FIT1043 Assignment 3
Semester 2, 2018

Due: Monday, 22 October 2018 (23:59)

There are three parts in this assignment. Students that complete only parts A and B1-B5 can only get a maximum of Distinction. Students that attempt part B6 and part C will be showing critical analysis skills and a deeper understanding of the task at hand and can achieve the highest grade.

Hand in Requirements:

1) Please submit 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 bash command lines and R scripts you use. If your answer is wrong, you may still get half marks if your command line or script is close to correct.
2) Late submission will be given a **penalty of 5% per day**. Any submission after one week will not be accepted and will result in **zero** marks.

NOTE: The data set for this assignment is in the Google shared drive:
https://drive.google.com/open?id=1tpcOQIlmg7damXO6SmpolQJ8JXuoeiV6

It is large, so you may need to download it while in the lab/studio and do the assignment there. You will need to use either a Linux machine for this or a Mac terminal or Cygwin on a Windows machine.

Part A: Investigating the Twitter Data in the Shell

Download the file Twitter_Data_1.gz from the link above. This is 1 GB 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?

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

Part B: Graphing the Data in 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 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/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) Based on the histogram in (3), choose one date/day that has the highest mention of Obama. Plot another histogram for this date/day, to show the frequency for every hour for that date/day. Can you deduce anything from observing the histogram?

[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.

6) 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: K-means Clustering on Other Data in R

We have demonstrated k-means clustering algorithm in Tutorial week 9 in Python. Your task in this part is to find an interesting dataset and apply k-mean clustering on a dataset using R. Kaggle, a private company which runs data science competitions, provides a list of their publicly available datasets:

https://www.kaggle.com/datasets

In particular you need to choose two numerical features in your dataset and apply k-mean clustering on your data into k clusters in R, where k>=2. Then visualise the data as well as the results of the k-means clustering. Ideally each cluster is shown in a different colour.

Here is a tutorial about k-mean clustering in R (link - https://www.datacamp.com/community/tutorials/k-means-clustering-r). Please note you cannot use the same data set used in this tutorial or tutorial 9 in this unit.

Please include a link to your dataset in your report. You may wish to:

1. provide the direct link to the public dataset from the internet, or
2. place the data file in your Monash student - google drive and provide its link in the submission.

Good Luck!