FIT5145 Assignment 3
Semester 2, 2019

Due: Monday 30th September 2019, 11:55pm

Hand in Requirements:

1) 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 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.

Data:

The dataset for this assignment is in the Google shared drive:
https://drive.google.com/open?id=1frjdZrDBGLo_gkQLF5QlSAuZMniDER3h

The dataset contains Facebook posts from 15 of the top mainstream media sources (e.g., ABC, BBC, etc.) from 2012 to 2016.
Note: This is a large file, so your best bet is to download them 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.

Assignment Tasks:

There are two tasks that you need to complete for this assignment. Students that complete **only Tasks A1-A10 AND B1-B2** can only get a **maximum of Distinction**. Students that **attempt tasks A11-A12 and B3** will be showing critical analysis skills and a deeper understanding of the task at hand and can achieve the **highest grade**. You need to use unix shell and R to complete the tasks.

Task A: Investigating Facebook Data using shell commands

Download the file FB_Dataset.csv.zip from the link above. 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 2nd column is the unique identifier for a Facebook post. What are the other columns?

4) How many Facebook posts are there in the file?

5) What is the date range for Facebook posts in this file? (Assume that the data is in order)

6) How many unique pages are there?

7) How many unique posts are there? [Hint: one page can have multiple posts]

8) When was the first mention in the file regarding “Italian Dishes” and what was the post?

9) How many times is “Barack Obama” mentioned in the file? How did you find this? (Do not ignore the case)

10) What about “Donald Trump”? Who is more popular on Facebook, Obama or Trump? (Do not ignore the case)

11) Select the posts where “Trump” (Ignore the case) is mentioned in the post content and number of likes for those posts are greater than 100. And generate a new file with post_id and **sorted** like_count and name it “trump.txt”. (In the output, you need to show the headers as well) [Hint: Find Trump in message column, i.e., 5th column]. Then copy and paste the first 5 lines of trump.txt in your answer.

12) Find the total number of love_count and angry_count for “Donald Trump” and “Barack Obama” separately. Who has more positive feeling among people? Justify your answer.
[Hint 1: you will need to search online to find how to sum a column of numbers using awk.
Hint 2: You will need to consider both love and angry count when justifying your answer.]

Task B: Graphing the Data in R

1) How many times does the term 'Trump' appear in the post content? (use shell to answer to this question)

2) We want to consider how the amount of discussion regarding Donald Trump varies over the time period covered by the data file. To answer this question, you will need to extract the timestamps for all posts referring to Trump using shell. 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://www.rdocumentation.org/packages/base/versions/3.6.1/topics/strptime
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?

    1. Once you have converted the timestamps, use the hist() function to plot the data in R.

    2. The plot has a bit of an unusual shape. Describe the pattern you see.

3) In this question, we want to investigate the Facebook posts of a few top media sources. To answer this question, you will need to extract the facebook posts made on the pages of "abc-news", "cnn" and "fox-news" from your original Facebook dataset.

    1. Use the unix shell to first generate a file containing all the records belonging to "abc-news", "cnn" and "fox-news" only. Then read the resulting file in R.

    2. Background: We now want to see if any relationship exists between the number of times a post is shared on Facebook and the number of likes it generates. Task: Use appropriate R code to generate a plot showing the relationship between the number of shares and the number of likes in your dataset. Do you see any relationship?

    3. Fit a linear regression model using R to the above data (i.e., shares_count and likes_count) and plot the linear fit. Does it look like a good fit to you?

    4. Use the linear fit to predict the number of likes a post will generate if it is shared 0 times, 1000 times, 10000 times and 100000 times on Facebook.

Good Luck!