

Lab Assignment 5 and 6:

Question 1: Spark and Smartphone/Watch Application:

Implement a smart application with big data analytics related to your project showing the collaboration between Spark and Smart Apps. Implement Twitter Streaming and perform word count on it and publish the results and showcase it in your Smart Phone/Watch Application

Description:

I have created a streaming context to get the twitter stream data and performed Map Reduce framework on the twitter stream to determine the word count stream. By establishing the socket connection I connected spark to smartphone and shown the Twitter stream word count in the phone.

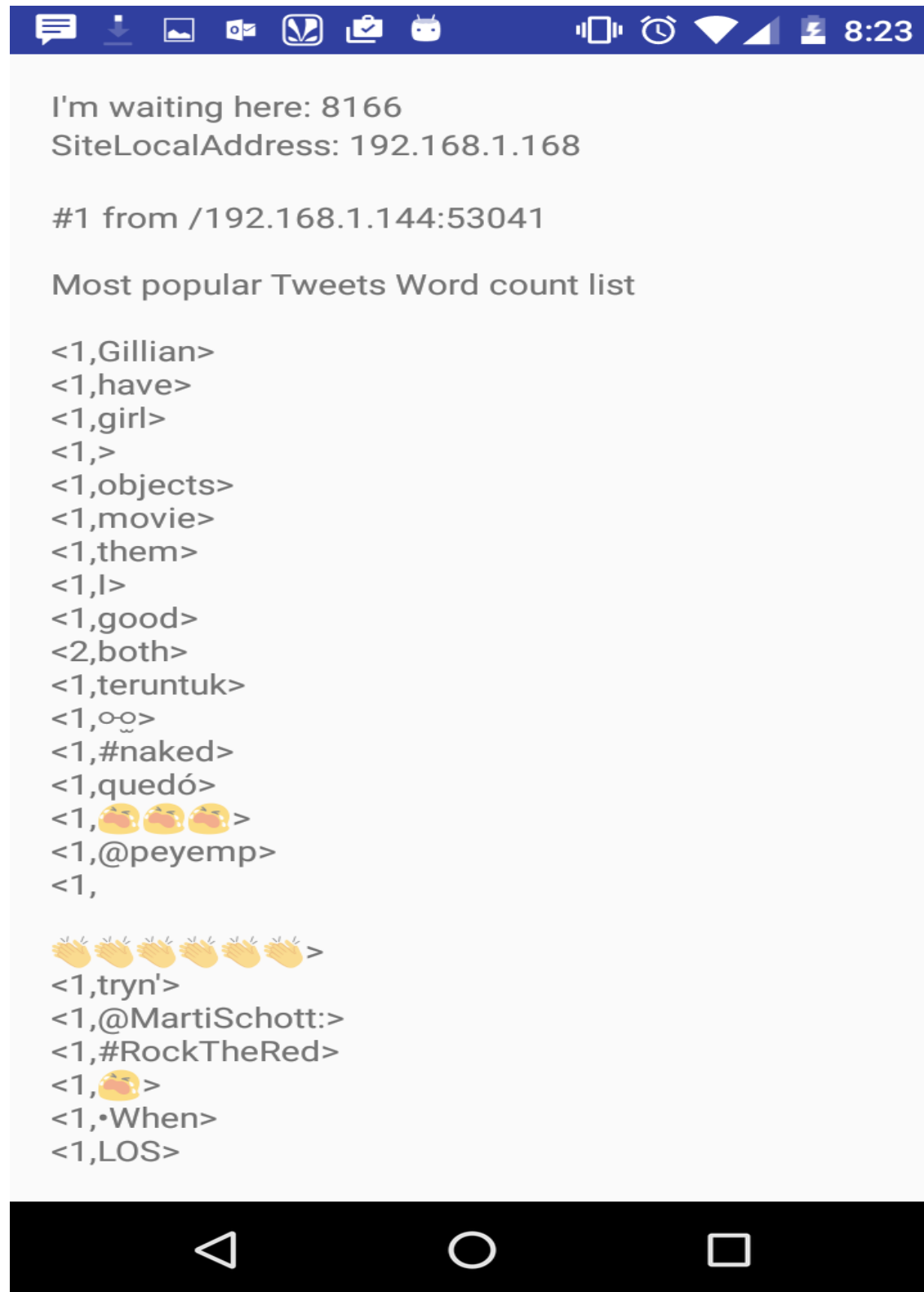
Screenshots:



I'm waiting here: 8166

SiteLocalAddress: 192.168.1.168





Question 2: Spark ML Lib Application

Perform a machine learning algorithm with the Twitter Streaming data to categorize each Tweet

- 1) Training datasets: Collect different categories of Tweets related to your project.(Categories can be based on Hashtags /Subjects etc.)
- 2) Test data: the upcoming twitter stream.

Description:

I have used *Naïve Bayes model* to categorize different tweets.

- 1) Training set: I collected the tweets based on 3 key words and are popular trends currently. Three keywords are:
 - a) "health"
 - b) "android"
 - c) "jobs"
- 2) I tested the data by sending the live twitter stream and the results are shown below in the screenshots

Screenshots:

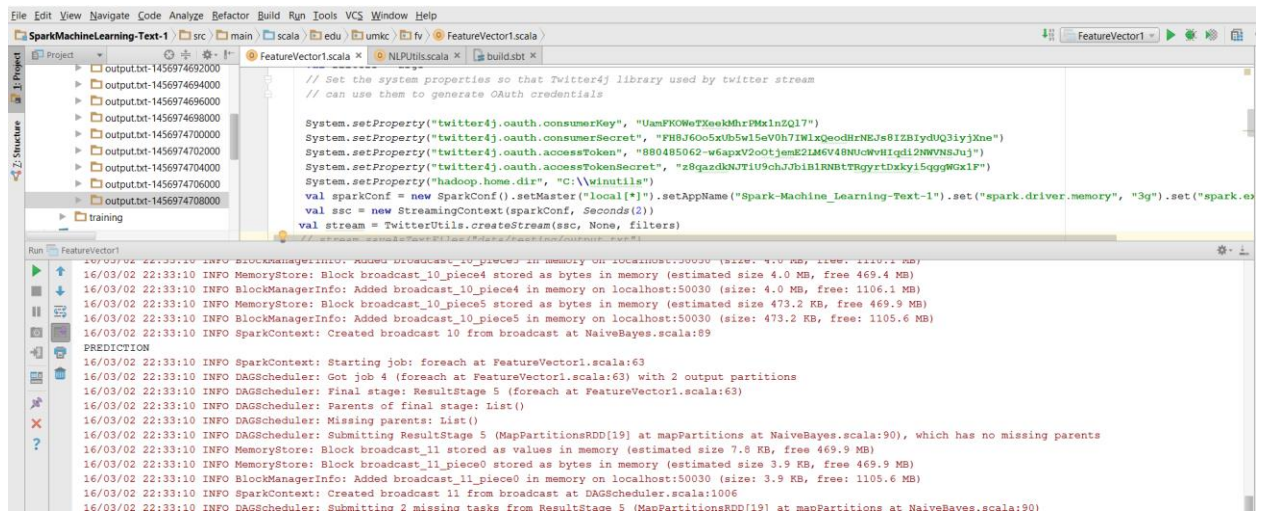
Tokenizing and lemmatizing the training data:

```

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SparkMachineLearning-Text-1 src main scala edu umkc fv FeatureVector.scala
Project Structure
  output.txt-1456974692000
  output.txt-1456974694000
  output.txt-1456974696000
  output.txt-1456974698000
  output.txt-1456974700000
  output.txt-1456974702000
  output.txt-1456974704000
  output.txt-1456974706000
  output.txt-1456974708000
  training
Run FeatureVector
  // Set the system properties so that Twitter4j library used by twitter stream
  // can use them to generate OAuth credentials
  System.setProperty("twitter4j.oauth.consumerKey", "UamFK0W6TXeekMhrPMxlnZQ17")
  System.setProperty("twitter4j.oauth.consumerSecret", "FH8J60o5xUb5w15eV0h7IwLxQeodHrNEJ58IZBIydUQ3iyjXne")
  System.setProperty("twitter4j.oauth.accessToken", "880485062-w6apxV2oOtJemE2IM6V48NUdWVHIqd12NwVNSJuj")
  System.setProperty("twitter4j.oauth.accessTokenSecret", "z8qazdKNJT1U9chJUb1B1NNtTRgyrTDxky15qggWGX1F")
  System.setProperty("hadoop.home.dir", "C:\\winutils")
  val sparkConf = new SparkConf().setMaster("local[*]").setAppName("Spark-Machine_Learning-Text-1").set("spark.driver.memory", "3g").set("spark.executor.memory", "1g")
  val ssc = new StreamingContext(sparkConf, Seconds(2))
  val stream = TwitterUtils.createStream(ssc, None, filters)
  s null usermentionentity usermentionentityjsonimpl name 何となくしてくれそうな奴らコピ〜bot screenname urlentity hashtagentity mediaentity symbolentity currentuserretweetid user userjs
  Adding annotator tokenize
  Adding annotator ssplit
  Adding annotator pos
  Adding annotator lemma
  16/03/02 22:32:44 WARN PTBTokenizer: Untokenizable: (U+3010, decimal: 12304)
  statusjsonimpl createdat wed mar cst text fayetteville plane crash source twitter web client istruncate false inreplytostatusid inreplytouserid isfavorited false isretweeted fa
  Adding annotator tokenize
  Adding annotator ssplit
  Adding annotator pos
  16/03/02 22:32:45 WARN PTBTokenizer: Untokenizable: (U+300C, decimal: 12300)
  statusjsonimpl createdat wed mar cst text listen step child source tweetdeck istruncate false inreplytostatusid inreplytouserid isfavorited false isretweeted false favoritecour
  Adding annotator tokenize
  Adding annotator ssplit
  Adding annotator pos
  Adding annotator lemma
  16/03/02 22:32:45 WARN PTBTokenizer: Untokenizable: (U+A4AA, decimal: 42154)
  statusjsonimpl createdat wed mar cst text libertad puede alcanzarse por medio tirania source ifttt istruncate false inreplytostatusid inreplytouserid isfavorited false isretwee

```

Predicting the Output:



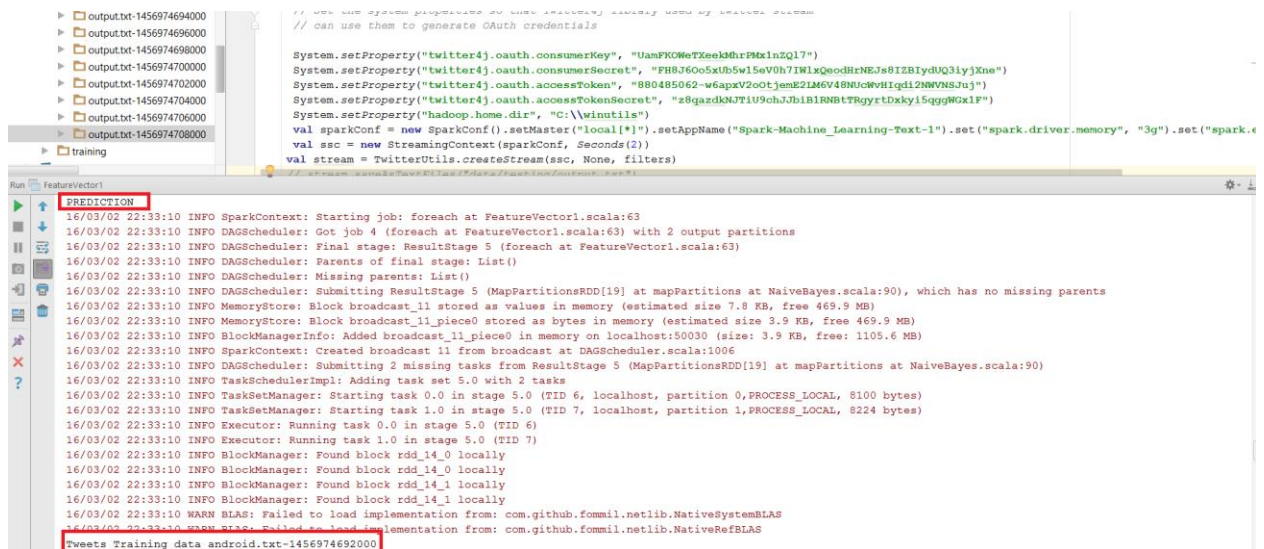
```
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SparkMachineLearning-Text-1 [src] main scala edu umkc fv FeatureVector1.scala
// Set the system properties so that Twitter4j library used by twitter stream
// can use them to generate OAuth credentials

System.setProperty("twitter4j.oauth.consumerKey", "UasFK0WtXeeKdhrPMxlnZql7")
System.setProperty("twitter4j.oauth.consumerSecret", "FHB360c5xth5w15eV0h7IwLxQeodHrNEJse8IZBlydUQ3ijjXne")
System.setProperty("twitter4j.oauth.accessToken", "880485062-w6apxV2o0tjemK2LmGV48NucwVHlqdl2NwVNBuJj")
System.setProperty("twitter4j.oauth.accessTokenSecret", "z8qazdKJUTiU9chJb1B1RNBtTRuyrTdxky15qggWGXlF")
System.setProperty("hadoop.home.dir", "C:\\winutils")
val sparkConf = new SparkConf().setMaster("local[*]").setAppName("Spark-Machine_Learning-Text-1").set("spark.driver.memory", "3g").set("spark.ex
val ssc = new StreamingContext(sparkConf, Seconds(2))
val stream = TwitterUtils.createStream(ssc, None, filters)

16/03/02 22:33:10 INFO BlockManagerInfo: Added broadcast_10 pieces in memory on localhost:50030 (size: 4.0 MB, free: 1106.1 MB)
16/03/02 22:33:10 INFO MemoryStore: Block broadcast_10_piece4 stored as bytes in memory (estimated size 4.0 MB, free 469.4 MB)
16/03/02 22:33:10 INFO BlockManagerInfo: Added broadcast_10_piece4 in memory on localhost:50030 (size: 4.0 MB, free: 1106.1 MB)
16/03/02 22:33:10 INFO MemoryStore: Block broadcast_10_piece5 stored as bytes in memory (estimated size 473.2 KB, free 469.9 MB)
16/03/02 22:33:10 INFO BlockManagerInfo: Added broadcast_10_piece5 in memory on localhost:50030 (size: 473.2 KB, free: 1105.6 MB)
16/03/02 22:33:10 INFO SparkContext: Created broadcast 10 from broadcast at NaiveBayes.scala:89

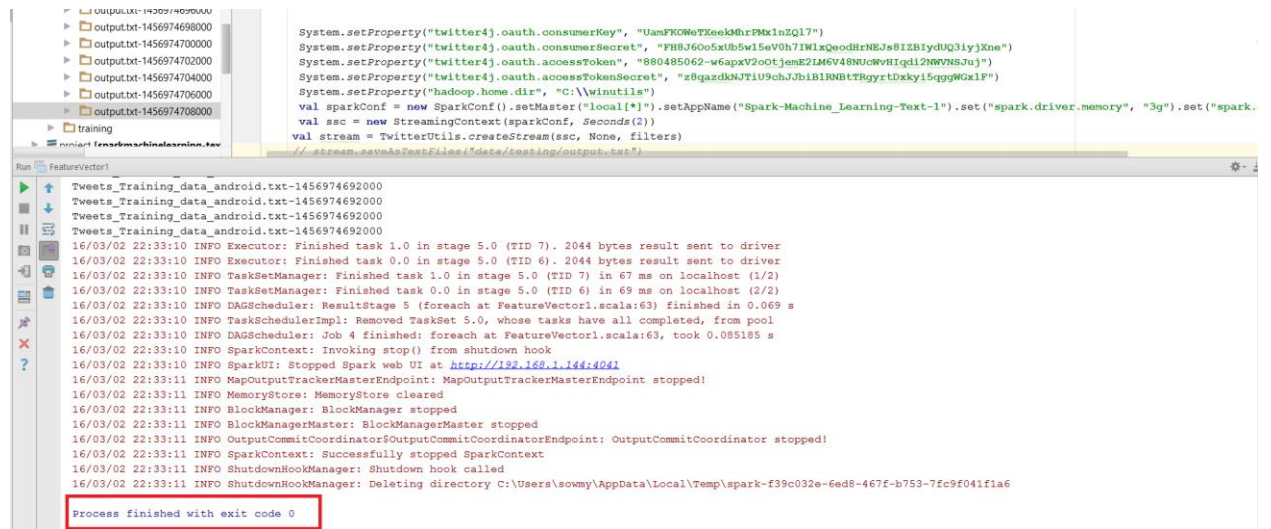
PREDICTION
16/03/02 22:33:10 INFO SparkContext: Starting job: foreach at FeatureVector1.scala:63
16/03/02 22:33:10 INFO DAGScheduler: Got job 4 (foreach at FeatureVector1.scala:63) with 2 output partitions
16/03/02 22:33:10 INFO DAGScheduler: Final stage: ResultStage 5 (foreach at FeatureVector1.scala:63)
16/03/02 22:33:10 INFO DAGScheduler: Parents of final stage: List()
16/03/02 22:33:10 INFO DAGScheduler: Missing parents: List()
16/03/02 22:33:10 INFO DAGScheduler: Submitting ResultStage 5 (MapPartitionsRDD[19] at mapPartitions at NaiveBayes.scala:90), which has no missing parents
16/03/02 22:33:10 INFO MemoryStore: Block broadcast_11 stored as values in memory (estimated size 7.8 KB, free 469.9 MB)
16/03/02 22:33:10 INFO MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 3.9 KB, free 469.9 MB)
16/03/02 22:33:10 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on localhost:50030 (size: 3.9 KB, free: 1105.6 MB)
16/03/02 22:33:10 INFO SparkContext: Created broadcast 11 from broadcast at DAGScheduler.scala:1006
16/03/02 22:33:10 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 5 (MapPartitionsRDD[19] at mapPartitions at NaiveBayes.scala:90)
```

Prediction of Output :



```
16/03/02 22:33:10 INFO SparkContext: Starting job: foreach at FeatureVector1.scala:63
16/03/02 22:33:10 INFO DAGScheduler: Got job 4 (foreach at FeatureVector1.scala:63) with 2 output partitions
16/03/02 22:33:10 INFO DAGScheduler: Final stage: ResultStage 5 (foreach at FeatureVector1.scala:63)
16/03/02 22:33:10 INFO DAGScheduler: Parents of final stage: List()
16/03/02 22:33:10 INFO DAGScheduler: Missing parents: List()
16/03/02 22:33:10 INFO DAGScheduler: Submitting ResultStage 5 (MapPartitionsRDD[19] at mapPartitions at NaiveBayes.scala:90), which has no missing parents
16/03/02 22:33:10 INFO MemoryStore: Block broadcast_11 stored as values in memory (estimated size 7.8 KB, free 469.9 MB)
16/03/02 22:33:10 INFO MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 3.9 KB, free 469.9 MB)
16/03/02 22:33:10 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on localhost:50030 (size: 3.9 KB, free: 1105.6 MB)
16/03/02 22:33:10 INFO SparkContext: Created broadcast 11 from broadcast at DAGScheduler.scala:1006
16/03/02 22:33:10 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 5 (MapPartitionsRDD[19] at mapPartitions at NaiveBayes.scala:90)
16/03/02 22:33:10 INFO TaskSchedulerImpl: Adding task set 5.0 with 2 tasks
16/03/02 22:33:10 INFO TaskSetManager: Starting task 0.0 in stage 5.0 (TID 6, localhost, partition 0, PROCESS_LOCAL, 8100 bytes)
16/03/02 22:33:10 INFO TaskSetManager: Starting task 1.0 in stage 5.0 (TID 7, localhost, partition 1, PROCESS_LOCAL, 8224 bytes)
16/03/02 22:33:10 INFO Executor: Running task 0.0 in stage 5.0 (TID 6)
16/03/02 22:33:10 INFO Executor: Running task 1.0 in stage 5.0 (TID 7)
16/03/02 22:33:10 INFO BlockManager: Found block rdd_14_0 locally
16/03/02 22:33:10 INFO BlockManager: Found block rdd_14_0 locally
16/03/02 22:33:10 INFO BlockManager: Found block rdd_14_1 locally
16/03/02 22:33:10 INFO BlockManager: Found block rdd_14_1 locally
16/03/02 22:33:10 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
16/03/02 22:33:10 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
Tweets_Training_data_android.txt-1456974692000
```

Execution completed successfully:



```
System.setProperty("twitter4j.oauth.consumerKey", "UamFK0WtXeeK3hrPMXlnZQl7")
System.setProperty("twitter4j.oauth.consumerSecret", "FHBj6Oo5xUb5w15eV0h7IW1xQeodHrNEJs81ZBIydUQ3iyjXne")
System.setProperty("twitter4j.oauth.accessToken", "880485062-w6apxV2oOtjemE2IM6V48MucWvHIqdi2NwVNSJuJ")
System.setProperty("twitter4j.oauth.accessTokenSecret", "z8qazdKNJtIU9chJb1B1RNBtTRgyrtDxkyi5gggWGx1F")
System.setProperty("hadoop.home.dir", "C:\\winutils")
val sparkConf = new SparkConf().setMaster("local[*]").setAppName("Spark-Machine_Learning-Text-1").set("spark.driver.memory", "3g").set("spark.
val ssc = new StreamingContext(sparkConf, Seconds(2))
val stream = TwitterUtils.createStream(ssc, None, filters)
// stream.saveAsTextFiles("data/testing/output.txt")
```

Run FeatureVector1

Tweets_Training_data_android.txt-1456974692000
Tweets_Training_data_android.txt-1456974692000
Tweets_Training_data_android.txt-1456974692000

16/03/02 22:33:10 INFO Executor: Finished task 1.0 in stage 5.0 (TID 7). 2044 bytes result sent to driver
16/03/02 22:33:10 INFO Executor: Finished task 0.0 in stage 5.0 (TID 6). 2044 bytes result sent to driver
16/03/02 22:33:10 INFO TaskSetManager: Finished task 1.0 in stage 5.0 (TID 7) in 67 ms on localhost (1/2)
16/03/02 22:33:10 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 6) in 69 ms on localhost (2/2)
16/03/02 22:33:10 INFO DAGScheduler: ResultStage 5 (foreach at FeatureVector1.scala:63) finished in 0.069 s
16/03/02 22:33:10 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
16/03/02 22:33:10 INFO DAGScheduler: Job 4 finished: foreach at FeatureVector1.scala:63, took 0.085185 s
16/03/02 22:33:10 INFO SparkContext: Invoking stop() from shutdown hook
16/03/02 22:33:10 INFO SparkUI: Stopped Spark web UI at <http://192.168.1.144:4041>
16/03/02 22:33:11 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
16/03/02 22:33:11 INFO MemoryStore: MemoryStore cleared
16/03/02 22:33:11 INFO BlockManager: BlockManager stopped
16/03/02 22:33:11 INFO BlockManagerMaster: BlockManagerMaster stopped
16/03/02 22:33:11 INFO OutputCommitCoordinator\$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
16/03/02 22:33:11 INFO SparkContext: Successfully stopped SparkContext
16/03/02 22:33:11 INFO ShutdownHookManager: Shutdown hook called
16/03/02 22:33:11 INFO ShutdownHookManager: Deleting directory C:\Users\sowmy\AppData\Local\Temp\spark-f39c032e-6ed8-467f-b753-7fc9f041fa6

Process finished with exit code 0