Home Work 3 (Case Study 1) – Collecting, Manipulating and Blending Data from Twitter

DS501 - Introduction to Data Science

Introduction

- Go to https://dev.twitter.com/apps/new and log in, if necessary
- Enter your Application Name, Description and your website address.
- Set the callback URL http://127.0.0.1:1410
- Accept the TOS, and solve the CAPTCHA.
- Submit the form by clicking the Create your Twitter Application
- Copy the consumer key (API key) and consumer secret from the screen into your application
- Download twitter package from https://github.com/geoffjentry/twitteR

Problem 1: Sampling Twitter Data with Streaming API about a certain topic

- Select a topic that you are interested in, for example, "#WPI" or "#DataScience"
- Use Twitter Streaming API to sample a collection of tweets about this topic in real time. (It would be recommended that the number of tweets should be larger than 50, but smaller than 500.
- Store the tweets you downloaded into a local file (csv file)

```r
library(rtweet)
library(stringr)
## authenticate via web browser
#token <- create_token(
#  app = "Twitter Analysis Case Study",
#  consumer_key = consumerKey,
 # consumer_secret = consumerSecret)
#tweets = search_tweets(q = '#nytimes', n=110)
#write_as_csv(tweets, "tweets_file.csv")
tweetsDF <- read.csv("tweets_file.csv")
```

Report some statistics about the tweets you collected

- The topic of interest: #nytimes
- The total number of tweets collected: 110

Problem 2: Analyzing Tweets and Tweet Entities with Frequency Analysis

1. Word Count:

- Use the tweets you collected in Problem 1, and compute the frequencies of the words being used in these tweets.

```r
# Your R code here
tweet_words <- str_split(str_trim(str_squish(tweetsDF$text)), " ")
tweet_words_anumeric <- unlist(tweet_words)
```

```
# strip all non-alphanumeric characters
tweet_words_anumeric <- str_replace_all(tweet_words_anumeric, "[^[:alnum:]]", " ")
tweetFreqs <- table(tweet_words_anumeric)
```

- Display a table of the top 30 words with their counts

```
#  Your R code here
tweetFreqsSorted <- sort(tweetFreqs, decreasing = TRUE)
tweetFreqsFinal <- head(tweetFreqsSorted, 30)
tweetFreqsFinal <- as.data.frame(tweetFreqsFinal)
colnames(tweetFreqsFinal) <- c("Word","Count")
tweetFreqsFinal
```

```
##                 Word Count
## 1          nytimes    58
## 2              wsj    42
## 3              CBD    38
## 4        marijuana    38
## 5         business    37
## 6         cannabis    36
## 7           forbes    36
## 8          newyork    36
## 9          foxnews    35
## 10         reuters    35
## 11              ad    34
## 12       bloomberg    33
## 13             cnn    33
## 14         latimes    33
## 15             bet    30
## 16         bitcoin    24
## 17      blockchain    24
## 18          crypto    24
## 19             the    24
## 20         NYTimes    23
## 21          nasdaq    22
## 22 IHub StockPosts    21
## 23              is    21
## 24            weed    20
## 25      robbreport    19
## 26         Chicago    17
## 27              to    16
## 28             and    15
## 29              by    14
## 30             The    14
```

2. Find the most popular tweets in your collection of tweets

- Please display a table of the top 10 tweets that are the most popular among your collection, i.e., the tweets with the largest number of retweet counts.

```r
# Your R code here
# Make sure each tweet is only counted once
tweetsDfUniqTweets <- tweetsDF[!duplicated(tweetsDF[,"text"]),]
# sort by retweet count
retweets <- head(sort(tweetsDfUniqTweets$retweet_count, decreasing = TRUE), 10)
retweets <- as.data.frame(retweets)
retweets
```

```
##    retweets
## 1        18
## 2         3
## 3         2
## 4         2
## 5         2
## 6         2
## 7         1
## 8         1
## 9         1
## 10        1
```

3. Find the most popular Tweet Entities in your collection of tweets

Please display a table of the top 10 hashtags, top 10 user mentions that are the most popular in your collection of tweets.

```r
# Your R code here
ul_tweet_words <- unlist(tweet_words)
ul_hash_tags <- ul_tweet_words[startsWith(ul_tweet_words, "#")]
ul_user_ments <- ul_tweet_words[startsWith(ul_tweet_words, "@")]
hash_tag_tbl <- table(ul_hash_tags)
user_ment_tbl <- table(ul_user_ments)
# sort the hash tags and user mentions in decreasing order
hash_tag_tbl <- sort(hash_tag_tbl, decreasing = TRUE)
user_ment_tbl <- sort(user_ment_tbl, decreasing = TRUE)
# only report the top 10 of each
top_hash_tags <- as.data.frame(head(hash_tag_tbl, 10))
top_user_ments <- as.data.frame(head(user_ment_tbl, 10))
tweet_entities <- cbind(top_hash_tags, top_user_ments)
colnames(tweet_entities) <- c("Hash Tag", "Frequency", "User", "Frequency")
tweet_entities
```

```
##      Hash Tag Frequency             User Frequency
## 1    #nytimes        56         @nytimes         2
## 2        #wsj        42 @realDonaldTrump         2
## 3        #CBD        38    @belcherjody1         1
## 4  #marijuana        38    @CarlTrump007         1
## 5   #business        37     @ChuckRossDC         1
## 6   #cannabis        36      @Colt_Coeur         1
## 7     #forbes        36      @fireboydml         1
## 8    #newyork        36      @FlashTweet         1
## 9    #foxnews        35      @JSpector23         1
## 10   #reuters        35      @kallywally         1
```

Problem 3: Getting "All" friends and "All" followers of a popular user in twitter

- Choose a popular twitter user who has many followers, such as @hadleywickham.
- Get the list of all friends and all followers of the twitter user.
- Display 20 out of the followers, Display their ID numbers and screen names in a table.
- Display 20 out of the friends (if the user has more than 20 friends), Display their ID numbers and screen names in a table.
- Compute the mutual friends within the two groups, i.e., the users who are in both friend list and follower list, Display their ID numbers and screen names in a table

```r
# 20 friends section
friends_of_user <- get_friends("@rihanna")
friends_of_user_orig <- friends_of_user
# only care to compute on 20 friends
friends_of_user <- friends_of_user[1:20,]
# Get enough info about the users to get their screen names
friends_info <- lookup_users(friends_of_user$user_id)
# remove the user column because only the friend user ids and
# friend screen names matter
friends_of_user <- subset(friends_of_user, select = c(-user))
friends_of_user <-
  cbind(friends_of_user, friends_info$screen_name)
colnames(friends_of_user) <- c("friend_id","friend_screen_name")
print(friends_of_user)
```

```
##              friend_id friend_screen_name
## 1  1113788945018519552      FentyOfficial
## 2  1013623332359557121          fentyfest
## 3            558606074               LVMH
## 4   927594927298551808       SavageXFenty
## 5             90034386      loneamorphous
## 6             32464360             Oppong
## 7   732027915894824960     alissa_ashleyy
## 8   884671087417544704          Acondria2
## 9           1898785885     ClaraLionelFdn
## 10          3306418615         FENTYXPUMA
## 11           250176361        stunningrih
## 12          1976143068     EmmanuelMacron
## 13  704881118000971777        fentybeauty
## 14            59500818         ItsMeBriaJ
## 15            78525538            IssaRae
## 16            35615827        indiachanel_
## 17           602993818     MsSarahPaulson
## 18  804922493681205249       talkthatcunt
## 19           596893898           GlblCtzn
## 20            20708202         Hughcevans
```

```r
# 20 followers section
followers_of_user <- get_followers("@rihanna")
followers_of_user_orig <- followers_of_user
followers_of_user <- followers_of_user[1:20,]
followers_info <- lookup_users(followers_of_user$user_id)
followers_of_user <- cbind(followers_of_user, followers_info$screen_name)
```

```r
colnames(followers_of_user) <- c("follower_id","follower_screen_name")
followers_of_user
```

```
##            follower_id follower_screen_name
## 1  1173009638633222146      Jessica67429600
## 2  1173011388383604737          Armandoper4
## 3  1173011000100106240         Gaby16636169
## 4  1173009937175371776         King53426173
## 5  1087380498119344129               AnijiF
## 6  1173009307945885696        tanya_rebecca
## 7  1173010632070836224            SamiHk12
## 8  1173008905754091520          LakeaHunter
## 9  1173009484685426690            pazonotes
## 10 1173010692166868992      daniela82225750
## 11 1173009692777492481      megabit56522157
## 12  935078124588593152      angiee_angiie27
## 13 1173009899158028289               OrajKy
## 14          2647392447          _Ogochenour
## 15 1173010661728804864           marionn972
## 16 1018971387321638912         outoftownmom
## 17 1173009190010466304             awichika
## 18           118579875            Rigojra10
## 19 1173008910447448064            EvanMary6
## 20 1173008480254517250       Unjour76351764
```

```r
# Mutual friends/followers section
common_users <- intersect(followers_of_user_orig$user_id, friends_of_user_orig$user_id)
if(length(common_users)) {
  common_users_info <- lookup_users(common_users)
  common_users <-
    cbind(common_users, common_users_info$screen_name)
  colnames(common_users) <-
    c("mutual_user_id","mutual_screenname")
} else {
  print("There are no mutual user IDs in the friends and followers lists for @rihanna")
}
```

```
## [1] "There are no mutual user IDs in the friends and followers lists for @rihanna"
```

Problem 4 (Optional): Explore the data

Run some additional experiments with your data to gain familiarity with the twitter data and twitter API

Done

All set!

What do you need to submit?

Report: please prepare a report (less than 10 pages) to report what you found in the data.

- What data you collected? In this data analysis, there were statistics collected about tweets related to the New York Times news organization. There was also data collected about friends and followers of the music artist Rihanna. There is a table showing the top 30 words found in tweets related to the New York Times and there is another table showing the most popular tweets related to the New York Times. For Rihanna, there are tables showing 20 friends and followers as well as their user IDs and screen names.

- Why this topic is interesting or important to you? (Motivations) The New York Times was selected as a topic for analysis because it contains well written articles about a variety of topics and it is effective about keeping its readers informed about key current events. Rihanna was selected as the subject of analysis for her friends and followers because she is a talented musician who has plenty of friends and followers on Twitter.

- How did you analyze the data? The rtweet developer API and its associated functions were used to collect the data from Twitter related to the New York Times and Rihanna. For the analysis, the main approach used was to filter the data obtained from the API to obtain the desired value categories. Then for organizing the values into the appropriate tables, a combination of data frame functions and functions related to the base R table type were used.

- What did you find in the data? (please include figures or tables in the report) For the analysis connecting to the New York Times twitter topic, almost all of the most frequent words in related tweets were related either to other news organizations such as the Wall Street or CNN, or they were words commonly discussed in news articles (e.g. POTUS/President of the United States). This is not surprising since the New York Times is an organization devoted to news reporting. A similar trend was found for the most popular Tweet entities.

Please create an R Markdown PDF including the R code in a report format.

How to submit: - Submit on Course Webpage on Canvas and/or - Send an email to ndingari@wpi.edu with the subject: “DS501 Case study 1”.