CSE – 587 Data Intensive Computing Spring 2018

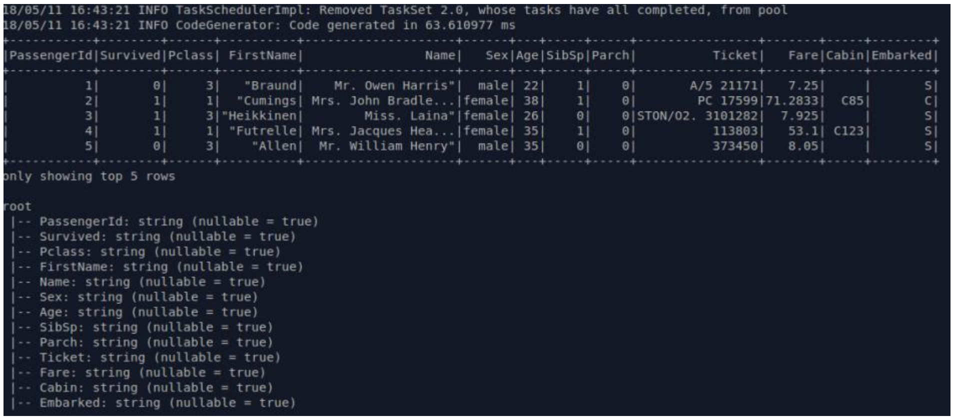
Report

Lab3: Data Analysis Pipeline Using Apache Spark

Submitted By:
Rahul Mallu (50245594)
Likhith Kumar Miryala (50249280)

1. Understand Apache Spark with Titanic Data Analysis

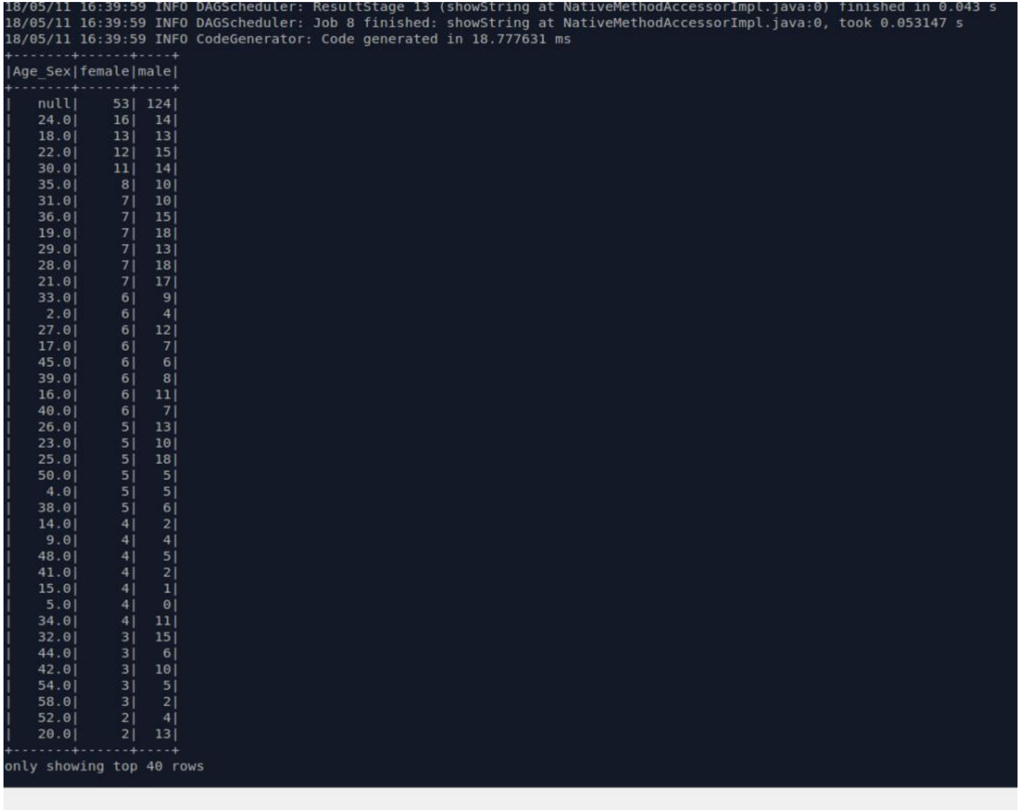
a. We have loaded the titanic data and parsed the RDD to DataFrame.
b. We removed the header from the data and separated each row by "," and converted it to a tuple.
c. Gave names as each column by the header. Figure 1 is the screenshot of the data frame



```
18/05/11 16:43:21 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
18/05/11 16:43:21 INFO CodeGenerator: Code generated in 63.610977 ms
+-----------+--------+------+----------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+
|PassengerId|Survived|Pclass| FirstName|                Name|   Sex|Age|SibSp|Parch|          Ticket|   Fare|Cabin|Embarked|
+-----------+--------+------+----------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+
|          1|       0|     3|  "Braund|    Mr. Owen Harris"|  male| 22|    1|    0|       A/5 21171|   7.25|     |       S|
|          2|       1|     1|  "Cumings| Mrs. John Bradle...|female| 38|    1|    0|        PC 17599|71.2833|  C85|       C|
|          3|       1|     3|"Heikkinen|        Miss. Laina"|female| 26|    0|    0|STON/O2. 3101282|  7.925|     |       S|
|          4|       1|     1| "Futrelle| Mrs. Jacques Hea...|female| 35|    1|    0|          113803|   53.1| C123|       S|
|          5|       0|     3|    "Allen|  Mr. William Henry"|  male| 35|    0|    0|          373450|   8.05|     |       S|
+-----------+--------+------+----------+--------------------+------+---+-----+-----+----------------+-------+-----+--------+
only showing top 5 rows

root
 |-- PassengerId: string (nullable = true)
 |-- Survived: string (nullable = true)
 |-- Pclass: string (nullable = true)
 |-- FirstName: string (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: string (nullable = true)
 |-- SibSp: string (nullable = true)
 |-- Parch: string (nullable = true)
 |-- Ticket: string (nullable = true)
 |-- Fare: string (nullable = true)
 |-- Cabin: string (nullable = true)
 |-- Embarked: string (nullable = true)
```

Figure: 1

d. After converting to DataFrame, we have manipulated the data by converting some of the features such as age, fare to numeric and Figure 2 is the screenshot of the DataFrame after this manipulation.



```
18/05/11 16:39:59 INFO DAGScheduler: ResultStage 13 (showString at NativeMethodAccessorImpl.java:0) finished in 0.043 s
18/05/11 16:39:59 INFO DAGScheduler: Job 8 finished: showString at NativeMethodAccessorImpl.java:0, took 0.053147 s
18/05/11 16:39:59 INFO CodeGenerator: Code generated in 18.777631 ms
+-------+------+----+
|Age_Sex|female|male|
+-------+------+----+
|   null|    53| 124|
|   24.0|    16|  14|
|   18.0|    13|  13|
|   22.0|    12|  15|
|   30.0|    11|  14|
|   35.0|     8|  10|
|   31.0|     7|  10|
|   36.0|     7|  15|
|   19.0|     7|  18|
|   29.0|     7|  13|
|   28.0|     7|  18|
|   21.0|     7|  17|
|   33.0|     6|   9|
|    2.0|     6|   4|
|   27.0|     6|  12|
|   17.0|     6|   7|
|   45.0|     6|   6|
|   39.0|     6|   8|
|   16.0|     6|  11|
|   40.0|     6|   7|
|   26.0|     5|  13|
|   23.0|     5|  10|
|   25.0|     5|  18|
|   50.0|     5|   5|
|    4.0|     5|   5|
|   38.0|     5|   6|
|   14.0|     4|   2|
|    9.0|     4|   4|
|   48.0|     4|   5|
|   41.0|     4|   2|
|   15.0|     4|   1|
|    5.0|     4|   0|
|   34.0|     4|  11|
|   32.0|     3|  15|
|   44.0|     3|   6|
|   42.0|     3|  10|
|   54.0|     3|   5|
|   58.0|     3|   2|
|   52.0|     2|   4|
|   20.0|     2|  13|
+-------+------+----+
only showing top 40 rows
```

Figure: 2

2. Collect and Clean data

a. We have collected the news articles for four categories: Politics, Sports, Business and Technology.
b. We formatted the data to store it in a .csv file so that we can use this for further processing down the line.
c. The further explanations and details have been mentioned in the Jupyter Notebook

3. Feature Engineering

a. We have extracted the words characterizing the category and also cleaned the data by removing stop words, punctuations, numbers, etc.
b. This was done by building a model pipeline in Spark which has six stages:
    i. We have used regexTokenizer which takes articles text as input and gives the words as output
    ii. We have created a custom list of stop words and used the StopWordsRemover function to remove those list of stop words
    iii. We have used the countVectorizer function to compute the term frequency
    iv. We have used HashingTF to map words to their frequencies
    v. We have used the IDF function to remove the sparse terms and it gives the TF-IDF
    vi. String Indexer function has been used to encode a string column of labels to a column of label indices. Here we convert the category to label.
c. By the end of this step, we have characterized all the categories

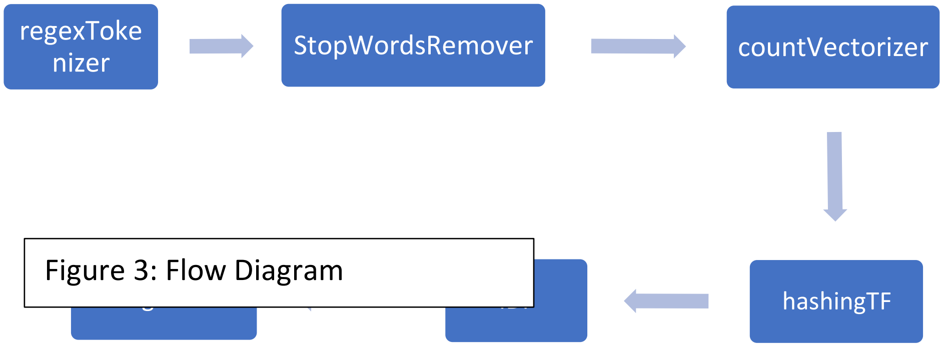
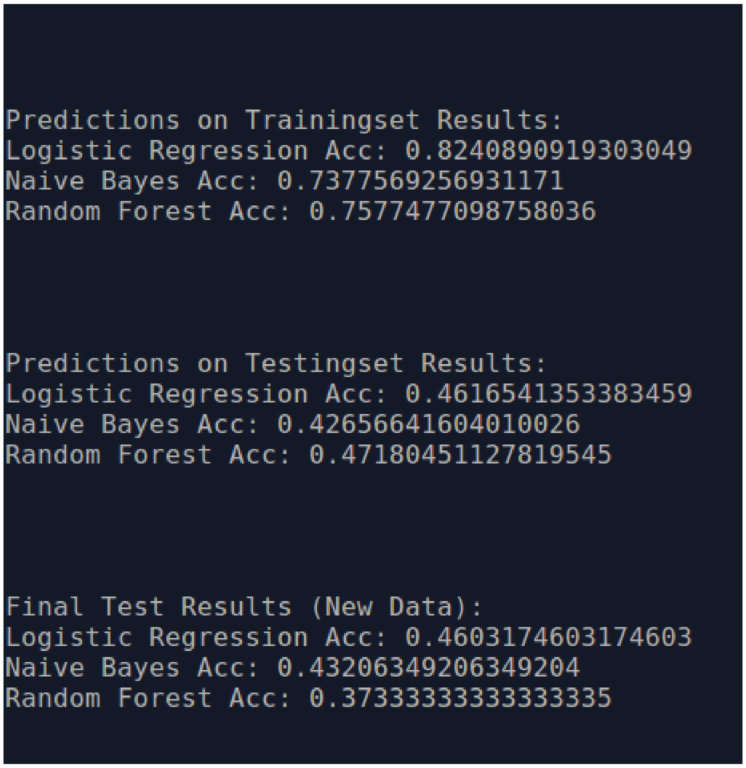


Figure 3: Flow Diagram

4. Multi-class Classification

a. We have done the classification using three different classifying algorithms: Logistic Regression, Naive Bayes and Random Forest.
b. We have made predictions and score on the test and training set and calculated the accuracy using multi-class classification evaluator
c. The accuracies obtained for train, test and random data are as follows:

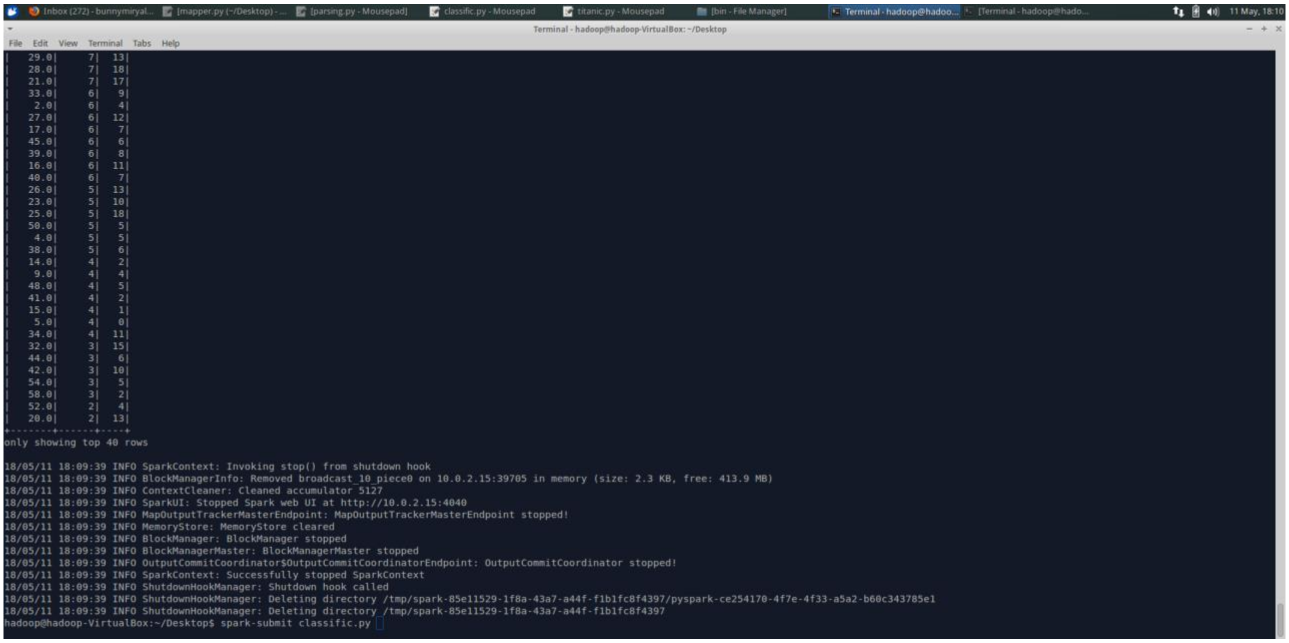


```
Predictions on Trainingset Results:
Logistic Regression Acc: 0.8240890919303049
Naive Bayes Acc: 0.7377569256931171
Random Forest Acc: 0.7577477098758036



Predictions on Testingset Results:
Logistic Regression Acc: 0.4616541353383459
Naive Bayes Acc: 0.42656641604010026
Random Forest Acc: 0.47180451127819545



Final Test Results (New Data):
Logistic Regression Acc: 0.4603174603174603
Naive Bayes Acc: 0.43206349206349204
Random Forest Acc: 0.37333333333333335
```

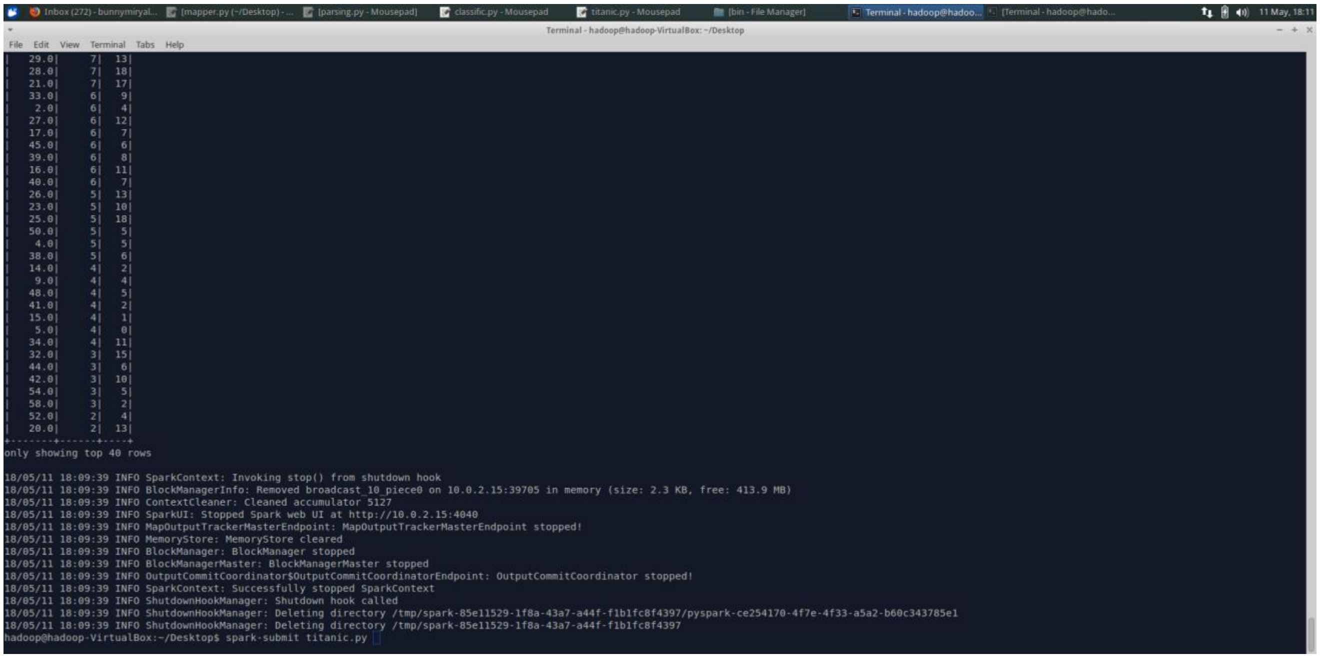
Figure: 4

d. From the above results, we understand that random forest is robust and versatile but for high dimensional sparse data, logistic regression model is preferable
e. Since there is discrepancy in the accuracy between the training, test and random data and as we have used cross validation we can say that the statistical properties of test, train and random data are different

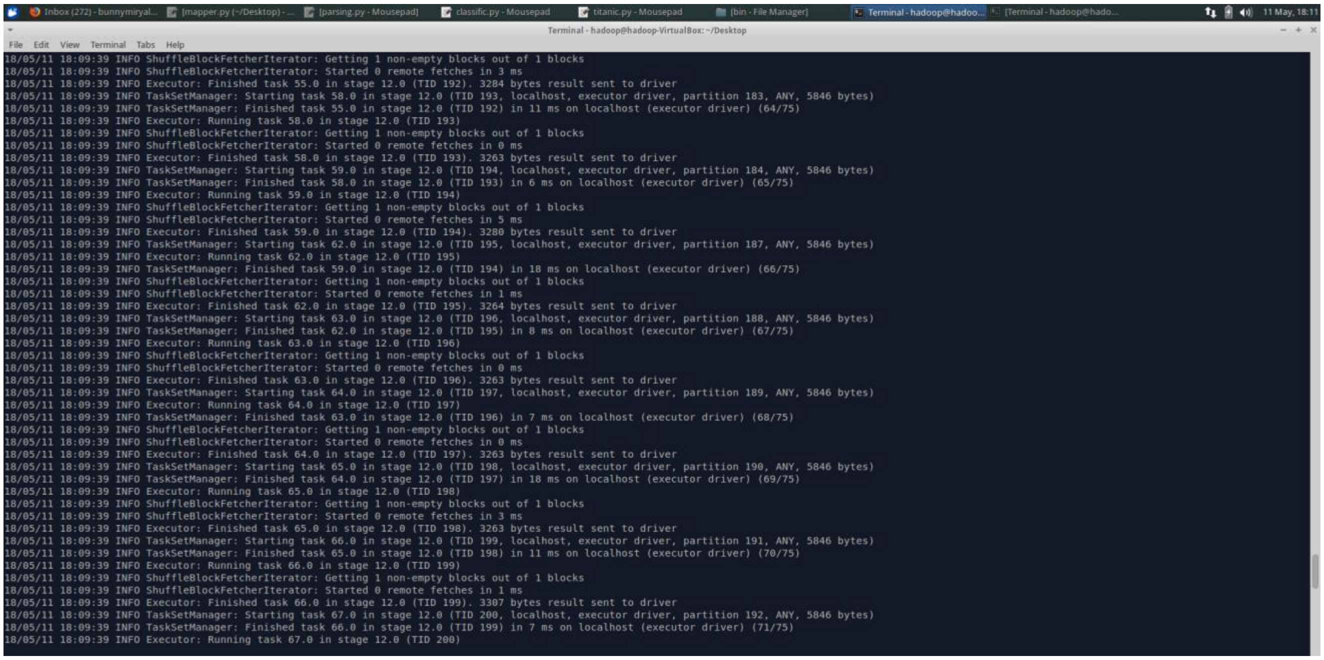
5. Screenshots of Working Program


Inbox (272) - bunnymiryal...
[mapper.py (~/Desktop) - ...
[parsing.py - Mousepad]
classific.py - Mousepad
titanic.py - Mousepad
[bin - File Manager]
Terminal - hadoop@hadoo...
[Terminal - hadoop@hado...
11 May, 18:10
Terminal - hadoop@hadoop-VirtualBox: ~/Desktop
File
Edit
View
Terminal
Tabs
Help

```
|   29.0|     7|  13|
|   28.0|     7|  18|
|   21.0|     7|  17|
|   33.0|     6|   9|
|    2.0|     6|   4|
|   27.0|     6|  12|
|   17.0|     6|   7|
|   45.0|     6|   6|
|   39.0|     6|   8|
|   16.0|     6|  11|
|   40.0|     6|   7|
|   26.0|     5|  13|
|   23.0|     5|  10|
|   25.0|     5|  18|
|   50.0|     5|   5|
|    4.0|     5|   5|
|   38.0|     5|   6|
|   14.0|     4|   2|
|    9.0|     4|   4|
|   48.0|     4|   5|
|   41.0|     4|   2|
|   15.0|     4|   1|
|    5.0|     4|   0|
|   34.0|     4|  11|
|   32.0|     3|  15|
|   44.0|     3|   6|
|   42.0|     3|  10|
|   54.0|     3|   5|
|   58.0|     3|   2|
|   52.0|     2|   4|
|   20.0|     2|  13|
+-------+------+----+
only showing top 40 rows

18/05/11 18:09:39 INFO SparkContext: Invoking stop() from shutdown hook
18/05/11 18:09:39 INFO BlockManagerInfo: Removed broadcast_10_piece0 on 10.0.2.15:39705 in memory (size: 2.3 KB, free: 413.9 MB)
18/05/11 18:09:39 INFO ContextCleaner: Cleaned accumulator 5127
18/05/11 18:09:39 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/05/11 18:09:39 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/11 18:09:39 INFO MemoryStore: MemoryStore cleared
18/05/11 18:09:39 INFO BlockManager: BlockManager stopped
18/05/11 18:09:39 INFO BlockManagerMaster: BlockManagerMaster stopped
18/05/11 18:09:39 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/05/11 18:09:39 INFO SparkContext: Successfully stopped SparkContext
18/05/11 18:09:39 INFO ShutdownHookManager: Shutdown hook called
18/05/11 18:09:39 INFO ShutdownHookManager: Deleting directory /tmp/spark-85e11529-1f8a-43a7-a44f-f1b1fc8f4397/pyspark-ce254170-4f7e-4f33-a5a2-b60c343785e1
18/05/11 18:09:39 INFO ShutdownHookManager: Deleting directory /tmp/spark-85e11529-1f8a-43a7-a44f-f1b1fc8f4397
hadoop@hadoop-VirtualBox:~/Desktop$ spark-submit classific.py
```


Inbox (272) - bunnymiryal...
[mapper.py (~/Desktop) - ...
[parsing.py - Mousepad]
classific.py - Mousepad
titanic.py - Mousepad
[bin - File Manager]
Terminal - hadoop@hadoo...
[Terminal - hadoop@hado...
11 May, 18:11
Terminal - hadoop@hadoop-VirtualBox: ~/Desktop
File
Edit
View
Terminal
Tabs
Help

```
|   29.0|     7|  13|
|   28.0|     7|  18|
|   21.0|     7|  17|
|   33.0|     6|   9|
|    2.0|     6|   4|
|   27.0|     6|  12|
|   17.0|     6|   7|
|   45.0|     6|   6|
|   39.0|     6|   8|
|   16.0|     6|  11|
|   40.0|     6|   7|
|   26.0|     5|  13|
|   23.0|     5|  10|
|   25.0|     5|  18|
|   50.0|     5|   5|
|    4.0|     5|   5|
|   38.0|     5|   6|
|   14.0|     4|   2|
|    9.0|     4|   4|
|   48.0|     4|   5|
|   41.0|     4|   2|
|   15.0|     4|   1|
|    5.0|     4|   0|
|   34.0|     4|  11|
|   32.0|     3|  15|
|   44.0|     3|   6|
|   42.0|     3|  10|
|   54.0|     3|   5|
|   58.0|     3|   2|
|   52.0|     2|   4|
|   20.0|     2|  13|
+-------+------+----+
only showing top 40 rows

18/05/11 18:09:39 INFO SparkContext: Invoking stop() from shutdown hook
18/05/11 18:09:39 INFO BlockManagerInfo: Removed broadcast_10_piece0 on 10.0.2.15:39705 in memory (size: 2.3 KB, free: 413.9 MB)
18/05/11 18:09:39 INFO ContextCleaner: Cleaned accumulator 5127
18/05/11 18:09:39 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/05/11 18:09:39 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/11 18:09:39 INFO MemoryStore: MemoryStore cleared
18/05/11 18:09:39 INFO BlockManager: BlockManager stopped
18/05/11 18:09:39 INFO BlockManagerMaster: BlockManagerMaster stopped
18/05/11 18:09:39 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
18/05/11 18:09:39 INFO SparkContext: Successfully stopped SparkContext
18/05/11 18:09:39 INFO ShutdownHookManager: Shutdown hook called
18/05/11 18:09:39 INFO ShutdownHookManager: Deleting directory /tmp/spark-85e11529-1f8a-43a7-a44f-f1b1fc8f4397/pyspark-ce254170-4f7e-4f33-a5a2-b60c343785e1
18/05/11 18:09:39 INFO ShutdownHookManager: Deleting directory /tmp/spark-85e11529-1f8a-43a7-a44f-f1b1fc8f4397
hadoop@hadoop-VirtualBox:~/Desktop$ spark-submit titanic.py
```


Inbox (272) - bunnymiryal...
[mapper.py (~/Desktop) - ...
[parsing.py - Mousepad]
classific.py - Mousepad
titanic.py - Mousepad
[bin - File Manager]
Terminal - hadoop@hadoo...
[Terminal - hadoop@hado...
11 May, 18:11
Terminal - hadoop@hadoop-VirtualBox: ~/Desktop
File
Edit
View
Terminal
Tabs
Help

```
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 3 ms
18/05/11 18:09:39 INFO Executor: Finished task 55.0 in stage 12.0 (TID 192). 3284 bytes result sent to driver
18/05/11 18:09:39 INFO TaskSetManager: Starting task 58.0 in stage 12.0 (TID 193, localhost, executor driver, partition 183, ANY, 5846 bytes)
18/05/11 18:09:39 INFO TaskSetManager: Finished task 55.0 in stage 12.0 (TID 192) in 11 ms on localhost (executor driver) (64/75)
18/05/11 18:09:39 INFO Executor: Running task 58.0 in stage 12.0 (TID 193)
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/11 18:09:39 INFO Executor: Finished task 58.0 in stage 12.0 (TID 193). 3263 bytes result sent to driver
18/05/11 18:09:39 INFO TaskSetManager: Starting task 59.0 in stage 12.0 (TID 194, localhost, executor driver, partition 184, ANY, 5846 bytes)
18/05/11 18:09:39 INFO TaskSetManager: Finished task 58.0 in stage 12.0 (TID 193) in 6 ms on localhost (executor driver) (65/75)
18/05/11 18:09:39 INFO Executor: Running task 59.0 in stage 12.0 (TID 194)
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 5 ms
18/05/11 18:09:39 INFO Executor: Finished task 59.0 in stage 12.0 (TID 194). 3280 bytes result sent to driver
18/05/11 18:09:39 INFO TaskSetManager: Starting task 62.0 in stage 12.0 (TID 195, localhost, executor driver, partition 187, ANY, 5846 bytes)
18/05/11 18:09:39 INFO Executor: Running task 62.0 in stage 12.0 (TID 195)
18/05/11 18:09:39 INFO TaskSetManager: Finished task 59.0 in stage 12.0 (TID 194) in 18 ms on localhost (executor driver) (66/75)
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/05/11 18:09:39 INFO Executor: Finished task 62.0 in stage 12.0 (TID 195). 3264 bytes result sent to driver
18/05/11 18:09:39 INFO TaskSetManager: Starting task 63.0 in stage 12.0 (TID 196, localhost, executor driver, partition 188, ANY, 5846 bytes)
18/05/11 18:09:39 INFO TaskSetManager: Finished task 62.0 in stage 12.0 (TID 195) in 8 ms on localhost (executor driver) (67/75)
18/05/11 18:09:39 INFO Executor: Running task 63.0 in stage 12.0 (TID 196)
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/11 18:09:39 INFO Executor: Finished task 63.0 in stage 12.0 (TID 196). 3263 bytes result sent to driver
18/05/11 18:09:39 INFO TaskSetManager: Starting task 64.0 in stage 12.0 (TID 197, localhost, executor driver, partition 189, ANY, 5846 bytes)
18/05/11 18:09:39 INFO Executor: Running task 64.0 in stage 12.0 (TID 197)
18/05/11 18:09:39 INFO TaskSetManager: Finished task 63.0 in stage 12.0 (TID 196) in 7 ms on localhost (executor driver) (68/75)
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
18/05/11 18:09:39 INFO Executor: Finished task 64.0 in stage 12.0 (TID 197). 3263 bytes result sent to driver
18/05/11 18:09:39 INFO TaskSetManager: Starting task 65.0 in stage 12.0 (TID 198, localhost, executor driver, partition 190, ANY, 5846 bytes)
18/05/11 18:09:39 INFO TaskSetManager: Finished task 64.0 in stage 12.0 (TID 197) in 18 ms on localhost (executor driver) (69/75)
18/05/11 18:09:39 INFO Executor: Running task 65.0 in stage 12.0 (TID 198)
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 3 ms
18/05/11 18:09:39 INFO Executor: Finished task 65.0 in stage 12.0 (TID 198). 3263 bytes result sent to driver
18/05/11 18:09:39 INFO TaskSetManager: Starting task 66.0 in stage 12.0 (TID 199, localhost, executor driver, partition 191, ANY, 5846 bytes)
18/05/11 18:09:39 INFO TaskSetManager: Finished task 65.0 in stage 12.0 (TID 198) in 11 ms on localhost (executor driver) (70/75)
18/05/11 18:09:39 INFO Executor: Running task 66.0 in stage 12.0 (TID 199)
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/05/11 18:09:39 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
18/05/11 18:09:39 INFO Executor: Finished task 66.0 in stage 12.0 (TID 199). 3307 bytes result sent to driver
18/05/11 18:09:39 INFO TaskSetManager: Starting task 67.0 in stage 12.0 (TID 200, localhost, executor driver, partition 192, ANY, 5846 bytes)
18/05/11 18:09:39 INFO TaskSetManager: Finished task 66.0 in stage 12.0 (TID 199) in 7 ms on localhost (executor driver) (71/75)
18/05/11 18:09:39 INFO Executor: Running task 67.0 in stage 12.0 (TID 200)
```

6. References

i. https://datascienceplus.com/multi-class-text-classification-with-pyspark

ii. https://6chaoran.wordpress.com/2016/08/13/__trashed/