Mohamed Gharibi (7)

1). Collect tweets using twitter streaming

```scala
System.setProperty("twitter4j.oauth.consumerKey", "R2v2WMKrF7UGipifRcMkOyjT1"
System.setProperty("twitter4j.oauth.consumerSecret", "InkVkLJfUsJPQyA17GzGks9
System.setProperty("twitter4j.oauth.accessToken", "3630687739-9y2qw6YKOMgeApm
System.setProperty("twitter4j.oauth.accessTokenSecret", "IBjoDz21BTBaXwnJ13jy

//Create a spark configuration with a custom name and master
// For more master configuration see  https://spark.apache.org/docs/1.2.0/sub
val sparkConf = new SparkConf().setAppName("STweetsApp").setMaster("local[*]"
//Create a Streaming Context with 2 second window
val ssc = new StreamingContext(sparkConf, Seconds(2))
//Using the streaming context, open a twitter stream (By the way you can also
//Stream generates a series of random tweets
val stream = TwitterUtils.createStream(ssc, None, filters)
stream.print()
//Map : Retrieving text
val getText = stream.flatMap(status => status.getText.split(" "))
```

2). Do MapReduce on the collected data.

```scala
//Finding the top word used on 6 second window
val topCounts3 = getText.map((_, 1)).reduceByKeyAndWindow(_ + _, Seconds(6))
  .map{case (word, count) => (count, word)}
  .transform(_.sortByKey(false))
```

3). Store results in a string and send it to Android phone.

Just get the top 20 results

```scala
// Print popular words
topCounts3.foreachRDD(rdd => {
  val topList = rdd.take(20)
  println("\nPopular words used in last 6 seconds (%s total):".format(rdd.count()))
  topList.foreach{case (count, word) => println("%s (%s times)".format(word, count))}


  var s:String="Popular words used in last 6 seconds (%s total): \nWords:Count \n"
  topList.foreach{case(count,word)=>{

    s+=word+" : "+count+"\n"

  }}
  SocketClient.sendCommandToRobot(s)
})
```

Screenshots:

Run Log:

```
Popular words used in last 6 seconds (415 total):
RT (26 times)
 (6 times)
que (6 times)
the (6 times)
mi (4 times)
de (4 times)
is (4 times)
me (3 times)
eu (3 times)
un (3 times)
for (3 times)
y (3 times)
- (3 times)
a (3 times)
I (3 times)
& (3 times)
you (3 times)
and (3 times)
los (2 times)
circo (2 times)
```

Android:

```
I'm waiting here: 1234
SiteLocalAddress: 192.168.1.10

#1 from /192.168.1.4:62214
Popular words used in last 6 seconds:
Words:Count
replayed: Hello from Android, you are #1
#2 from /192.168.1.4:62217
Popular words used in last 6 seconds:
Words:Count
RT : 26
 : 6
que : 6
the : 6
mi : 4
de : 4
is : 4
me : 3
eu : 3
un : 3
for : 3
y : 3
- : 3
a : 3
I : 3
& : 3
you : 3
and : 3
los : 2
circo : 2
replayed: Hello from Android, you are #2
#3 from /192.168.1.4:62218
Popular words used in last 6 seconds:
Words:Count
RT : 62
 : 19
que : 14
the : 10
de : 10
a : 9
- : 7
is : 7
for : 6
of : 6
me : 5
mi : 5
x : 5
The : 5
I : 5
y : 5
& : 5
```

Question 2:

1. Get the trained data from Twitter using: GetTrainingData

```scala
val trainingFoodStream = stream.filter(_.getHashtagEntities.mkString.contains("food")).map(_.getText)
val trainingSportStream = stream.filter(_.getHashtagEntities.mkString.contains("sport")).map(_.getText)
val trainingOthersStream = stream.filter(!_.getHashtagEntities.mkString.contains("food")).filter(!_.getHashtagEntities.mkString.contains("sport")).m

val trainingFood = trainingFoodStream.foreachRDD(rdd =>
  { val count = rdd.count()
    if (count > 0){
      rdd.repartition(1).saveAsTextFile("data/training/hashtag.food")
    }
  })
val trainingSport = trainingSportStream.foreachRDD(rdd =>
  { val count = rdd.count()
    if (count > 0) {
      rdd.repartition(1).saveAsTextFile("data/training/hashtag.sport")
    }
  })
val trainingOthers = trainingOthersStream.foreachRDD((rdd,time) =>
  { val count = rdd.count()
    if (count > 0) {
      rdd.repartition(1).saveAsTextFile("data/training/hashtag.other")
      val temp = count
    }
  })
```

While the program is running the testing data will be stored

```scala
val stream = TwitterUtils.createStream(ssc, None, filters)
    stream.print()

//Testing data
val testing = stream.filter(!_.getHashtagEntities.isEmpty).map(_.getText)

val testingData = testing.foreachRDD(rdd =>
{ val count = rdd.count()
  if (count > 0){
     rdd.saveAsTextFile("data/testing/")
  }
})
```

Train the trained data and do the predictions on the testing data

```scala
//Train tweets datasets
val sc = ssc.sparkContext
val stopWords = sc.broadcast(loadStopWords("/stopwords.txt")).value
val labelToNumeric = createLabelMap("data/training/")
println(labelToNumeric)
var model: NaiveBayesModel = null

val training = sc.wholeTextFiles("data/training/*")
  .map(rawText => createLabeledDocument(rawText, labelToNumeric, stopWords))
val X_train = tfidfTransformer(training)
X_train.foreach(vv => println(vv))

model = NaiveBayes.train(X_train, lambda = 1.0)


val lines=sc.wholeTextFiles("data/testing/*")
val data = lines.map(line => {

  val test = createLabeledDocumentTest(line._2, labelToNumeric, stopWords)
  println(test.body)
  test

})


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)
  }
  }
})
```
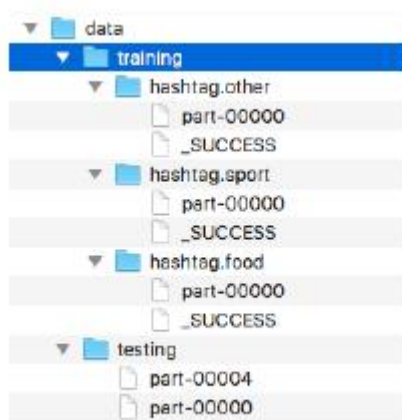
Folder Structure:

Example of trained data:

```
● ● ●                    📄 part-00000 ⌄
Noite da Pizza... #pizza #food #foodpic #instafood #instagood #yummi #nacasadosprimo #familia…
https://t.co/yefiUF0y4P
```

```
● ● ●                    📄 part-00000 ⌄
#tickets #deals NASCAR SPRINT CUP SERIES/ AUTO CLUB 400, 2 tickets 3/20/2016 https://t.co/
arnNMUG8Cg #sports #cheap #ebay
```

```
● ● ●                    📄 part-00000 ⌄
RT @ethiopianpeter: If it comes between @realDonaldTrump vs @HillaryClinton for president, I
will chose not to vote #SouthCarolinaPrimary
I'm trying to binge watch all of Fuller House since we're getting rid of Netflix tomorrow but
the wifi decided to die.😭
RT @la_iguanatv: Muere niña víctima del hambre en Colombia  https://t.co/nNMAXU9S0L https://
t.co/sEZtw8Wwix
*
RT @EthanDolan: So close to meeting you all😍😩😊😊
RT @SyahmirulAmir: Drama melayu..asal budak u je kehulu hilir pegang buku kat tangan..dia xtau
ke real life student g class kertas pon mint…
Miauuu
RT @EthanDolan: So close to meeting you all😍😩😊😊
RT @sorelatabIe: When your phone autocorrects to its ducking lit https://t.co/pLbFMaIMQv
@MEI_KMYD
おはよぉ!
RT @EthanDolan: So close to meeting you all😍😩😊😊
RT @oh_jessJessJESS: RT @iDntWearCondoms Ladies u just won 10 MILL... Your man in jail, bail 6
mill... How much do u have? https://t.co/Dhe…
I have really bad taste in guys sometimes
@sweet_minga1 KKKKKKK ja perdoou?
RT @kplove8226: These Things that Mean More to Men Than "I Love You" https://t.co/WuKJ1zyQ9j
https://t.co/BnmAtVPVFn
En un rato a lo de Rodri
RT @HmadiFatma: #ihabamir
#ايهاب_مغربي_حتى_النخاع
مغربي حر بحبه وأخلاقه جعل علم المغرب
يرفع في الجزائر والجزائري يرفع في المغرب
```

Example of testing data:

```
● ● ●                    📄 part-00001 ⌄
*LIVE NOW*
*Husband & WIFEY! play RL*

https://t.co/TSEKWPRpkE

#RocketLeague #twitch @dirkened @RocketLeague https://t.co/Va06xixmq0
```