FIT 5145

Assessment 3

Mukul Gupta

29873150

TABLE OF CONTENTS

FIT5145 Assessment - 3

Part A: Investigating the Twitter Data in the Shell

Question 1: Decompress the Twitter_Data_1.gz file. How big is it?

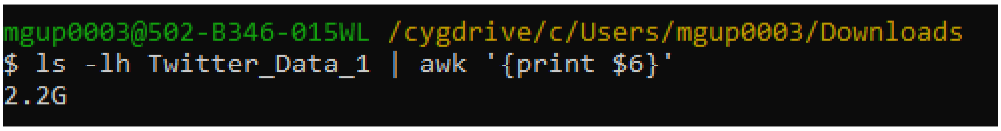
Ans: File is 2.2 GB

Shell command:

ls -lh Twitter_Data_1

OR

ls -lh Twitter_Data_1 | awk '{print $6}'

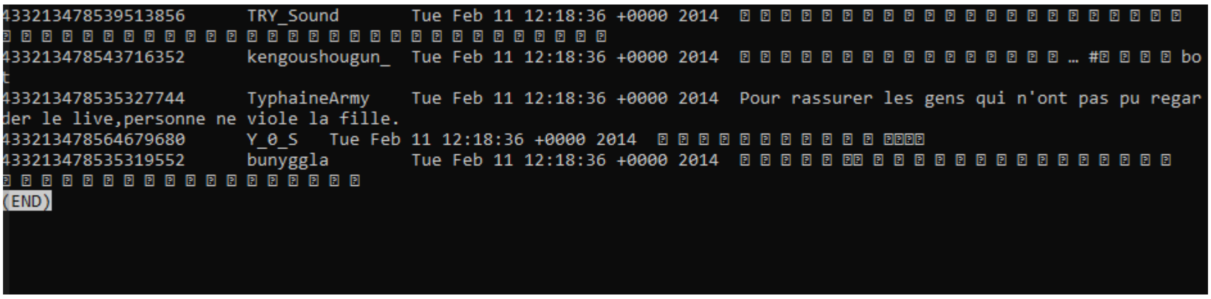


```
mgup0003@502-B346-015WL /cygdrive/c/Users/mgup0003/Downloads
$ ls -lh Twitter_Data_1 | awk '{print $6}'
2.2G
```

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

Ans: Delimiter used in '\t'. Columns are tab separated. There are 4 columns

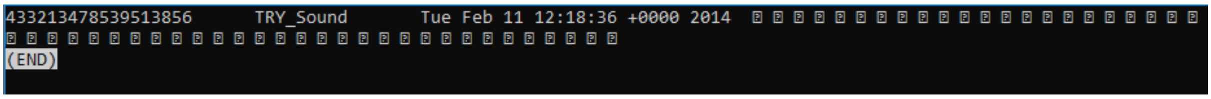
Shell command: head -5 Twitter_Data_1 | less



```
433213478539513856      TRY_Sound       Tue Feb 11 12:18:36 +0000 2014  □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □
□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □
433213478543716352      kengoushougun_  Tue Feb 11 12:18:36 +0000 2014  □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ … #□ □ □ □ bo
t
433213478535327744      TyphaineArmy    Tue Feb 11 12:18:36 +0000 2014  Pour rassurer les gens qui n'ont pas pu regar
der le live,personne ne viole la fille.
433213478564679680      Y_0_S   Tue Feb 11 12:18:36 +0000 2014  □ □ □ □ □ □ □ □ □ □ □ □□□□
433213478535319552      bunyggla        Tue Feb 11 12:18:36 +0000 2014  □ □ □ □ □ □□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □
□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □
(END)
```

There seems to be uneven spacing, which suggests it may be a tab character.

Shell command: head -1 Twitter_Data_1 | less



```
433213478539513856      TRY_Sound       Tue Feb 11 12:18:36 +0000 2014  □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □
□ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ □
(END)
```

All the tabs in the line light up by using /<tab>. Tab is indeed the delimiter of this file.

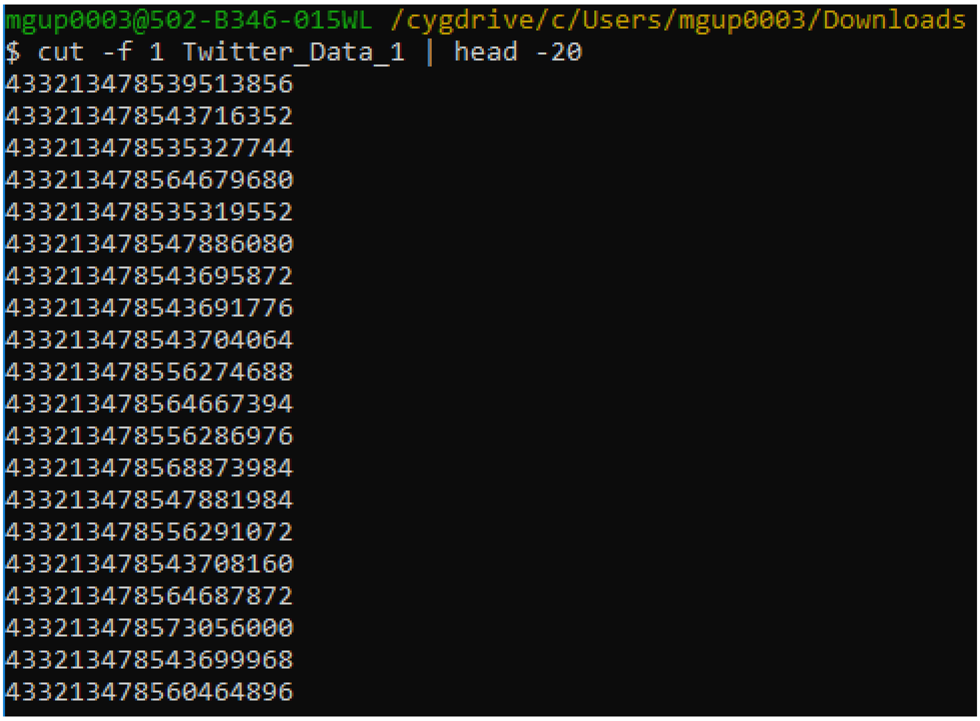
Question 3: The first column is a unique identifier for a Tweet. What are the other columns?

Ans: After running the following shell commands, it can be said:

- First column is the unique identifier of the tweet
- Second column is the username of the user who wrote the tweet

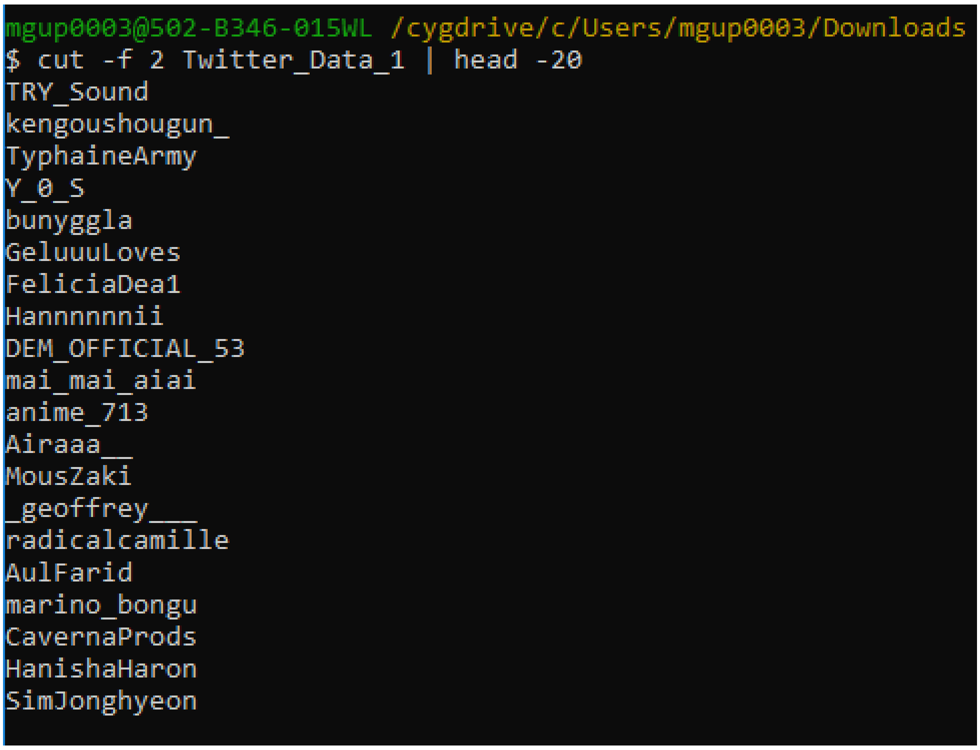
- Third column is the time when the tweet was written
- Fourth column contains the text of the tweet. It is in different languages.

Shell command: cut -f 1 Twitter_Data_1 | head -20



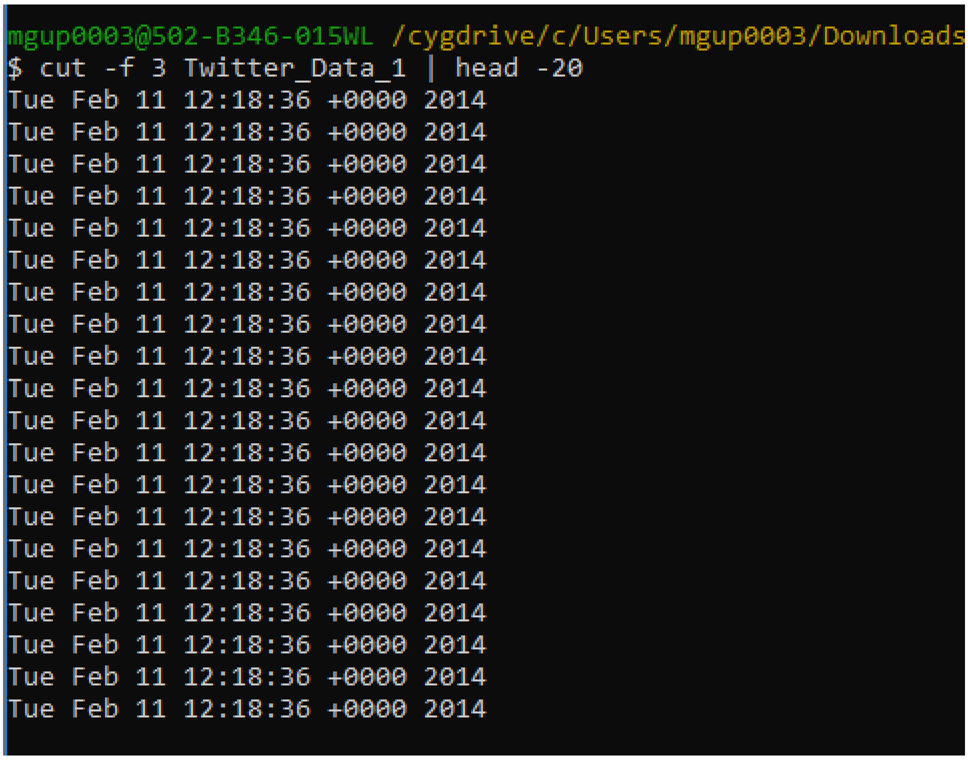
```
mgup0003@502-B346-015WL /cygdrive/c/Users/mgup0003/Downloads
$ cut -f 1 Twitter_Data_1 | head -20
433213478539513856
433213478543716352
433213478535327744
433213478564679680
433213478535319552
433213478547886080
433213478543695872
433213478543691776
433213478543704064
433213478556274688
433213478564667394
433213478556286976
433213478568873984
433213478547881984
433213478556291072
433213478543708160
433213478564687872
433213478573056000
433213478543699968
433213478560464896
```

Shell command: cut -f 2 Twitter_Data_1 | head -20



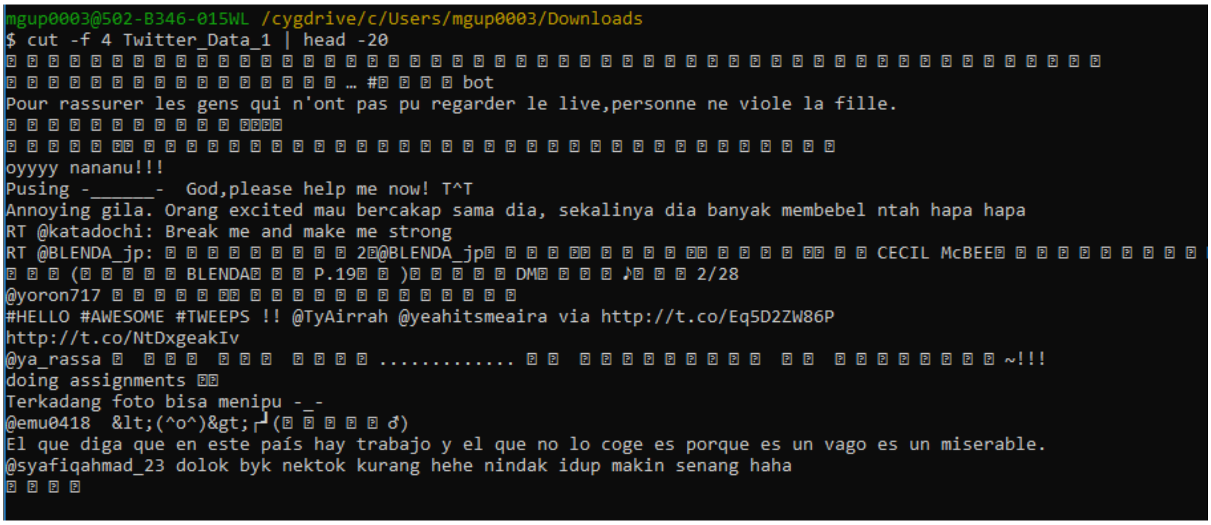
```
mgup0003@502-B346-015WL /cygdrive/c/Users/mgup0003/Downloads
$ cut -f 2 Twitter_Data_1 | head -20
TRY_Sound
kengoushougun_
TyphaineArmy
Y_0_S
bunyggla
GeluuuLoves
FeliciaDea1
Hannnnnnii
DEM_OFFICIAL_53
mai_mai_aiai
anime_713
Airaaa__
MousZaki
_geoffrey___
radicalcamille
AulFarid
marino_bongu
CavernaProds
HanishaHaron
SimJonghyeon
```

Shell command: cut -f 3 Twitter_Data_1 | head -20



```
mgup0003@502-B346-015WL /cygdrive/c/Users/mgup0003/Downloads
$ 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
Tue Feb 11 12:18:36 +0000 2014
```

Shell command: cut -f 4 Twitter_Data_1 | head -20

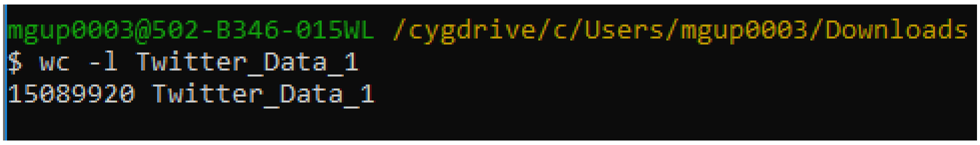


```
mgup0003@502-B346-015WL /cygdrive/c/Users/mgup0003/Downloads
$ cut -f 4 Twitter_Data_1 | head -20
▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯
▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ … #▯ ▯ ▯ ▯ bot
Pour rassurer les gens qui n'ont pas pu regarder le live,personne ne viole la fille.
▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯▯▯▯
▯ ▯ ▯ ▯ ▯ ▯▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯
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 @BLENDA_jp: ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ 2▯@BLENDA_jp▯ ▯ ▯ ▯ ▯▯ ▯ ▯ ▯ ▯ ▯▯ ▯ ▯ ▯ ▯ ▯▯ ▯ ▯ CECIL McBEE▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯
▯ ▯ ▯ (▯ ▯ ▯ ▯ ▯ BLENDA▯ ▯ ▯ P.19▯ ▯ )▯ ▯ ▯ ▯ ▯ DM▯ ▯ ▯ ▯ ♪▯ ▯ ▯ 2/28
@yoron717 ▯ ▯ ▯ ▯ ▯ ▯▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯
#HELLO #AWESOME #TWEEPS !! @TyAirrah @yeahitsmeaira via http://t.co/Eq5D2ZW86P
http://t.co/NtDxgeakIv
@ya_rassa ▯  ▯ ▯ ▯  ▯ ▯ ▯  ▯ ▯ ▯ ▯ ............. ▯ ▯  ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯  ▯ ▯  ▯ ▯ ▯ ▯ ▯ ▯ ▯ ▯ ~!!!
doing assignments ▯▯
Terkadang foto bisa menipu -_-
@emu0418  &lt;(^o^)&gt;┌┘(▯ ▯ ▯ ▯ ▯ ♂)
El que diga que en este país hay trabajo y el que no lo coge es porque es un vago es un miserable.
@syafiqahmad_23 dolok byk nektok kurang hehe nindak idup makin senang haha
▯ ▯ ▯ ▯
```

Question 4: How many Tweets are there in the file?

Ans: 15089920 tweets

Shell command: wc -l Twitter_Data_1



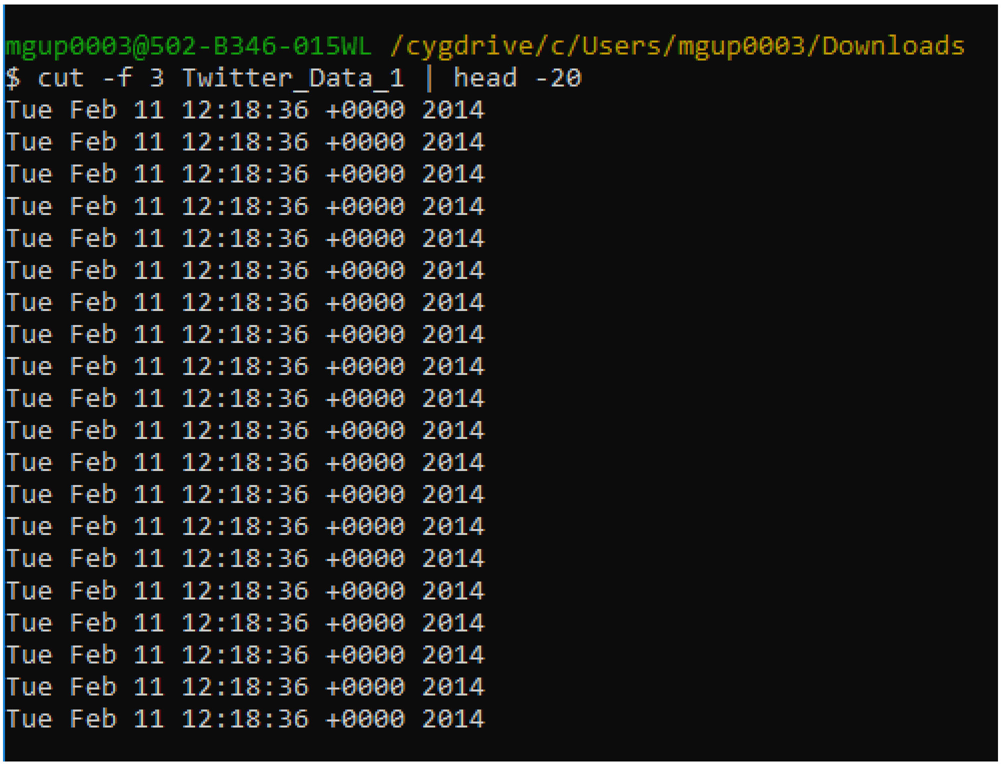
```
mgup0003@502-B346-015WL /cygdrive/c/Users/mgup0003/Downloads
$ wc -l Twitter_Data_1
15089920 Twitter_Data_1
```

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

Ans: Range is 8 days from 11 February 2014 to 18 February 2014

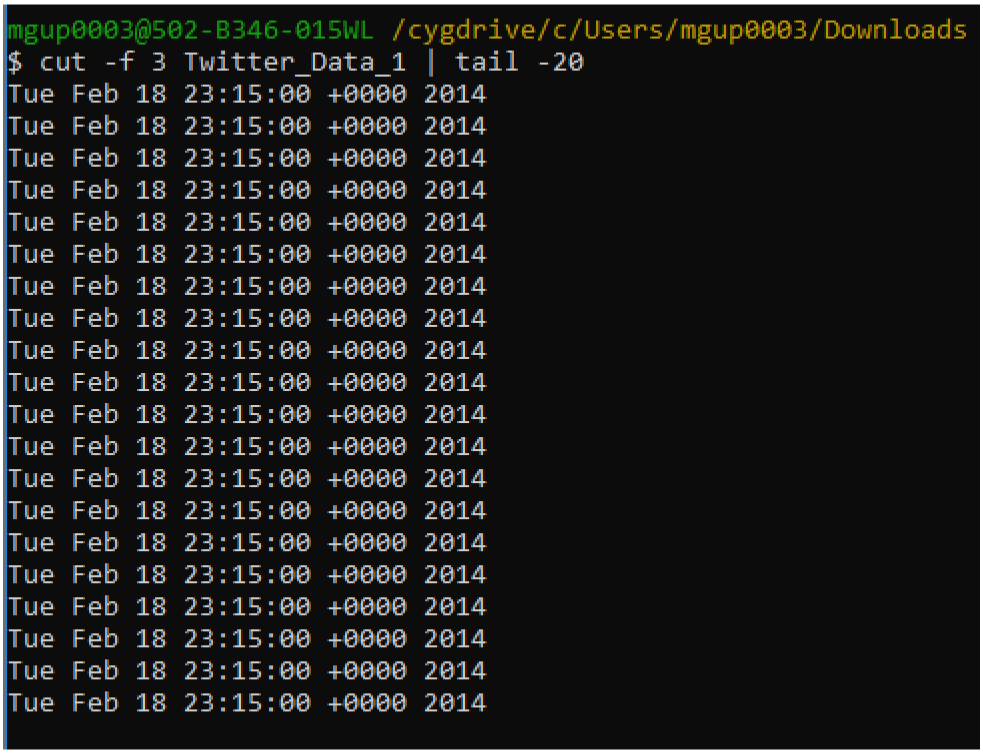
Found by reading the head and tail of the tweet time column

Shell command: cut -f 3 Twitter_Data_1 | head -20



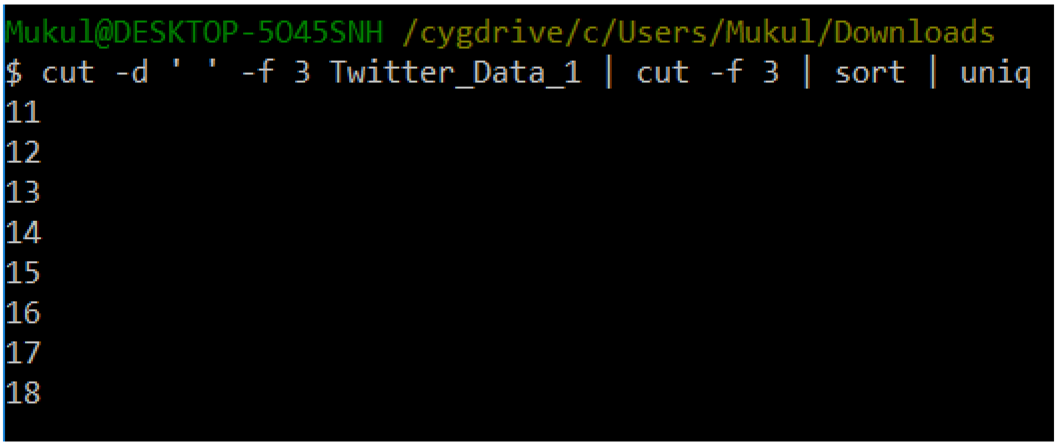
```
mgup0003@502-B346-015WL /cygdrive/c/Users/mgup0003/Downloads
$ cut -f 3 Twitter_Data_1 | head -20
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +0000 2014
Tue Feb 11 12:18:36 +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
Tue Feb 11 12:18:36 +0000 2014
```

Shell command: cut -f 3 Twitter_Data_1 | tail -20



```
mgup0003@502-B346-015WL /cygdrive/c/Users/mgup0003/Downloads
$ cut -f 3 Twitter_Data_1 | tail -20
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
Tue Feb 18 23:15:00 +0000 2014
```

To ensure the dates, I cut Twitter data timestamp column with space (‘ ‘) and sorted the dates. The date range was from 11 to 18.

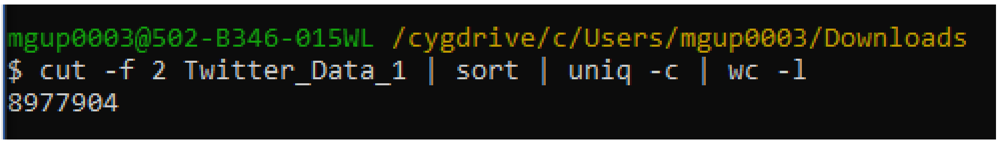


```
Mukul@DESKTOP-5O45SNH /cygdrive/c/Users/Mukul/Downloads
$ cut -d ' ' -f 3 Twitter_Data_1 | cut -f 3 | sort | uniq
11
12
13
14
15
16
17
18
```

Question 6: How many unique users are there?

Ans: 8977904 unique users.

Shell command: cut -f 2 Twitter_Data_1 | sort | uniq -c | wc –l



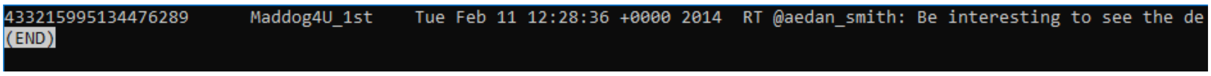
```
mgup0003@502-B346-015WL /cygdrive/c/Users/mgup0003/Downloads
$ cut -f 2 Twitter_Data_1 | sort | uniq -c | wc -l
8977904
```

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

Ans: First mention in the file of “Donald Trump” was on Tue Feb 11 12:28:36 +0000 2014.

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…

Shell command: grep -m1 -i "Donald Trump" Twitter_Data_1 | less -s

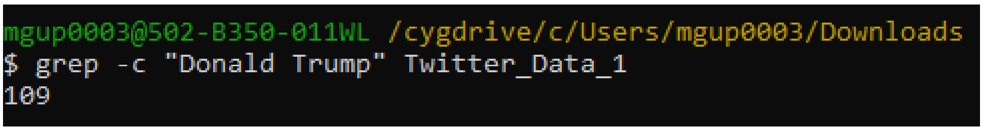


```
433215995134476289      Maddog4U_1st    Tue Feb 11 12:28:36 +0000 2014  RT @aedan_smith: Be interesting to see the de
(END)
```

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

Ans: Number of tweets in which “Donald Trump” was mentioned are 109. I have assumed that I have to count the tweets in which “Donald Trump” was mentioned and not the occurrences of “Donald Trump” keyword which can multiple in a tweet. I have used –c to count the number of times. This is case sensitive.

Shell command: grep -c "Donald Trump" Twitter_Data_1

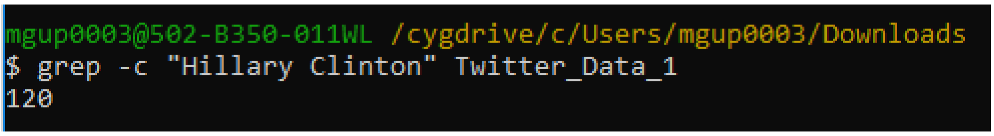


```
mgup0003@502-B350-011WL /cygdrive/c/Users/mgup0003/Downloads
$ grep -c "Donald Trump" Twitter_Data_1
109
```

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

Ans: Number of tweets in which “Hillary Clinton” was mentioned are 109. I have assumed that I have to count the tweets in which “Hillary Clinton” was mentioned and not the occurrences of “Hillary Clinton” keyword which can multiple in a tweet. I have used –c to count the number of times. This is case sensitive.

Shell command: grep -c "Hillary Clinton" Twitter_Data_1



```
mgup0003@502-B350-011WL /cygdrive/c/Users/mgup0003/Downloads
$ grep -c "Hillary Clinton" Twitter_Data_1
120
```

As per the number of tweets, Hillary Clinton (120) is more popular than Donald Trump (109). So, it can be said Hillary Clinton is more popular than Donald Trump.

Question 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?

Ans: No we have not captured all references to Donald and Hillary. We have only searched the tweets for the words “Hillary Clinton” and “Donald Trump” which are case sensitive. This means words like “hillary clinton” or “donald trump” would be ignored. People may refer to Hillary Clinton in their tweets with similar words like Hillary Diane Rodham Clinton or Mrs Clinton or just Hillary including others. Some people may refer to Donald Trump in their tweets as Mr Trump or Agent Orange or Donald Chump including others.

We can try using “Trump”, “Hillary” as strings. But there can be instances in which people are not referring to Hillary Clinton and Donald Trump specifically but some other individuals.

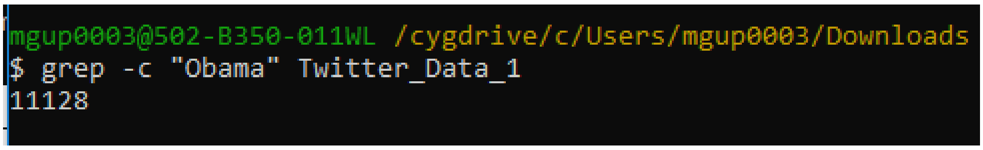
It is hard to tell whom people are referring to in their tweets as they can come with their own nicknames. Also, there may be spelling mistakes in the tweets. Language can also differ in various regions around the globe. People may also refer to other individuals and not Hillary Clinton and Donald Trump.

Part B: Graphing the Data in R

Question 1. How many times does the term ‘Obama’ appear in tweets?

Ans: “Obama” was mentioned 11128 tweets. I have assumed that I have to count the tweets in which Obama was mentioned and not the occurrences of Obama keyword which can multiple in a tweet.

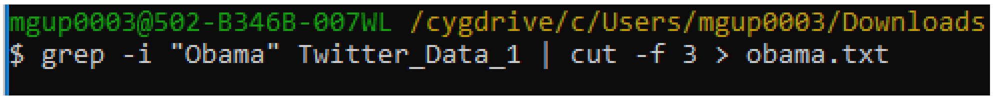
Shell command: grep -c "Obama" Twitter_Data_1



```
mgup0003@502-B350-011WL /cygdrive/c/Users/mgup0003/Downloads
$ grep -c "Obama" Twitter_Data_1
11128
```

Question 2. 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.

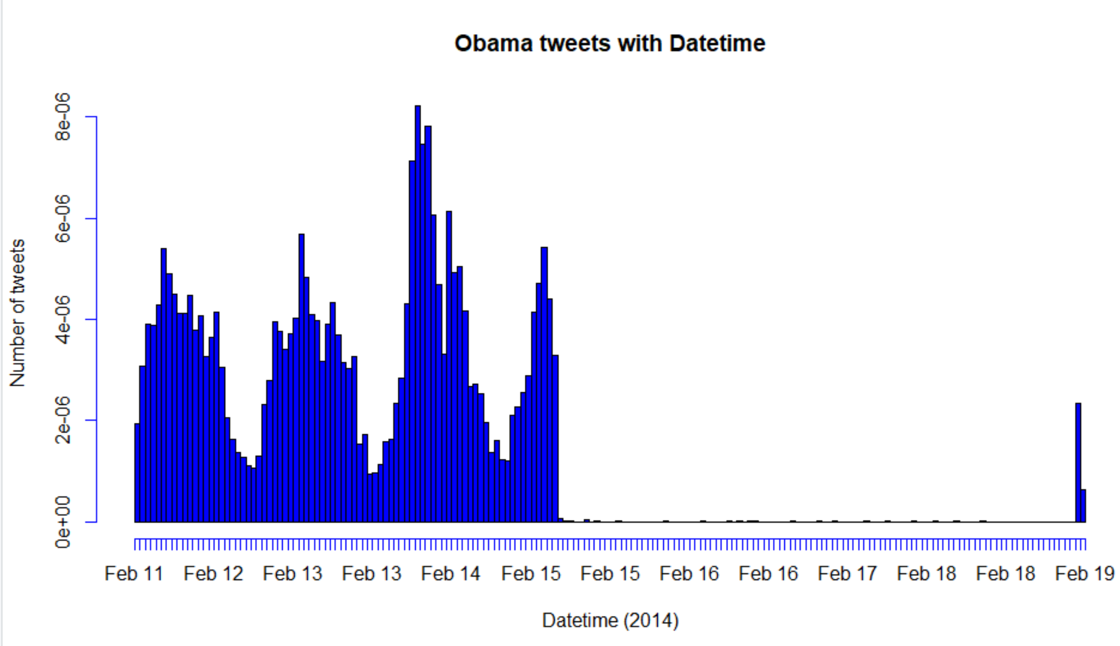
Ans: Shell command: grep -i "Obama" Twitter_Data_1 | cut -f 3 > obama.txt



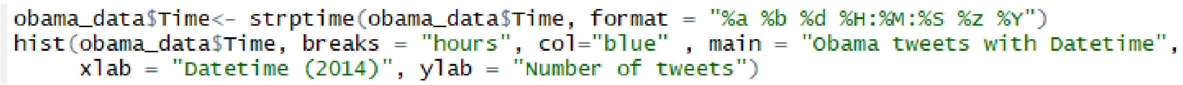
```
mgup0003@502-B346B-007WL /cygdrive/c/Users/mgup0003/Downloads
$ grep -i "Obama" Twitter_Data_1 | cut -f 3 > obama.txt
```

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

Ans:



R code:



```
obama_data$Time<- strptime(obama_data$Time, format = "%a %b %d %H:%M:%S %z %Y")
hist(obama_data$Time, breaks = "hours", col="blue" , main = "Obama tweets with Datetime",
     xlab = "Datetime (2014)", ylab = "Number of tweets")
```

obama_data<- read.csv("C:\\Users\\mgup0003\\Downloads\\obama.txt", fill = TRUE, stringsAsFactors = FALSE)

obama_data$Time<- strptime(obama_data$Time, format = "%a %b %d %H:%M:%S %z %Y")

hist(obama_data$Time, breaks = "hours", col="blue" , main = "Obama tweets with Datetime",

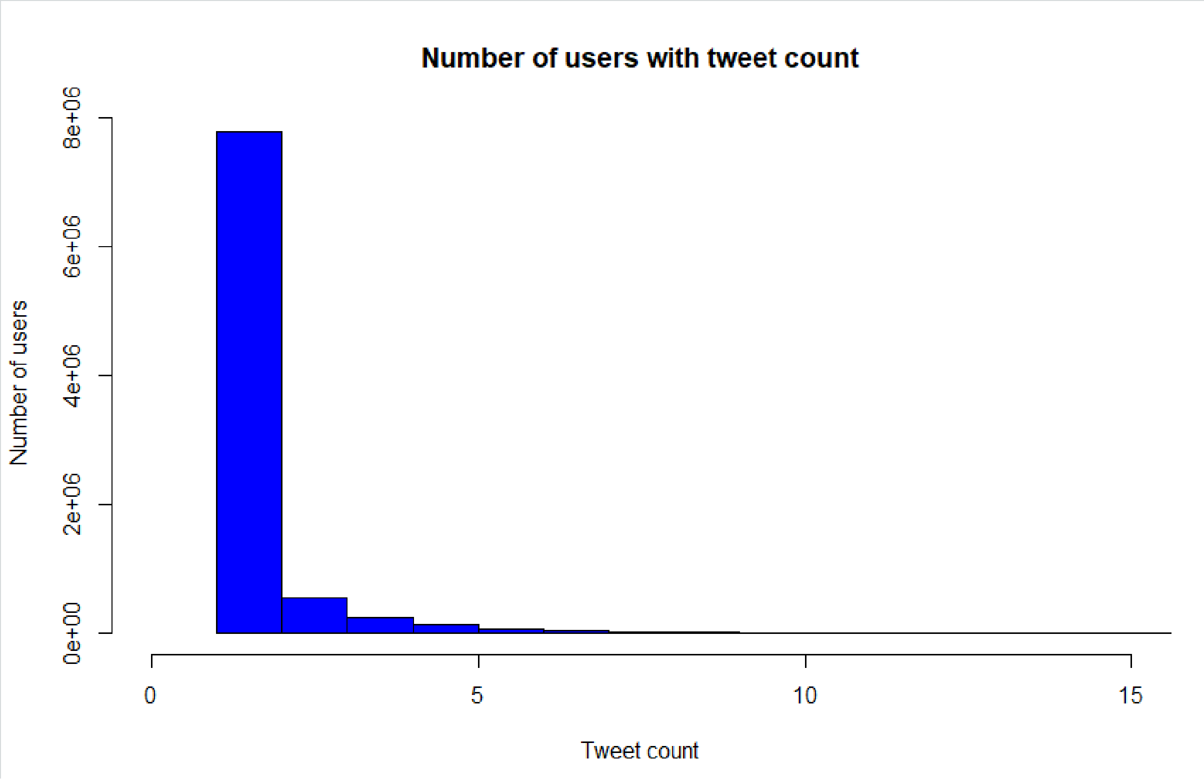
xlab = "Datetime (2014)", ylab = "Number of tweets")

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

Ans: Before Feb 15, Obama is mentioned a lot of times in the tweets. After that, there was a sudden drop in tweets mentioning Obama. On 18th February, there was again mention of Obama in the tweets.

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

Ans: Large number of people has tweeted once. For tweets more than 1, number of tweets per person decreases.



Shell code:

```
cut -f 2 Twitter_Data_1 | sort | uniq -c > tweets_author.txt
```

```
cut -f 2 Twitter_Data_1 | sort | uniq -c | sed -e 's/^[ \t]*//' > tweets_author.txt
```

(Karthik, 2015)

R code:

```
tweets_per_author<- read.table("C:\\Users\\mgup0003\\Downloads\\tweets_author.txt", sep = " ", fill = TRUE, stringsAsFactors = FALSE)
```

```
tweets_per_author_final<- tweets_per_author[complete.cases(tweets_per_author), ]
```

hist(tweets_per_author_final$V1, breaks = 200, col="blue" , main = "Number of users with tweet count",xlab = "Tweet count", ylab = "Number of users", xlim = c(0,15))

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

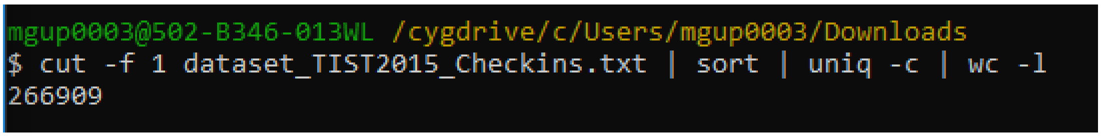
Question 1: Open the zipfile and have a look at the files it contains. One is a readme file giving the metadata. One is a log of user check-ins. How many check-ins are there and how many users?

Ans:

Shell code:unzip dataset_TIST2015.zip

There are 266,909 users according to the code and the readme file.

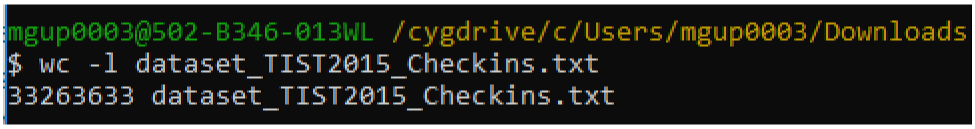
Shell code: cut -f 1 dataset_TIST2015_Checkins.txt | sort | uniq -c | wc -l



```
mgup0003@502-B346-013WL /cygdrive/c/Users/mgup0003/Downloads
$ cut -f 1 dataset_TIST2015_Checkins.txt | sort | uniq -c | wc -l
266909
```

Shell code: cut -f 1 dataset_TIST2015_Checkins.txt | sort | uniq -c | wc -l

There are 33,263,633 check-ins according to the code but there are .33,278,683 check-ins according to the readme file.



```
mgup0003@502-B346-013WL /cygdrive/c/Users/mgup0003/Downloads
$ wc -l dataset_TIST2015_Checkins.txt
33263633 dataset_TIST2015_Checkins.txt
```

Question 2.

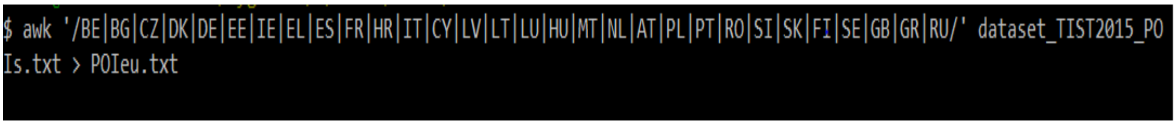
Question A. Submit the created POIeu.txt along with your PDF file

Ans: File is attached.

I have taken Russia (RU) in Europe as major part of Russia (77% of the population) is in Europe.

Shell code:

awk '/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|GB|GR|RU/' dataset_TIST2015_POIs.txt > POIeu.txt



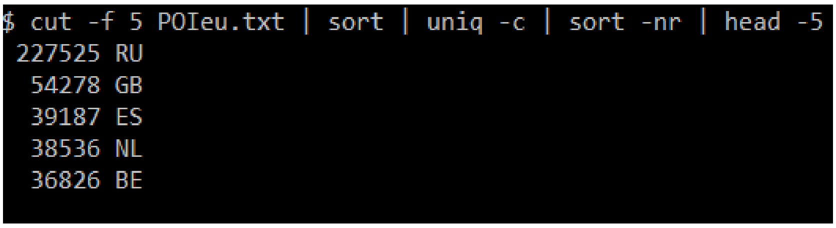
```
$ awk '/BE|BG|CZ|DK|DE|EE|IE|EL|ES|FR|HR|IT|CY|LV|LT|LU|HU|MT|NL|AT|PL|PT|RO|SI|SK|FI|SE|GB|GR|RU/' dataset_TIST2015_PO
Is.txt > POIeu.txt
```

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

Ans: RU (Russia) has the most venues. EE (Estonia) has the least venues.

Shell code to see which country has most venues:

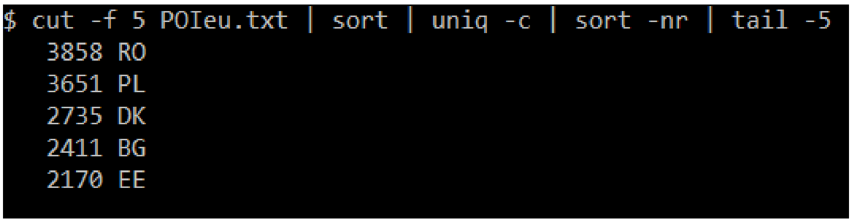
cut -f 5 POIeu.txt | sort | uniq -c | sort -nr | head -5



```
$ cut -f 5 POIeu.txt | sort | uniq -c | sort -nr | head -5
 227525 RU
  54278 GB
  39187 ES
  38536 NL
  36826 BE
```

Shell code to see which country has least venues:

cut -f 5 POIeu.txt | sort | uniq -c | sort -nr | tail -5

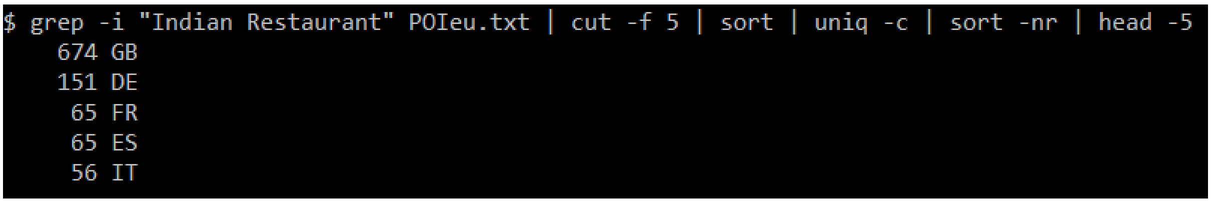


```
$ cut -f 5 POIeu.txt | sort | uniq -c | sort -nr | tail -5
   3858 RO
   3651 PL
   2735 DK
   2411 BG
   2170 EE
```

Question C. Who has the most Indian restaurants?

Ans: GB (Great Britain) has the most Indian restaurants.

Shell code: grep -i "Indian Restaurant" POIeu.txt | cut -f 5 | sort | uniq -c | sort -nr | head -5



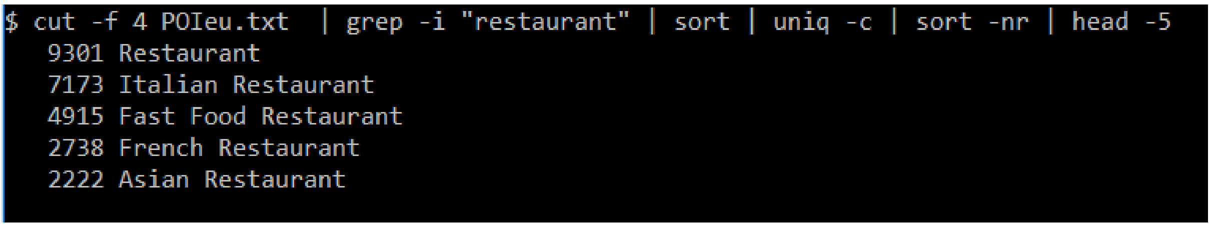
```
$ grep -i "Indian Restaurant" POIeu.txt | cut -f 5 | sort | uniq -c | sort -nr | head -5
    674 GB
    151 DE
     65 FR
     65 ES
     56 IT
```

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

Ans: Most common class of restaurant in Europe is: Italian Restaurant

I have assumed that the Restaurant class which is appearing with the most type of restaurants is uncategorical. Therefore, it can be ignored. That is why, Italian Restaurant is most common.

Shell code: cut -f 4 POIeu.txt | grep -i "restaurant" | sort | uniq -c | sort -nr | head -5

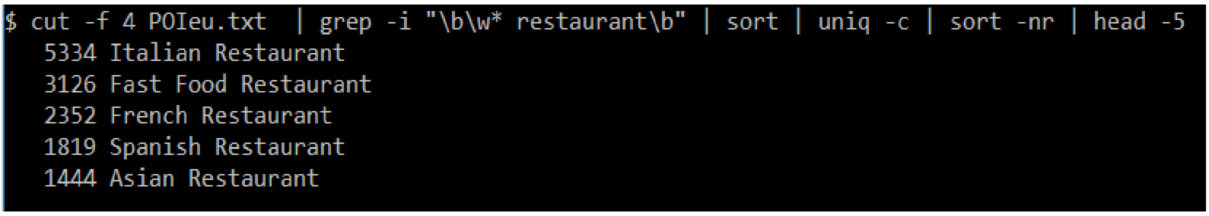


```
$ cut -f 4 POIeu.txt  | grep -i "restaurant" | sort | uniq -c | sort -nr | head -5
   9301 Restaurant
   7173 Italian Restaurant
   4915 Fast Food Restaurant
   2738 French Restaurant
   2222 Asian Restaurant
```

Another way to do this question is by using the following code in which I have ignored the “Restaurant” class:

cut -f 4 POIeu.txt | grep -i "\b\w* restaurant\b" | sort | uniq -c | sort -nr | head -5

(Patashu, 2013)



```
$ cut -f 4 POIeu.txt  | grep -i "\b\w* restaurant\b" | sort | uniq -c | sort -nr | head -5
   5334 Italian Restaurant
   3126 Fast Food Restaurant
   2352 French Restaurant
   1819 Spanish Restaurant
   1444 Asian Restaurant
```

References

Karthik. (2015, June 29). Retrieved from https://unix.stackexchange.com/questions/212925/using-sed-to-replace-special-characters

Patashu. (2013, May 9). Retrieved from https://stackoverflow.com/questions/16472430/grep-for-words-ending-in-ing-immediately-after-a-comma