FIT1043 Assignment3

Name: Lianzheng Xie

Student ID: 32068611

**Task A : Investigating User global-scale check-in data collected from**

**Foursquare Data in the Shell**

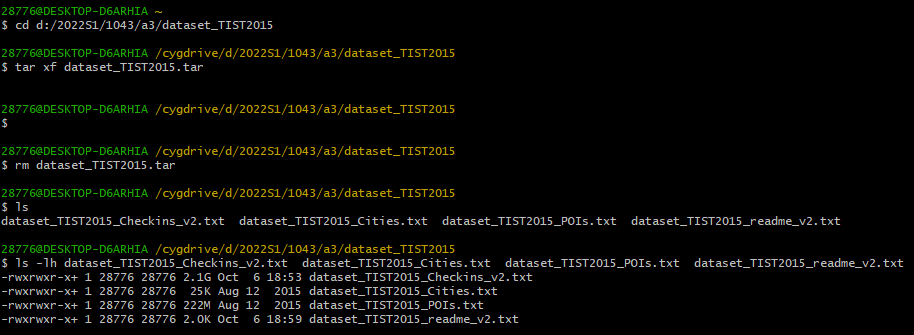
cd **open the folder where the dataset resides**

tar xf dataset\_TIST2015.tar **decompress the compressed file,**

ls **get the file，**

ls -lh **get each file size.**

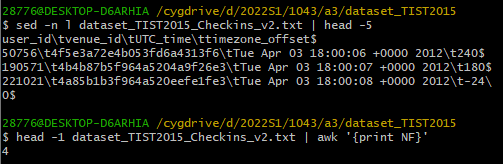
1. There are 4 files in the tar file, dataset\_TIST2015\_Checkins\_v2.txt is 2.1G, dataset\_TIST2015\_readme\_v2.txt is 2.0k, dataset\_TIST2015\_Cities.txt is 222M, dataset\_TIST2015\_POIs.txt is 25k.



1. The delimiter of dataset\_TIST2015\_Checkins\_v2.txt is \t(\<tab>), and this file has 4 columns.

sed -n l dataset\_TIST2015\_Checkins\_v2.txt | head -5 **Look at the separator in the first five lines of the file**

head -1 dataset\_TIST2015\_Checkins\_v2.txt | awk '{print NF}' **get the number of columns of first line of file**

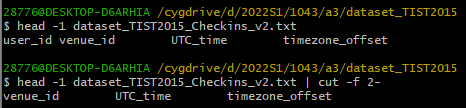


1. Other columns are venue\_id, UTC\_time and timezone\_offset

head -1 dataset\_TIST2015\_Checkins\_v2.txt Show the first line of the file.

head -1 dataset\_TIST2015\_Checkins\_v2.txt | cut -f 2-

Show all columns in the first row except the first column, cut is cut from the second column.



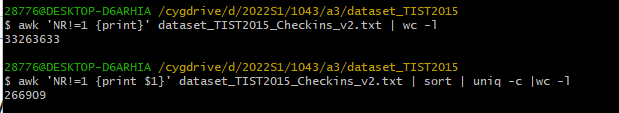
1. There are 33263633 Checkins, and 266909 users in the file.

awk 'NR!=1 {print}' dataset\_TIST2015\_Checkins\_v2.txt | wc -l

Get all the data in the file except the first line and then get the number of lines

awk 'NR!=1 {print $1}' dataset\_TIST2015\_Checkins\_v2.txt|sort| uniq -c |wc -l

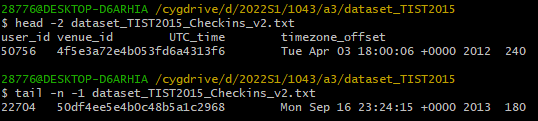
Get the first column in the file except the first line, then sort, remove the repetition and count the number of repetitions, and finally get the number of rows.



1. First date is Tue Apr 03 18:00:06, last date is Mon Sep 16 23:24:15.

head -2 dataset\_TIST2015\_Checkins\_v2.txt Show the first two line of the file

tail -n -1 dataset\_TIST2015\_Checkins\_v2.txt Show the last line of the file



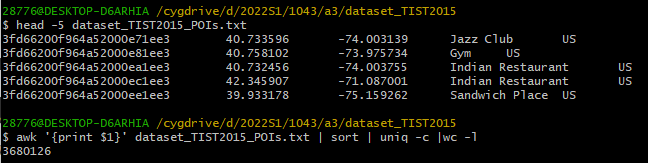
1. There are 3680126 unique venue IDs in the file.

head -5 dataset\_TIST2015\_POIs.txt

Show the first 5 rows to find venue IDs in the first column

awk '{print $1}' dataset\_TIST2015\_POIs.txt | sort | uniq -c |wc -l

Get the first column of data in the file, then sort, remove the repetition and count the number of repetitions, and finally get the number of rows.



1. France contains 384 unique Venue categories in the file.

grep "FR" dataset\_TIST2015\_POIs.txt| cut -f 4 | sort | uniq -c| wc -l

grep finds all the rows containing FR, intercepts the fourth column representing the site category, then sort, remove the repetition and count the number of repetitions, and finally get the number of rows.



1. A. awk -F '\t' '$2>=36 && $2<=71.08 && $3>=-9.31 && $3<=66.10' dataset\_TIST2015\_POIs.txt> POIeu.txt

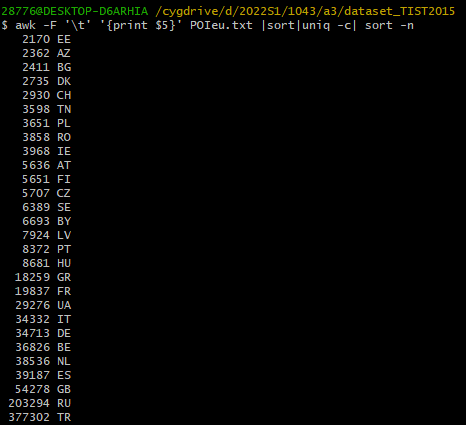
According longitude and latitude range of Europe(36,71.08&-9.31,66.10) get the all data that meets the conditions, then print as a txt.



B. awk -F '\t' '{print $5}' POIeu.txt |sort | uniq -c| sort -n

According to ‘\t’ get the fifth column(country) of txt(A8.A), then sort, remove the repetition and count the number of repetitions, and finally get the number of rows.

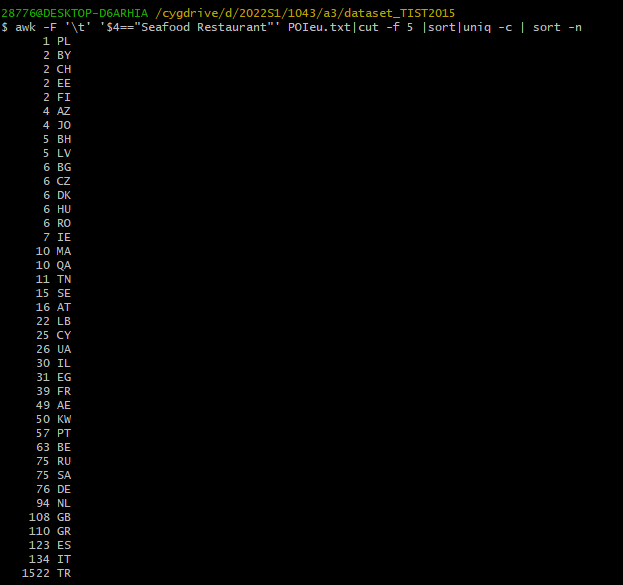
Most venues is Turkey(TR) with 377302, the least venues is Estonia(EE) with 2170.



C. awk -F '\t' '$4=="Seafood Restaurant"' POIeu.txt|cut -f 5 |sort|uniq -c | sort -n

Put out all the rows in the file whose fourth column is Seafood Restaurant, and the fifth column will be captured, then sort, remove the repetition and count the number of repetitions, and finally get the number of rows.

Turkey(TR) has most Seafood restaurants with 1522.



D. grep "Restaurant" POIeu.txt |awk -F '\t' '{print $4}' |sort|uniq -c|sort -n

Found out all restaurant type and print them out with numbers

“Restaurant” is most common class of restaurant in Europe with 16838.

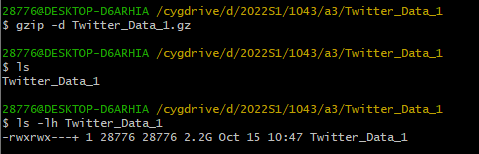


**Task B: Investigating the Twitter Data in the Shell and Graphing in R**

1. It appeared 116 times.

gzip -d Twitter\_Data\_1.gz,ls,ls -lh Unzip the files, look up file, get file size

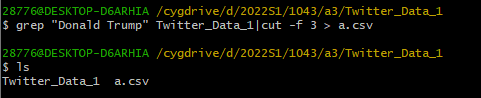
grep -o "Donald Trump" Twitter\_Data\_1 | wc -l Look for "Donald Trump" in the file, that is, the grep command uses the -o parameter to convert rows into columns, and then statistics.





1. grep finds the line containing "Donald Trump" in the file, intercepts the third column(timestamps) and exports it to csv.

grep "Donald Trump" Twitter\_Data\_1 | cut -f 3 > a.csv



Read the data in a.csv with R and change it according to the format of the timestamp.

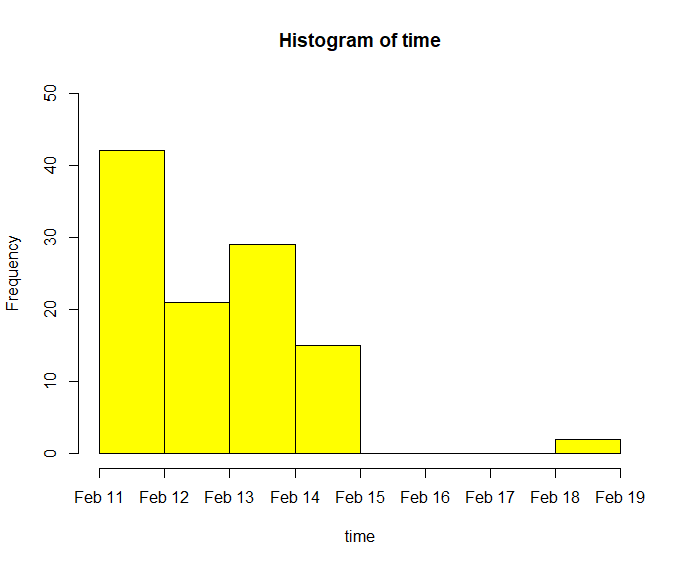
setwd("D:/2022S1/1043/a3/Twitter\_Data\_1")

Sys.setlocale("LC\_TIME", "C")

twitter <- read.csv("a.csv", header = F)

twitter$V1 <- strptime(twitter$V1, format = "%a %b %d %H:%M:%S %z %Y", tz = 'UTC')

1. hist(twitter$V1,"days",xlab = "time",col = "yellow",freq = T,ylim = c(0,50))



1. It can be seen from the figure (Q3) that the data before February 15 had the largest number of occurrences on February 11 (more than 40 times), followed by February 13 (less than 30 times), February 12 (more than 20 times), February 14 (less than 20 times). There were no tweets about Donald Trump for three days(Feb 15,16,17), and then two more tweets about him on February 18.

5)

Pull out all the user data in the second column of twitter data, then sort, remove the repetition and count the number of repetitions, and finally get the number of rows. Finally, put the data into b.txt.

awk -F '\t' '{print $2}' Twitter\_Data\_1 |sort | uniq -c > b.txt



numtwitter <- read.table("b.txt",fill = TRUE, head = FALSE) # read the txt as a table

names(numtwitter)<- c("number\_twitter","id")# rename of each columns

max(numtwitter$number\_twitter) # find the max freq

hist(numtwitter$number\_twitter,breaks = 243,freq = T, xlim = c(0,10)) # create histogram

Change the names of the two columns in the txt to “number-twitter” and “id”, max() get the max number of twitter.

