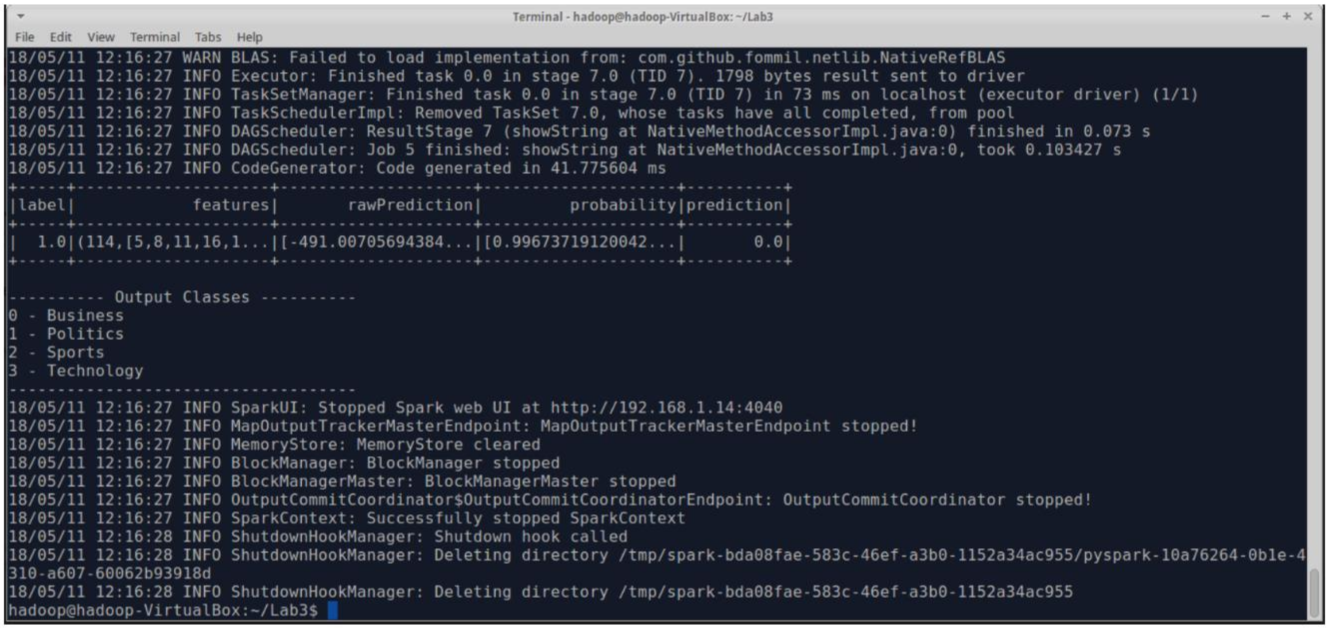
  spark-submit ClassifyFile_MapReduce.py features_file/ Article.txt fileclassify

- 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.



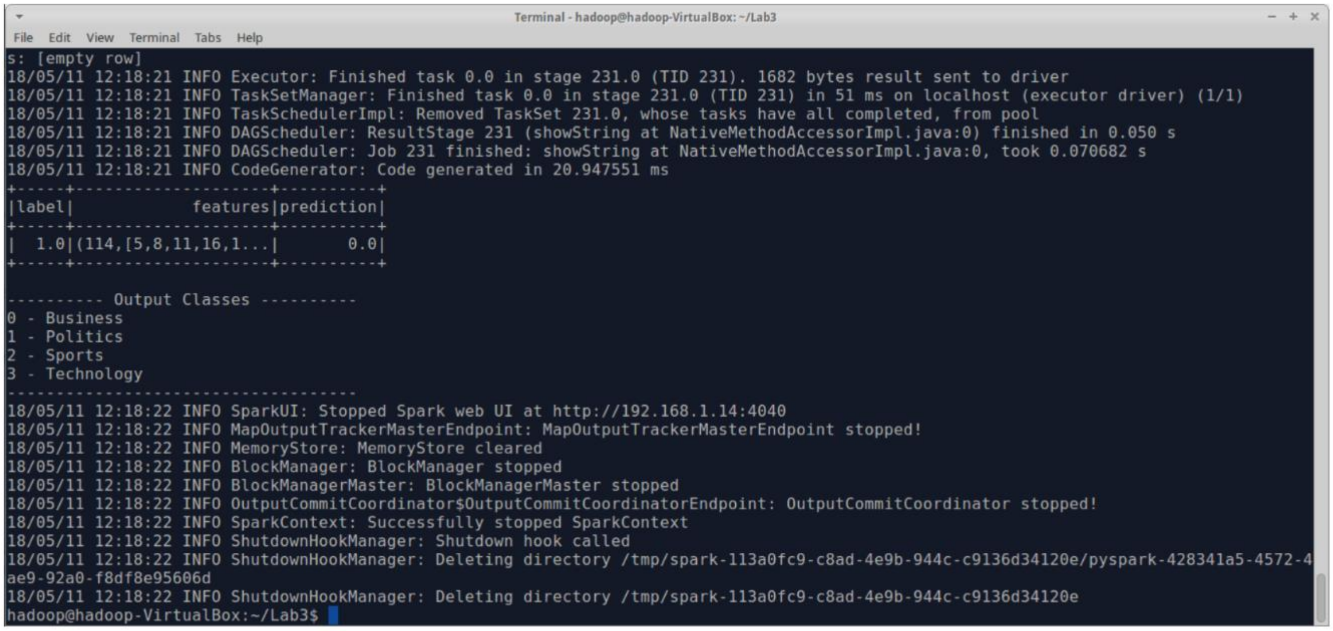
```
18/05/11 12:16:27 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
18/05/11 12:16:27 INFO Executor: Finished task 0.0 in stage 7.0 (TID 7). 1798 bytes result sent to driver
18/05/11 12:16:27 INFO TaskSetManager: Finished task 0.0 in stage 7.0 (TID 7) in 73 ms on localhost (executor driver) (1/1)
18/05/11 12:16:27 INFO TaskSchedulerImpl: Removed TaskSet 7.0, whose tasks have all completed, from pool
18/05/11 12:16:27 INFO DAGScheduler: ResultStage 7 (showString at NativeMethodAccessorImpl.java:0) finished in 0.073 s
18/05/11 12:16:27 INFO DAGScheduler: Job 5 finished: showString at NativeMethodAccessorImpl.java:0, took 0.103427 s
18/05/11 12:16:27 INFO CodeGenerator: Code generated in 41.775604 ms
+-----+--------------------+--------------------+--------------------+----------+
|label|            features|       rawPrediction|         probability|prediction|
+-----+--------------------+--------------------+--------------------+----------+
|  1.0|(114,[5,8,11,16,1...|[-491.00705694384...|[0.99673719120042...|       0.0|
+-----+--------------------+--------------------+--------------------+----------+

---------- Output Classes ----------
0 - Business
1 - Politics
2 - Sports
3 - Technology
------------------------------------
18/05/11 12:16:27 INFO SparkUI: Stopped Spark web UI at http://192.168.1.14:4040
18/05/11 12:16:27 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/11 12:16:27 INFO MemoryStore: MemoryStore cleared
18/05/11 12:16:27 INFO BlockManager: BlockManager stopped
18/05/11 12:16:27 INFO BlockManagerMaster: BlockManagerMaster stopped
18/05/11 12:16:27 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/05/11 12:16:27 INFO SparkContext: Successfully stopped SparkContext
18/05/11 12:16:28 INFO ShutdownHookManager: Shutdown hook called
18/05/11 12:16:28 INFO ShutdownHookManager: Deleting directory /tmp/spark-bda08fae-583c-46ef-a3b0-1152a34ac955/pyspark-10a76264-0b1e-4
310-a607-60062b93918d
18/05/11 12:16:28 INFO ShutdownHookManager: Deleting directory /tmp/spark-bda08fae-583c-46ef-a3b0-1152a34ac955
hadoop@hadoop-VirtualBox:~/Lab3$
```

- **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'


Terminal - hadoop@hadoop-VirtualBox: ~/Lab3
File Edit View Terminal Tabs Help

```
s: [empty row]
18/05/11 12:18:21 INFO Executor: Finished task 0.0 in stage 231.0 (TID 231). 1682 bytes result sent to driver
18/05/11 12:18:21 INFO TaskSetManager: Finished task 0.0 in stage 231.0 (TID 231) in 51 ms on localhost (executor driver) (1/1)
18/05/11 12:18:21 INFO TaskSchedulerImpl: Removed TaskSet 231.0, whose tasks have all completed, from pool
18/05/11 12:18:21 INFO DAGScheduler: ResultStage 231 (showString at NativeMethodAccessorImpl.java:0) finished in 0.050 s
18/05/11 12:18:21 INFO DAGScheduler: Job 231 finished: showString at NativeMethodAccessorImpl.java:0, took 0.070682 s
18/05/11 12:18:21 INFO CodeGenerator: Code generated in 20.947551 ms
+-----+--------------------+----------+
|label|            features|prediction|
+-----+--------------------+----------+
|  1.0|(114,[5,8,11,16,1...|       0.0|
+-----+--------------------+----------+

---------- Output Classes ----------
0 - Business
1 - Politics
2 - Sports
3 - Technology
------------------------------------
18/05/11 12:18:22 INFO SparkUI: Stopped Spark web UI at http://192.168.1.14:4040
18/05/11 12:18:22 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/11 12:18:22 INFO MemoryStore: MemoryStore cleared
18/05/11 12:18:22 INFO BlockManager: BlockManager stopped
18/05/11 12:18:22 INFO BlockManagerMaster: BlockManagerMaster stopped
18/05/11 12:18:22 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/05/11 12:18:22 INFO SparkContext: Successfully stopped SparkContext
18/05/11 12:18:22 INFO ShutdownHookManager: Shutdown hook called
18/05/11 12:18:22 INFO ShutdownHookManager: Deleting directory /tmp/spark-113a0fc9-c8ad-4e9b-944c-c9136d34120e/pyspark-428341a5-4572-4
ae9-92a0-f8df8e95606d
18/05/11 12:18:22 INFO ShutdownHookManager: Deleting directory /tmp/spark-113a0fc9-c8ad-4e9b-944c-c9136d34120e
hadoop@hadoop-VirtualBox:~/Lab3$
```