

## Lab Assignment 5 &amp; 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:** Here I implemented word count(hash count) on twitter streaming data and published the results on my smart phone. Here basically we have android client and spark server where data from spark is pushed on to android phone. The results on the android phone will be stopped only when we terminate spark process. The following are the screen shots

```

TwitterSparkStreaming - [F:/BAA/Tutorial ppt/Spark Streaming/TwitterSparkStreaming] - [root] - .../src/main/scala/TwitterStreaming.scala - IntelliJ IDEA 15.0.1
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
TwitterSparkStreaming | src | main | scala | TwitterStreaming.scala
Project | TwitterSparkStreaming [root] | F:/BAA/Tutorial ppt/... | val sparkConf = new SparkConf().setAppName("Deeppuapp1").setMaster("local[*]")
Run | TwitterStreaming
Popular topics in last 60 seconds (175 total):
#HeartAwards (6 tweets)
#KCA (6 tweets)
#BestFanArmy (4 tweets)
# (3 tweets)
#BestCover (2 tweets)
#NicomaineAt21 (2 tweets)
#Hadith (2 tweets)
#KaraSevda (2 tweets)
#bekleeb... (2 tweets)
#GGMD (2 tweets)
#TeenWolf (2 tweets)
#MothersDay (2 tweets)
#GAYNZ (1 tweets)
#ThabAmir (1 tweets)
#iLoveMorocco (1 tweets)
#MDFC (1 tweets)
#naturism (1 tweets)
#해외축구토토배당좋은사이트 (1 tweets)
#bloggers (1 tweets)
#RenewableEnergy (1 tweets)
#AES2016 (1 tweets)
#VotaSebastianVillalobos (1 tweets)
#UnitedStates (1 tweets)
#الجزائر (1 tweets)
#Nudiat (1 tweets)
#iLoveHawaina (1 tweets)
#VoteSM (1 tweets)
#Art (1 tweets)
#ElyTALENTOSAotriz.carisma (1 tweets)
#MySibilingWeird (1 tweets)
16/03/02 16:09:28 INFO BlockManagerInfo: Removed broadcast_74_piece0 on localhost:64429 in memory (size: 781.0 B, free: 490.5 MB)
Run Run TODO Terminal Event Log
Compilation completed successfully in 8s 324ms (2 minutes ago) 4867.25 LF: UTF-8

```

```

TwitterSparkStreaming - [F:/BAA/Tutorial ppt/Spark Streaming/TwitterSparkStreaming] - [root] - .../src/main/scala/TwitterStreaming.scala - IntelliJ IDEA 15.0.1
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
TwitterSparkStreaming | src | main | scala | TwitterStreaming.scala
Project | TwitterSparkStreaming [root] | F:/BAA/Tutorial ppt/... | val sparkConf = new SparkConf().setAppName("Deeppuapp1").setMaster("local[*]")
Run | TwitterStreaming
16/03/02 16:09:30 INFO JobScheduler: Finished job streaming job 1456956570000 ms.1 from job set of time 1456956570000 ms
16/03/02 16:09:30 INFO JobScheduler: Total delay: 0.620 s for time 1456956570000 ms (execution: 0.363 s)
16/03/02 16:09:30 INFO ShuffledRDD: Removing RDD 130 from persistence list
Popular topics in last 60 seconds (190 total):
#KCA (7 tweets)
#HeartAwards (6 tweets)
#NicomaineAt21 (4 tweets)
#BestFanArmy (4 tweets)
# (3 tweets)
#BestCover (2 tweets)
#Hadith (2 tweets)
#KaraSevda (2 tweets)
#bekleeb... (2 tweets)
#GGMD (2 tweets)
#TeenWolf (2 tweets)
#MothersDay (2 tweets)
#GAYNZ (1 tweets)
#ThabAmir (1 tweets)
#iLoveMorocco (1 tweets)
#MDFC (1 tweets)
#naturism (1 tweets)
#해외축구토토배당좋은사이트 (1 tweets)
#bloggers (1 tweets)
#RenewableEnergy (1 tweets)
#AES2016 (1 tweets)
#VotaSebastianVillalobos (1 tweets)
#UnitedStates (1 tweets)
#الجزائر (1 tweets)
#Vijay (1 tweets)
#andnang (1 tweets)
#Nudiat (1 tweets)
Run Run TODO Terminal Event Log
Compilation completed successfully in 8s 324ms (a minute ago) 5404.1 LF: UTF-8

```

## Lab Assignment 5 &amp; 6

|  |   |
|--|---|
| <p>I'm waiting here: 1234<br/>SiteLocalAddress: 192.168.1.135</p> <p>#1 from /192.168.1.171:54935<br/>Popular topics used in last 10 seconds:<br/>Words:Count</p> <p>#2 from /192.168.1.171:54939<br/>Popular topics used in last 10 seconds:<br/>Words:Count</p> <p>#3 from /192.168.1.171:54943<br/>Popular topics used in last 10 seconds:<br/>Words:Count</p> <p>#4 from /192.168.1.171:54959<br/>Popular topics used in last 10 seconds:<br/>Words:Count<br/>#고양오피 : 2<br/>#강남오피 : 2<br/>#광주오피 : 2<br/>#군포오피 : 2<br/>#nike : 1<br/>#carolinablue : 1<br/>#LilWayneALegend : 1<br/>#How : 1<br/>#20AnosSemMamonas<br/>Eu : 1<br/># 시화오피 : 1<br/>#Question : 1<br/>#lacrosse : 1<br/>#Bulls : 1<br/>#YearInSpace : 1<br/>#valorchristianhighschool : 1<br/>#KCA : 1<br/>#ConstruyendoJuntosElFuturo : 1<br/>#MueresMoviendoMexico.@gobiernohidalgo : 1<br/># 해운대오피 : 1<br/>#HowTo<br/>https://t.co/yHJ1TW6zSU : 1</p> | <p>I'm waiting here: 1234<br/>SiteLocalAddress: 192.168.1.135</p> <p>#4 from /192.168.1.171:54959<br/>Popular topics used in last 10 seconds:<br/>Words:Count<br/>#고양오피 : 2<br/>#강남오피 : 2<br/>#광주오피 : 2<br/>#군포오피 : 2<br/>#nike : 1<br/>#carolinablue : 1<br/>#LilWayneALegend : 1<br/>#How : 1<br/>#20AnosSemMamonas<br/>Eu : 1<br/># 시화오피 : 1<br/>#Question : 1<br/>#lacrosse : 1<br/>#Bulls : 1<br/>#YearInSpace : 1<br/>#valorchristianhighschool : 1<br/>#KCA : 1<br/>#ConstruyendoJuntosElFuturo : 1<br/>#MueresMoviendoMexico.@gobiernohidalgo : 1<br/># 해운대오피 : 1<br/>#HowTo<br/>https://t.co/yHJ1TW6zSU : 1</p> <p>#5 from /192.168.1.171:54967<br/>Popular topics used in last 10 seconds:<br/>Words:Count<br/>#KCA : 1</p> |
|--|---|

|   |  |
|---|--|
| <p>I'm waiting here: 1234<br/>SiteLocalAddress: 192.168.1.135</p> <p>#14 from /192.168.1.171:55060<br/>Popular topics used in last 10 seconds:<br/>Words:Count<br/>#KCA : 14<br/>#TkmSeLaComeElFandomSeLaDa : 4<br/>#NicomaineAt21 : 3<br/>#VotaSebastianVillalobos : 2<br/>#고양오피 : 2<br/>#Comedy : 2<br/>#VotaLaliEsposito : 2<br/>#Question : 2<br/>#TuitUtil : 2<br/>#WishesToCamila : 2<br/>#광주오피 : 2<br/>#강남오피 : 2<br/>#CriminalMinds : 2<br/>#군포오피 : 2<br/>#MahaRehmokarmDiwas : 1<br/>#0405 : 1<br/>#Ch : 1<br/>#Sagitario : 1<br/>#Makai : 1<br/>#Antipasto : 1<br/>#진영 : 1<br/>#comic : 1<br/>#escuchoeltrasnochow : 1<br/>#322 : 1<br/>#엔터K : 1<br/>#nike : 1<br/>#carolinablue : 1<br/>#B612 : 1<br/>#Caps : 1</p> | <p>I'm waiting here: 1234<br/>SiteLocalAddress: 192.168.1.135</p> <p>#6 from /192.168.1.171:54977<br/>Popular topics used in last 10 seconds:<br/>Words:Count<br/>#KCA : 5<br/>#고양오피 : 2<br/>#VotaLaliEsposito : 2<br/>#Question : 2<br/>#TuitUtil : 2<br/>#강남오피 : 2<br/>#광주오피 : 2<br/>#군포오피 : 2<br/>#MahaRehmokarmDiwas : 1<br/>#0405 : 1<br/>#Ch : 1<br/>#Makai : 1<br/>#nike : 1<br/>#carolinablue : 1<br/>#VotaSebastianVillalobos : 1<br/>#OurThreeBrothers : 1<br/>#feat_xiumin : 1<br/>#OurThreeBoys : 1<br/>#DE : 1<br/>#blacklivesmatter : 1<br/>#LilWayneALegend : 1<br/>#CALLYOUBAE : 1<br/>#How : 1<br/>#三月生誕祭 : 1<br/>#Saginaw : 1<br/>#20AnosSemMamonas<br/>Eu : 1<br/>#playlist : 1<br/>#qandapril : 1<br/>#leina : 1</p> |
|---|--|

## Lab Assignment 5 &amp; 6

**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:** we first classify the tweets based on hashtags like "Oscar", "Trump", "Music" etc into training folder. Now the upcoming stream is dumped into testing folder and prediction is made saying to which category hashtag it belongs to

**Screen Shots:**

```

twitterdj - [F:/BAA/twitterdj] - [twitterdj] - ...src/main/scala/TwitterStream.scala - IntelliJ IDEA 15.0.1
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
twitterdj src main scala TwitterStream.scala
Run StatusStreamer
* Created by DEEPU on 3/2/2016.

Retweet and spread.
KATHNIEL AsiasEmergingLoveTeam
#VoteKathrynFPP #KCA https://t.co/dDfoFBdKQ
RT @TATAKLizQuen: Ha-surprise pero maganda pa rin. Ikaw na ang pinagpala.

DolceAmore Surprise
#VoteEnriqueFPP #KCA https://t.co/IUCiv0iVod
And all I can do is be true to myself
I don't need permission from nobody else ☺ @

#VoteKathrynFPP #KCA 30/50 https://t.co/FSageEHYVF
RT @ianderwincg: Whoah! Iba na talaga ang Chemistry nina @bernardokath @imdanieldpadilla Mr Trabaho
#VoteKathrynFPP #KCA https://t.co/E2Du...
RT @primadonilo: @imdanieldpadilla WTF

#VoteKathrynFPP #KCA
RT @mariobautista_: Que dicen votamos un ratito o andan cansadonas ? #KCA #VoteMarioBautista ☺
#VoteKathrynFPP #KCA https://t.co/GIG2n6QD1M
RT @BarbieAbilon: Smile everyday :)
#VoteKathrynFPP #KCA https://t.co/P5584NgJ5x
RT @Bautisteragdl_: Recuerden dar RT RT y más RT para juntar más y más votos☺☺☺

#KCA #VoteMarioBautista
RT @JaDine_Mommies: Bash pa more! Para lalong sumikat ang JaDine!

VOTE NADINE FOR MYX
#VoteJamesFPP #KCA
RT @imdanieldpadilla: #VoteKathrynFPP #KCA https://t.co/ciQTw012gK
Hello there KN! ☺ #VoteKathrynFPP #KCA https://t.co/n3Mx2K5Kv
#VoteDragMeDownUK #KCA https://t.co/QC9ITS5ub5
RT @lauraruiz_14: Querida luna, tú que puedes verlo dile cuánto deseo volver a sentirme segura en sus brazos. #KCA #VoteSebastiánVillalobos

Run TODO Terminal Event Log
Compilation completed successfully in 4s 392ms (moments ago)
841:1 CRLF UTF-8

```

## Lab Assignment 5 & 6

The image shows the IntelliJ IDEA 15.0.1 IDE interface. The title bar at the top reads "FeatureExtractionText1 - [F:/BAA/Tutorial ppt/SparkMachineLearning-Text-1 - .../src/main/scala/edu/umkc/fv/FeatureVector1.scala - IntelliJ IDEA 15.0.1". The menu bar includes File, Edit, View, Navigate, Code, Analyze, Refactor, Build, Run, Tools, VCS, Window, and Help. The toolbar contains icons for saving, running, and other IDE functions. The left sidebar shows the "Project" view with a tree structure of the project "sparkmachinelearning-text-1-build". The tree includes folders like "data", "testing", "training", "CRICKET", "KCA", "MUSIC", "OSCAR", and "TRUMP". The "src" folder is expanded, showing "main", "java", "resources", "scala", and "edu.umkc.fv". The "scala" folder is selected, and the "FeatureVector1.scala" file is open in the main editor. The code in the editor is as follows:

```
package edu.umkc.fv

import edu.umkc.fv.NLPUtils._
import edu.umkc.fv.Utils._
import org.apache.spark.SparkConf
import org.apache.spark.mllib.classification.{NaiveBayes, NaiveBayesModel}
import org.apache.spark.streaming.twitter.TwitterUtils
import org.apache.spark.streaming.{Seconds, StreamingContext}

/**
 * Created by Deepu on 02-Mar-16.
 */
object FeatureVector1 {

  def main(args: Array[String]) {
    System.setProperty("hadoop.home.dir", "F:\\winutils")

    val filters = args

    // Set the system properties so that Twitter4j library used by twitter stream
    // can use them to generate OAuth credentials

    System.setProperty("twitter4j.oauth.consumerKey", "cOQ8HCgri8t5HCbyhb8HDOZRW5")
    System.setProperty("twitter4j.oauth.consumerSecret", "41ZbQlsAI6V8HCfHXORFSPmSWNlPaAYqaaE0lftyt598FsrJ")
    System.setProperty("twitter4j.oauth.accessToken", "4565837185-yKlZgmTW3Ea2r4p05hRgpp53pipeTjv590Q4")
    System.setProperty("twitter4j.oauth.accessTokenSecret", "HsR84cLkRkFBgjcnaH6Wae6on50E7DeoYr6H6hpj0rT8")

    //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().setAppName("Deepuappl").setMaster("local[*"]).setAppName("FeatureVector1").set("spark.driver.memory", "1g")
    //Create a StreamingContext with 30 second window
    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 tweets = TwitterStreamSource(ssc, sparkConf).newStream()
  }
}
```

The bottom status bar shows "Run FeatureVector1", "Run", "TODO", "Terminal", and "Event Log". The status bar also indicates "All files are up-to-date (2 minutes ago)" and "85:1 CRLF UTF-8".

The screenshot shows the IntelliJ IDEA 15.0.1 IDE. The top toolbar includes menus for File, Edit, View, Navigate, Code, Analyze, Refactor, Build, Run, Tools, VCS, Window, and Help. The left sidebar shows the project structure for 'SparkMachineLearning-Text-1', with folders like 'data', 'testing', 'training', 'CRICKET', 'KCA', 'MUSIC', 'OSCAR', and 'TRUMP'. The main editor window displays a Scala file named 'FeatureVector1.scala' containing a list of tweets. The bottom status bar indicates the file is up-to-date (2 minutes ago).

## Lab Assignment 5 &amp; 6

## Processing data in test folder:stemming and lemmentation

```

package edu.umkc.fv

import edu.umkc.fv.NLPUtils._
import edu.umkc.fv.Utils._
import org.apache.spark.SparkConf
import org.apache.spark.mllib.classification.{NaiveBayes, NaiveBayesModel}
import org.apache.spark.streaming.twitter.TwitterUtils
import org.apache.spark.streaming.{Seconds, StreamingContext}

/**
 * Created by Deepu on 02-Mar-16.
 */
object FeatureVector1 {

  def main(args: Array[String]) {
    // ... (code continues)
  }
}

```

Run FeatureVector1

```

Adding annotator pos
Adding annotator lemma
16/03/02 23:12:35 WARN PTBTokenizer: Untokenizable: (U+FFFD, decimal: 65533)
retweeted jus hope leave lll legacy behind future hooper rock mean something number mmm mmm mmmextra kellog superstar card reggie jackson kung sagutin kaya yan bday now lol st
Adding annotator tokenize
Adding annotator split
Adding annotator pos
Adding annotator lemma
16/03/02 23:12:36 WARN PTBTokenizer: Untokenizable: (U+FFFD, decimal: 65533)
soakoh vaarey vehumuge sababun feydhoo gethakah libifaivaas gehlun ballavalavanee kajriwal liberal anti national element political score bjp government centre people watch year
16/03/02 23:12:36 INFO MemoryStore: ensureFreeSpace(2304) called with curMem=9137943, maxMem=499415777
16/03/02 23:12:36 INFO BlockManagerInfo: Added rdd_12_1 in memory on localhost:57399 (size: 2.3 KB, free: 475.8 MB)
16/03/02 23:12:36 INFO Executor: Finished task 1.0 in stage 4.0 (TID 9). 8432427 bytes result sent to driver
16/03/02 23:12:36 INFO TaskSetManager: Finished task 1.0 in stage 4.0 (TID 9) in 749 ms on localhost (1/2)
someone happy and just watch thing people anxiety want friend know via themighcyate bore get marry will wife giant minnie mouse head brighten day thank original friend script
16/03/02 23:12:36 INFO MemoryStore: ensureFreeSpace(2904) called with curMem=9140247, maxMem=499415777
16/03/02 23:12:36 INFO MemoryStore: Block rdd_12_0 stored as values in memory (estimated size 2.8 KB, free 467.6 MB)
16/03/02 23:12:36 INFO BlockManagerInfo: Added rdd_12_0 in memory on localhost:57399 (size: 2.8 KB, free: 475.8 MB)

```

```

package edu.umkc.fv

import edu.umkc.fv.NLPUtils._
import edu.umkc.fv.Utils._
import org.apache.spark.SparkConf
import org.apache.spark.mllib.classification.{NaiveBayes, NaiveBayesModel}
import org.apache.spark.streaming.twitter.TwitterUtils
import org.apache.spark.streaming.{Seconds, StreamingContext}

/**
 * Created by Deepu on 02-Mar-16.
 */
object FeatureVector1 {

  def main(args: Array[String]) {
    // ... (code continues)
  }
}

```

Run FeatureVector1

```

16/03/02 23:12:37 INFO MemoryStore: Block broadcast_10 stored as values in memory (estimated size 80.0 MB, free 379.2 MB)
16/03/02 23:12:37 INFO MemoryStore: ensureFreeSpace(4026210) called with curMem=101817428, maxMem=499415777
16/03/02 23:12:37 INFO MemoryStore: Block broadcast_10_piece0 stored as bytes in memory (estimated size 3.8 MB, free 375.3 MB)
16/03/02 23:12:37 INFO BlockManagerInfo: Added broadcast_10_piece0 in memory on localhost:57399 (size: 3.8 MB, free: 471.6 MB)
16/03/02 23:12:37 INFO SparkContext: Created broadcast 10 from broadcast at NaiveBayes.scala:89
PREDICTION
16/03/02 23:12:39 INFO SparkContext: Starting job: foreach at FeatureVector1.scala:67
16/03/02 23:12:39 INFO DAGScheduler: Got job 4 (foreach at FeatureVector1.scala:67) with 2 output partitions
16/03/02 23:12:39 INFO DAGScheduler: Final stage: ResultStage 5(foreach at FeatureVector1.scala:67)
16/03/02 23:12:39 INFO DAGScheduler: Parents of final stage: List()
16/03/02 23:12:39 INFO DAGScheduler: Submitting ResultStage 5 (MapPartitionsRDD[17] at mapPartitions at NaiveBayes.scala:90), which has no missing parents
16/03/02 23:12:39 INFO MemoryStore: ensureFreeSpace(6016) called with curMem=105843638, maxMem=499415777
16/03/02 23:12:39 INFO MemoryStore: Block broadcast_11 stored as values in memory (estimated size 5.9 KB, free 375.3 MB)
16/03/02 23:12:39 INFO MemoryStore: ensureFreeSpace(3600) called with curMem=105849654, maxMem=499415777
16/03/02 23:12:39 INFO MemoryStore: Block broadcast_11_piece0 stored as bytes in memory (estimated size 3.5 KB, free 375.3 MB)
16/03/02 23:12:39 INFO BlockManagerInfo: Added broadcast_11_piece0 in memory on localhost:57399 (size: 3.5 KB, free: 471.6 MB)
16/03/02 23:12:39 INFO SparkContext: Created broadcast 11 from broadcast at DAGScheduler.scala:861
16/03/02 23:12:39 INFO DAGScheduler: Submitting 2 missing tasks from ResultStage 5 (MapPartitionsRDD[17] at mapPartitions at NaiveBayes.scala:90)

```

## Lab Assignment 5 &amp; 6

## Predicted results:

```
package edu.umkc.fv

import edu.umkc.fv.NLPUtils._
import edu.umkc.fv.Utils._
import org.apache.spark.SparkConf
import org.apache.spark.mllib.classification.{NaiveBayes, NaiveBayesModel}
import org.apache.spark.streaming.twitter.TwitterUtils
import org.apache.spark.streaming.{Seconds, StreamingContext}

/**
 * Created by Deepu on 02-Mar-16.
 */
object FeatureVector1 {

  def main(args: Array[String]) {
```

Run FeatureVector1

```
16/03/02 23:12:39 INFO Executor: Running task 1.0 in stage 5.0 (TID 11)
16/03/02 23:12:39 INFO Executor: Running task 0.0 in stage 5.0 (TID 10)
16/03/02 23:12:39 INFO BlockManager: Found block rdd_12_1 locally
16/03/02 23:12:39 INFO BlockManager: Found block rdd_12_1 locally
16/03/02 23:12:39 INFO BlockManager: Found block rdd_12_0 locally
16/03/02 23:12:39 INFO BlockManager: Found block rdd_12_0 locally
16/03/02 23:12:39 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeSystemBLAS
16/03/02 23:12:39 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
TRUMP
TRUMP
TRUMP
16/03/02 23:12:39 INFO Executor: Finished task 0.0 in stage 5.0 (TID 10). 2044 bytes result sent to driver
16/03/02 23:12:39 INFO Executor: Finished task 1.0 in stage 5.0 (TID 11). 2044 bytes result sent to driver
16/03/02 23:12:39 INFO TaskSetManager: Finished task 0.0 in stage 5.0 (TID 10) in 122 ms on localhost (1/2)
16/03/02 23:12:39 INFO TaskSetManager: Finished task 1.0 in stage 5.0 (TID 11) in 121 ms on localhost (2/2)
16/03/02 23:12:39 INFO DAGScheduler: ResultStage 5 (foreach at FeatureVector1.scala:67) finished in 0.123 s
16/03/02 23:12:39 INFO TaskSchedulerImpl: Removed TaskSet 5.0, whose tasks have all completed, from pool
16/03/02 23:12:39 INFO DAGScheduler: Job 4 finished: foreach at FeatureVector1.scala:67, took 0.141441 s
16/03/02 23:12:40 INFO ReceiverTracker: Starting 1 receivers
```

119 chars, 4 lines 329:1 CRLF: UTF-8: [T]

All files are up-to-date (4 minutes ago)