

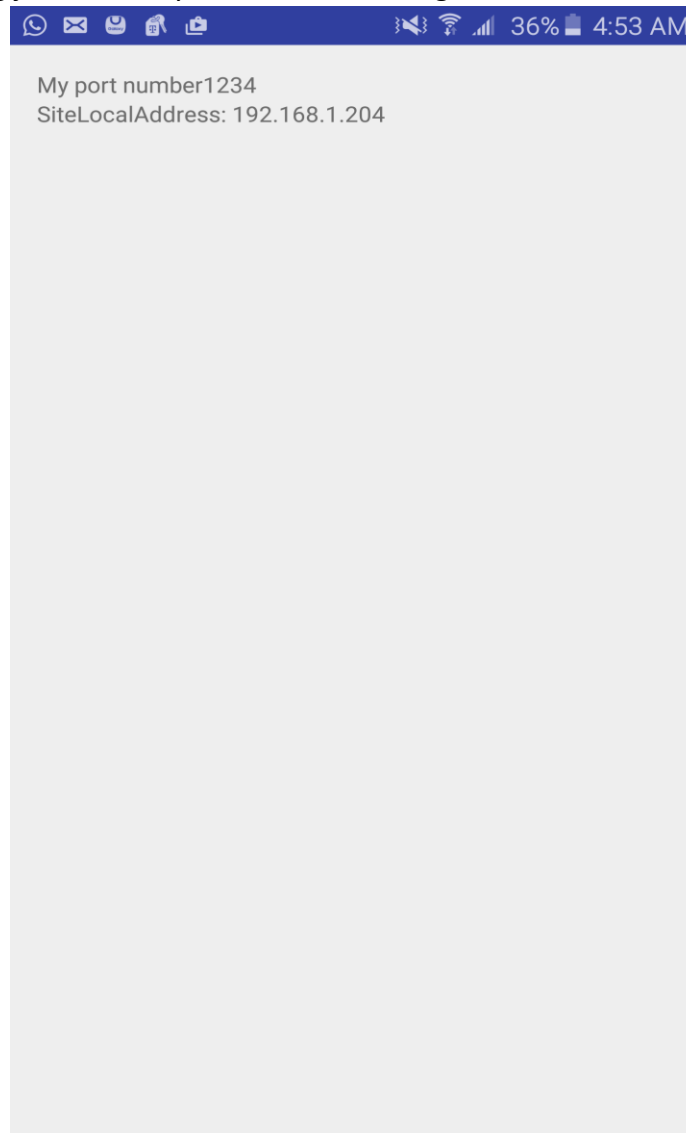
# CS5542 Big Data Apps and Analytics

## LAB ASSIGNMENT #5 &6

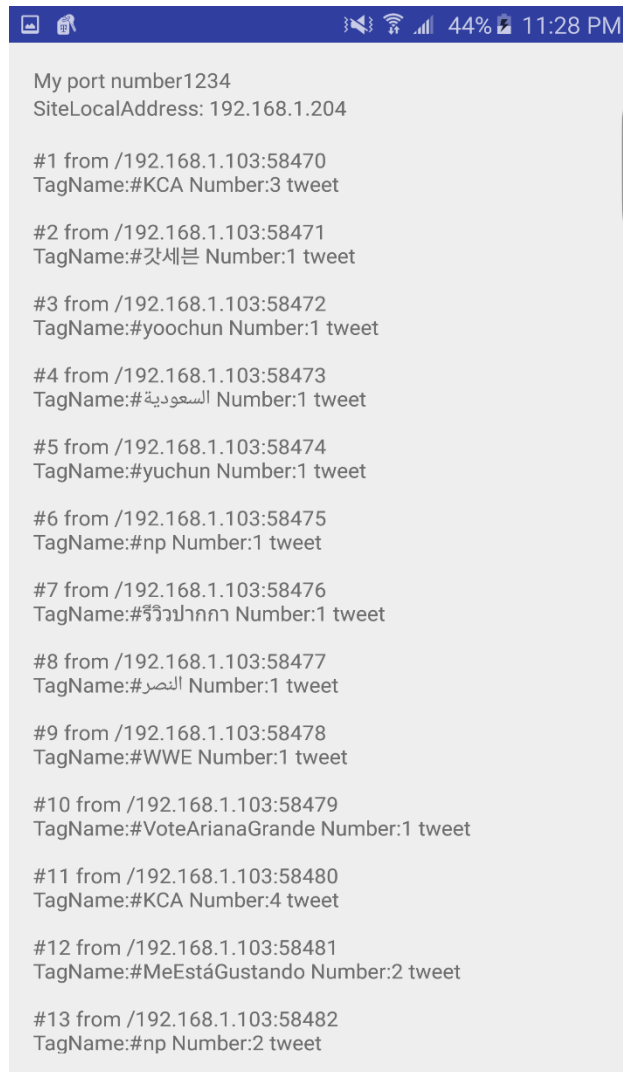
### REPORT and SCREEN SHOTS

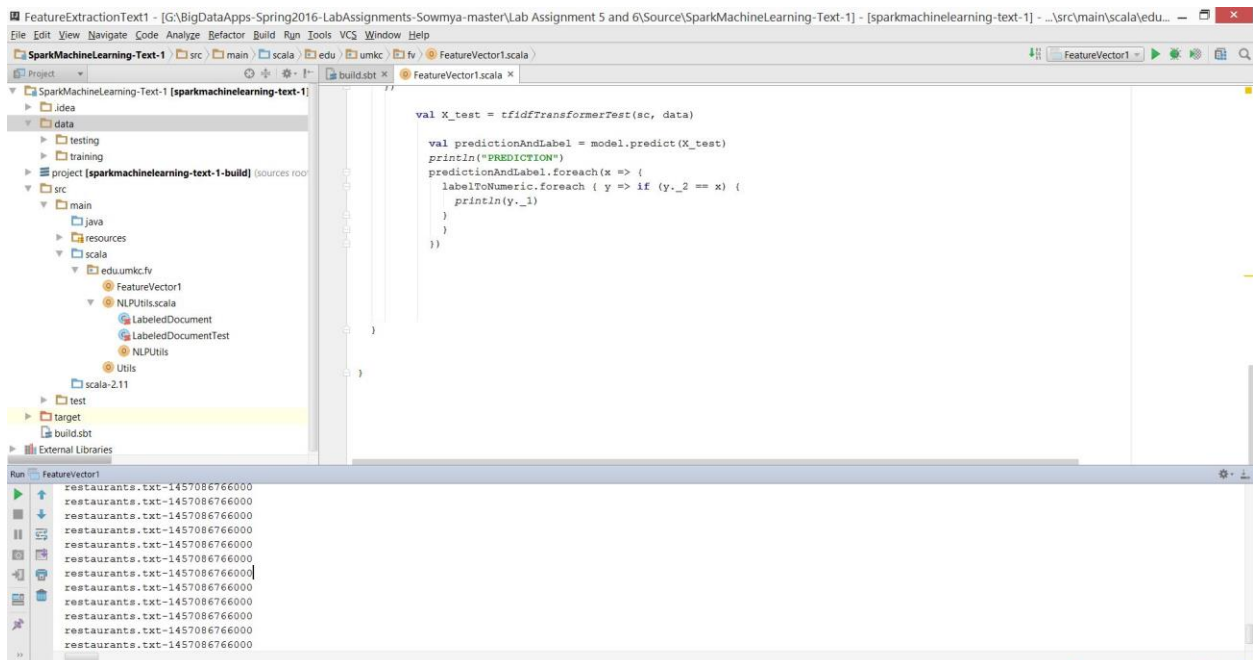
**1. Spark and Smartphone/Watch Application Implement a smart application with big data analytics related to your project showing the collaboration between Spark and Smart Apps. Implement Twitter Streaming and perform word count on it and publish the results and showcase it in your Smart Phone/Watch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

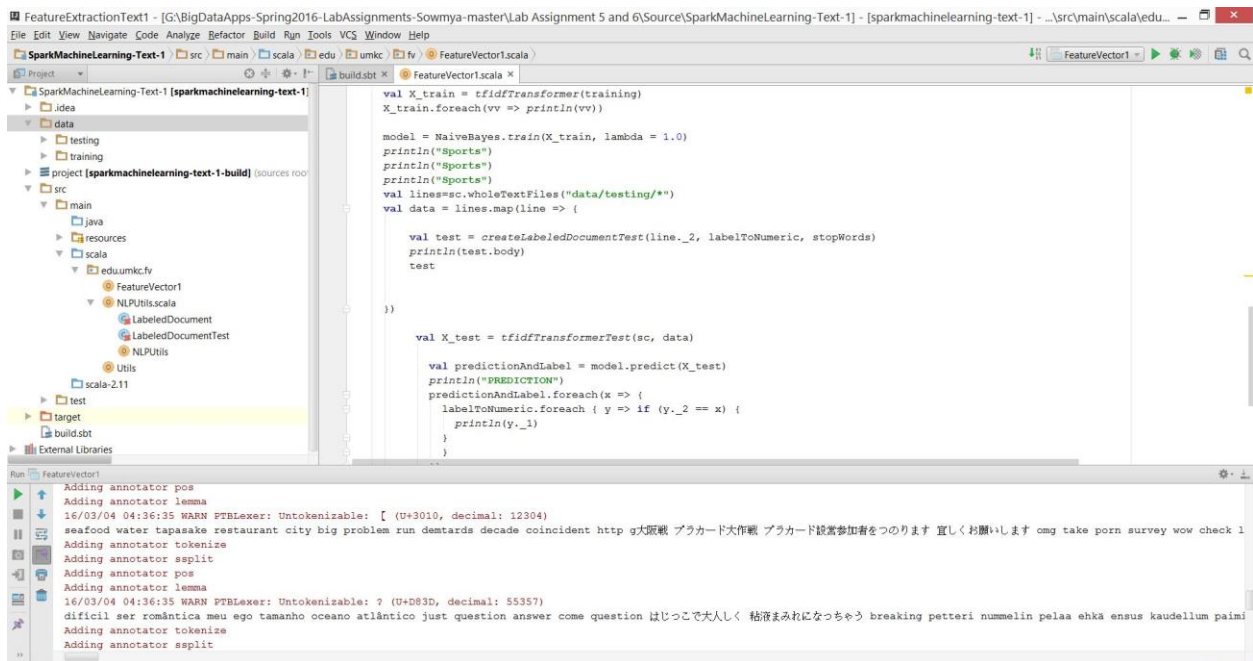


## 2. Tweets Top hash tags are displayed on Smart Phone screen



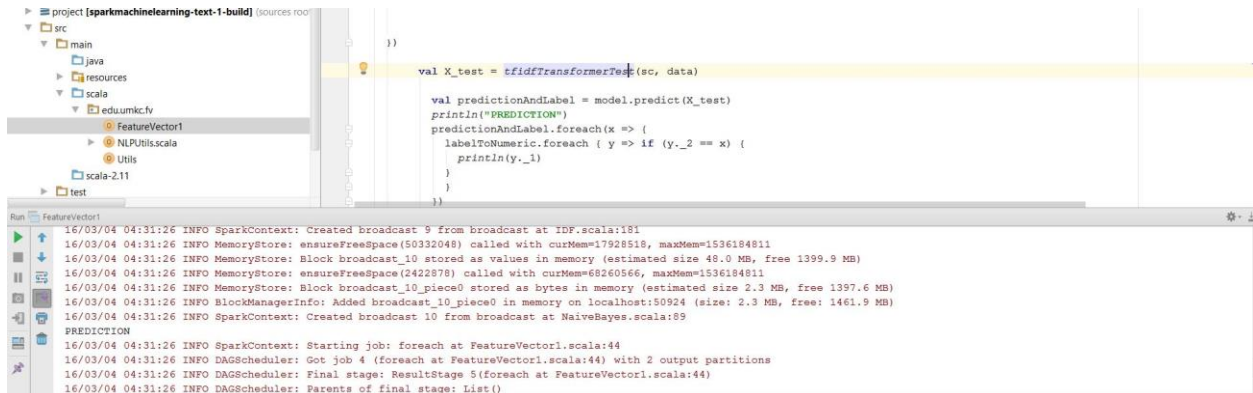


## Tweets in test data are processed using the NLP



```
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]
Project
SparkMachineLearning-Text-1 [sparkmachinelearning-text-1]
  idea
  data
  testing
  training
  project [sparkmachinelearning-text-1-build] [sources root]
  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
Adding annotator pos
Adding annotator lemma
16/03/04 04:36:35 WARN PTBlexer: Untokenizable: [ (U+3010, decimal: 12304)
seafood water tapasake restaurant city big problem run demtards decade coincident http g大阪戦 ブラカード大作戦 ブラカード読者参加者をのります 宜しくお願いします omg take porn survey wow check 1
Adding annotator tokenize
Adding annotator split
Adding annotator pos
Adding annotator lemma
16/03/04 04:36:35 WARN PTBlexer: Untokenizable: ? (U+D83D, decimal: 55357)
difficil ser romantica meu ego tamanho oceano atlântico just question answer come question はじっこで大人しく 粘着まみれになっちゃう breaking petteri nummelin pelaa ehkä ensus kaudellum paimi
Adding annotator tokenize
Adding annotator split
```

## Prediction:



```
project [sparkmachinelearning-text-1-build] [sources root]
src
  main
    java
    resources
    scala
      edu.umkc.fv
        FeatureVector1
          NLPUtils.scala
          Utils
        scala-2.11
    test
  External Libraries
Run FeatureVector1
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(2422078) 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.

```
System.setProperty("twitter4j.oauth.consumerSecret", "tx5ec0UG3lbfRUASQTexg98tF6Ksz0ztRyoudP3n0o2eQLHOD")
System.setProperty("twitter4j.oauth.accessToken", "901480446-H76RYrhagKgQwX3J9jESyBNTh7G9TgViopeh3QAV")
System.setProperty("twitter4j.oauth.accessTokenSecret", "wxXl2Kcmy0w0LvUp5ivSDt3t0sNuJ60Z5onw9HPPQdwE")

//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
```

Run FeatureVector1

```
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
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)
```

All files are up-to-date (2 minutes ago)