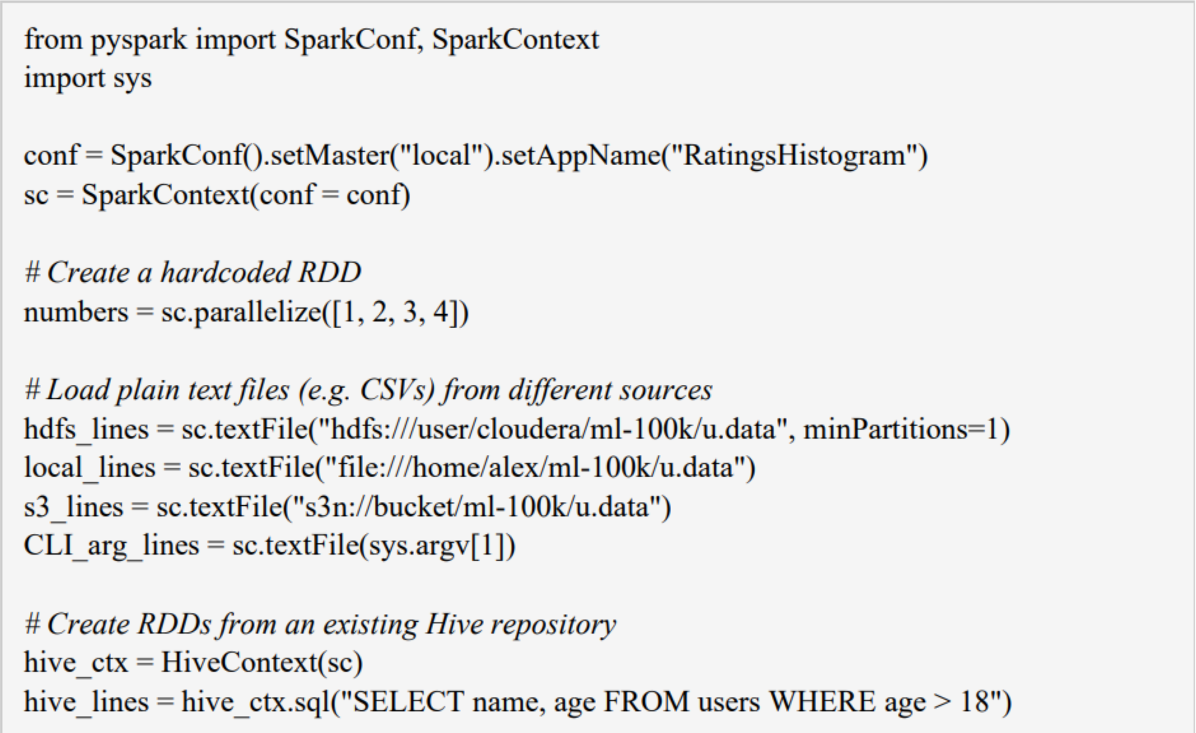
PRACTICAL-11

AIM:

To perform Sentiment Analysis using Twitter data, Scala and Spark.

CODE:

Reading data:



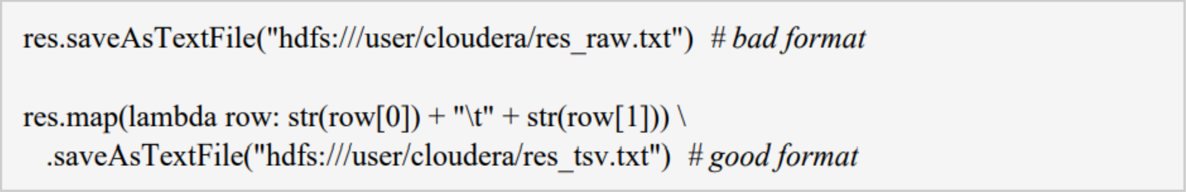
```
from pyspark import SparkConf, SparkContext
import sys

conf = SparkConf().setMaster("local").setAppName("RatingsHistogram")
sc = SparkContext(conf = conf)

# Create a hardcoded RDD
numbers = sc.parallelize([1, 2, 3, 4])

# Load plain text files (e.g. CSVs) from different sources
hdfs_lines = sc.textFile("hdfs:///user/cloudera/ml-100k/u.data", minPartitions=1)
local_lines = sc.textFile("file:///home/alex/ml-100k/u.data")
s3_lines = sc.textFile("s3n://bucket/ml-100k/u.data")
CLI_arg_lines = sc.textFile(sys.argv[1])

# Create RDDs from an existing Hive repository
hive_ctx = HiveContext(sc)
hive_lines = hive_ctx.sql("SELECT name, age FROM users WHERE age > 18")
```

Notice that this can not be run with the standard Python interpreter. Instead, you use the spark-submit to submit it as a batch task, or name the Shell Pyspark

Writing of data:

● The RDD class has the method saveAsTextFile. However, this saves the representation of each variable in a string. In Python, the resulting text file will contain lines such as (1949, 111).

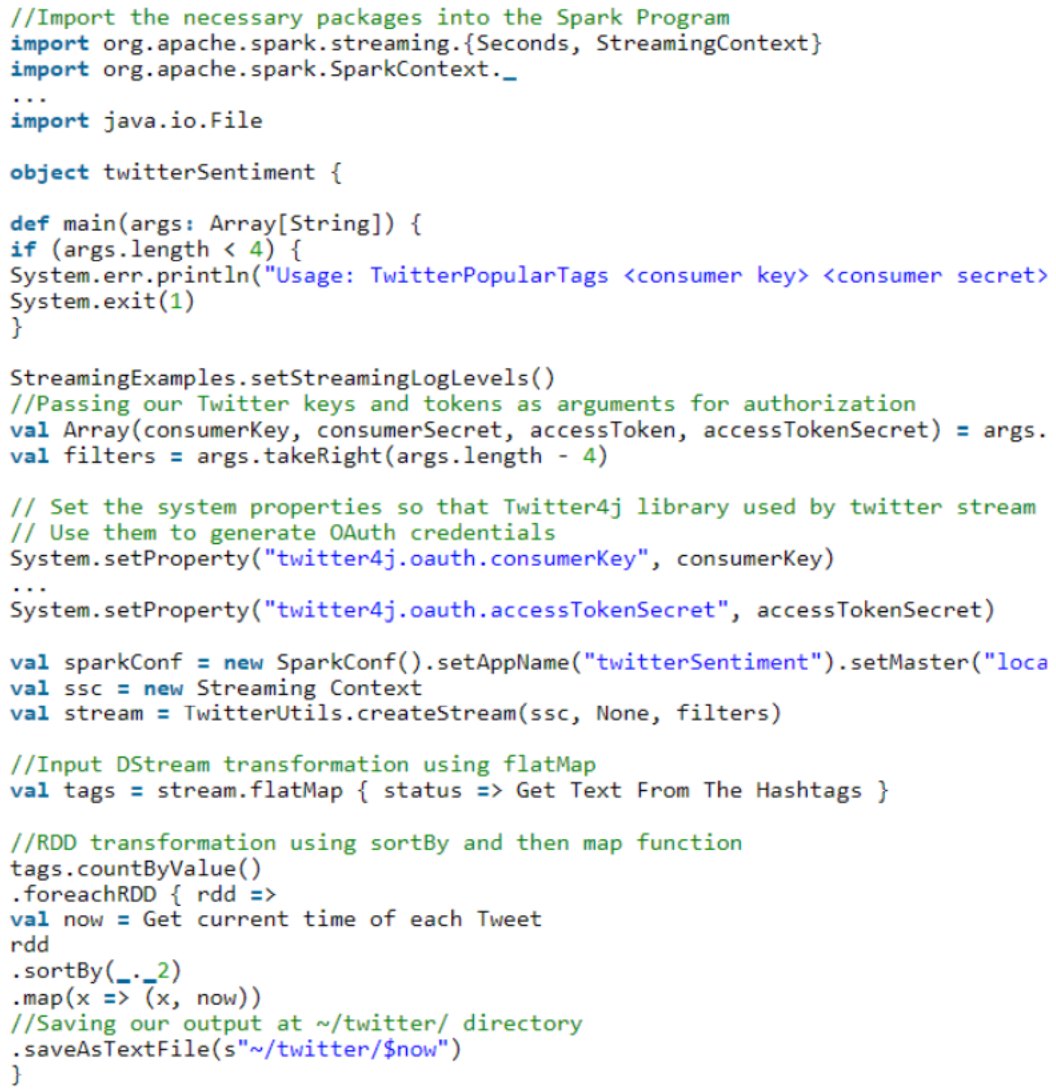
● If you want to save the data in CSV or TSV format, you can either use Python's StringIO and csv modules or simply map each variable (vector) into a single string, e.g. as follows:



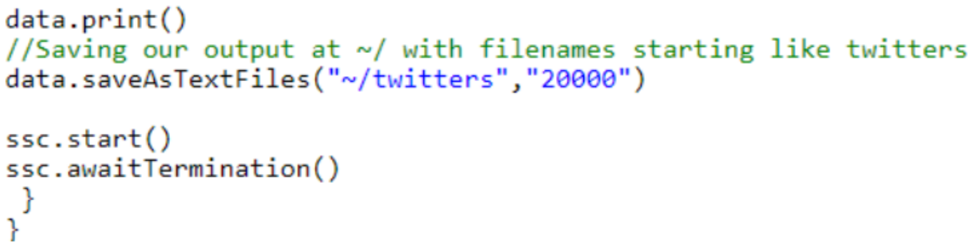
```
res.saveAsTextFile("hdfs:///user/cloudera/res_raw.txt")  # bad format

res.map(lambda row: str(row[0]) + "\t" + str(row[1])) \
  .saveAsTextFile("hdfs:///user/cloudera/res_tsv.txt")  # good format
```

Sentiment Analysis using Scala:

Spark Streaming Implementation:

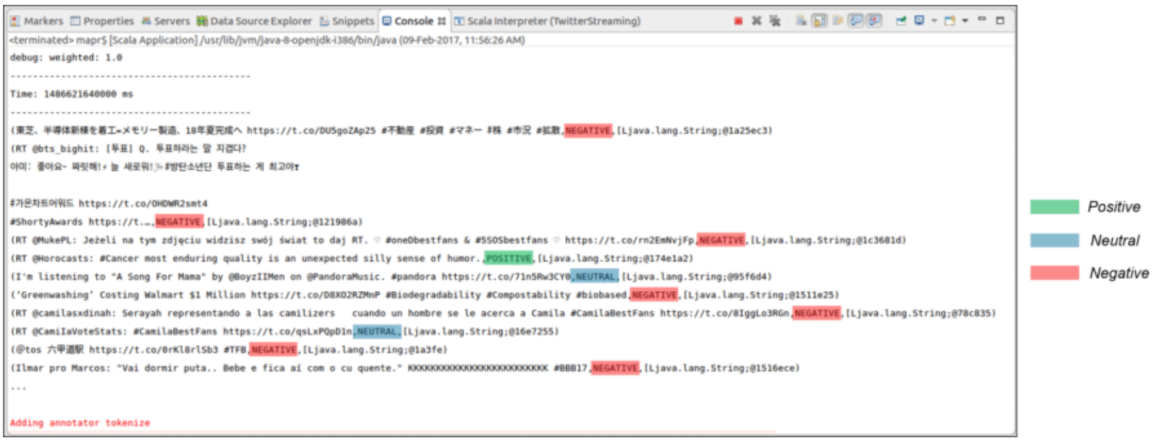


```
//Import the necessary packages into the Spark Program
import org.apache.spark.streaming.{Seconds, StreamingContext}
import org.apache.spark.SparkContext._
...
import java.io.File

object twitterSentiment {

def main(args: Array[String]) {
if (args.length < 4) {
System.err.println("Usage: TwitterPopularTags <consumer key> <consumer secret>
System.exit(1)
}

StreamingExamples.setStreamingLogLevels()
//Passing our Twitter keys and tokens as arguments for authorization
val Array(consumerKey, consumerSecret, accessToken, accessTokenSecret) = args.
val filters = args.takeRight(args.length - 4)

// Set the system properties so that Twitter4j library used by twitter stream
// Use them to generate OAuth credentials
System.setProperty("twitter4j.oauth.consumerKey", consumerKey)
...
System.setProperty("twitter4j.oauth.accessTokenSecret", accessTokenSecret)

val sparkConf = new SparkConf().setAppName("twitterSentiment").setMaster("loca
val ssc = new Streaming Context
val stream = TwitterUtils.createStream(ssc, None, filters)

//Input DStream transformation using flatMap
val tags = stream.flatMap { status => Get Text From The Hashtags }

//RDD transformation using sortBy and then map function
tags.countByValue()
.foreachRDD { rdd =>
val now = Get current time of each Tweet
rdd
.sortBy(_._2)
.map(x => (x, now))
//Saving our output at ~/twitter/ directory
.saveAsTextFile(s"~/twitter/$now")
}
```



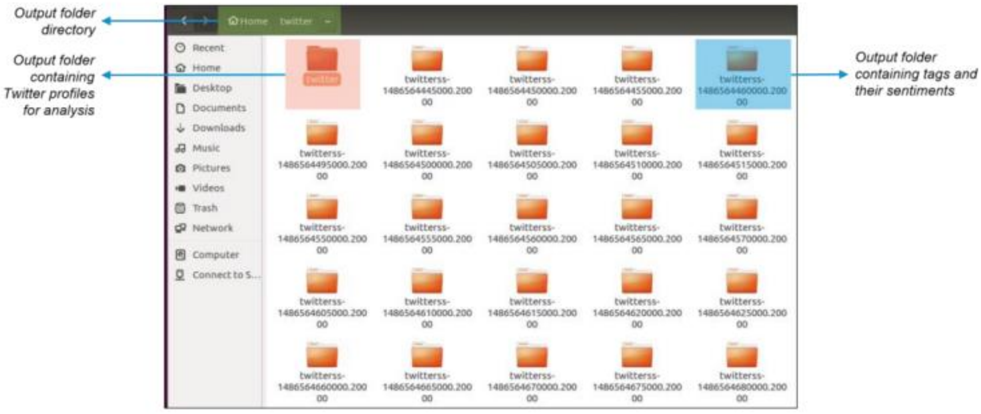
```
data.print()
//Saving our output at ~/ with filenames starting like twitters
data.saveAsTextFiles("~/twitters","20000")

ssc.start()
ssc.awaitTermination()
 }
}
```

OUTPUT:

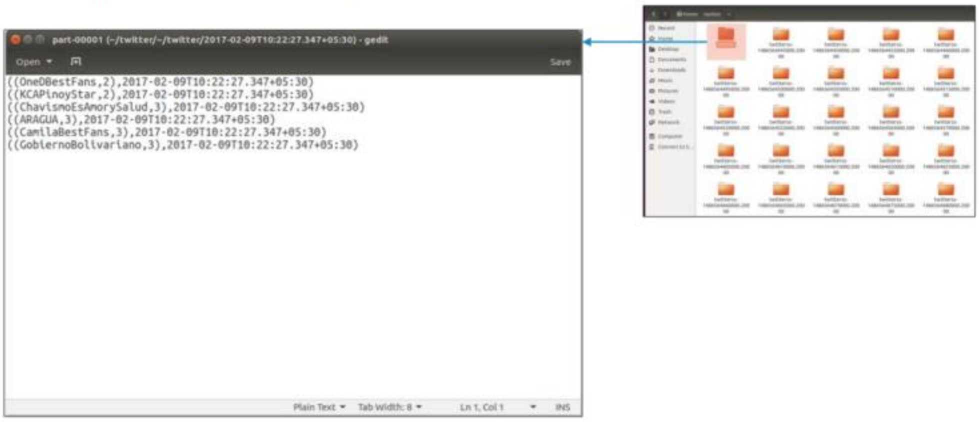


```
<terminated> mapr$ [Scala Application] /usr/lib/jvm/java-8-openjdk-i386/bin/java (09-Feb-2017, 11:56:26 AM)
debug: weighted: 1.0
-------------------------------------------
Time: 1486621640000 ms
-------------------------------------------
(東芝、半導体新棟を着工=メモリー製造、18年夏完成へ https://t.co/DU5goZAp25 #不動産 #投資 #マネー #株 #市況 #拡散,NEGATIVE,[Ljava.lang.String;@1a25ec3)
(RT @bts_bighit: [투표] Q. 투표하라는 말 지겹다?
아미: 좋아요~ 짜릿해!⚡ 늘 새로워!✨#방탄소년단 투표하는 게 최고야♥

#가온차트어워드 https://t.co/OHDWR2smt4
#ShortyAwards https://t.…,NEGATIVE,[Ljava.lang.String;@121986a)
(RT @MukePL: Jeżeli na tym zdjęciu widzisz swój świat to daj RT. ♡ #oneDbestfans & #5SOSbestfans ♡ https://t.co/rn2EmNvjFp,NEGATIVE,[Ljava.lang.String;@1c3681d)
(RT @Horocasts: #Cancer most enduring quality is an unexpected silly sense of humor.,POSITIVE,[Ljava.lang.String;@174e1a2)
(I'm listening to "A Song For Mama" by @BoyzIIMen on @PandoraMusic. #pandora https://t.co/71n5Rw3CY0,NEUTRAL,[Ljava.lang.String;@95f6d4)
('Greenwashing' Costing Walmart $1 Million https://t.co/D8XO2RZMnP #Biodegradability #Compostability #biobased,NEGATIVE,[Ljava.lang.String;@1511e25)
(RT @camilasxdinah: Serayah representando a las camilizers   cuando un hombre se le acerca a Camila #CamilaBestFans https://t.co/8IggLo3RGn,NEGATIVE,[Ljava.lang.String;@78c835)
(RT @CamiIaVoteStats: #CamilaBestFans https://t.co/qsLxPQpD1n,NEUTRAL,[Ljava.lang.String;@16e7255)
(@tos 六甲道駅 https://t.co/0rKl8rlSb3 #TFB,NEGATIVE,[Ljava.lang.String;@1a3fe)
(Ilmar pro Marcos: "Vai dormir puta.. Bebe e fica aí com o cu quente." KKKKKKKKKKKKKKKKKKKKKKKKK #BBB17,NEGATIVE,[Ljava.lang.String;@1516ece)
...

Adding annotator tokenize
```

- As we can see from the screenshot, all tweets are categorised as Positive, Neutral and Negative according to the feelings of the contents of the tweets.
- The output of the Sentiments of the Tweets is stored in folders and files, depending on the time they were created. This output can be stored on the local file system or HDFS as appropriate. The output directory looks like this one:

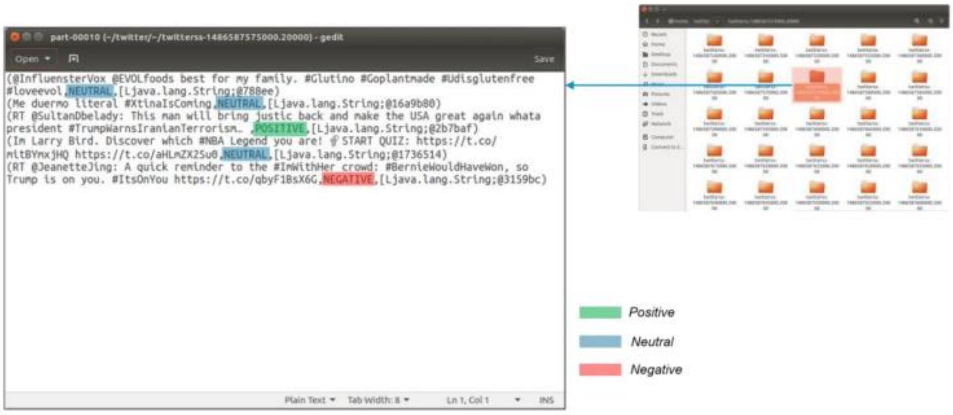

Output folder directory
Output folder containing Twitter profiles for analysis
Output folder containing tags and their sentiments
Home
twitter
Recent
Home
Desktop
Documents
Downloads
Music
Pictures
Videos
Trash
Network
Computer
Connect to S...
twitter
twitterss-1486564445000.20000
twitterss-1486564450000.20000
twitterss-1486564455000.20000
twitterss-1486564460000.20000
twitterss-1486564495000.20000
twitterss-1486564500000.20000
twitterss-1486564505000.20000
twitterss-1486564510000.20000
twitterss-1486564515000.20000
twitterss-1486564550000.20000
twitterss-1486564555000.20000
twitterss-1486564560000.20000
twitterss-1486564565000.20000
twitterss-1486564570000.20000
twitterss-1486564605000.20000
twitterss-1486564610000.20000
twitterss-1486564615000.20000
twitterss-1486564620000.20000
twitterss-1486564625000.20000
twitterss-1486564660000.20000
twitterss-1486564665000.20000
twitterss-1486564670000.20000
twitterss-1486564675000.20000
twitterss-1486564680000.20000

- Here, within the twitter folders, we will find the usernames of the Twitter users and the timestamp for. tweet as seen below:


part-00001 (~/twitter/~/twitter/2017-02-09T10:22:27.347+05:30) - gedit
Open
Save
Plain Text
Tab Width: 8
Ln 1, Col 1
INS

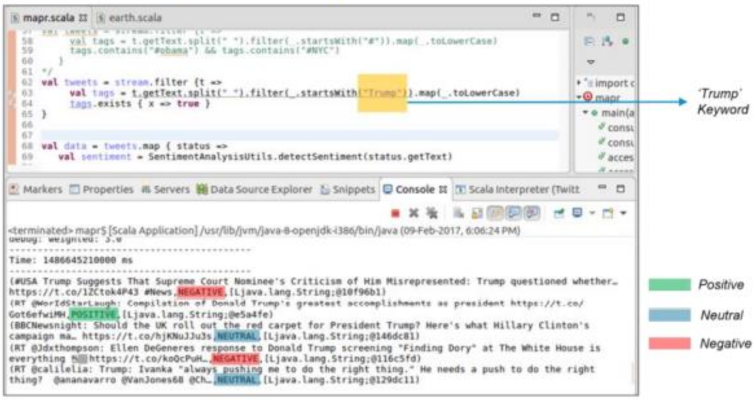
```
((OneDBestFans,2),2017-02-09T10:22:27.347+05:30)
((KCAPinoyStar,2),2017-02-09T10:22:27.347+05:30)
((ChavismoEsAmorySalud,3),2017-02-09T10:22:27.347+05:30)
((ARAGUA,3),2017-02-09T10:22:27.347+05:30)
((CamilaBestFans,3),2017-02-09T10:22:27.347+05:30)
((GobiernoBolivariano,3),2017-02-09T10:22:27.347+05:30)
```

- Now that we have Twitter usernames and timestamps, let 's look at the Sentiments and Tweets saved in the main directory. Here, every tweet is accompanied by a feeling of emotion. This Sentiment, which is processed, is further used to examine a broad variety of business observations.



```
(@InfluensterVox @EVOLfoods best for my family. #Glutino #Goplantmade #Udisglutenfree
#loveevol,NEUTRAL,[Ljava.lang.String;@788ee)
(Me duermo literal #XtinaIsComing,NEUTRAL,[Ljava.lang.String;@16a9b80)
(RT @SultanDbelady: This man will bring justic back and make the USA great again whata
president #TrumpWarnsIranianTerrorism… ,POSITIVE,[Ljava.lang.String;@2b7baf)
(Im Larry Bird. Discover which #NBA Legend you are! START QUIZ: https://t.co/
mltBYmxjHQ https://t.co/aHLmZX2Su0,NEUTRAL,[Ljava.lang.String;@1736514)
(RT @JeanetteJing: A quick reminder to the #ImWithHer crowd: #BernieWouldHaveWon, so
Trump is on you. #ItsOnYou https://t.co/qbyF1BsX6G,NEGATIVE,[Ljava.lang.String;@3159bc)
```

- Tweaking Code: Now, let's modify our code a little to get feelings about specific hashtags (topics). Donald Trump, the President of the United States, is now travelling through television outlets and online social media. Let us look at the emotions associated with the keyword 'Trump.'



```
        val tags = t.getText.split(" ").filter(_.startsWith("#")).map(_.toLowerCase)
        tags.contains("#obama") && tags.contains("#NYC")
    }
*/
val tweets = stream.filter {t =>
        val tags = t.getText.split(" ").filter(_.startsWith("Trump")).map(_.toLowerCase)
        tags.exists { x => true }
}


val data = tweets.map { status =>
    val sentiment = SentimentAnalysisUtils.detectSentiment(status.getText)
```

```
<terminated> mapr$ [Scala Application] /usr/lib/jvm/java-8-openjdk-i386/bin/java (09-Feb-2017, 6:06:24 PM)
-------------------------------------------
Time: 1486645210000 ms
-------------------------------------------
(#USA Trump Suggests That Supreme Court Nominee's Criticism of Him Misrepresented: Trump questioned whether…
https://t.co/1ZCtok4P43 #News,NEGATIVE,[Ljava.lang.String;@10f96b1)
(RT @WorldStarLaugh: Compilation of Donald Trump's greatest accomplishments as president https://t.co/
Got6efwiMH,POSITIVE,[Ljava.lang.String;@e5a4fe)
(BBCNewsnight: Should the UK roll out the red carpet for President Trump? Here's what Hillary Clinton's
campaign ma… https://t.co/hjKNuJJu3s,NEUTRAL,[Ljava.lang.String;@146dc81)
(RT @Jdxthompson: Ellen DeGeneres response to Donald Trump screening "Finding Dory" at The White House is
everything https://t.co/koQcPuH…,NEGATIVE,[Ljava.lang.String;@116c5fd)
(RT @calilelia: Trump: Ivanka "always pushing me to do the right thing." He needs a push to do the right
thing?  @ananavarro @VanJones68 @Ch…,NEUTRAL,[Ljava.lang.String;@129dc11)
```

CONCLUSION:

In this practical, we learnt about spark and installed it and configured it. We explored the spark shell as well.