Session 20

Assignment 2

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 20 – Assignment 2 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 20 – Assignment 2** |
|  |  |
| **Start Date:** | 20/10/2017 |
|  |  |
|  |  |



# 

# Contents

[Contents 2](#_Toc496383359)

[Change History 3](#_Toc496383360)

[1. Problem Statement 4](#_Toc496383361)

[2. Datasets 5](#_Toc496383362)

[3. Solution 6](#_Toc496383363)

[4. Results 7](#_Toc496383364)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 21/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

**Follow the blog and iterate the same results.**

**Sentiment analysis on demonetization**

Let us find out the views of different people on the demonetization by analysing the tweets from twitter. Here is the dataset where twitter tweets are gathered in CSV format. I downloaded the dataset from the below link.

[*https://drive.google.com/open?id=0ByJLBTmJojjzNkRsZWJiY1VGc28*](https://drive.google.com/open?id=0ByJLBTmJojjzNkRsZWJiY1VGc28)

**The code to implement the blog**

val tweets = sc.textFile("file:///home/kiran/Documents/datasets/demonetization-tweets.csv").map(x => x.split(",")).filter(x=>x.length>=2).map(x => (x(0).replaceAll("\"",""),x(1).replaceAll("\"","").toLowerCase)).map(x => (x.\_1,x.\_2.split(" "))).toDF("id","words")

tweets.registerTempTable("tweets")

val explode = spark.sql("select id as id,explode(words) as word from tweets").registerTempTable("tweet\_word")

val afinn = sc.textFile("file:///home/kiran/Documents/datasets/AFINN.txt").map(x => x.split("\t")).map(x => (x(0),x(1))).toDF("word","rating").registerTempTable("afinn")

val join = spark.sql("select t.id,AVG(a.rating) as rating from tweet\_word t join afinn a on t.word=a.word group by t.id order by rating desc").show

# Datasets

There are 2 datasets the following show an extract from the files to show the an example of the contents.

**AFINN.txt**

abandon    -2

abandoned    -2

abandons    -2

abducted    -2

abduction    -2

abductions    -2

abhor    -3

abhorred    -3

abhorrent    -3

abhors    -3

abilities    2

ability    2

aboard    1

absentee    -1

absentees    -1

.

.

youthful    2

yucky    -2

yummy    3

zealot    -2

zealots    -2

zealous    2

**demonenetization-tweets.csv**

"","text","favorited","favoriteCount","replyToSN","created","truncated","replyToSID","id","replyToUID","statusSource","screenName","retweetCount","isRetweet","retweeted"

"1","RT @rssurjewala: Critical question: Was PayTM informed about #Demonetization edict by PM? It's clearly fishy and requires full disclosure &amp;…",FALSE,0,NA,"2016-11-23 18:40:30",FALSE,NA,"801495656976318464",NA,"<a

href=""<http://twitter.com/download/android>"" rel=""nofollow"">Twitter for Android</a>","HASHTAGFARZIWAL",331,TRUE,FALSE

"2","RT @Hemant\_80: Did you vote on #Demonetization on Modi survey

app?",FALSE,0,NA,"2016-11-23

18:40:29",FALSE,NA,"801495654778413057",NA,"<a

href=""<http://twitter.com/download/android>"" rel=""nofollow"">Twitter for Android</a>","PRAMODKAUSHIK9",66,TRUE,FALSE

"3","RT @roshankar: Former FinSec, RBI Dy Governor, CBDT Chair + Harvard Professor lambaste #Demonetization.

Here is what we are taking

"State" ,"Persons","Males" ,"Females" ,"Growth\_1991\_2001" ,"Rural" ,"Urban" ,"Scheduled\_Caste\_population" ,"Percentage\_SC\_to\_total" ,"Number\_of\_households" ,"Household\_size\_per\_household" ,"Sex\_ratio\_females\_per\_1000\_males " ,"Sex\_ratio\_0\_6\_years" ,"Scheduled\_Tribe\_population" ,"Percentage\_to\_total\_population\_ST" ,"Persons\_literate" ,"Males\_Literate" ,"Females\_Literate" ,"Persons\_literacy\_rate" ,"Males\_Literatacy\_Rate" ,"Females\_Literacy\_Rate" ,"Total\_Educated"

# Solution

I have set the environment and followed the blog as follows.

**package** com.duncb.spark

**import** org.apache.spark.\_

**import** org.apache.spark.SparkContext.\_

**import** org.apache.log4j.\_

**import** org.apache.spark.sql.\_

**import** org.apache.spark.sql.functions.\_

**object** SAD {

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

// Create a SparkContext using every core of the local machine

**val** sc = **new** SparkContext("local[\*]", "deploy1")

**val** sqlContext = **new** org.apache.spark.sql.SQLContext(sc)

**import** sqlContext.implicits.\_

**val** spark = SparkSession

.builder

.appName("SparkSQL")

.master("local[1]")

.config("spark.sql.warehouse.dir", "file:///C:/temp") // Necessary to work around a Windows bug in Spark 2.0.0; omit if you're not on Windows.

.getOrCreate()

Logger.getLogger("org").setLevel(Level.ERROR)

**val** tweets = sc.textFile("file:///N:/Datasets/demonetization-tweets.csv").map(x => x.split(",")).filter(x=>x.length>=2).map(x => (x(0).replaceAll("\"",""),x(1).replaceAll("\"","").toLowerCase)).map(x => (x.\_1,x.\_2.split(" "))).toDF("id","words")

tweets.~~registerTempTable~~("tweets")

**val** explode = spark.sql("select id as id,explode(words) as word from tweets").~~registerTempTable~~("tweet\_word")

**val** afinn = sc.textFile("file:///N:/Datasets/AFINN.txt").map(x => x.split("\t")).map(x => (x(0),x(1))).toDF("word","rating").~~registerTempTable~~("afinn")

**val** join = spark.sql("select t.id,AVG(a.rating) as rating from tweet\_word t join afinn a on t.word=a.word group by t.id order by rating desc").show

println("Sentiment analysis on demonetization results")

println("--------------------------------------------------------")

}

}

# Results

+----+------+

| id|rating|

+----+------+

|5733| 4.0|

|7281| 4.0|

|6610| 4.0|

|7994| 4.0|

|6546| 4.0|

|4185| 4.0|

|3822| 4.0|

|7025| 4.0|

| 308| 3.5|

|2696| 3.0|

|1500| 3.0|

|5473| 3.0|

|2654| 3.0|

| 938| 3.0|

|6491| 3.0|

|4484| 3.0|

|5829| 3.0|

|1497| 3.0|

|3494| 3.0|

|4144| 3.0|

+----+------+

only showing top 20 rows

Sentiment analysis on demonetization results

--------------------------------------------------------