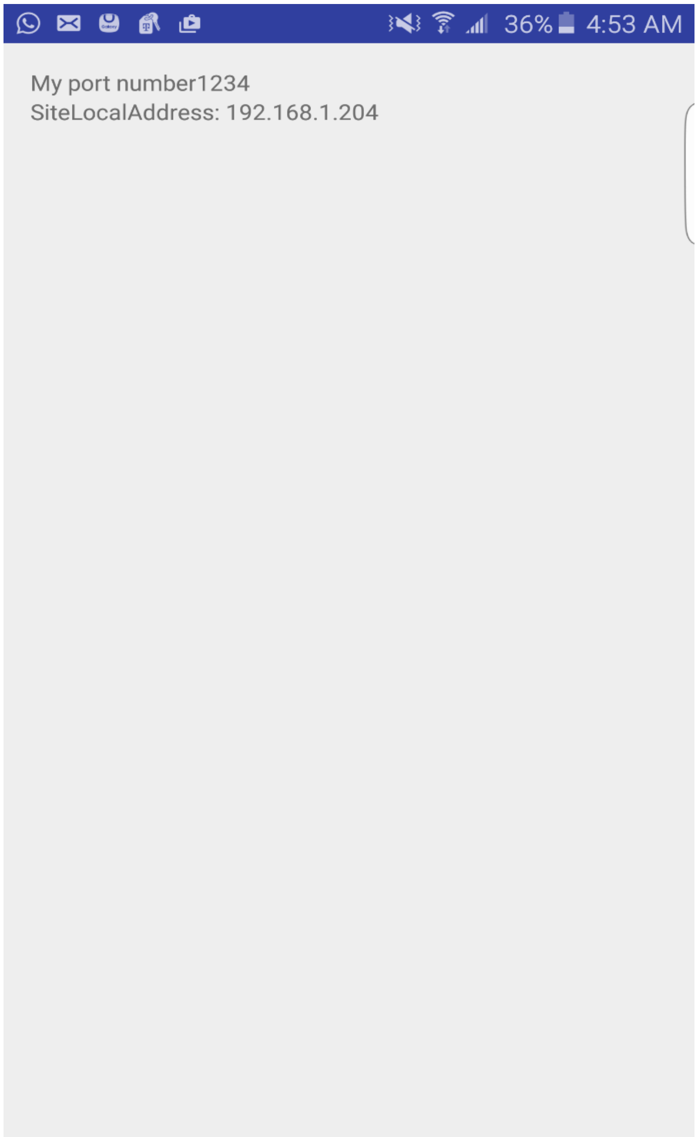
CS5542 Big Data Apps and Analytics
LAB ASSIGNMENT #5 &6
REPORT and SCREEN SHOTS

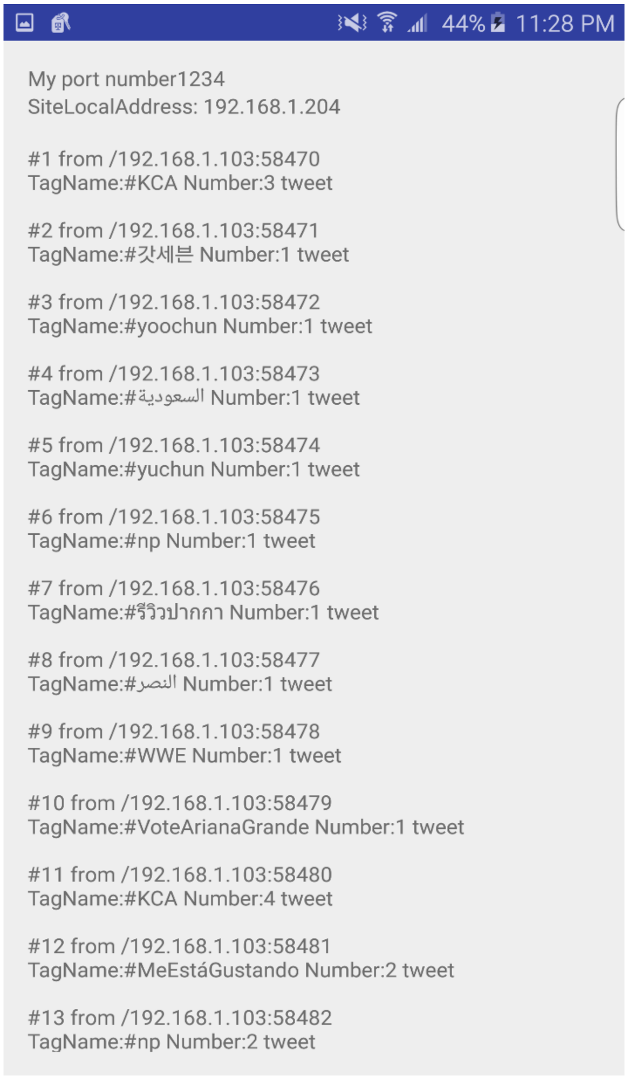
1. Spark and Smartphone/Watch Application Implement a smart application with big data analytics related to your project showing the collaboration between Spark and Smart Apps. Implement Twitter Streaming and perform word count on it and publish the results and showcase it in your Smart Phone/Watch Application.

1. We open a socket connection between the smart phone and spark . when we run the twitter streaming job, for every 10 seconds hashtags are sent to device.



My port number1234
SiteLocalAddress: 192.168.1.204

2. Tweets Top hash tags are displayed on Smart Phone screen



```
My port number1234
SiteLocalAddress: 192.168.1.204

#1 from /192.168.1.103:58470
TagName:#KCA Number:3 tweet

#2 from /192.168.1.103:58471
TagName:#갓세븐 Number:1 tweet

#3 from /192.168.1.103:58472
TagName:#yoochun Number:1 tweet

#4 from /192.168.1.103:58473
TagName:#السعودية Number:1 tweet

#5 from /192.168.1.103:58474
TagName:#yuchun Number:1 tweet

#6 from /192.168.1.103:58475
TagName:#np Number:1 tweet

#7 from /192.168.1.103:58476
TagName:#รีวิวปากกา Number:1 tweet

#8 from /192.168.1.103:58477
TagName:#النصر Number:1 tweet

#9 from /192.168.1.103:58478
TagName:#WWE Number:1 tweet

#10 from /192.168.1.103:58479
TagName:#VoteArianaGrande Number:1 tweet

#11 from /192.168.1.103:58480
TagName:#KCA Number:4 tweet

#12 from /192.168.1.103:58481
TagName:#MeEstáGustando Number:2 tweet

#13 from /192.168.1.103:58482
TagName:#np Number:2 tweet
```

2. Spark ML Lib ApplicationPerform a machine learning algorithm with the Twitter Streaming data tocategorize each Tweet

1)Training datasets: Collect different categories of Tweets related to your project.(Categoriescan be based on HashTags /Subjects etc.)

2)Test data:theupcoming twitter stream

1) I have collected the different categories of tweets such as Tweets related to restaurants, movies, sports.

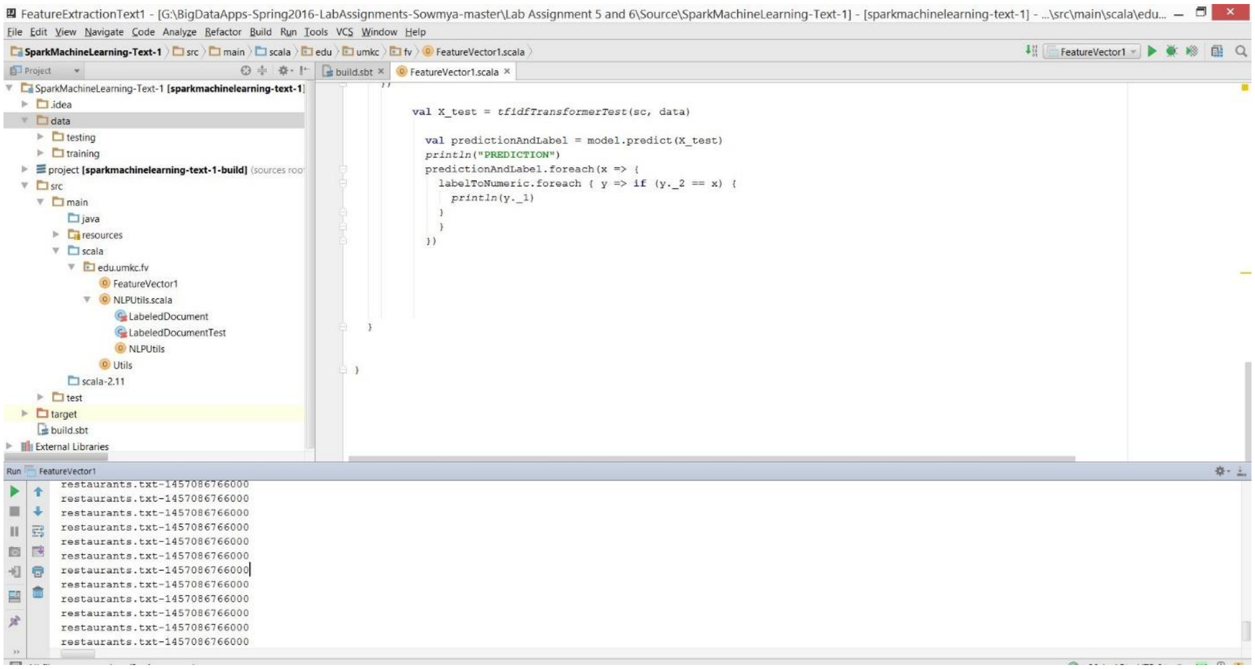
I trained my model to predict these categories using the Naïve Bayes Machine learning algorithm

The collected data is Divided into training data and Test data.

Training Data consists of Tweets related to three different categories. I collected the training data using the filters.

Testing data is collection of streaming tweets and we tried to predict the tweets which are more related to which category using classification algorithm.

Training Data and Test Data under Data tab

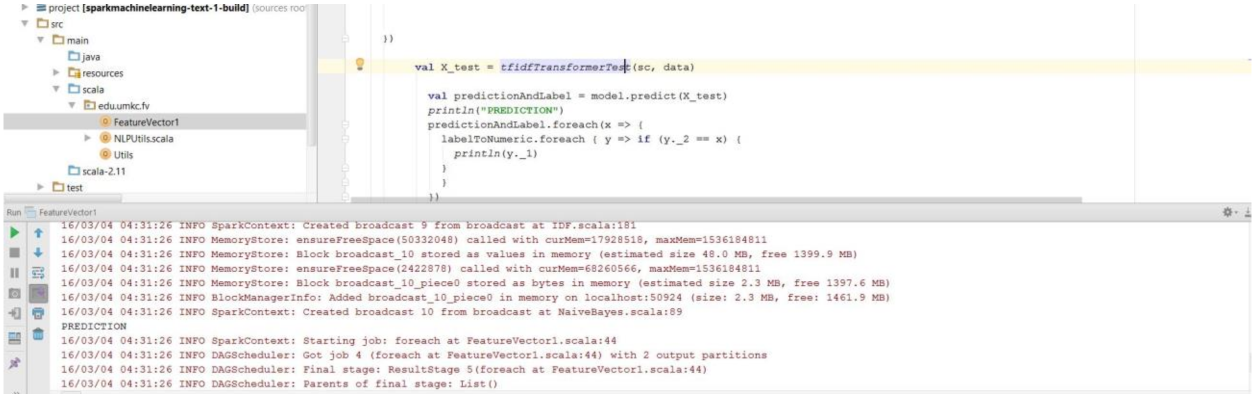

FeatureExtractionText1 - [G:\BigDataApps-Spring2016-LabAssignments-Sowmya-master\Lab Assignment 5 and 6\Source\SparkMachineLearning-Text-1] - [sparkmachinelearning-text-1] - ...\src\main\scala\edu...
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SparkMachineLearning-Text-1 src main scala edu umkc fv FeatureVector1.scala
FeatureVector1
Project
build.sbt
FeatureVector1.scala
SparkMachineLearning-Text-1 [sparkmachinelearning-text-1]
.idea
data
testing
training
project [sparkmachinelearning-text-1-build] (sources roo
src
main
java
resources
scala
edu.umkc.fv
FeatureVector1
NLPUtils.scala
LabeledDocument
LabeledDocumentTest
NLPUtils
Utils
scala-2.11
test
target
build.sbt
External Libraries
Run
FeatureVector1

```
      val X_test = tfidfTransformerTest(sc, data)

       val predictionAndLabel = model.predict(X_test)
       println("PREDICTION")
       predictionAndLabel.foreach(x => {
         labelToNumeric.foreach { y => if (y._2 == x) {
           println(y._1)
         }
         }
       })

  }

}
```

```
restaurants.txt-1457086766000
restaurants.txt-1457086766000
restaurants.txt-1457086766000
restaurants.txt-1457086766000
restaurants.txt-1457086766000
restaurants.txt-1457086766000
restaurants.txt-1457086766000
restaurants.txt-1457086766000
restaurants.txt-1457086766000
restaurants.txt-1457086766000
restaurants.txt-1457086766000
restaurants.txt-1457086766000
```

Tweets in test data are processed using the NLP

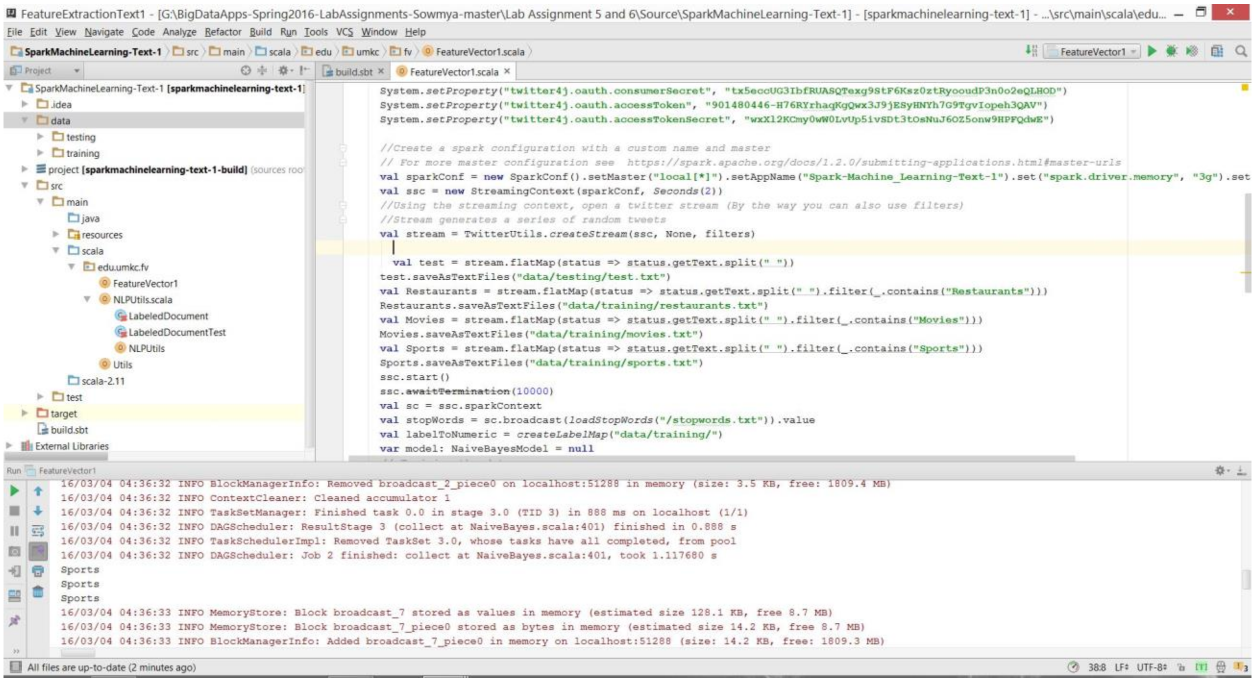
Prediction:



```scala
})

      val X_test = tfidfTransformerTest(sc, data)

        val predictionAndLabel = model.predict(X_test)
        println("PREDICTION")
        predictionAndLabel.foreach(x => {
          labelToNumeric.foreach { y => if (y._2 == x) {
            println(y._1)
          }
          }
        })
```

```
16/03/04 04:31:26 INFO SparkContext: Created broadcast 9 from broadcast at IDF.scala:181
16/03/04 04:31:26 INFO MemoryStore: ensureFreeSpace(50332048) called with curMem=17928518, maxMem=1536184811
16/03/04 04:31:26 INFO MemoryStore: Block broadcast_10 stored as values in memory (estimated size 48.0 MB, free 1399.9 MB)
16/03/04 04:31:26 INFO MemoryStore: ensureFreeSpace(2422878) called with curMem=68260566, maxMem=1536184811
16/03/04 04:31:26 INFO MemoryStore: Block broadcast_10_piece0 stored as bytes in memory (estimated size 2.3 MB, free 1397.6 MB)
16/03/04 04:31:26 INFO BlockManagerInfo: Added broadcast_10_piece0 in memory on localhost:50924 (size: 2.3 MB, free: 1461.9 MB)
16/03/04 04:31:26 INFO SparkContext: Created broadcast 10 from broadcast at NaiveBayes.scala:89
PREDICTION
16/03/04 04:31:26 INFO SparkContext: Starting job: foreach at FeatureVector1.scala:44
16/03/04 04:31:26 INFO DAGScheduler: Got job 4 (foreach at FeatureVector1.scala:44) with 2 output partitions
16/03/04 04:31:26 INFO DAGScheduler: Final stage: ResultStage 5(foreach at FeatureVector1.scala:44)
16/03/04 04:31:26 INFO DAGScheduler: Parents of final stage: List()
```

Predicting the given test data is more falls under the category of Sports Tweets.


FeatureExtractionText1 - [G:\BigDataApps-Spring2016-LabAssignments-Sowmya-master\Lab Assignment 5 and 6\Source\SparkMachineLearning-Text-1] - [sparkmachinelearning-text-1] - ...\src\main\scala\edu...
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SparkMachineLearning-Text-1 › src › main › scala › edu › umkc › fv › FeatureVector1.scala
FeatureVector1
Project
build.sbt ×
FeatureVector1.scala ×
SparkMachineLearning-Text-1 [sparkmachinelearning-text-1]
.idea
data
testing
training
project [sparkmachinelearning-text-1-build] (sources roo
src
main
java
resources
scala
edu.umkc.fv
FeatureVector1
NLPUtils.scala
LabeledDocument
LabeledDocumentTest
NLPUtils
Utils
scala-2.11
test
target
build.sbt
External Libraries
Run
FeatureVector1
All files are up-to-date (2 minutes ago)
38:8 LF‡ UTF-8‡

```scala
System.setProperty("twitter4j.oauth.consumerSecret", "tx5eccUG3IbfRUASQTexg9StF6Ksz0ztRyooudP3n0o2eQLHOD")
System.setProperty("twitter4j.oauth.accessToken", "901480446-H76RYrhagKgQwx3J9jESyHNYh7G9TgvIopeh3QAV")
System.setProperty("twitter4j.oauth.accessTokenSecret", "wxXl2KCmy0wW0LvUp5ivSDt3tOsNuJ6OZ5onw9HPFQdwE")

//Create a spark configuration with a custom name and master
// For more master configuration see  https://spark.apache.org/docs/1.2.0/submitting-applications.html#master-urls
val sparkConf = new SparkConf().setMaster("local[*]").setAppName("Spark-Machine_Learning-Text-1").set("spark.driver.memory", "3g").set
val ssc = new StreamingContext(sparkConf, Seconds(2))
//Using the streaming context, open a twitter stream (By the way you can also use filters)
//Stream generates a series of random tweets
val stream = TwitterUtils.createStream(ssc, None, filters)

  val test = stream.flatMap(status => status.getText.split(" "))
test.saveAsTextFiles("data/testing/test.txt")
val Restaurants = stream.flatMap(status => status.getText.split(" ").filter(_.contains("Restaurants")))
Restaurants.saveAsTextFiles("data/training/restaurants.txt")
val Movies = stream.flatMap(status => status.getText.split(" ").filter(_.contains("Movies")))
Movies.saveAsTextFiles("data/training/movies.txt")
val Sports = stream.flatMap(status => status.getText.split(" ").filter(_.contains("Sports")))
Sports.saveAsTextFiles("data/training/sports.txt")
ssc.start()
ssc.~~awaitTermination~~(10000)
val sc = ssc.sparkContext
val stopWords = sc.broadcast(loadStopWords("/stopwords.txt")).value
val labelToNumeric = createLabelMap("data/training/")
var model: NaiveBayesModel = null
```

```
16/03/04 04:36:32 INFO BlockManagerInfo: Removed broadcast_2_piece0 on localhost:51288 in memory (size: 3.5 KB, free: 1809.4 MB)
16/03/04 04:36:32 INFO ContextCleaner: Cleaned accumulator 1
16/03/04 04:36:32 INFO TaskSetManager: Finished task 0.0 in stage 3.0 (TID 3) in 888 ms on localhost (1/1)
16/03/04 04:36:32 INFO DAGScheduler: ResultStage 3 (collect at NaiveBayes.scala:401) finished in 0.888 s
16/03/04 04:36:32 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
16/03/04 04:36:32 INFO DAGScheduler: Job 2 finished: collect at NaiveBayes.scala:401, took 1.117680 s
Sports
Sports
Sports
16/03/04 04:36:33 INFO MemoryStore: Block broadcast_7 stored as values in memory (estimated size 128.1 KB, free 8.7 MB)
16/03/04 04:36:33 INFO MemoryStore: Block broadcast_7_piece0 stored as bytes in memory (estimated size 14.2 KB, free 8.7 MB)
16/03/04 04:36:33 INFO BlockManagerInfo: Added broadcast_7_piece0 in memory on localhost:51288 (size: 14.2 KB, free: 1809.3 MB)
```