# FIT1043 Assignment 2
# Semester 2, 2016

*Due: Week 12, Sunday, 11.55pm.*

**Hand in Requirements:** Please hand in a PDF file containing your answers to all the questions, numbered correspondingly.

- You can use Word or other word processing software to format your submission. Just save the final copy to a PDF before submitting.
- Make sure to include screenshots/images of the graphs you generate in order to justify your answers to all the questions.
- Make sure to include copies of all the scripts and bash command lines you use. If your answer is wrong, you may still get half marks if your command line is close to correct.

**NOTE:** Two data sets for this assignment are in the Google shared drive:
https://drive.google.com/open?id=0B4wDFbhepc1faDhZY25JeE9MejQ
Both are large, so your best bet is to download them while in the lab/studio and do the assignment there. You will need to use a Linux machine for this, though a Mac should also work.

**Part A:** Investigating the 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?

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

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

4) How many Tweets are there in the file?

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

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

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

---

[1] If you don’t want to be patient, redirect the output of the command to a file and run the command “in the background” by typing an ampersand character "&" at the end.

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

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

*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?

**Part B:** Graphing the data R

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

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 recognise the strings as timestamps automatically, so you’ll nee 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/R-devel/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?

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.]

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

5) (Hard) 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”. ]

**Part C:** Investigating User Check-in Data

Download the file dataset_TIST2015.zip, which contains user check-in data from Foursquare (https://foursquare.com/).

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?

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. What country has the most venues and what the least, with how many?
   B. Who has the most Indian restaurants?
   C. What is the most common (as in, how many venues) class of restaurant in Europe?

3) *Background:* How would you select check-ins from Europe? We have a list of venues in Europe, constructed using the above question in the file "POIeu.txt". In database terms, to get European check-ins we have to "join" the check-ins and the venues (POI) files on the common key "Venue ID", but we are only interested in the case where the venue is in Europe. Now we can try running the command "join":
   join -11 -22 POIeu.txt dataset_TIST2015_Checkins.txt (1)
   This joins on key 1 of the first file and key 2 of the second file. But this returns an error:
   join: dataset_TIST2015_Checkins.txt:2: is not sorted: …
   Reading the user manual for "join" you will see the comment:
   Important: FILE1 and FILE2 must be sorted on the join fields.
   So, which files are not sorted?
   Moreover, once fixing this, you will find the output looks unorganised because "join" has converted all the field separators, previously TABs, to spaces. reading up the manual again, you see the command line option:
   -t CHAR

use CHAR as input and output field separator

However, use of this requires you enter a TAB as a single character.

*Questions:* Show how to run the join command correctly to create a file with check-ins for Europe. This combines columns from the check-ins and the venues files:

A. What should you do to fix up the join command (1) above? Which file should be sorted and how?
B. Use Google to find out how to enter a single TAB as your separator so you can run "join", *i.e.*, what should "XXX" be in
   join -11 -22 -t XXX ….
   [HINT: this requires you make a Google search, say "linux join tab separator", browse through the solutions and attempt to get a working solution.]
C. Run the join. The 4th column of this result file is the venue type, and the 5th column is the two letter country code. Now find out the five most popular (as in, how many check-ins) classes of restaurants in Italy, Germany and France.