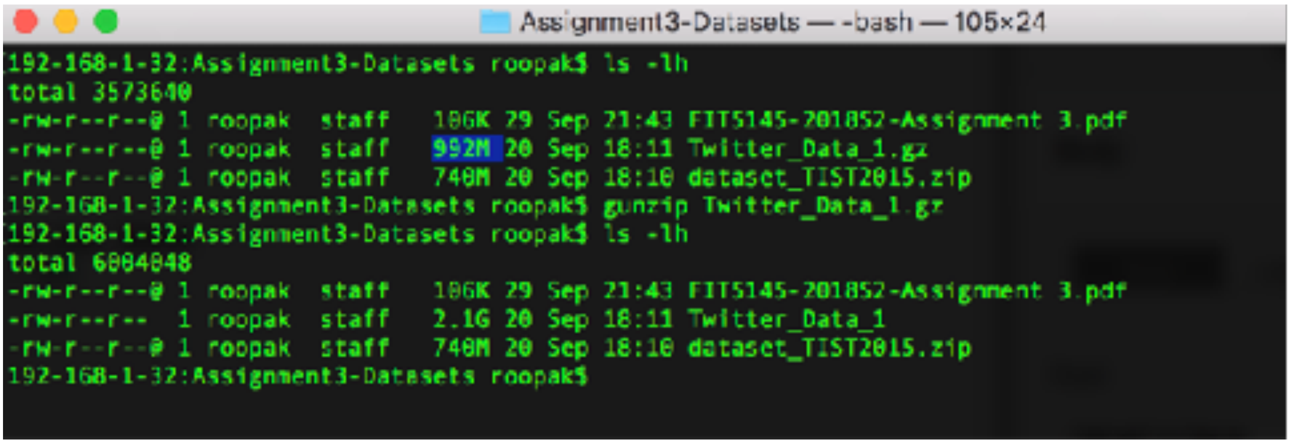
Roopak Thiyyathuparambil Jayachandran StudentId: 29567467

Data Science Assignment-3

PART-A

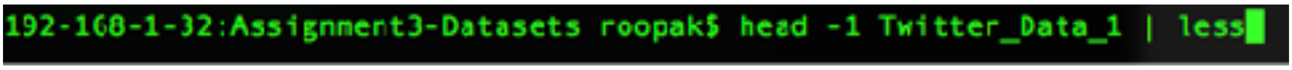
1. Decompress the file. How big is it?

Initial given data is Twitter_Data_1.gz which is the compressed by the standard GNU zip.
Twitter_Data_1 initial .gz file size : **992M** (MegaBytes)
After de compressing : **2.1 G (Giga Bytes)**

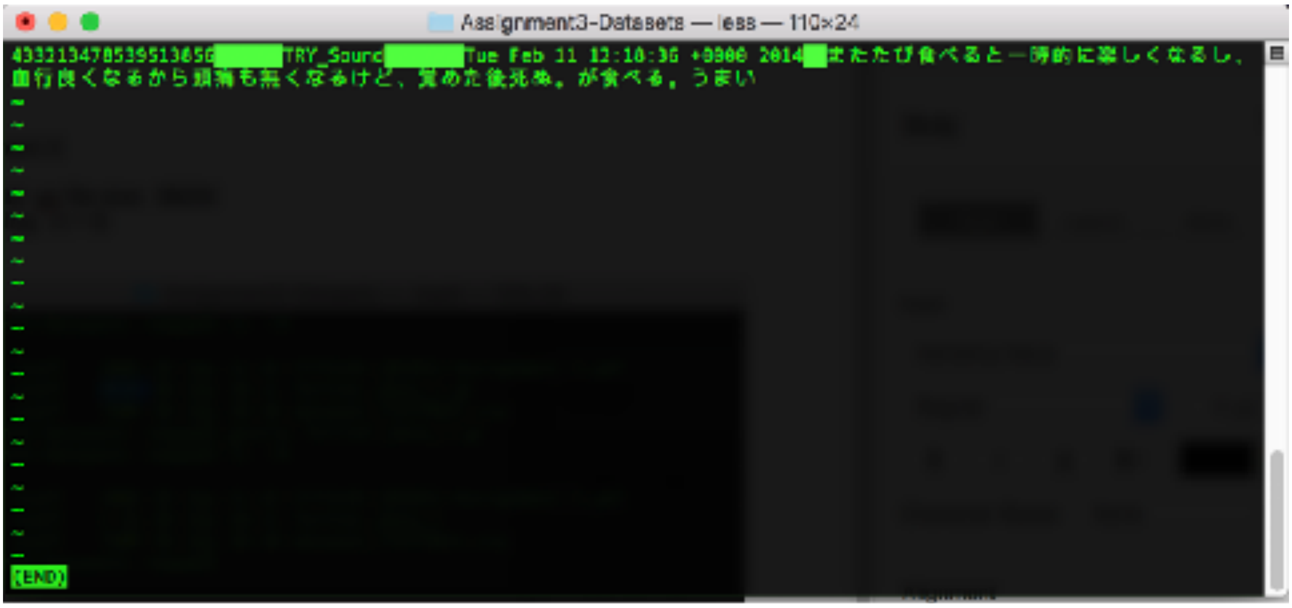


```
192-168-1-32:Assignment3-Datasets roopak$ ls -lh
total 3573640
-rw-r--r--@ 1 roopak  staff   186K 29 Sep 21:43 FIT5145-201852-Assignment 3.pdf
-rw-r--r--@ 1 roopak  staff   992M 20 Sep 18:11 Twitter_Data_1.gz
-rw-r--r--@ 1 roopak  staff   740M 20 Sep 18:10 dataset_TIST2015.zip
192-168-1-32:Assignment3-Datasets roopak$ gunzip Twitter_Data_1.gz
192-168-1-32:Assignment3-Datasets roopak$ ls -lh
total 6084848
-rw-r--r--@ 1 roopak  staff   186K 29 Sep 21:43 FIT5145-201852-Assignment 3.pdf
-rw-r--r--  1 roopak  staff   2.1G 20 Sep 18:11 Twitter_Data_1
-rw-r--r--@ 1 roopak  staff   740M 20 Sep 18:10 dataset_TIST2015.zip
192-168-1-32:Assignment3-Datasets roopak$
```

2. What delimiter is used to separate the columns in the file and how many columns are there?
There are 4 columns and the delimiter used is **tab.** This can be identified by less. Use less command for the first 2 lines of the dataset file. After opening the file in less, searching the pattern /<tab>, we can see that space between each of the column is highlighted. Hence we can conclude that **Tab is used as the delimiter**



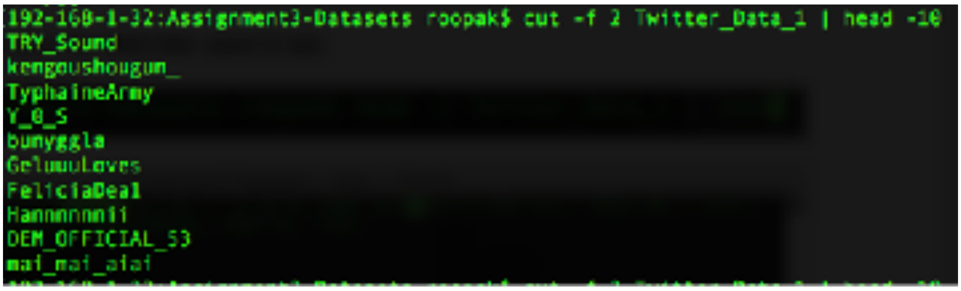
```
192-168-1-32:Assignment3-Datasets roopak$ head -1 Twitter_Data_1 | less
```



```
433213478539513856	TRY_Sourc	Tue Feb 11 12:10:36 +0000 2014	またたび食べると一時的に楽しくなるし、
血行良くなるから頭痛も無くなるけど、覚めた後死ぬ。が食べる，うまい
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
~
(END)
```

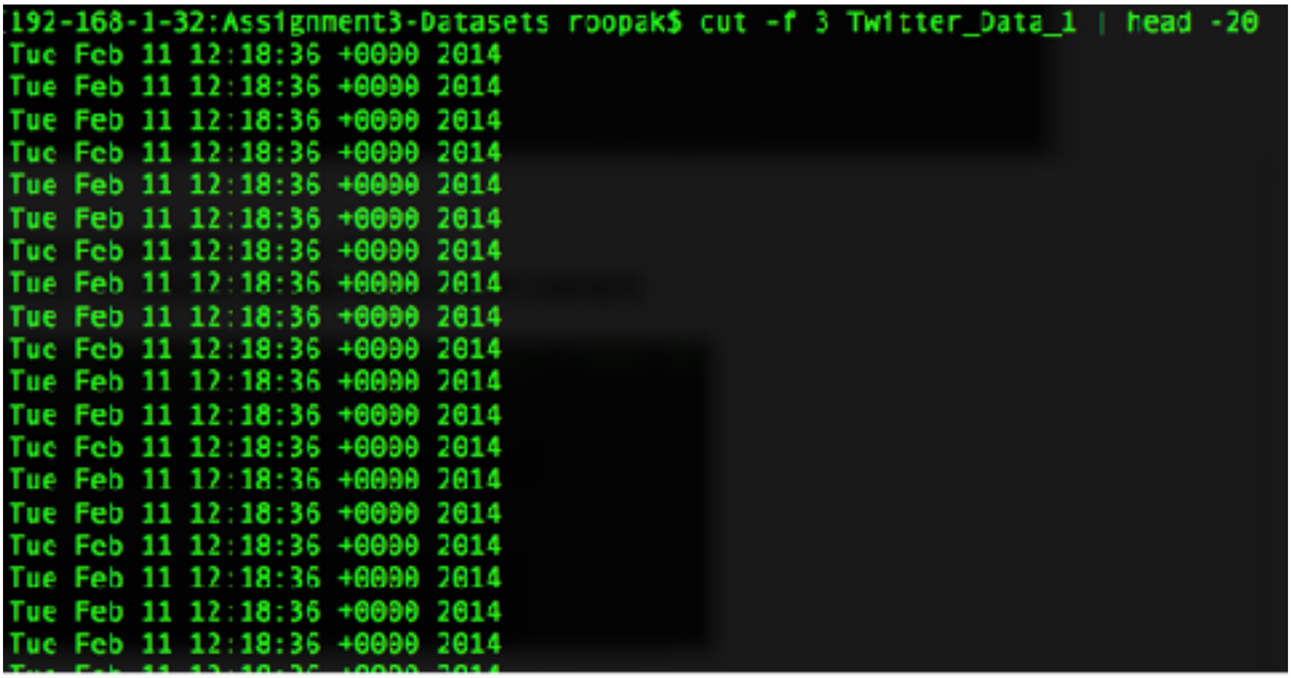
3. The first column is a unique identifier for a Tweet. What are the other columns?

First column is the **unique identifier** of the tweet.
Second column represents the **twitter handle or account owners**



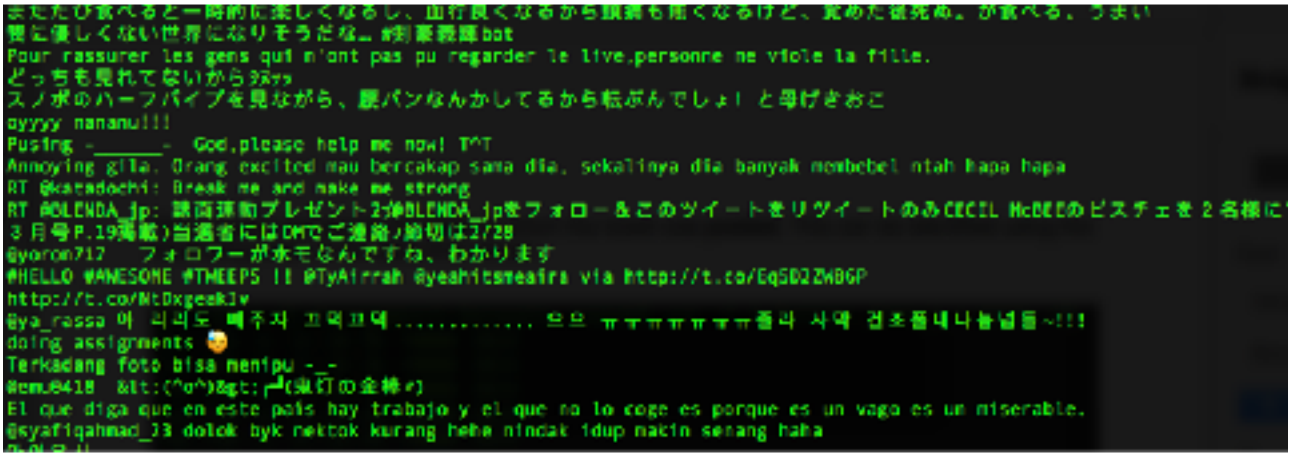
```
192-168-1-32:Assignment3-Datasets roopak$ cut -f 2 Twitter_Data_1 | head -10
TRY_Sound
kengoushougun_
TyphaineArmy
Y_8_S
bunyggla
GelmuuLoves
FeliciaDeal
Hannnnnnnii
DEM_OFFICIAL_53
mai_mai_aiai
```

Third column represents the **time** at which the tweet was **posted.** This can be identified using the ascending order of the data time values



```
192-168-1-32:Assignment3-Datasets roopak$ cut -f 3 Twitter_Data_1 | head -20
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
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
Tue Feb 11 12:18:36 +0000 2014
```

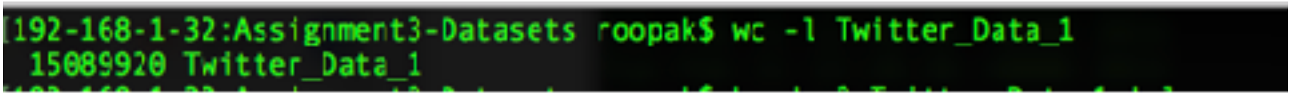
Fourth column is the actual **tweet content**



```
またたび食べると一時的に楽しくなるし、血行良くなるから頭痛も酷くなるけど、覚めた後死ぬ。が食べる。うまい
猫に優しくない世界になりそうだな… #刹那義羅bot
Pour rassurer les gens qui n'ont pas pu regarder le live,personne ne viole la fille.
どっちも見れてないから変ｯｯ
スノボのハーフパイプを見ながら、腰パンなんかしてるから転ぶんでしょ！と母げきおこ
oyyyy nananu!!!
Pusing -_______-  God,please help me now! T^T
Annoying gila. Orang excited mau bercakap sama dia. sekalinya dia banyak membebel ntah hapa hapa
RT @katadochi: Break me and make me strong
RT @LENDA_jp: 読者謝恩プレゼント2弾@LENDA_jpをフォロー＆このツイートをリツイートのみCECIL McBEEのビスチェを２名様に
３月号P.19掲載)当選者にはDMでご連絡♪締切は2/28
@yoron717  フォロワーが永そなんですね、わかります
#HELLO #AWESOME #THEEPS !! @TyAirrah @yeahitsmeaira via http://t.co/Eq5D2ZWBGP
http://t.co/MtDxgeekJv
@ya_rassa 에 [illegible] ............... ㅡㅡ ㅠㅠㅠㅠㅠㅠㅠ [illegible]~!!!
doing assignments 😓
Terkadang foto bisa menipu -_-
@emu0418  &lt;(^o^)&gt;┏(鬼灯の金棒♯)
El que diga que en este país hay trabajo y el que no lo coge es porque es un vago es un miserable.
@syafiqahmad_33 dolok byk nektok kurang hehe nindak idup makin senang haha
```

4. How many Tweets are there in the file?

There are **15089920 tweets** in this file since each row is a separate tweet posted. We can use word count command as used below

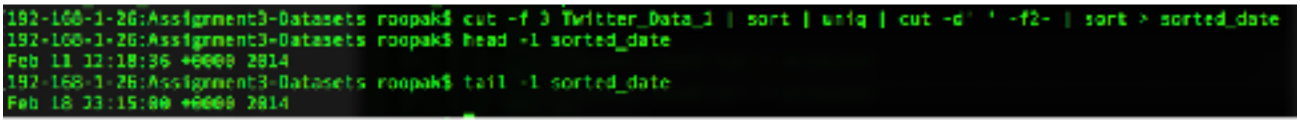


```
192-168-1-32:Assignment3-Datasets roopak$ wc -l Twitter_Data_1
 15089920 Twitter_Data_1
```

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

To get the date range, we will first extract the third column , from that remove the first column which represents the day of the week, get unique, sort it and move to a separate file. Since the data is of Feb month, we don't need to worry about month while sorting.

Command used: cut -f 3 Twitter_Data_1 | sort | uniq | cut -d' ' -f2- | sort > sorted_date
Starting date : head -1 sorted_date
Last date : tail -1 sorted_date

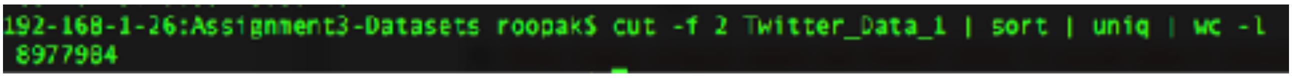


```
192-168-1-26:Assignment3-Datasets roopak$ cut -f 3 Twitter_Data_1 | sort | uniq | cut -d' ' -f2- | sort > sorted_date
192-168-1-26:Assignment3-Datasets roopak$ head -1 sorted_date
Feb 11 12:18:36 +0000 2014
192-168-1-26:Assignment3-Datasets roopak$ tail -1 sorted_date
Feb 18 23:15:00 +0000 2014
```

So the date range is from **11th feb 2014 12:18:36** to **18th feb 2014 23:15:00**.

6. How many unique users are there? [Hint: It could take 5 minutes to sort such a big list, so be patient!]1

Command Used : **cut -f 2 Twitter_Data_1 | sort | uniq | wc -l**



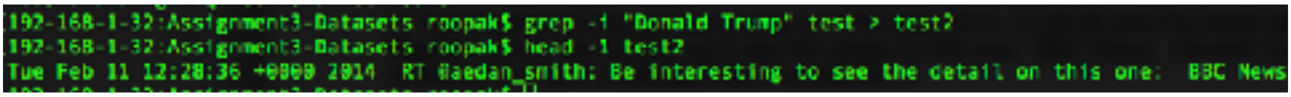
```
192-168-1-26:Assignment3-Datasets roopak$ cut -f 2 Twitter_Data_1 | sort | uniq | wc -l
 8977904
```

There are **8977904** unique users in the given data

7. When was the first mention in the file of "Donald Trump" and what was the tweet?

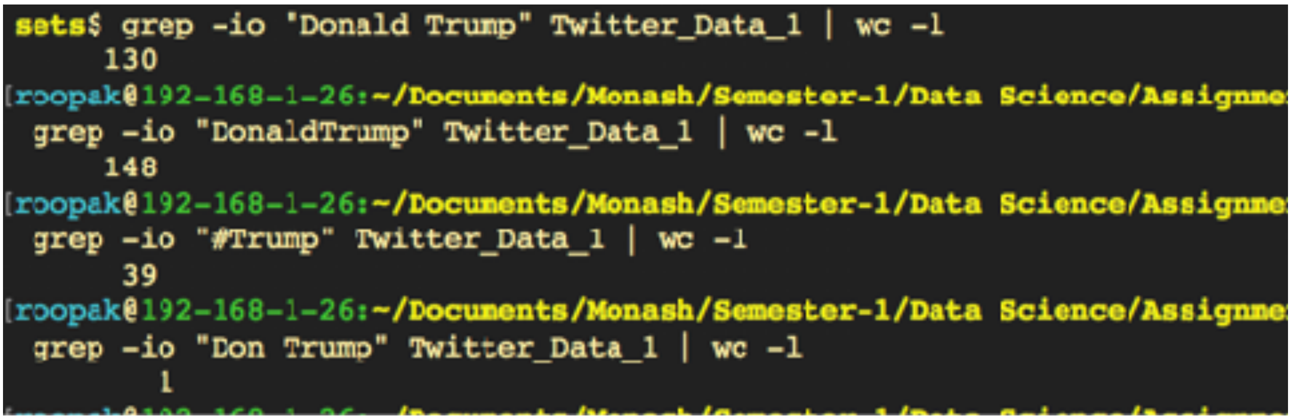
First mention of Donald Trump was on **11th Feb, 2014: 12:28:36** and the tweet was **"RT @aedan_smith: Be interesting to see the detail on this one: BBC News - Donald Trump loses offshore wind farm challenge http://t.co/qAcG…"** This is done using grep command ignoring cases

Command used : **grep -i "Donald Trump" test > test2,** where test is the file where column with posted time and tweets are extracted.



```
192-168-1-32:Assignment3-Datasets roopak$ grep -i "Donald Trump" test > test2
192-168-1-32:Assignment3-Datasets roopak$ head -1 test2
Tue Feb 11 12:28:36 +0000 2014  RT @aedan_smith: Be interesting to see the detail on this one:  BBC News
```

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

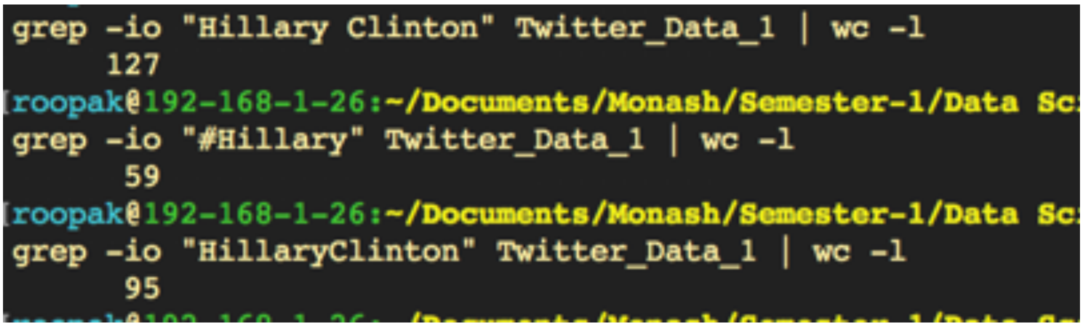


```
sets$ grep -io 'Donald Trump" Twitter_Data_1 | wc -l
     130
[roopak@192-168-1-26:~/Documents/Monash/Semester-1/Data Science/Assignme
 grep -io "DonaldTrump" Twitter_Data_1 | wc -l
     148
[roopak@192-168-1-26:~/Documents/Monash/Semester-1/Data Science/Assignme
 grep -io "#Trump" Twitter_Data_1 | wc -l
      39
[roopak@192-168-1-26:~/Documents/Monash/Semester-1/Data Science/Assignme
 grep -io "Don Trump" Twitter_Data_1 | wc -l
       1
```

Trump has been mentioned 318 times

Explanation: Donald Trump has been referred in the data in different ways like “Donald Trump”, “DonaldTrump”, “#Trump”, “Don Trump”. To get this, we have extracted column 4 and moved it to a different file. So doing a grep on each of the references ignoring the case will give the number of times Donald Trump has been referred.
“Donald Trump” was mentioned 130 times
“DonaldTrump” was mentioned 148 times
“#Trump” was mentioned 39 times
“Don Trump” was mentioned 1 time

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

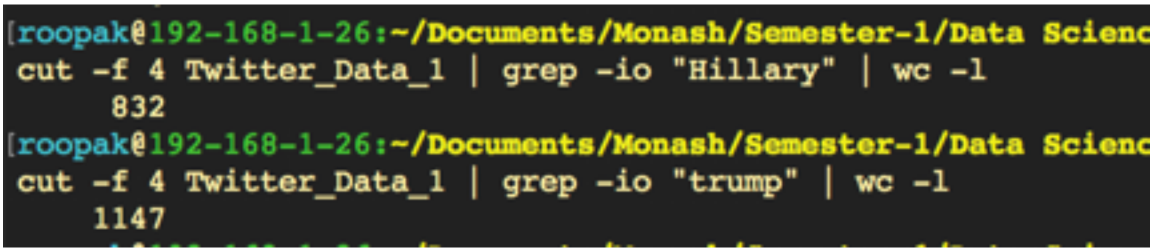


```
grep -io "Hillary Clinton" Twitter_Data_1 | wc -l
     127
[roopak@192-168-1-26:~/Documents/Monash/Semester-1/Data Sc
grep -io "#Hillary" Twitter_Data_1 | wc -l
      59
[roopak@192-168-1-26:~/Documents/Monash/Semester-1/Data Sc
grep -io "HillaryClinton" Twitter_Data_1 | wc -l
      95
```

Hillary Clinton is referred using terms like “HillaryClinton”,”#Hillary”. All these words have the term Hillary. So checking the occurrences of term “Hillary”. We get **281** such occurrences. If we compare occurrences of both Hillary Clinton and Donald Trump, we can conclude that **Donald Trump was more popular in Twitter for the given dataset**

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?

Though we have captured most of the occurrences of Donald Trump and Hillary Clinton, but some occurrences were not covered. These include non usage of complete name like "Don Trump",misspelled names or words that were used for hashtags like "#MrsClinton" etc.
Since the usage can not be generalised to get the actual count, we can only get the rough estimate only. Even if we search the occurrence of "Trump" or "Hillary" this will not be accurate since this could be in reference with other meanings.



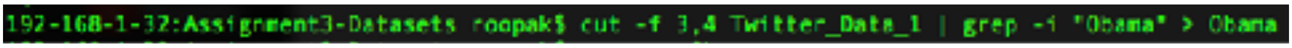
```
[roopak@192-168-1-26:~/Documents/Monash/Semester-1/Data Scienc
 cut -f 4 Twitter_Data_1 | grep -io "Hillary" | wc -l
      832
[roopak@192-168-1-26:~/Documents/Monash/Semester-1/Data Scienc
 cut -f 4 Twitter_Data_1 | grep -io "trump" | wc -l
     1147
```

Here trump reference can also be like "**@ABandquotes How many trumpets does it take to change a lightbulb?   One, but they have to stand on top of their ego to reach it ~TSL**"

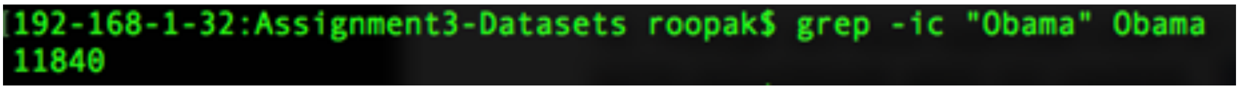
PART-B

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

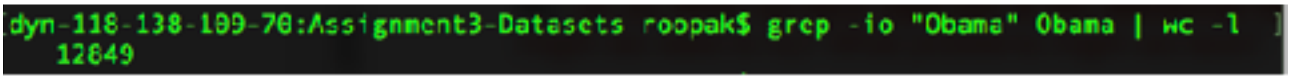
1. Number of tweets of Obama in the given dataset: **11840.** This can be found using grep command to search for the word "Obama" ignoring cases
   Number of occurrence of Obama in total file ignoring case are : **12849**



```
192-168-1-32:Assignment3-Datasets roopak$ cut -f 3,4 Twitter_Data_1 | grep -i "Obama" > Obama
```



```
[192-168-1-32:Assignment3-Datasets roopak$ grep -ic "Obama" Obama
11840
```



```
[dyn-118-138-199-70:Assignment3-Datasets roopak$ grep -io "Obama" Obama | wc -l  ]
   12849
```

2. 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?

The following format string will be used to convert the current time stamp column to parse it

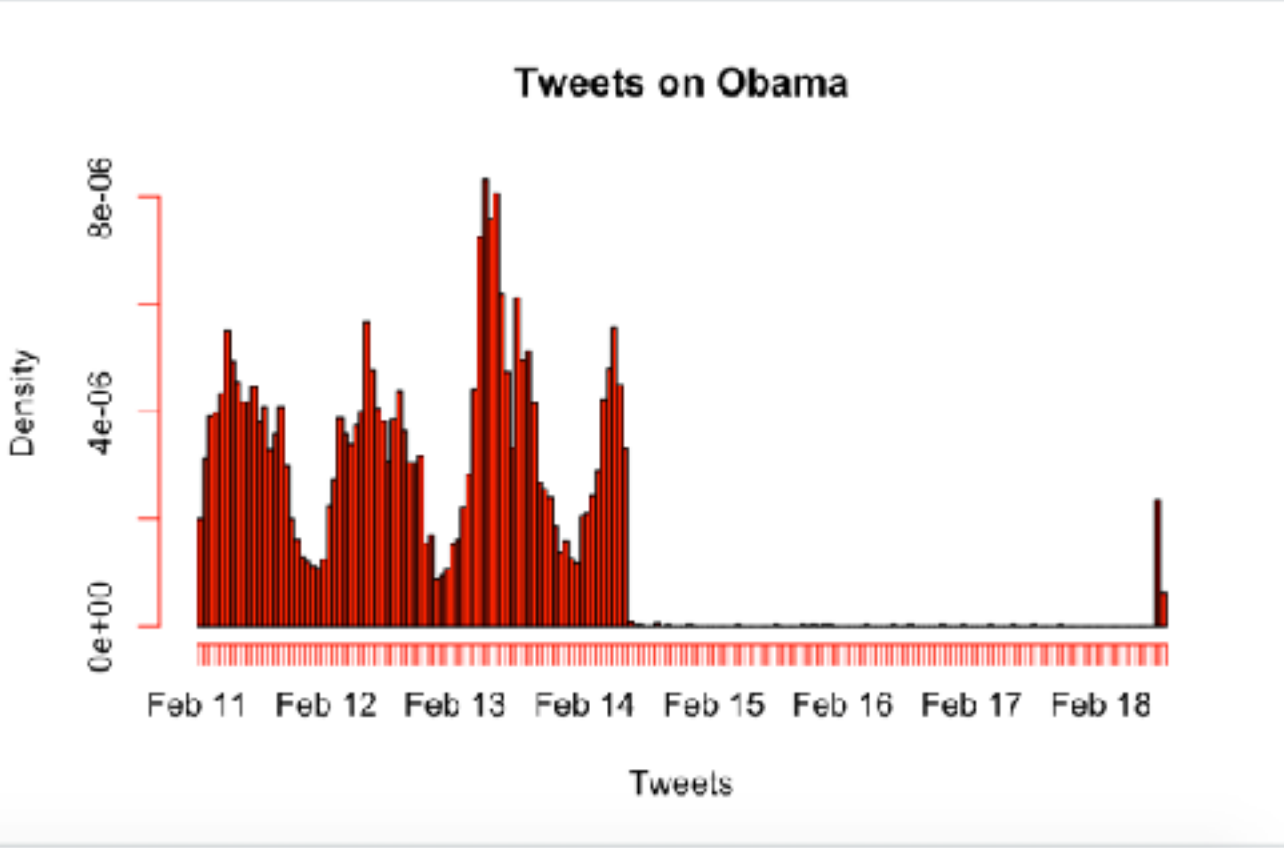
Column value example: Tue Feb 11 12:19:39 +0000 2014

A <-strptime(obama[1,], format = "%a %b %d %H:%M:%S %z %Y")

%a : Abbreviated weekday name
%b : Abbreviated month name
%d : Date of the month
%H : Hour in 24hr format
%M : Minute
%S : Second
%z : Timezone offset
%Y : Year in 4 digit format

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

```
A <-strptime(obama[1,], format = "%a %b %d %H:%M:%S %z %Y”)
hist(A, breaks = "hours", xlab = "Tweets", col = "red", main = "Tweets on Obama”)
```



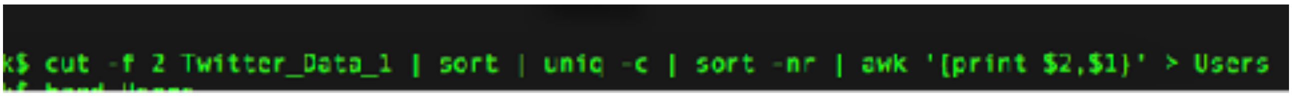
4. The plot has a bit of an unusual shape. Can you see a pattern before Feb 15 and what happens after that?

From the plot, it is clear that the tweets on “Obama” are very high from feb 11th to feb 14th and then it drastically reduces to almost zero (not exactly 0) from 15th to 18th. Again on 18th we can see that tweets are high. Even if we consider each day from 11th to 14th, tweet rises to peak by noon and again reduces by night, again rising in the next day noon.

5. Number of Tweets by each unique author in the Twitter file giving a file with two columns “user” and “twitter count”.

Following is the histogram representing tweet count for the number of users. Since most of the users have done only one tweet, which is very high comparing to other count and also the data is sorted in reverse order, we will get this plot.

Terminal Code: **cut -f 2 Twitter_Data_1 | sort | uniq -c | sort -nr | awk ‘{print $2,$1}’ > Users**

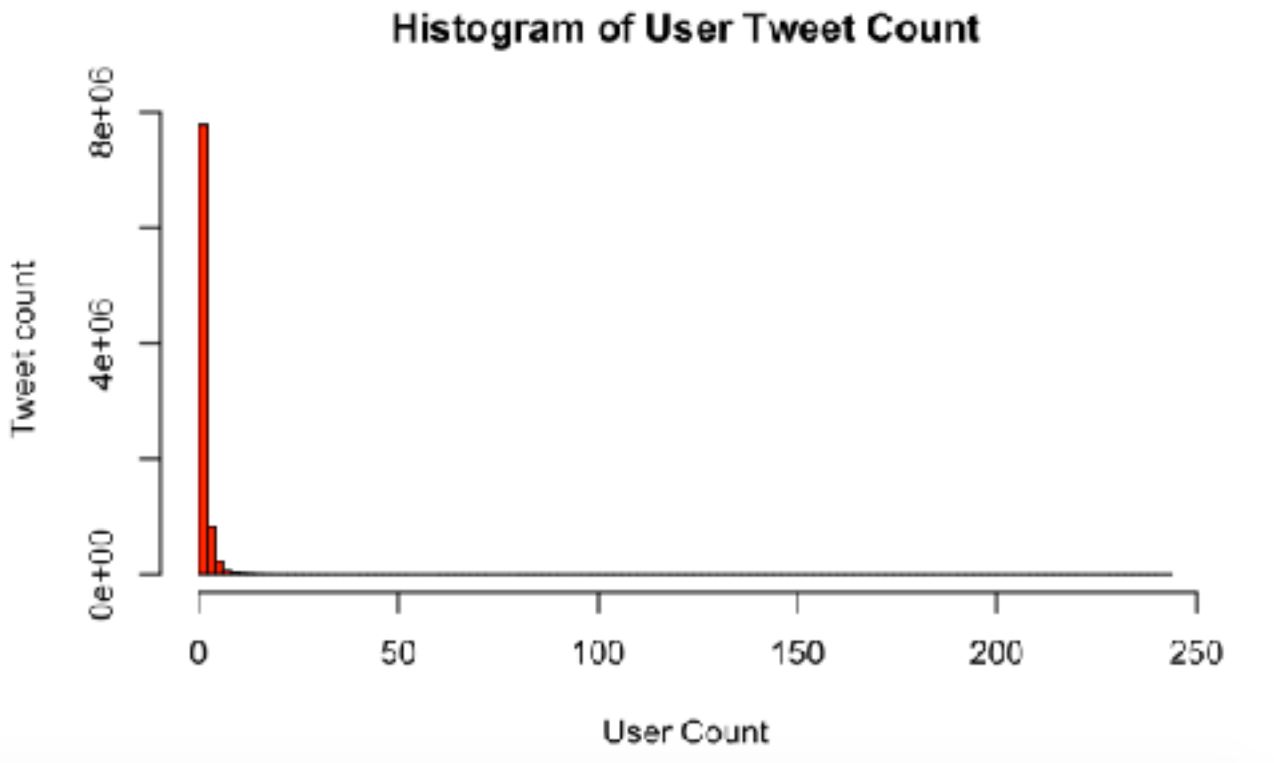


```
k$ cut -f 2 Twitter_Data_1 | sort | uniq -c | sort -nr | awk '{print $2,$1}' > Users
```

R Code:

> user = read.csv("/Users/roopak/Documents/Monash/Semester-1/Data Science/Assignment3-Datasets/Users", sep = " ", header = FALSE)

> hist(user$V2, breaks = 100, col = 'red', xlab = "User Count", ylab = "Tweet count", main = "Histogram of User Tweet Count")

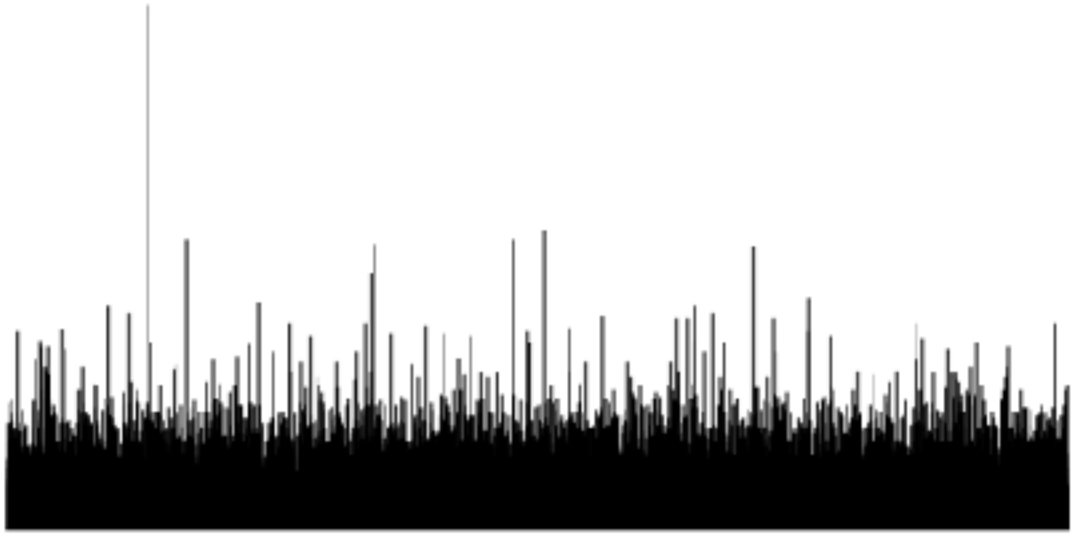


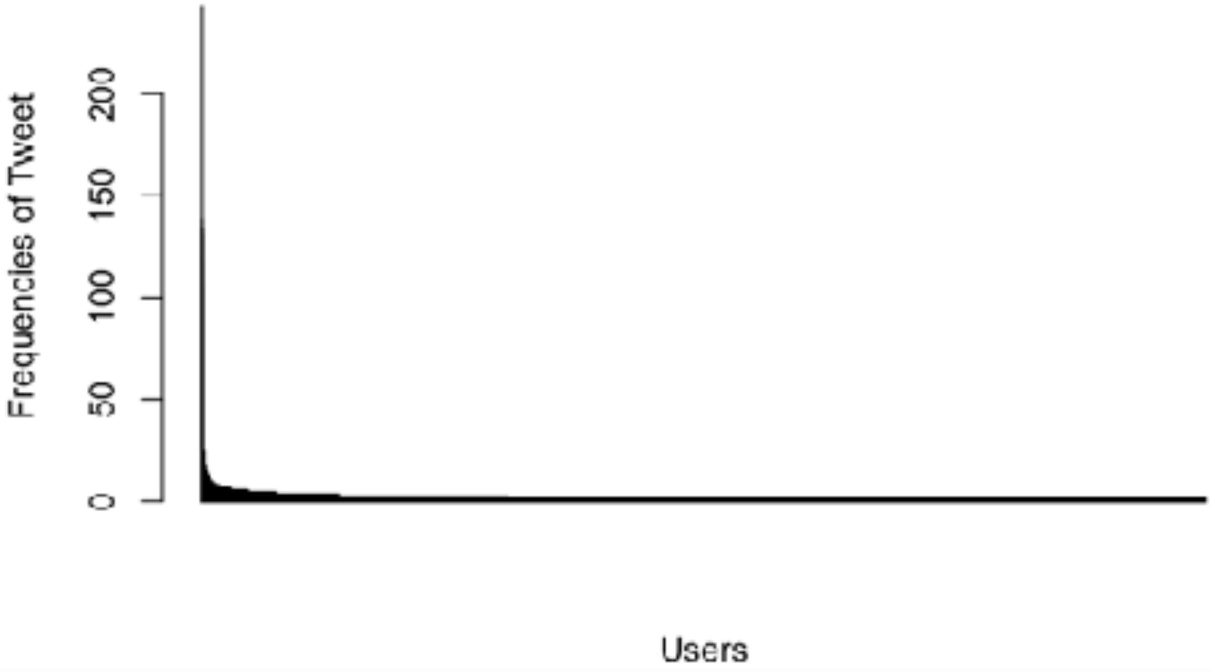
Plot a second histogram, but this time showing the distribution over number of tweets per author in the file.

R code to get the frequency of tweets per user

```
> mp <- barplot(user$V2, beside = FALSE, col = "green",xlab = "Users", ylab = "Frequencies of Tweet", main = "Tweets per User")
> axis(1,at=mp,labels=user$V1)
```

Plotting the unsorted list first. Then we will try the same after sorting





Since the data was sorted in descending order according to the frequency of tweets, Due to large number of users, the x-ticks are not visible in this(which was hanging the system)

If we try to reduce the size of the data, considering just first 200 data, we can plot the histogram for the same using the following command:

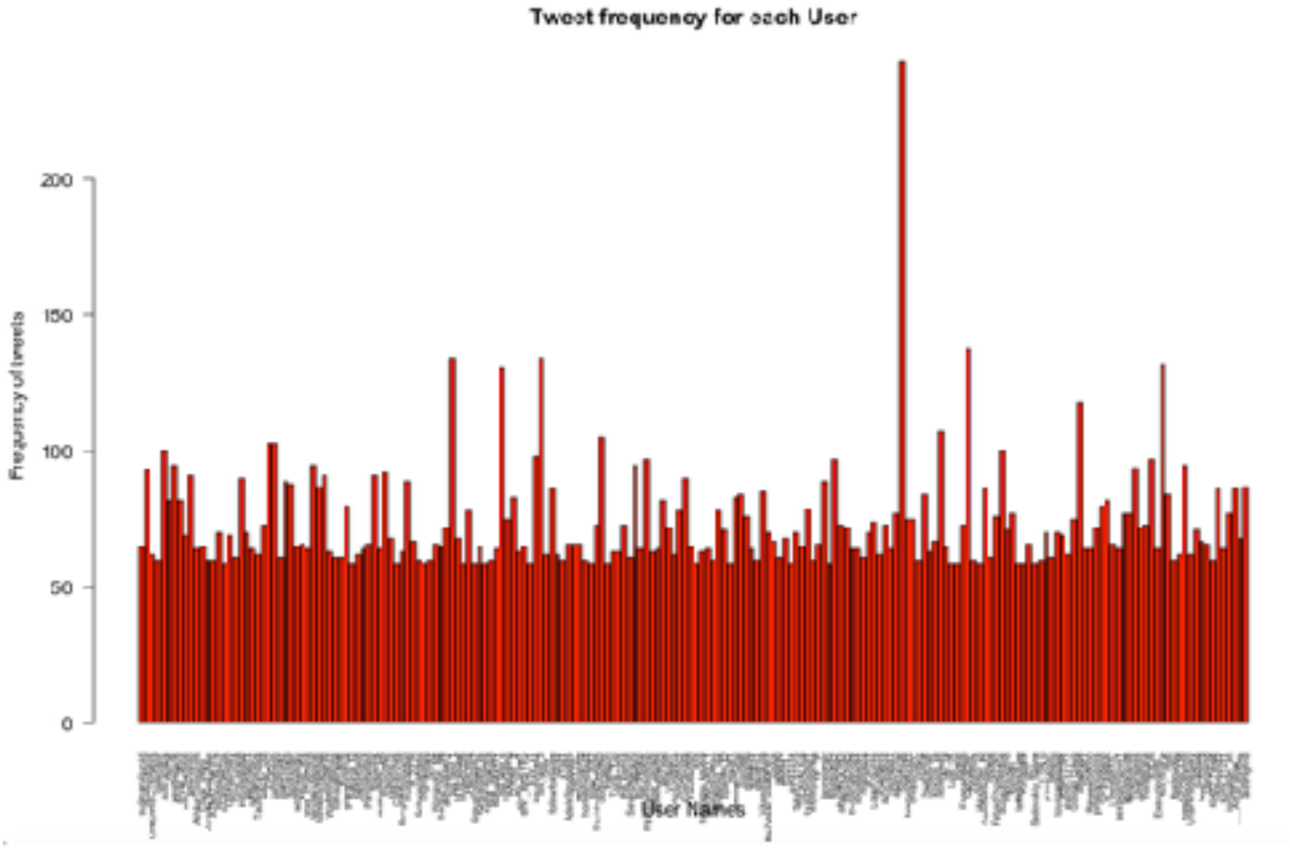
Terminal command to get first 200 values : **head -200 Users_test | sort -R > Users_testb**

R Code : First read the file

user = read.csv("/Users/roopak/Documents/Monash/Semester-1/Data Science/Assignment3-Datasets/Users_testb", sep = " ", header = FALSE)

Plot the bar chart of names and frequencies.

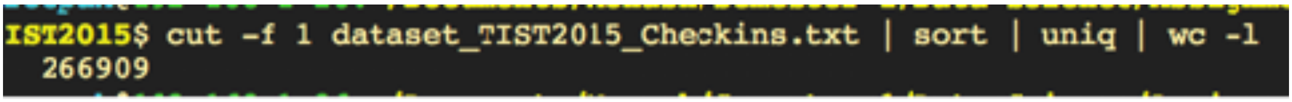
barplot(user$V2, names.arg = user$V1, xlab = "User Names", ylab = "Frequency of tweets", col = "red", main = "Tweet frequency for each User", cex.names=0.5, las=2)



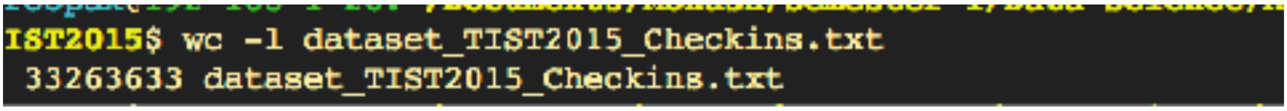
PART-C

1. Open the zipfile 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?Since the first column in dataset_TIST2015_Checkins is userId, so doing a unique to that gives the number of unique users. So there are **266909** unique users

Taking the line count of dataset_TIST2015_Checkins will give the number of login, since each rows in the dataset represents a separate login. So there are **33263633** separate logins



```
IST2015$ cut -f 1 dataset_TIST2015_Checkins.txt | sort | uniq | wc -l
  266909
```



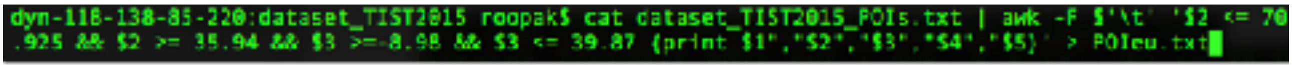
```
IST2015$ wc -l dataset_TIST2015_Checkins.txt
 33263633 dataset_TIST2015_Checkins.txt
```

2. Create an awk script to create a European subset of the POI file, and name the subset file "POIeu.txt".

AWK script:

From the graph of Europe Topmost latitude is 70.925 and southernmost latitude is 35.946883 and rightmost longitude is 39.871697 and leftmost longitude is -8.987678. So using this to filter Europe

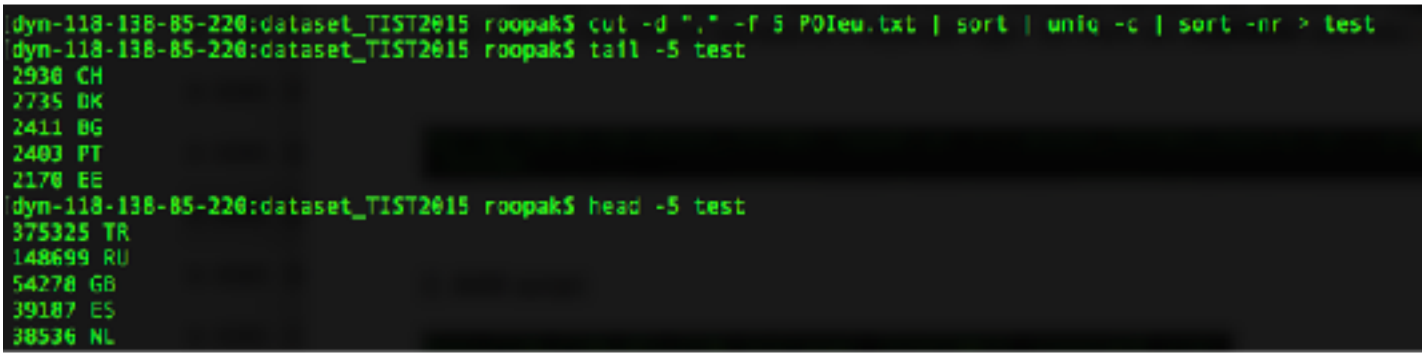
cat dataset_TIST2015_POIs.txt | awk -F $'\t' '$2 <= 70.925 && $2 >= 35.94 && $3 >=-8.98 && $3 <= 39.87 {print $1","$2","$3","$4","$5}' > POIeu.txt



```
dyn-118-138-85-220:dataset_TIST2015 roopak$ cat dataset_TIST2015_POIs.txt | awk -F $'\t' '$2 <= 70
.925 && $2 >= 35.94 && $3 >=-8.98 && $3 <= 39.87 {print $1","$2","$3","$4","$5}' > POIeu.txt
```

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

.Country code with '**TR'(TURKEY) has the highest(375325)** places of POI and country code with **'EE'(ESTONIA) has the least(2170) POI**

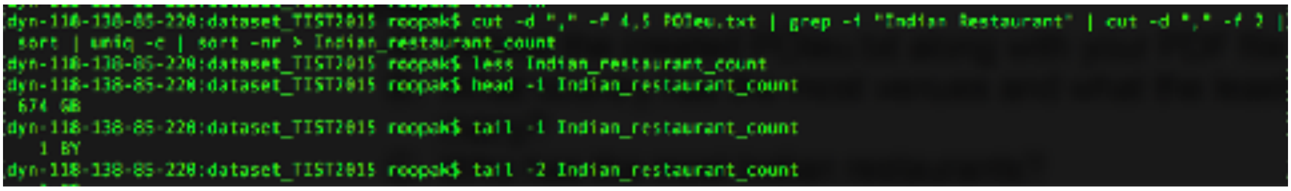


```
dyn-118-138-85-220:dataset_TIST2015 roopak$ cut -d "," -f 5 POIeu.txt | sort | uniq -c | sort -nr > test
dyn-118-138-85-220:dataset_TIST2015 roopak$ tail -5 test
2930 CH
2735 DK
2411 BG
2403 PT
2170 EE
dyn-118-138-85-220:dataset_TIST2015 roopak$ head -5 test
375325 TR
148699 RU
54278 GB
39187 ES
38536 NL
```

C. Who has the most Indian restaurants?

Command to get country with most Indian Restaurants

cut -d "," -f 4,5 POIeu.txt | grep -i "Indian Restaurant" | cut -d "," -f 2 | sort | uniq -c | sort -nr > Indian_restaurant_count

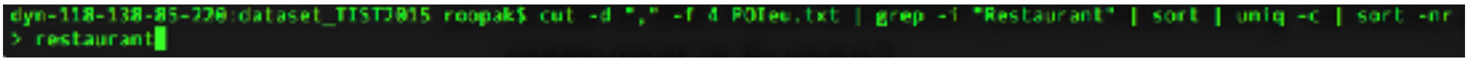


```
dyn-118-138-85-220:dataset_TIST2015 roopak$ cut -d "," -f 4,5 POIeu.txt | grep -i "Indian Restaurant" | cut -d "," -f 2 |
 sort | uniq -c | sort -nr > Indian_restaurant_count
dyn-118-138-85-220:dataset_TIST2015 roopak$ less Indian_restaurant_count
dyn-118-138-85-220:dataset_TIST2015 roopak$ head -1 Indian_restaurant_count
 674 GB
dyn-118-138-85-220:dataset_TIST2015 roopak$ tail -1 Indian_restaurant_count
   1 BY
dyn-118-138-85-220:dataset_TIST2015 roopak$ tail -2 Indian_restaurant_count
```

So Country with code “**GB**”(UNITED KINGDOM) has the maximum Indian Restaurants

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

Restaurants are the most common POIs but this does not belong to a particular class, so we can consider the next one which is Turkish restaurants.
Turkish restaurants are present in **9965** venues making it the most common class of restaurant in Europe



```
dyn-118-138-85-220:dataset_TIST2015 roopak$ cut -d "," -f 4 POIeu.txt | grep -i "Restaurant" | sort | uniq -c | sort -nr
> restaurant
```