MONASH University

FIT5145 Assignment 3

Reynaldo Bonita, Jr.

29999804

Part A: Investigating the Twitter Data in the Shell

1) Decompress the file. How big is it?

The file is 2.1GB or `2271087104 bytes.`

```
gunzip Twitter_Data_1.gz
ls -lh
```

2) a) What delimiter is used to separate the columns in the file.

The delimiter used is `tab.` I found out by searching for the tab character:

```
head Twitter_Data_1 | less
/<hit tab>
```

b) and how many columns are there?

There are 4 columns.

3) The first column is a unique identifier for a Tweet. What are the other columns?

The next columns are the Twitter Username, Tweet Date and the Tweet Text.

4) How many Tweets are there in the file?

There are 15,089,920 tweets.

```
wc -l Twitter_Data_1
15,089,920 Twitter_Data_1
```

5) What is the date range for Tweets in this file?

The date range is from Tue Feb 11 12:18:36 +0000 2014 to Tue Feb 18 23:15:00 +0000 2014.

```
head Twitter_Data_1
Tue Feb 11 12:18:36 +0000 2014

tail Twitter_Data_1
Tue Feb 18 23:15:00 +0000 2014

# I validated using:
cut -f 2 Twitter_Data_1 | sort | uniq -c
```

6) How many unique users are there?

There are 8,977,904 unique users.

```
cut -f 2 Twitter_Data_1 | sort | uniq | wc -l
```

7) When was the first mention in the file of “Donald Trump” and what was the tweet?

```
cat Twitter_Data_1 | grep "Donald Trump" | less
```

```
433215995134476289      Maddog4U_1st    Tue Feb 11 12:28:36
+0000 2014  RT @aedan_smith: Be interesting to see the detail
on this one:  BBC News - Donald Trump loses offshore wind farm
challenge http://t.co/qAcG…
```

8) How many times has he been mentioned in the file? How did you find this?

There were 130 tweets that mentioned “Donald Trump” based on a non-case sensitive search.

```
grep "Donald Trump" -i -o Twitter_Data_1 | wc -l
130
```

There were 116 tweets that mentioned “Donald Trump” based on case sensitive search.

```
grep "Donald Trump" -o Twitter_Data_1 | wc -l
116
```

9) What about “Hillary Clinton”? Who is a more popular on Twitter, Donald or Hillary?

There were 127 tweets that mentioned “Hillary Clinton” based on a non-case sensitive search.

```
grep "Hillary Clinton" -i -o Twitter_Data_1 | wc -l
127
```

There were 120 tweets that mentioned “Hillary Clinton” based on a case sensitive search.

```
grep "Hillary Clinton" -o Twitter_Data_1 | wc -l
120
```

If the basis are the case sensitive search results, Hillary is more popular. If we consider non-case sensitive search results, Donald is more popular.

10) Do you think we have captured all the references to Donald and Hillary? What other strings might we need to try? What problems might we face?

They are often referred to using their first name or last names only. However, if we use this, it might show tweets about “Melania Trump”. Also, not all tweets mention “Trump” with a capital T. However, if we choose to search it lowercase, we will encounter the of “trump” as a verb too. If we use his first name, “Donald”, it will return other popular collocations like “Donald Duck”. Similarly, Hillary’s name is common among celebrities too and it may yield tweets talking about other Hillarys. Searching for mere “Clinton” can return tweets about his husband too.

Part B: Graphing the Data in R

1) How many times does the term 'Obama' appear in tweets?

"Barack Obama" appears 482 times in the tweets based on a non-case sensitive search.

```
cut -f 4 Twitter_Data_1 | grep -o -i "Barack Obama" | wc -l
482
```

But doing a case sensitive search on "Barack Obama" yields 460 results only.

```
cut -f 4 Twitter_Data_1 | grep -o "Barack Obama" | wc -l
460
```

Moreover, doing a non-case sensitive search for "Obama" returns 12,849 results which may include "Michelle Obama", "Obamacare" etc.

```
cut -f 4 Twitter_Data_1 | grep -i -o "Obama" | wc -l
12,849
```

Doing a case sensitive search for "Obama" returns 11,736 results only.

```
cut -f 4 Twitter_Data_1 | grep -o "Obama" | wc -l
11,736
```

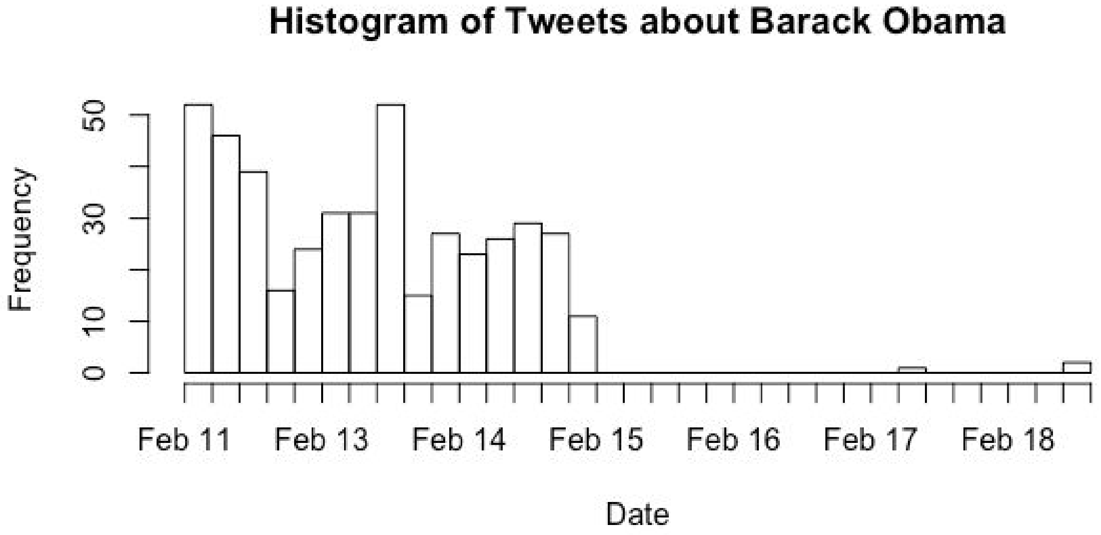
2) Background: We want to consider how the amount of discussion regarding Barack Obama varies over the time period covered by the data file.
Question: You will need to write a format string, starting with "%a %b" to tell the function how to parse the particular date/time format in your file. What format string do you need to use?

`"%a %b %d %H:%M:%S %z %Y"` is the format string I'll use.

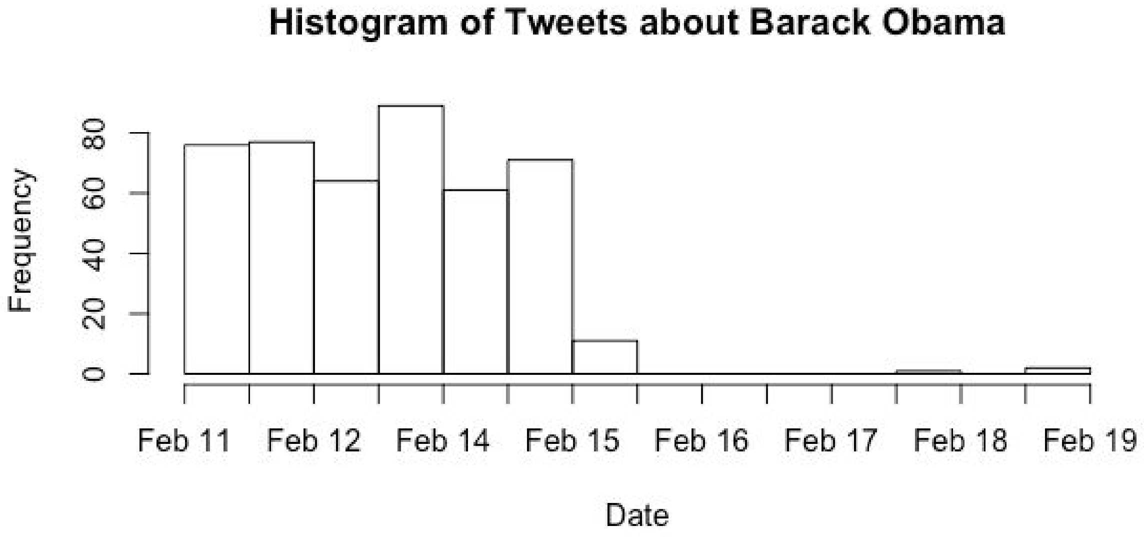
```
# Using Bash, his extracts the time stamps of the tweets that
refer to Obama
cat Twitter_Data_1 | grep "Barack Obama" | cut -f 3 >
barackobama_timestamps.txt

# In R Studio, this is my local working directory.
setwd("/Users/developer/Documents/Monash S2 2018/FIT5145
Intro to DS/Assessment 3")

# Using R, this loads the CSV into a list.
obama <- read.csv("barackobama_timestamps.txt", header =
FALSE)

# This formats the string timestamps into R datetime objects
obama <- strptime(obama[[1]][1:452], "%a %b %d %H:%M:%S %z
%Y")
```

3) Once you’ve converted the timestamps, use the hist() function to plot the data.

```
     hist(obama,breaks=40, xlab="Date", main = "Histogram of
Tweets about Barack Obama", freq=TRUE)
```



```
     # I tried a smaller bin
     hist(obama, breaks=10, xlab="Date", main = "Histogram of
Tweets about Barack Obama", freq=TRUE)
```

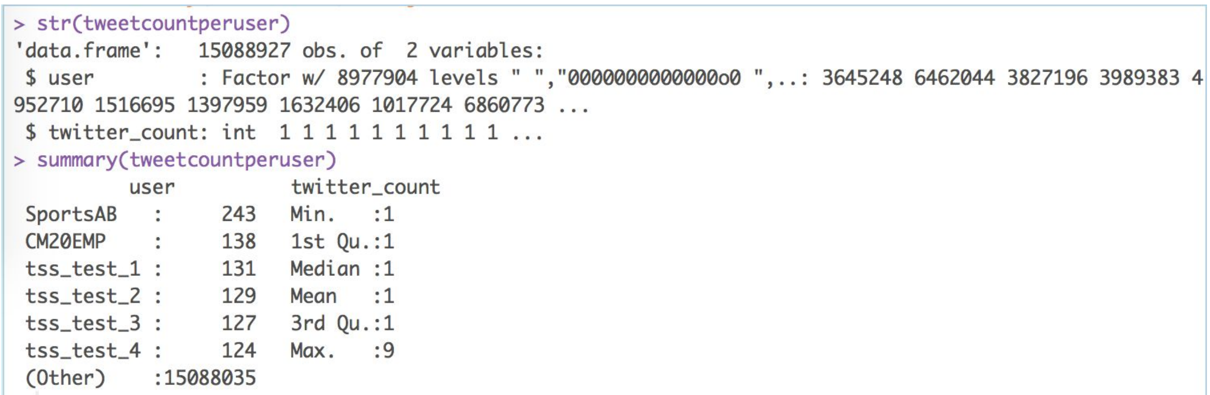


4) The plot has a bit of an unusual shape. Can you see a pattern before Feb 15 and what happens after that?

Tweets about Obama ranges from about 60 to 80 per day from Feb 11 to 15 of 2014. There were no tweets mentioning “Barack Obama” after the 15th apart from one on the 18th and another one on the 19th.

5) (Challenge) Plot a second histogram, but this time showing the distribution over number of tweets per author in the file.

```
        # Getting the count of tweets per user
        cut -f 2 Twitter_Data_1 | uniq -c > tweet_count_per_user.txt
        head tweet_count_per_user.txt

        # Reformatting as a CSV
        cat  tweet_count_per_user.txt | awk '{print $2,","$1}' >
        reformatted_tweet_count_per_user_v1-3.txt

        # Adding a header
        echo "user,twitter_count" | cat -
        reformatted_tweet_count_per_user_v1-3.txt >
        reformatted_tweet_count_per_user_v2-0.txt

        tweetcountperuser <-
read.csv("reformatted_tweet_count_per_user_v2-0.txt",header = TRUE)
        str(tweetcountperuser)

        # Most of the users only tweeted onceabout Obama.
        summary(tweetcountperuser)
```

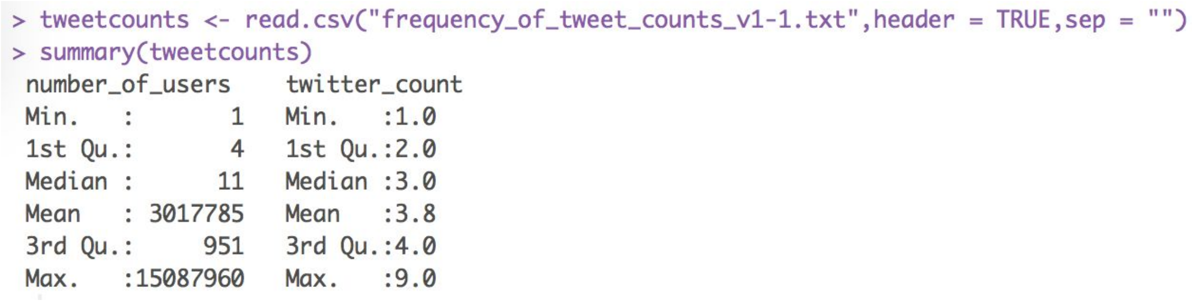


```
> str(tweetcountperuser)
'data.frame':   15088927 obs. of  2 variables:
 $ user         : Factor w/ 8977904 levels " ","0000000000000o0 ",..: 3645248 6462044 3827196 3989383 4
952710 1516695 1397959 1632406 1017724 6860773 ...
 $ twitter_count: int  1 1 1 1 1 1 1 1 1 1 ...
> summary(tweetcountperuser)
          user           twitter_count
 SportsAB   :     243   Min.   :1
 CM20EMP    :     138   1st Qu.:1
 tss_test_1 :     131   Median :1
 tss_test_2 :     129   Mean   :1
 tss_test_3 :     127   3rd Qu.:1
 tss_test_4 :     124   Max.   :9
 (Other)    :15088035
```

Then load them into R. This is a large file so you can also just isolate the counts, sort and count them to get a summary statistics file with columns “twitter count” and “number of users”.]

```
        # Reforming the data set so that we get the frequency of the
        tweet counts, ie. how many tweeted once, 2x, etc.

        awk '{print $1}' tweet_count_per_user.txt | uniq -c >
        frequency_of_tweet_counts.txt

        # adding a header
```

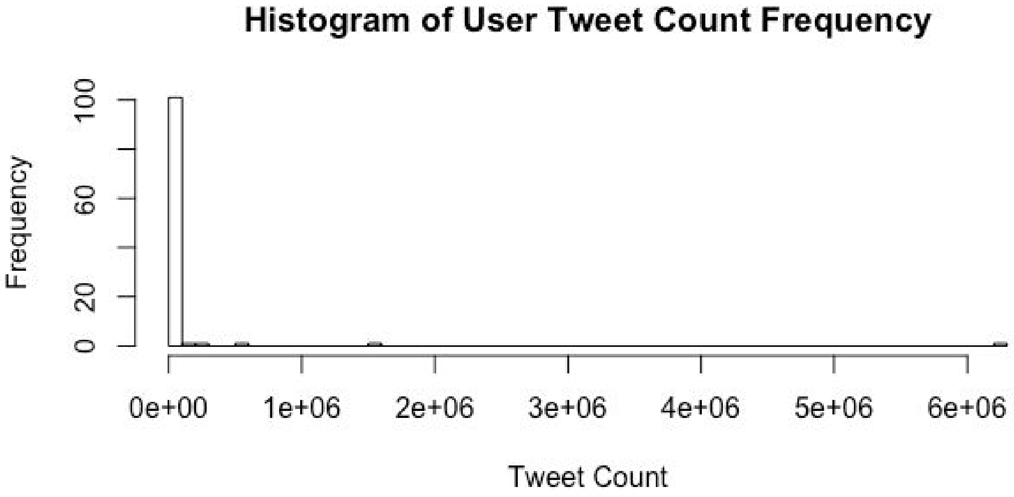
```
echo "number_of_users twitter_count" | cat -
frequency_of_tweet_counts.txt >
frequency_of_tweet_counts_v1-1.txt

tweetcounts <-
read.csv("frequency_of_tweet_counts_v1-1.txt",header =
TRUE,sep = "")

summary(tweetcounts)
```



```
> tweetcounts <- read.csv("frequency_of_tweet_counts_v1-1.txt",header = TRUE,sep = "")
> summary(tweetcounts)
 number_of_users     twitter_count 
 Min.   :       1   Min.   :1.0  
 1st Qu.:       4   1st Qu.:2.0  
 Median :      11   Median :3.0  
 Mean   : 3017785   Mean   :3.8  
 3rd Qu.:     951   3rd Qu.:4.0  
 Max.   :15087960   Max.   :9.0  
```

```
# Checking the structure of the parsed data
str(tweetcounts)

# Sorting the number of users
sort(tweetcounts$number_of_users)

# Sorting the values of the list
tweetcounts[order(sapply(tweetcounts, function(x) x[1],
simplify=TRUE), decreasing=TRUE)]

# It seems like the data set has several outliers. There are
6,260,301 users who only tweeted about Obama once. Those who
tweeted twice and thrice are 1,504,504 and 558,525
respectively while most of the other rows are less than 150.
When the histogram is graphed, the rest of the values will
not be seen.
```

```
hist(tweetcounts$number_of_users,breaks=50,xlab="Tweet
Count", main = "Histogram of User Tweet Count Frequency")
```



Part C: Investigating User Check-in Data in the Shell

1) Open the zip file and have a look at the files it contains. One is a readme file giving the metadata. One is a log of user check-ins. How many check-ins are there and how many users?

There are 33,263,633 lines in the file.

```
wc -l dataset_TIST2015_Checkins.txt
33,263,633
```

However, the ReadMe file says that "It contains 33,278,683 checkins…".

2) Background: How would you select venues from Europe? Question: Create an awk script to create a European subset of the POI file, and name the subset file "POIeu.txt". Investigate your European subset.

A. Submit the created `POIeu.txt` along with your PDF file.

I decided to create the `POIeu.txt` using the Wikipedia article enumerating the Member States of the European Union (https://en.wikipedia.org/wiki/Member_state_of_the_European_Union).

The country codes are:

```
"BE" "BG" "CZ" "DK" "DE" "EE" "IE" "EL" "ES" "FR" "HR" "IT"
"CY"  "LV" "LT" "LU" "HU" "MT" "NL" "AT" "PL" "PT" "RO" "SI"
"SK" "FI" "SE" "UK".
```

I filtered `dataset_TIST2015_POIs.txt` per country and outputted the results in a file. I, then, concatenated these text files to create `POIeu.txt`.

```
# Done for all member states. I am not including the
individual commands here because it might be too long. I'll
just show one:

awk -F\\t < dataset_TIST2015_POIs.txt '{if ($5 == "BE") print
$0}' > "BE.txt"

cat "BE.txt" "BG.txt" "CZ.txt" "DK.txt" "DE.txt" "EE.txt"
"IE.txt" "EL.txt" "ES.txt" "FR.txt" "HR.txt" "IT.txt"
"CY.txt" "LV.txt" "LT.txt" "LU.txt" "HU.txt" "MT.txt"
"NL.txt" "AT.txt" "PL.txt" "PT.txt" "RO.txt" "SI.txt"
"SK.txt" "FI.txt" "SE.txt" "UK.txt" > POIeu.txt
```

B. What country has the most venues and what the least, with how many?

Spain (ES) has the most venues with 39,187 records while Estonia (EE) has the least (2,170 records).
The following European countries don’t have records at all: El Salvador (EL), Croatia (HR), Lithuania (LT), Luxembourg (LU), Malta (MT), Slovenia (SI), Slovakia (SK), and United Kingdom (UK).

```
cut -f 5 POIeu.txt | uniq -c | sort -g

2170 EE
2411 BG
2735 DK
3651 PL
3858 RO
3968 IE
5636 AT
5651 FI
5707 CZ
6389 SE
6804 CY
7924 LV
8681 HU
8721 PT
19837 FR
34332 IT
34713 DE
36826 BE
38536 NL
39187 ES
```

C. Who has the most Indian restaurants?

Germany (DE) has the most Indian restaurants with 151 records while Romania has the least with 3 records.

```
awk -F\\t < POIeu.txt '{if ($4 == "Indian Restaurant") print
$5}' | uniq -c | sort -g

   3 RO
   5 LV
   6 EE
   6 PL
   7 CY
   9 DK
  12 HU
  14 CZ
  14 IE
  15 AT
  28 FI
```

```
 31 PT
 34 BE
 34 NL
 52 SE
 56 IT
 65 ES
 65 FR
151 DE
```

D. What is the most common (as in, how many venues) class of restaurant in Europe?

Among the 53 categories with the word "Restaurant", the category "Restaurant" is the most common in the listing with 5,863. Among the specialized restaurant records Italian Restaurant has the most records with 5,334 entries. The least is "Filipino Restaurant" with 5 venues.

```
awk -F\\t < POIeu.txt '{if ($4 ~ "Restaurant") print $0}' |
cut -f 4  | sort | uniq -c | sort -g

    5 Filipino Restaurant
    7 Afghan Restaurant
    7 Mongolian Restaurant
   10 Southern / Soul Food Restaurant
   13 Malaysian Restaurant
   18 Australian Restaurant
   19 Peruvian Restaurant
   20 Swiss Restaurant
   21 Dumpling Restaurant
   21 Ethiopian Restaurant
   21 Gluten-free Restaurant
   22 New American Restaurant
   24 Arepa Restaurant
   26 Cajun / Creole Restaurant
   26 Cuban Restaurant
   29 Dim Sum Restaurant
   33 Caribbean Restaurant
   52 Indonesian Restaurant
   52 Latin American Restaurant
   56 Molecular Gastronomy Restaurant
   70 South American Restaurant
   98 Moroccan Restaurant
  106 African Restaurant
  107 Korean Restaurant
  112 Brazilian Restaurant
  119 Paella Restaurant
  215 Argentinian Restaurant
```

```
 237 Vietnamese Restaurant
 264 Portuguese Restaurant
 290 Scandinavian Restaurant
 343 Vegetarian / Vegan Restaurant
 397 Turkish Restaurant
 453 Mexican Restaurant
 485 Falafel Restaurant
 485 Thai Restaurant
 493 American Restaurant
 506 Middle Eastern Restaurant
 522 Eastern European Restaurant
 533 Greek Restaurant
 607 Indian Restaurant
 689 Seafood Restaurant
 933 German Restaurant
 949 Sushi Restaurant
 981 Japanese Restaurant
1258 Mediterranean Restaurant
1330 Tapas Restaurant
1357 Chinese Restaurant
1444 Asian Restaurant
1819 Spanish Restaurant
2352 French Restaurant
3126 Fast Food Restaurant
5334 Italian Restaurant
5863 Restaurant
```