**Sentiment Analysis Using Twitter App**

*Author: Prashanth Shantigrama*

*Owner: Prashanth Shantigrama*

**Document History**

**Revision History**

|  |  |  |  |
| --- | --- | --- | --- |
| Revision Number | Revision Date | Summary of Changes | Revised By |
| 1.0 | 6/30/2015 | Document Creation | Prashanth Shantigrama |
| 1.1 | 09/29/2017 | Work Done for IBM Certification in Hadoop using FLUME, HIVE & R | Srinivas Bikki |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Table of Contents

**1.0 Flume Script** 3

**1.1** Hive Scripts 5

**1.2** R Script 6

**1.3** R Sentiment Analysis Score 7

**1.4** R Visual Representation 9

## **1.0 Flume Script**

**Step 1: Create and application in Twitter**

https://apps.twitter.com

**Step 2: Note the Consumer Key,Consumer Secret,Access Token & Access Token Secret**

**from the application created**

**Step 3: Download the jar file flume-sources-1.0-SNAPSHOT.jar and copy it into /usr/local/flume/lib**

This jar file is the source component for creating a Twitter Agent

( This is a flume configuration for connecting to Twitter Server and sinking the data into HDFS)

**Step 4. Build the Flume Agent configuration file.**

Here the Source is extracting the data from Twitter

Sink is writing data to HDFS

# describe and configure the source

TwitterAgent.sources = Twitter

TwitterAgent.channels = MemChannel

TwitterAgent.sinks = HDFS

TwitterAgent.sources.Twitter.type = com.cloudera.flume.source.TwitterSource

TwitterAgent.sources.Twitter.channels = MemChannel

TwitterAgent.sources.Twitter.consumerKey = dyxT0U8SLYXFkDJvCXSD5RHCz

TwitterAgent.sources.Twitter.consumerSecret = ErtB1AbJhsXhfJH7zYMA2z8nHsdUAjfT1Z4wz2TN6AsNwSuBqg

TwitterAgent.sources.Twitter.accessToken = 2788097081-MSrVLTejFhbHRN2AVtQ63LLiRFT5FymXkl7N4LL

TwitterAgent.sources.Twitter.accessTokenSecret = n2hpnOoc6TcmQ2EowKV924DhstuY5ZtTEQ6TPGOFSfL3U

# Twitter handles to search

TwitterAgent.sources.Twitter.keywords = hadoop, big data, analytics, bigdata, cloudera, data science, data scientiest, business intelligence, mapreduce, data warehouse, data warehousing, mahout, hbase, nosql, newsql, businessintelligence, cloudcomputing, iphone 7, iphone8, iphone x

# describe and configure the sink

TwitterAgent.sinks.HDFS.channel = MemChannel

TwitterAgent.sinks.HDFS.type = hdfs

TwitterAgent.sinks.HDFS.hdfs.path = /user/cloudera/twitter/sa1

TwitterAgent.sinks.HDFS.hdfs.fileType = DataStream

TwitterAgent.sinks.HDFS.hdfs.writeFormat = Text

TwitterAgent.sinks.HDFS.hdfs.batchSize = 1000

TwitterAgent.sinks.HDFS.hdfs.rollSize = 0

TwitterAgent.sinks.HDFS.hdfs.rollCount = 10000

TwitterAgent.sinks.HDFS.hdfs.rollInterval = 600

# describe and configure the channel

TwitterAgent.channels.MemChannel.type = memory

TwitterAgent.channels.MemChannel.capacity = 1000

TwitterAgent.channels.MemChannel.transactionCapacity = 10

**5. Run the Flume Agent to collect the data from Twitter**

flume-ng agent -n TwitterAgent -f TwitterAgent.conf

## **Hive Scripts**

add jar /home/hadoop/hive-serdes-1.0-SNAPSHOT.jar;

CREATE EXTERNAL TABLE tweets (

id BIGINT,

created\_at STRING,

source STRING,

favorited BOOLEAN,

retweeted\_status STRUCT<

text:STRING,

user:STRUCT<screen\_name:STRING,name:STRING>,

retweet\_count:INT>,

entities STRUCT<

urls:ARRAY<STRUCT<expanded\_url:STRING>>,

user\_mentions:ARRAY<STRUCT<screen\_name:STRING,name:STRING>>,

hashtags:ARRAY<STRUCT<text:STRING>>>,

text STRING,

user STRUCT<

screen\_name:STRING,

name:STRING,

friends\_count:INT,

followers\_count:INT,

statuses\_count:INT,

verified:BOOLEAN,

utc\_offset:INT,

time\_zone:STRING>,

in\_reply\_to\_screen\_name STRING

)

ROW FORMAT SERDE 'com.cloudera.hive.serde.JSONSerDe';

LOAD DATA INPATH '/user/cloudera/twitter/sa1' INTO TABLE tweets;

## **R Script**

# Reading the text into R

readLines("sentiFile.txt")

str(readLines("sentiFile.txt"))

paste(readLines("sentiFile.txt"), collapse=" ") # we get Vector length 1

# Now perform TEXT Mining by Cleaning text data in R

text <- paste(readLines("sentiFile.txt"), collapse=" ")

words <- gsub(pattern="\\W", replace=" ", text) #Removing punctuations

words <- gsub(pattern="\\d", replace=" ", words) # find digits and REPLACE with blank

words <- tolower(words) # convert text to lower case

# load package

library(tm)

words <- removeWords(words, stopwords()) # Removing stopwords

words <- gsub(pattern="\\b[A-z]\\b{1}", replace=" ", words) # clean Upper/Lower case of single length

words <- stripWhitespace(words) # clean whitespace

# load Packages

library(stringr)

library(wordcloud)

# Lets put it into bag of words

textwords <- str\_split(words, pattern="\\s+")

textwords <- unlist(textwords) # to be a character vector

class(textwords)

str(textwords)

# Next step we will check with "Positve Words" and "Negative Words"

# Reading the text into R

positivewords <- readLines("positive-words.txt")

str(positivewords)

match(textwords, positivewords) # we get NA when it didnt match with positive words

!is.na(match(textwords, positivewords))

sum(!is.na(match(textwords, positivewords))) # I want to sum this up as it will treat as binary. True will be treated as 1

negativewords <- readLines("negative-words.txt")

str(negativewords)

match(textwords, negativewords) # we get NA when it didnt match with negative words

!is.na(match(textwords, negativewords))

sum(!is.na(match(textwords, negativewords))) # I want to sum this up as it will treat as binary. True will be treated as 1

# The last step is to subtract the negative from positive

sum(!is.na(match(textwords, positivewords))) # I want to sum this up as it will treat as binary. True will be treated as 1

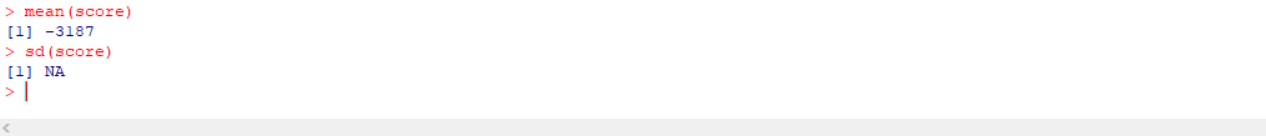
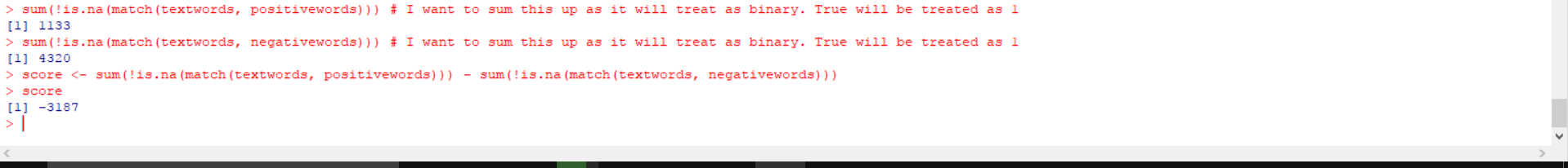
sum(!is.na(match(textwords, negativewords))) # I want to sum this up as it will treat as binary. True will be treated as 1

score <- sum(!is.na(match(textwords, positivewords))) - sum(!is.na(match(textwords, negativewords)))

score

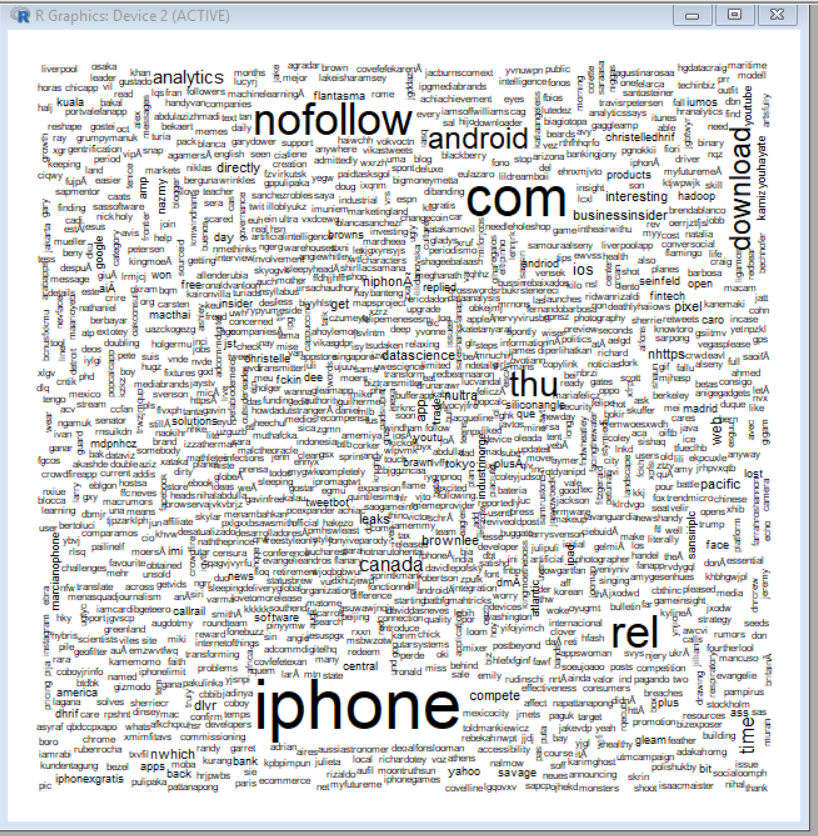
## **R Sentiment Analysis Score**

#### 



## **R Visual Representation**

wordcloud(textwords)



wordcloud(textwords, random.order=FALSE, scale=c(3,0.5), color=rainbow(3))

