FIT1043 Assignment 3
Semester 2, 2019

Hand in Requirer

1) Please han ntaining your answers to all the questions and, numbered c
2) Your report following cases:
   - The screenshots/images of the outputs/graphs you generate in order to justify your answers to all the questions.
   - 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.
3) Please be informed that you need to explain what each part of command does for all your answers. For instance, if the code you use is 'unzip tutorial_data.zip', you need to explain that the code is used to uncompress the zip file.
4) Please don't include the questions into the assignment (It has 5% penalty).

Data:

The dataset for this assignment is in the Google shared drive:
https://drive.google.com/file/d/15icdh6UTmJ7QJ1vtMk-gp0OT4G8ZnqSm/view?usp=sharing

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-A9** and **B1** can only get a **maximum of Distinction**. Students that attempt tasks **A10 and B2** (**Challenge Questions)** 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 the 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[illegible]llowing questions.

1) Decompres[illegible]is it?
2) What delim[illegible]parate the columns in the file and how many columns ar[illegible]
3) The 2nd column is the unique identifier for a Facebook post. Print out the name of other columns in the output?
4) How many unique pages are there?
5) What is the date range for Facebook posts in this file? (Assume that the data is in order)
6) When was the first mention in the file regarding "Italian Dishes" and what was the post name?
7) How many times is "Donald Trump" mentioned in the file? How did you find this? (Do not ignore the case, i.e., lower/upper case)
8) What about "Barack Obama"? Who is more popular on Facebook, Obama or Trump? (Do not ignore the case)
9) 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". (You need to add a screenshot of the first 5 rows and the column headers in your report).
10) **Challenge:** 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) We want to amount of discussion regarding Donald Trump varies over vered by the data file. To answer this question, you will ne mestamps for all posts referring to Trump using the shell. Y to read them into R and generate a histogram. [Hint: To re first generate a file containing only the timestamp column as e file into R as a CSV.] R will not recognise the strings as t tically, 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?

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

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

2) **Challenge:** In this question, we want to look at a specific content type that influences engagement on Facebook. To make this task easier, we will specifically look at the number of comments posted against each of the post type (event, link, photo, status and video) for "fox-news".

   A. Draw a boxplot to show the distribution of comments made against each type of post (event, link, photo, status and video) created by "fox-news". What can you infer from this plot? Which is the most engaging post type?
   B. You may have noticed that the presence of outliers affects the readability and interpretation of the data in the box plot. Redraw the boxplot by filtering out values (comments_count) greater than 10,000.
   C. Which type of post (event, link, photo, status or video) has on average been most effective for "fox-news". In other words, which post_type has the highest median comment count.

Good Luck!