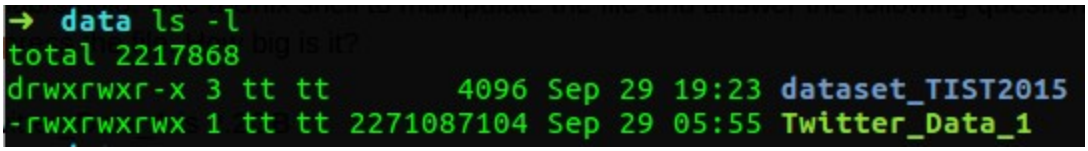
FIT5145 Assignment – 3

Jaimon Thyparambil Thomas
Student ID : 29566428
Email : jthy0001@monash.student.edu
Monash University
October 02, 2018

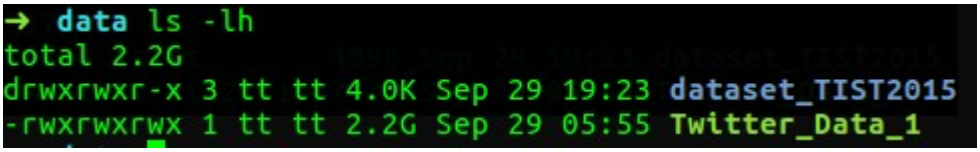
Part A: Investigating the Twitter Data in the Shell

Download the file Twitter_Data_1.gz from the link above. This is 1Gb file so do this at a Monash computer lab/studio. Use a Unix shell to manipulate the file and answer the following questions.
1)Decompress the file. How big is it?

Size after decompressing

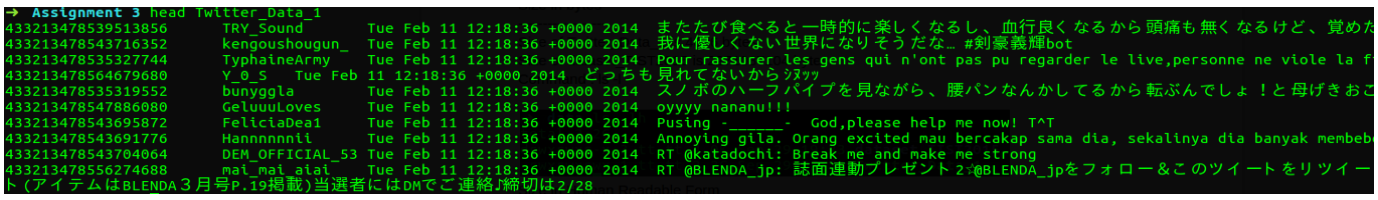


Size in bytes
Size of Twitter_Data_1 is **4096 bytes**
Size of dataset_TIST2015 is **2271087104 bytes**
Command: ls -l



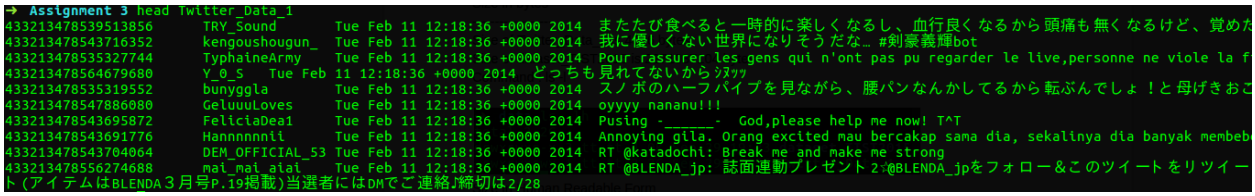
Size in Human Readable Form
Size of Twitter_Data_1 is **2.2GB**
Size of dataset_TIST2015 is **4.0KB**
Command: ls -lh

2)What delimiter is used to separate the columns in the file and how many columns are there?



'\t' is the delimiter used to separate the columns in the file.
There are **4** columns.

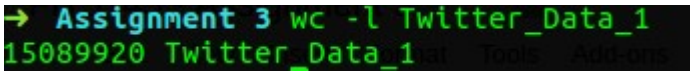
3)The first column is a unique identifier for a Tweet. What are the other Columns?



The values that each column represent is been defined below.
Column 1: Unique identifier for a Tweet
Column 2: Twitter Username or handle
Column 3: Time at which the tweet was posted.
(in format "%a %b %d %H:%M:%S %z %Y")
Column 4: Tweet Content

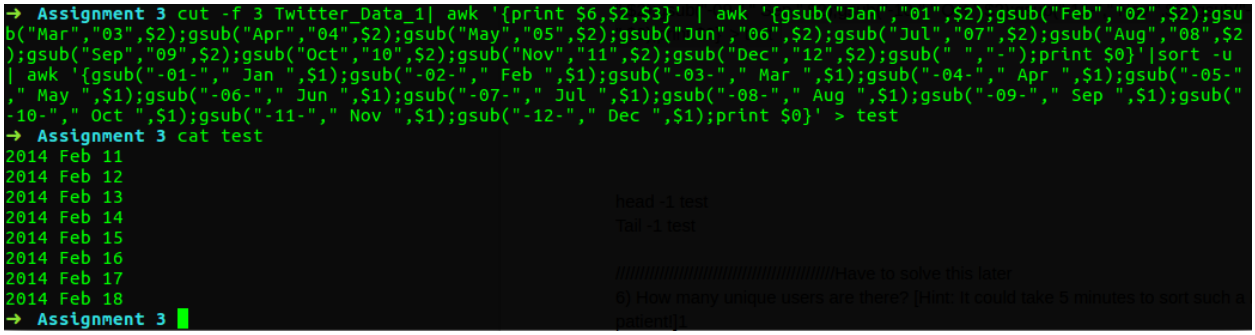
Where each symbol in UTC Time format is been explained below
%a - Abbreviated weekday name in the current locale on this platform.
%b - Abbreviated month name in the current locale on this platform
%d - Day of the month as decimal number (01–31).
%H - Hours as decimal number (00–23)
%M - Minute as decimal number (00–59).
%S - Second as integer (00–61), allowing for up to two leap-seconds
%z - Signed offset in hours and minutes from UTC, so -0800 is 8 hours behind UTC. Values up to +1400 are accepted
%Y - Year with century.

4) How many Tweets are there in the file?



15089920 tweets are there in the file.
Command: wc -l Twitter_Data_1

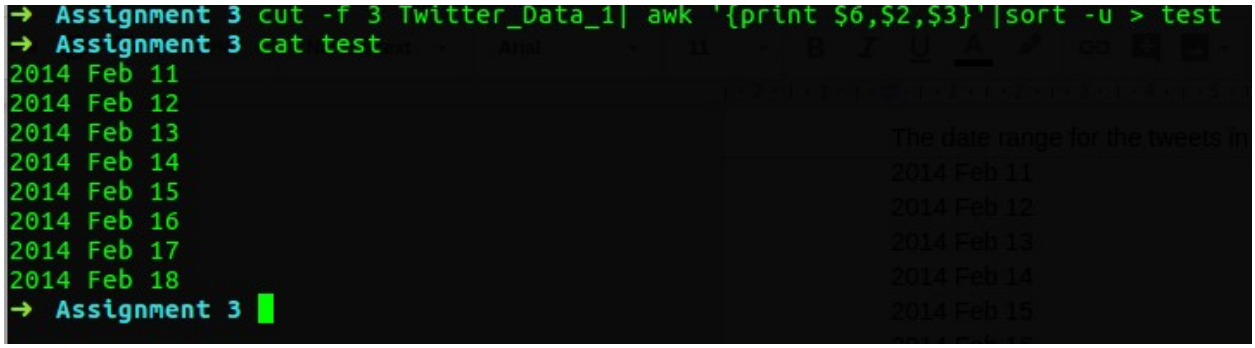
5) What is the date range for Tweets in this file?



The date range for the tweets in this file is
2014 Feb 11
2014 Feb 12
2014 Feb 13
2014 Feb 14
2014 Feb 15
2014 Feb 16
2014 Feb 17
2014 Feb 18

That is **from 2014-Feb-11 till 2014-Feb-18**

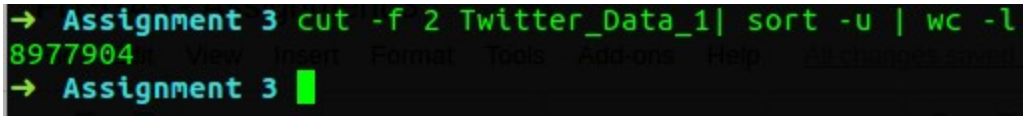
Command:
cut -f 3 Twitter_Data_1| awk '{print $6,$2,$3}' | awk '{gsub("Jan","01",$2);gsub("Feb","02",$2);gsub("Mar","03",$2);gsub("Apr","04",$2);gsub("May","05",$2);gsub("Jun","06",$2);gsub("Jul","07",$2);gsub("Aug","08",$2);gsub("Sep","09",$2);gsub("Oct","10",$2);gsub("Nov","11",$2);gsub("Dec","12",$2);gsub(" ","-");print $0}'|sort -u | awk '{gsub("-01-"," Jan ",$1);gsub("-02-"," Feb ",$1);gsub("-03-"," Mar ",$1);gsub("-04-"," Apr ",$1);gsub("-05-"," May ",$1);gsub("-06-"," Jun ",$1);gsub("-07-"," Jul ",$1);gsub("-08-"," Aug ",$1);gsub("-09-"," Sep ",$1);gsub("-10-"," Oct ",$1);gsub("-11-"," Nov ",$1);gsub("-12-"," Dec ",$1);print $0}' > test



Alternate command as only Feb is present as month:
cut -f 3 Twitter_Data_1| awk '{print $6,$2,$3}'|sort -u > test

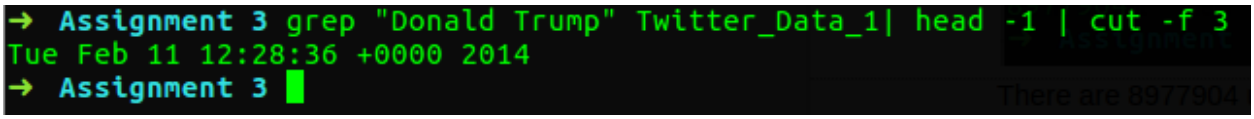
Entire date range available can be viewed using the command: ***cat test***
Starting date can be obtained by command: ***head -1 test***
Ending date can be obtained by command: ***tail -1 test***

6) How many unique users are there? [Hint: It could take 5 minutes to sort such a big list, so be patient!]

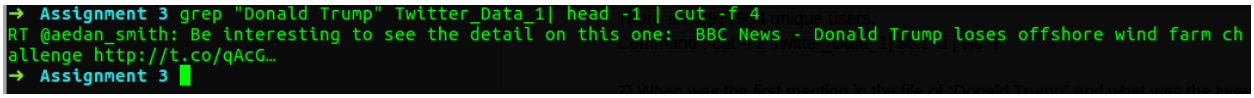


There are **8977904** unique users.
Command: cut -f 2 Twitter_Data_1| sort -u | wc -l

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



First mention of “Donald Trump” was on **11 Feb 2014 at 12:28:36**
Time: Tue Feb 11 12:28:36 +0000 2014
Command: grep "Donald Trump" Twitter_Data_1| head -1 | cut -f 3



The tweet in which “Donald Trump” was first mentioned is:
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…
Command: grep "Donald Trump" Twitter_Data_1| head -1 | cut -f 4

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

If we are searching just for the name “Donald Trump”. Then

When searched respective of case in the name we found that his name was present in **109 tweets** and total no of occurrences were **116**

```
→ Assignment 3 grep "Donald Trump" Twitter_Data_1| wc -l
109
```

```
→ Assignment 3 grep -o "Donald Trump" Twitter_Data_1| wc -l
116
→ Assignment 3
```

Command: ***grep "Donald Trump" Twitter_Data_1 | wc -l***
grep -o "Donald Trump" Twitter_Data_1| wc -l

When searched irrespective of case we found that his name was present in **122 tweets** and total no of occurrences were **130**

```
→ Assignment 3 grep -i "Donald Trump" Twitter_Data_1| wc -l
122
→ Assignment 3
```

```
→ Assignment 3 grep -io "Donald Trump" Twitter_Data_1| wc -l
130
→ Assignment 3
```

Command: ***grep -i "Donald Trump" Twitter_Data_1 | wc -l***
grep -io "Donald Trump" Twitter_Data_1| wc -l

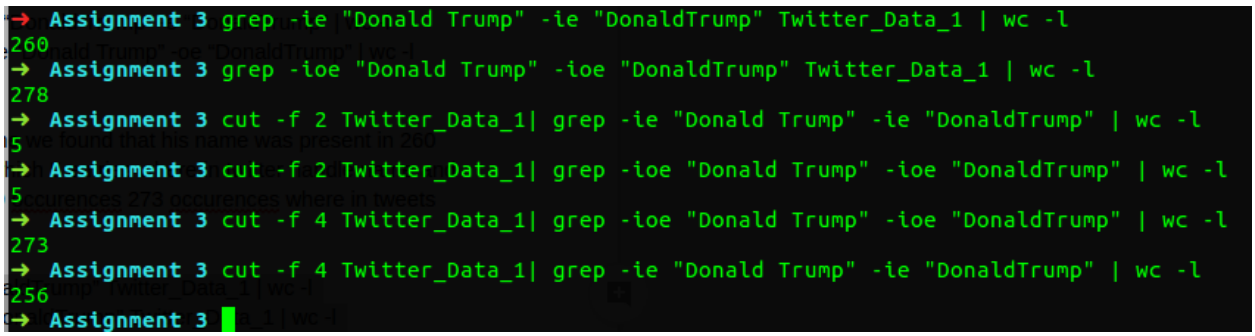
If we were searching for values "Donald Trump" and "DonaldTrump". Then

```
→ Assignment 3 grep -e "Donald Trump" -e "DonaldTrump" Twitter_Data_1 | wc -l
243
→ Assignment 3 grep -oe "Donald Trump" -oe "DonaldTrump" Twitter_Data_1 | wc -l
259
→ Assignment 3 cut -f 2 Twitter_Data_1| grep -e "Donald Trump" -e "DonaldTrump" | wc -l
5
→ Assignment 3 cut -f 2 Twitter_Data_1| grep -oe "Donald Trump" -oe "DonaldTrump" | wc -l
5
→ Assignment 3 cut -f 4 Twitter_Data_1| grep -e "Donald Trump" -e "DonaldTrump" | wc -l
239
→ Assignment 3 cut -f 4 Twitter_Data_1| grep -oe "Donald Trump" -oe "DonaldTrump" | wc -l
254
→ Assignment 3
```

When searched respective of case in the name we found that his name was present in **243** entries and a total of **259** occurrences. Out of which 5 entries where in twitter handle name and 239 entries were in tweets. Similarly, out of 259 occurrences 254 occurrences where in tweets and rest 5 entries were in twitter handle.

Command:
grep -e "Donald Trump" -e "DonaldTrump" Twitter_Data_1 | wc -l
grep -oe "Donald Trump" -oe "DonaldTrump" Twitter_Data_1 | wc -l

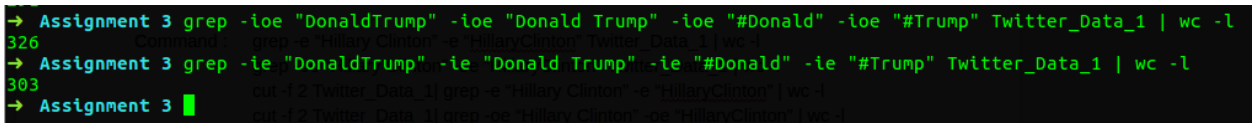
```
cut -f 2 Twitter_Data_1| grep -e "Donald Trump" -e "DonaldTrump" | wc -l
cut -f 2 Twitter_Data_1| grep -oe "Donald Trump" -oe "DonaldTrump" | wc -l
cut -f 4 Twitter_Data_1| grep -e "Donald Trump" -e "DonaldTrump" | wc -l
cut -f 4 Twitter_Data_1| grep -oe "Donald Trump" -oe "DonaldTrump" | wc -l
```



When searched irrespective of case in the name we found that his name was present in **260** entries and a total of **278** occurrences. Out of which 5 entries where in twitter handle name and 256 entries were in tweets. Similarly, out of 278 occurrences 273 occurrences where in tweets and rest 5 entries were in twitter handle.

Command:

```
grep -ie "Donald Trump" -ie "DonaldTrump" Twitter_Data_1 | wc -l
grep -ioe "Donald Trump" -ioe "DonaldTrump" Twitter_Data_1 | wc -l
cut -f 2 Twitter_Data_1| grep -ie "Donald Trump" -ie "DonaldTrump" | wc -l
cut -f 2 Twitter_Data_1| grep -ioe "Donald Trump" -ioe "DonaldTrump" | wc -l
cut -f 4 Twitter_Data_1| grep -ie "Donald Trump" -ie “DonaldTrump” | wc -l
cut -f 4 Twitter_Data_1| grep -ioe "Donald Trump" -ioe “DonaldTrump” | wc -l
```



When we searched irrespective of case for combinations “DonaldTrump”, ”Donald Trump”, ”#Donald” and ‘#Trump’. We found that there were **303** entries ad **326** occurrences.

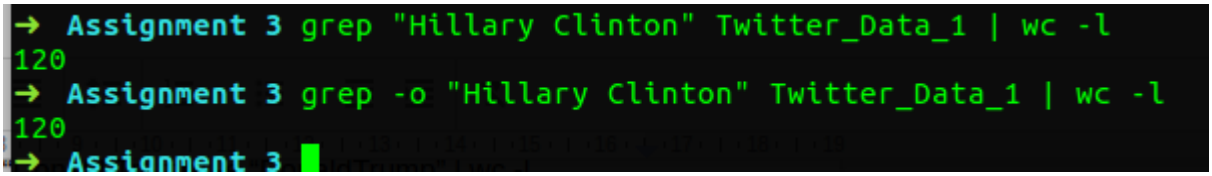
Command:

```
grep -ie "HillaryClinton" -ie "Hillary Clinton" -ie "#Hillary" -ie
"#Clinton" Twitter_Data_1 | wc -l
grep -ioe "HillaryClinton" -ioe "Hillary Clinton" -ioe "#Hillary" -ioe
"#Clinton" Twitter_Data_1 | wc -l
```

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

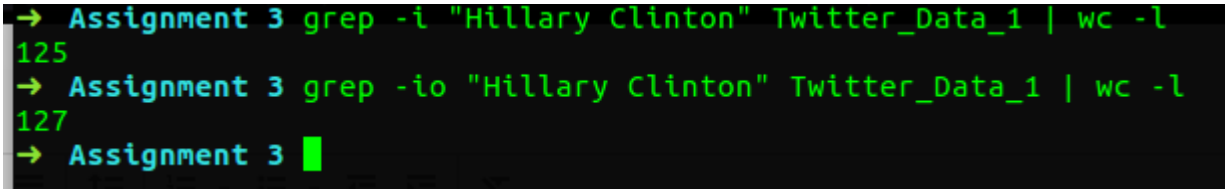
If we are searching just for the name "Hillary Clinton". Then

When searched respective of case in the name we found that her name was present in **120** tweets and total no of occurrences were **120**



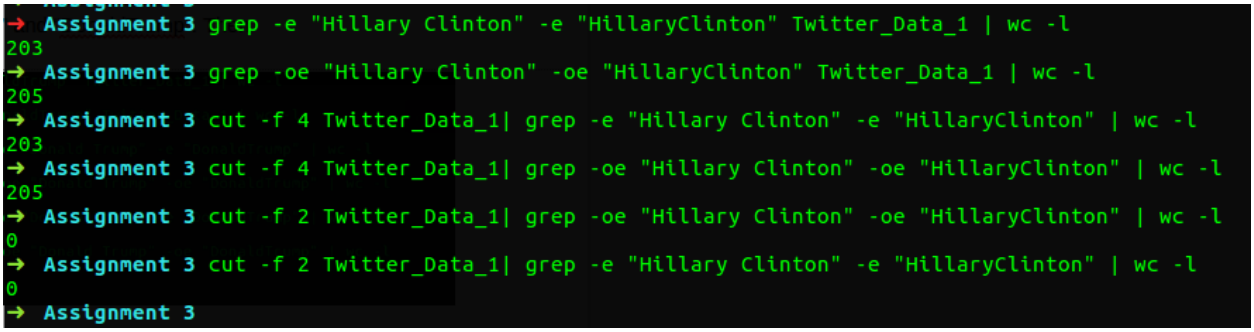
Command: ***grep "Hillary Clinton" Twitter_Data_1 | wc -l***
grep -o "Hillary Clinton" Twitter_Data_1| wc -l

When searched irrespective of case we found that her name was present in **125** tweets and total no of occurrences were **127**



Command: ***grep -i "Hillary Clinton" Twitter_Data_1 | wc -l***
grep -io "Hillary Clinton" Twitter_Data_1| wc -l

If we were searching for values "Hillary Clinton" and "HillaryClinton". Then

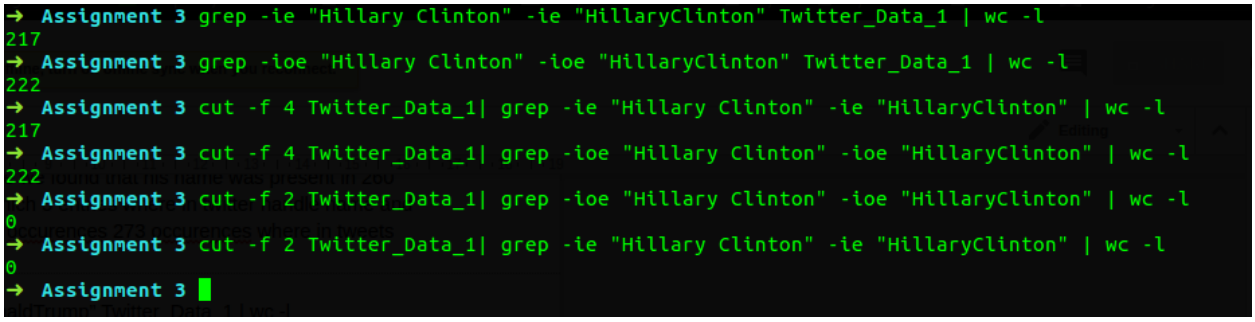


When searched respective of case in the name we found that her name was present in **203** entries and a total of **205** occurrences. Out of

which all 203 entries were in tweets. Similarly, out of which all 205 occurrences where in tweets.

Command:

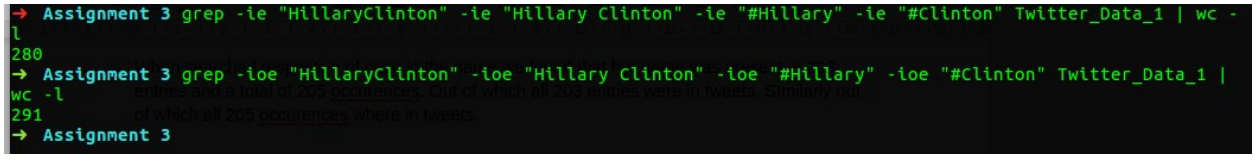
```
grep -e "Hillary Clinton" -e "HillaryClinton" Twitter_Data_1 | wc -l
grep -oe "Hillary Clinton" -oe "HillaryClinton" Twitter_Data_1 | wc -l
cut -f 2 Twitter_Data_1| grep -e "Hillary Clinton" -e "HillaryClinton"
| wc -l
cut -f 2 Twitter_Data_1| grep -oe "Hillary Clinton" -oe
"HillaryClinton" | wc -l
cut -f 4 Twitter_Data_1| grep -e "Hillary Clinton" -e "HillaryClinton"
| wc -l
cut -f 4 Twitter_Data_1| grep -oe "Hillary Clinton" -oe
"HillaryClinton" | wc -l
```



When searched irrespective of case in the name we found that his name was present in **217** entries and a total of **222** occurrences. Out of which all 217 entries were in tweets. Similarly, out of 222 occurrences all occurrences were in tweets.

Command:

```
grep -ie "Hillary Clinton" -ie "HillaryClinton" Twitter_Data_1 | wc -l
grep –ioe "Hillary Clinton" -ioe "HillaryClinton" Twitter_Data_1 | wc
-l
cut -f 2 Twitter_Data_1| grep -ie "Hillary Clinton" –ie
"HillaryClinton" | wc -l
cut -f 2 Twitter_Data_1| grep -ioe "Hillary Clinton" -ioe
"HillaryClinton" | wc -l
cut -f 4 Twitter_Data_1| grep -ie "Hillary Clinton" -ie
"HillaryClinton" | wc -l
cut -f 4 Twitter_Data_1| grep -ioe "Hillary Clinton" -ioe
"HillaryClinton" | wc -l
```



When we searched irrespective of case for combinations “Hillary Clinton”, ”HillaryClinton”, ”#Hillary” and ‘#Clinton’. We found that there were **280** entries and **291** occurrences.

Command:
grep -ie "HillaryClinton" -ie "Hillary Clinton" -ie "#Hillary" -ie "#Clinton" Twitter_Data_1 | wc -l
grep -ioe "HillaryClinton" -ioe "Hillary Clinton" -ioe "#Hillary" -ioe "#Clinton" Twitter_Data_1 | wc -l

If we Consider the exact name “Donald Trump” and “Hillary Clinton” and if we are searching irrespective of case sensitivity, then Hillary is more popular as it has more occurrences and tweets. But If we consider irrespective of case then the no of tweets is more for Hillary whereas no of occurrences is more for Donald Trump.

Now if we consider a combination of words that is “Donald Trump” and “DonaldTrump” and similarly for Hillary as “Hillary Clinton” and “HIllaryClinton”. In this case both respective and irrespective of case Donald trump has more number of tweets and occurrences and thus is more popular

Now if we consider a combination of words that is “Donald Trump”, “DonaldTrump” ,”#Trump” and ”#Donald”. Similarly for Hillary as “Hillary Clinton”, “HillaryClinton”, ”#Hillary” and “#Clinton” . In this case irrespective of case Donald trump has more number of tweets and occurrences and thus is more popular

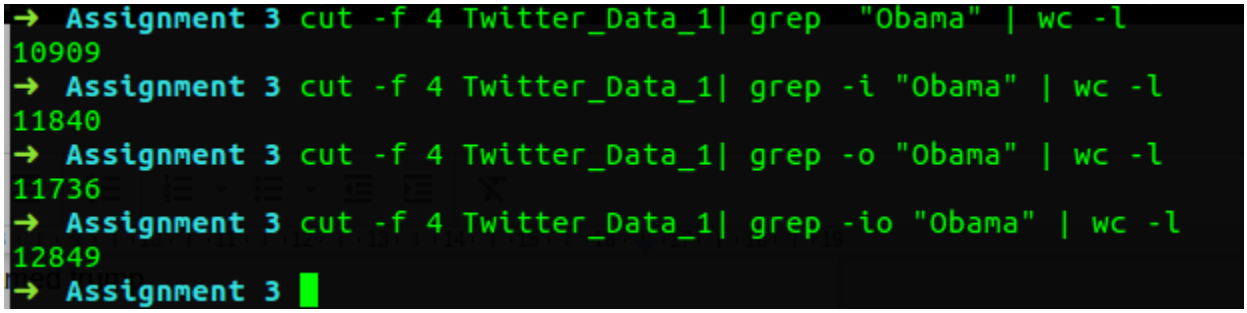
10) Do you think we have captured all the references to Donald and Hillary? What other strings might we need to try? What problems might we face?

No, I don’t think we have captured all the references as when we just added a new term in search that is “HillaryClinton” and “DonaldTrump” we saw that the results just turned. Similarly, there might be more key words which might represent Donald Trump and Hillary Clinton. Like #Trump and #Hillary which we might not have covered.
Also there are chances of misspelled words which might not show in the result but which was actually meant for Donald trump or Hillary Clinton. Those kind of tweets or occurrences we won’t be able to find out.

Same way if we search for just trump and hillary irrespective of case we will get a result of 1643 and 1417 respectively but the issue in this search is that we don't know out of these 1643 how many are actually for Donald Trump similarly out of 1417 how many are actually for Hillary Clinton as some occurrences may be for some other person named trump which we cannot determine.

Part B: Graphing the Data in R

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



Case Sensitive search cases
When searched for the case sensitive term "Obama". We found that it appeared in 10909 Tweets and the total no of occurrences was 11736.

Tweets: 10909
Command: cut -f 4 Twitter_Data_1| grep "Obama" | wc -l
Occurrences: 11736
Command: cut -f 4 Twitter_Data_1| grep -o "Obama" | wc -l

Irrespective of Case sensitivity search result
When searched Irrespective of case for the term "Obama". We found that it appeared in 11840 Tweets and the total no of occurrences was 12849.

Tweets: 11840
Command: cut -f 4 Twitter_Data_1| grep -i "Obama" | wc -l
Occurrences: 12849
Command: cut -f 4 Twitter_Data_1| grep -io "Obama" | wc -l

2) Background: We want to consider how the amount of discussion regarding Barack Obama varies over the time period covered by the data file. To answer this question, you will need to extract the timestamps for all tweets referring to Obama. You will then need to read them into R and generate a histogram. [Hint: To read the data into R, first generate a file containing only the timestamp column as text. Then read the file into R as a CSV.] R will not recognize the strings as timestamps automatically, so you'll need to convert them from text values using the strptime() function. Instructions on how to

use the function is available here: (https://stat.ethz.ch/R-manual/Rdevel/library/base/html/strptime.html). 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?

```
➜  Assignment 3 echo "time" > timeStamp.csv
➜  Assignment 3 cut -f 3 Twitter_Data_1 >> timeStamp.csv
➜  Assignment 3 head timeStamp.csv
time
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
➜  Assignment 3
```

Command in shell for Filtering Data:

```
echo "time" > timeStamp.csv
cut -f 3,4 Twitter_Data_1| grep -i "Obama" | cut -f 1 >>
timeStamp.csv
```

```
> setwd("~/Study Items/Monash/S1/FIT5145/Assignment\ 3")
> timeDetails <- read.csv(file = "timeStamp.csv",header = TRUE)
>
> timeDetails = strptime(timeDetails[,1],format = "%a %b %d %H:%M:%S %z %Y")
> head timeDetails
Error: unexpected symbol in "head timeDetails"
> head(timeDetails)
[1] "2014-02-11 23:18:36" "2014-02-11 23:18:36" "2014-02-11 23:18:36" "2014-02-11 23:18:36" "2014-02-11 23:18:36" "2014-02-11
 23:18:36"
>
```

Script In R:

```
setwd("~/Study Items/Monash/S1/FIT5145/Assignment\ 3")
timeDetails <- read.csv(file = "timeStamp.csv",header = TRUE)
timeDetails = strptime(timeDetails[,1],format = "%a %b %d %H:%M:
%S %z %Y")
```

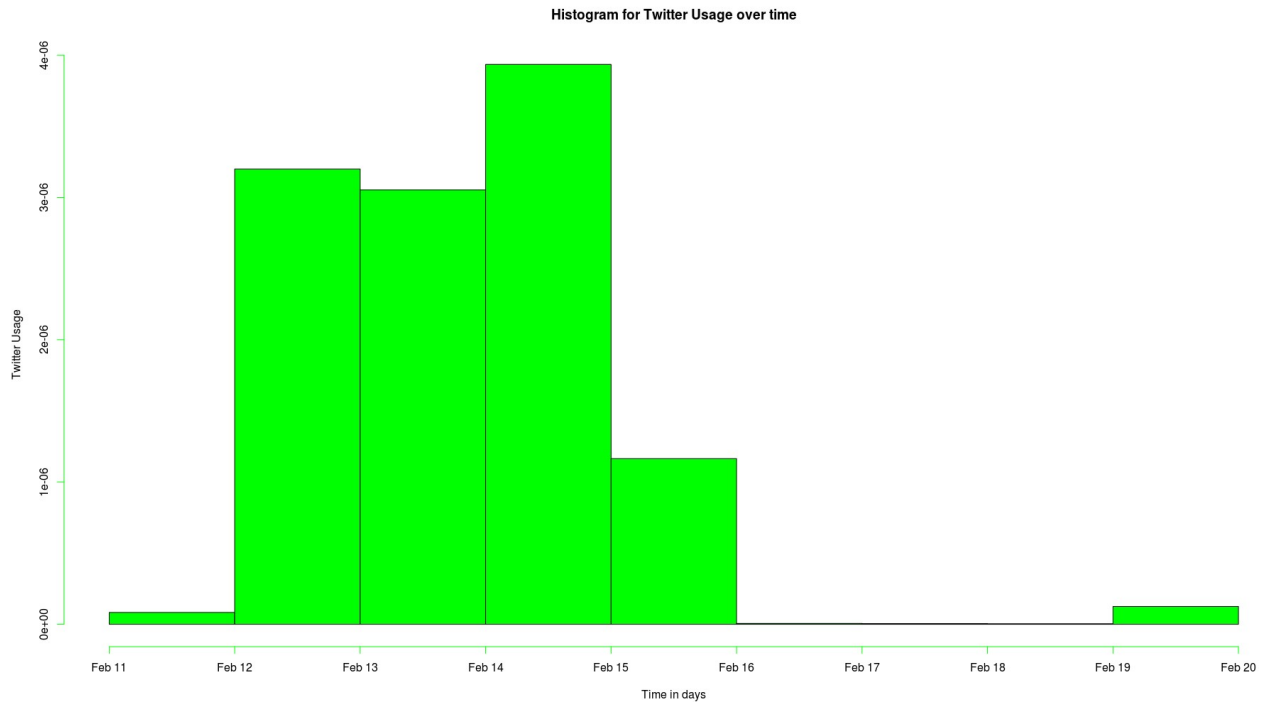
The Format String that I have used is: %a %b %d %H:%M:%S %z %Y

Where each symbol is been explained below

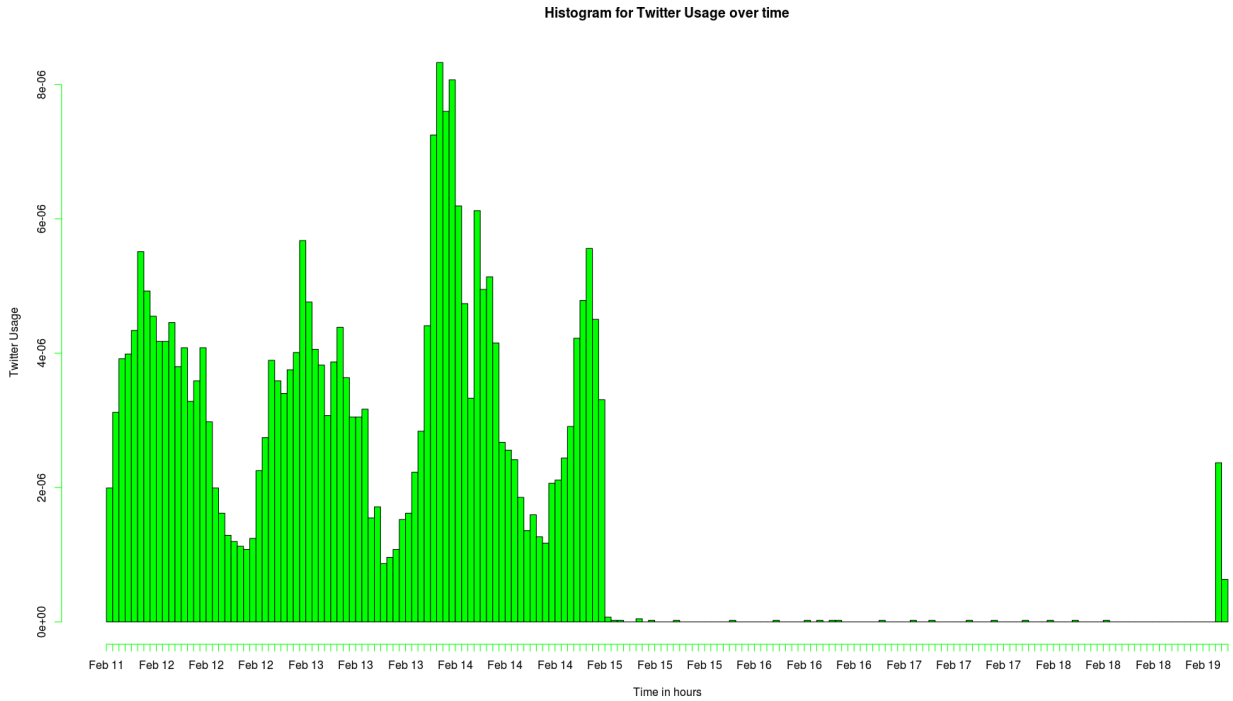
%a - Abbreviated weekday name in the current locale on this platform.
%b - Abbreviated month name in the current locale on this platform
%d - Day of the month as decimal number (01–31).
%H - Hours as decimal number (00–23)
%M - Minute as decimal number (00–59).
%S - Second as integer (00–61), allowing for up to two leap-seconds

`%z - Signed offset in hours and minutes from UTC, so -0800 is 8 hours behind UTC. Values up to +1400 are accepted`
`%Y - Year with century.`

3) Once you've converted the timestamps, use the hist() function to plot the data. [Hint: you will need to set the number of bins sufficiently high to see the variation over time well.]



Command:
hist(timeDetails,main="Histogram for Twitter Usage over time",xlab="Time in days",ylab="Twitter Usage",col="green",breaks = "day")



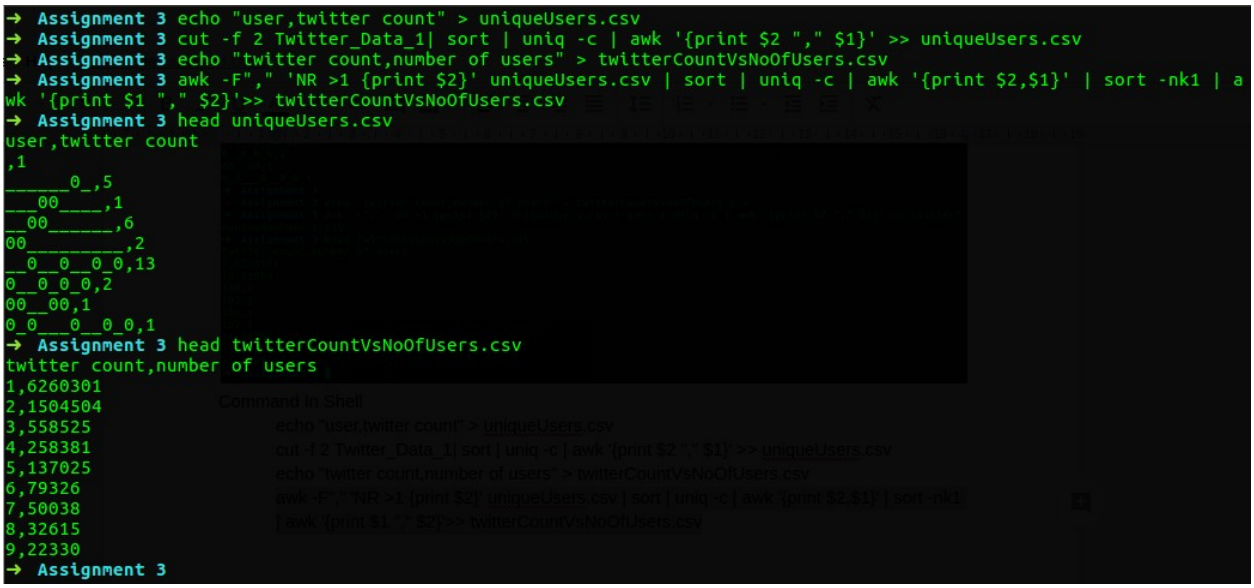
Command:

```
hist(timeDetails,main="Histogram for Twitter Usage over time",xlab="Time in hours",ylab="Twitter Usage",col="green",breaks = "hours")
```

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

Yes we can see a pattern ie every day from feb 12 till feb 15 every morning there seems to be having lot of activity and usually keeps on decreasing till evening with fews ups and downs and at night there is least activity. After feb 15 till feb 19 there is not much activity that is approximately equal to zero.

5) (Challenge) Plot a second histogram, but this time showing the distribution over number of tweets per author in the file. [Hint: You'll need to count up the number of Tweets by each unique author in the Twitter file giving a file with two columns "user" and "twitter count". 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".]



Commands for populating initial data for processing

Command In Shell:

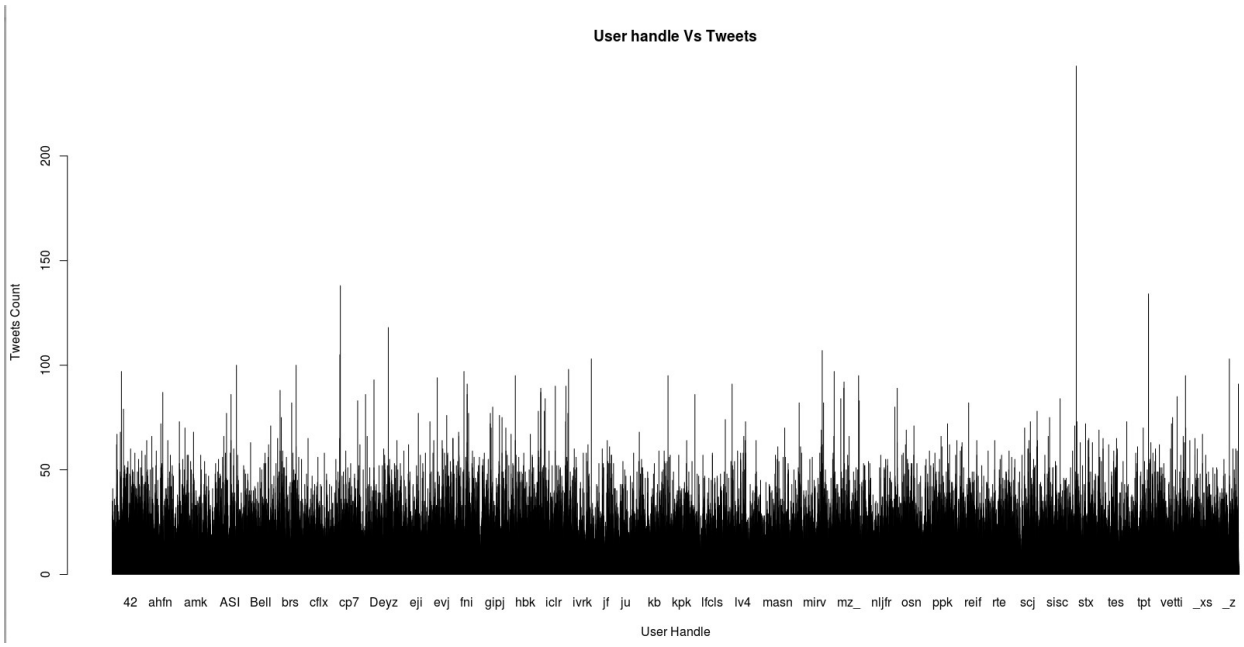
```
echo "user,twitter count" > uniqueUsers.csv
cut -f 2 Twitter_Data_1| sort | uniq -c | awk '{print $2 "," $1}' >>
uniqueUsers.csv
echo "twitter count,number of users" > twitterCountVsNoOfUsers.csv
awk -F"," 'NR >1 {print $2}' uniqueUsers.csv | sort | uniq -c | awk
'{print $2,$1}' | sort -nk1 | awk '{print $1 "," $2}'>>
twitterCountVsNoOfUsers.csv
echo "tweet" > tweetsHIst.csv
awk -F"," 'NR >1 {print $2}' uniqueUsers.csv | sort -n >>
tweetsHIst.csv
```

Command In R:

```
uniqUsers <- read.csv(file = "uniqueUsers.csv",header = TRUE,sep =
",")
tweets <- read.csv(file = "tweetsHIst.csv",header = TRUE)
tweetsVsUserCount <- read.csv(file =
"twitterCountVsNoOfUsers.csv",header = TRUE,sep = ",")
```

When I filtered the no of unique users the no of users count I got is too high like around 8977904 users which is a huge amount of data. So when I plotted the graph for users vs no of tweets I got a graph like one below but actually I don't think this is the actual graph as I believe it won't be able to plot around 8 million lines and represent those unique names in the x-axis. Also when I checked the mean of the no of tweets data it was around 1.6 something but the mean of this graph doesn't look like its 1.6 something. So I believe it is actually a subset of the actual graph

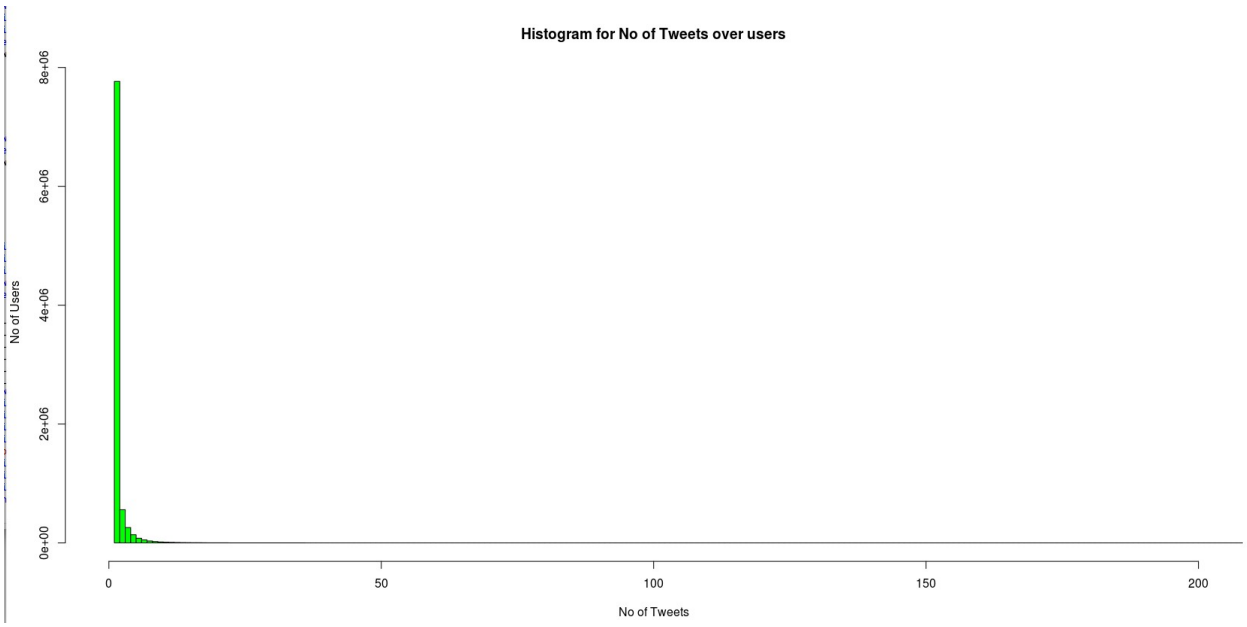
Plot graph of twitter user handle over tweets count



Command in R:

```
barplot(uniqUsers$twitter.count,names.arg = uniqUsers$user,col = "green", main="User handle Vs Tweets",xlab="User Handle",ylab="Tweets Count")
```

Plot a graph for no of tweets over users



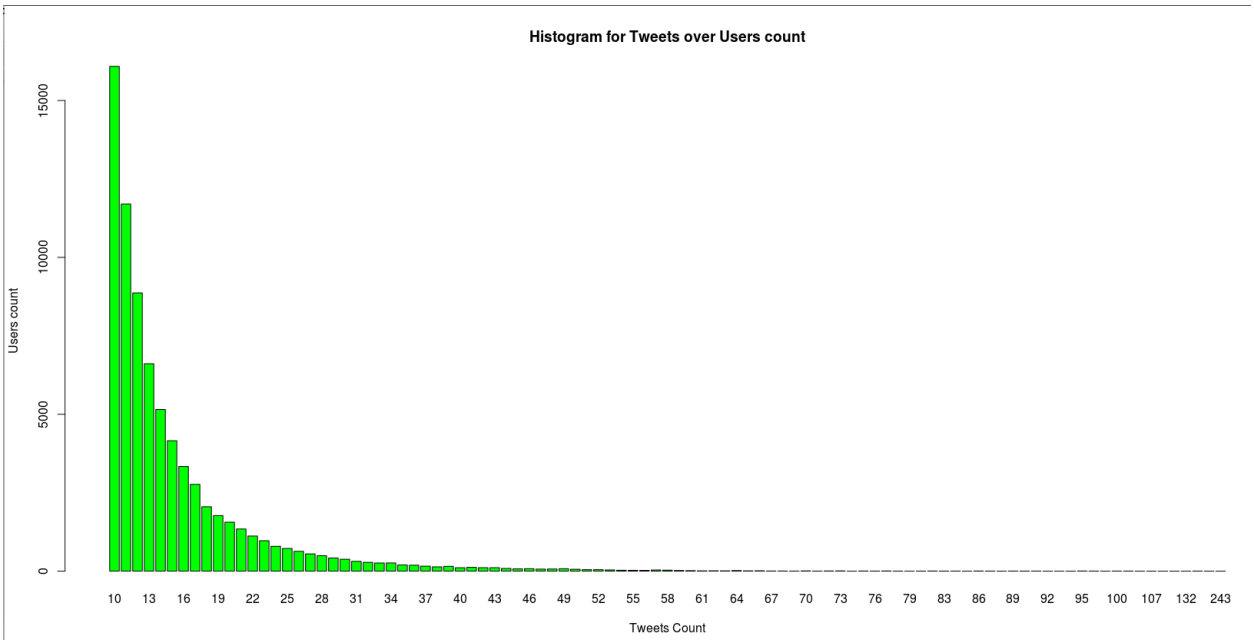
Command in R:

hist(tweets$tweet,xlim=c(0,200),main="Histogram for No of Tweets over users",xlab="No of Tweets",ylab="No of Users",col="green",breaks = 200)

In this data only the first 9 values are visible in the chart rest all has negligible values. In order to get the chart of data from 10 tweets onwards we can use the following commands

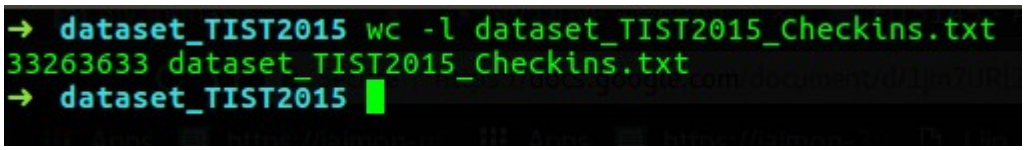
Command in R:
barplot(tweetsVsUserCount$number.of.users[10:length(tweetsVsUserCount$number.of.users)],names.arg = tweetsVsUserCount$twitter.count[10:length(tweetsVsUserCount$number.of.users)], col = "green", main="Histogram for Tweets over Users count",xlab="Tweets Count",ylab="Users count")

Plotting graph for no of tweets over users count (Note: Skipping first 9 entries of tweet vs users count)



Part C: Investigating User Check-in Data in the Shell Download the file dataset_TIST2015.zip, which contains user check-in data from Foursquare (https://foursquare.com/).

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 is about **33263633** checkin data.
Command: wc -l dataset_TIST2015_Checkins.txt

There are **266909** unique users in the data provided
Command: cut -f 1 dataset_TIST2015_Checkins.txt| sort -u | wc -l

2) Background: How would you select venues from Europe? Consider the structure of the data presented in the readme file. Check-ins are index by a Venue ID, and these are described separately in a separate file, the POI file. You can select European venues from the POI file in (at least) two ways: select items in a latitude longitude bounding box, or select items by country code. The first is easier for all or Europe, but if you want to select a single current use the country code, and if you want to count on country then count on country code. Look on a map to find latitude and longitude coordinates to bound Europe. Don't be too fussed by the exact locations (include or exclude Turkey, Ukraine, etc., that is OK either way). 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.ro

For finding the latitude longitude bounding box. I first found out latitudes and longitudes of each countries in Europe and wrote it in file Data (euLatLong.txt) was obtained from link http://www.allplacesmap.com/europe/lat-long.html.
The file looks like

```
➜ dataset_TIST2015 head euLatLong.txt
Albania 41.635543 19.712892
Austria 47.144935 9.775851
Belarus 54.749436 29.250870
Belgium 50.730000 5.420000
Bulgaria 41.721532 26.320317
Croatia 46.164650 15.867662
Czech 50.559430 15.416210
Denmark 55.785290 12.321330
Estonia 58.654500 25.037227
Finland 2.600000 25.733330
➜ dataset_TIST2015
```

First column shows the country name and second column indicates latitude and third column indicates the longitude value

Finding Latitude Bounding Box Values

From euLatLong.txt I sorted based on latitude and created another file latSorted.txt
Command: sort -nk2 euLatLong.txt > latSorted.txt

```
➜ dataset_TIST2015 head latSorted.txt
Malta 35.883324 14.494650
Portugal 38.850000 -7.580000
Turkey 38.963700 35.243300
Greece 40.260298 24.249458
Albania 41.635543 19.712892
Bulgaria 41.721532 26.320317
Marcedonia 41.910730 20.913320
Spain 42.351592 -0.730019
Montenegro 42.959727 19.093008
Monaco 43.740416 7.425678
```

From the data it is pretty evident that we can considered Malta's Coordinate as min coordinate value for latitude. From Site http://www.allplacesmap.com/europe/malta/lat-long.html we found that the min latitude value for Malta is 35.826116

```
→ dataset_TIST2015 tail latSorted.txt
Netherlands 53.272666 7.028446
Lithuania 54.236580 23.511890
Belarus 54.749436 29.250870
Denmark 55.785290 12.321330
Latvia 57.425010 25.900680
Estonia 58.654500 25.037227
Sweden 58.968180 16.200450
Norway 59.308102 4.881957
Finland 62.600000 25.733330
Iceland 64.006432 -22.563629
```

Similarly, when I took the tail Iceland looked like an outlier so we will find a bounding box for Iceland separately. So for the main bounding box I considered either value of Norway or Finland as the max coordinate. From Site http://www.allplacesmap.com/europe/norway/lat-long.html we found that the max latitude value for Norway is 71.165552 which is greater than the max value in Finland

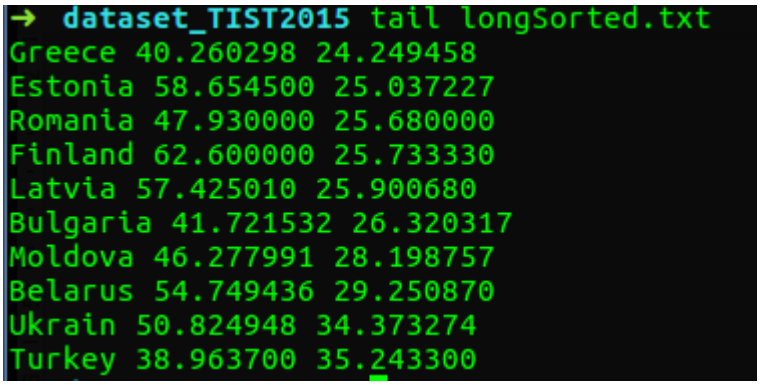
Finding Longitude bounding box values

From euLatLong.txt I sorted based on longitude and created another file longSorted.txt

Command: sort -nk3 euLatLong.txt > longSorted.txt

```
→ dataset_TIST2015 head longSorted.txt
Iceland 64.006432 -22.563629
Ireland 52.955368 -7.800643
Portugal 38.850000 -7.580000
UK 49.238740 -2.173634
Spain 42.351592 -0.730019
France 46.552664 2.422229
Norway 59.308102 4.881957
Belgium 50.730000 5.420000
Luxembourg 49.868953 6.158001
Netherlands 53.272666 7.028446
→ dataset_TIST2015
```

From the above fig it is evident that Iceland is an outlier. So we will take Coordinates of Ireland as the min coordinate for the main bounding box. From Site http://www.allplacesmap.com/europe/ireland/lat-long.html we found that the min longitude value is -6.041501



From the above figure we will take coordinates of Turkey as max longitude value for the current bounding box.
On referring https://www.latlong.net/place/ankara-ankara-province-turkey-2708.html we can see that 43.378143 is the max longitude value

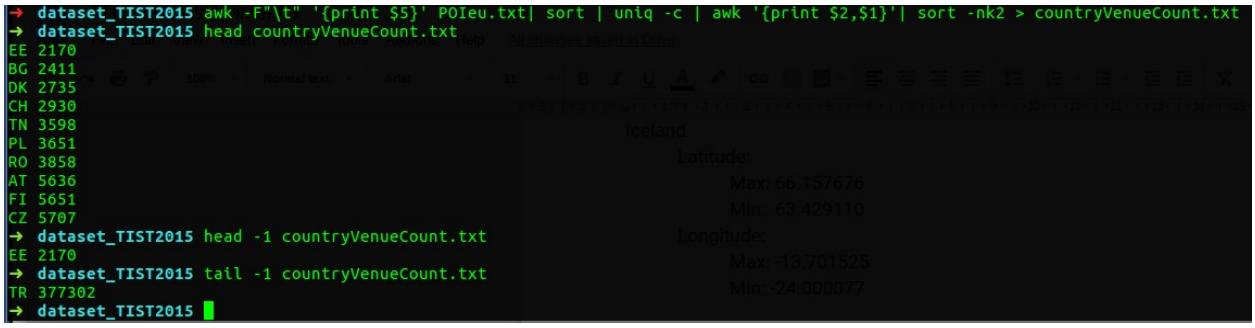
From http://www.allplacesmap.com/europe/iceland/lat-long.html we found out the coordinates of Iceland

Main bounding box
Latitude:
Max: 71.165552
Min: 35.826116
Longitude:
Max: 43.378143
Min: -6.041501

Iceland
Latitude:
Max: 66.157676
Min: 63.429110
Longitude:
Max: -13.701525
Min: -24.000077

Command used to filter POI is:
awk -F'\t' '($2>=35.826116 && $2 <=71.165552 && $3>=-6.041501 && $3<=43.378143) || ($2>=63.429110 && $2<=66.157676 && $3>=-24.000077 && $3<=-13.701525) {print $0}' dataset_TIST2015_POIs.txt> POIeu.txt

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



Command to extract country codes of country with max and min venues

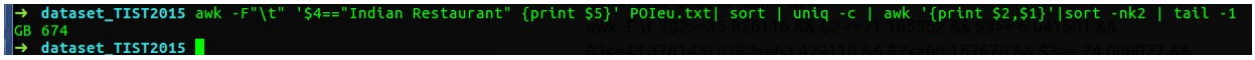
awk -F"\t" '{print $5}' POIeu.txt| sort | uniq -c | awk '{print $2,$1}'| sort -nk2 > countryVenueCount.txt
head -1 countryVenueCount.txt
tail -1 countryVenueCount.txt

Command to extract country name from country code:

awk -F"\t" '$4 =="EE" || $4=="TR" {print $4,$5}' dataset_TIST2015_Cities.txt| sort -k1 > tt
head -1 tt
tail -1 tt

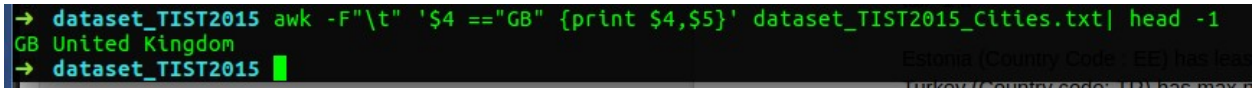
Estonia (Country Code: EE) has least no of venues that is **2170** venues
Turkey (Country code: TR) has max no of venues that is **377302** venues

C. Who has the most Indian restaurants?



Command to extract country code of country which has max Indian restraunts:

awk -F"\t" '$4=="Indian Restaurant" {print $5}' POIeu.txt| sort | uniq -c | awk '{print $2,$1}'|sort -nk2 | tail -1

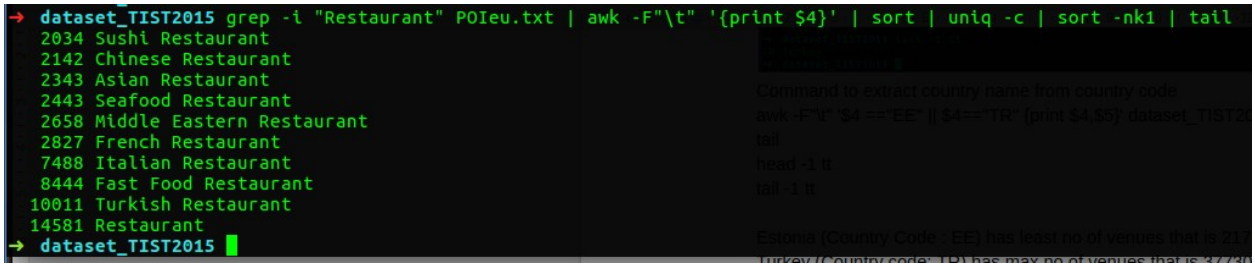


Command to extract country name from country code GB

awk -F"\t" '$4 =="GB" {print $4,$5}' dataset_TIST2015_Cities.txt| head -1

The Country with most number of Indian Restraunts is **United Kingdom** (Country code GB) with **674** restraunts

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



Command:

grep -i "Restaurant" POIeu.txt | awk -F"\t" '{print $4}' | sort | uniq -c | sort -nk1 | tail -1

So the most common class of Restaurant in Europe is "**Restaurant**" itself with **14581** venues. Second most common class of Restaurant in Europe is "Turkish Restaurant" with 10011 venues