FIT1043 Assignment 3: Specification

Due date: Friday 20th October 2023 - 11:55 pm

Aim

The objective of this assignment is to read and process a (*large*) data set using the BASH shell scripts and visualise the summarised data using plots generated by the R programming language. This assignment will test your ability to:

- Read a reasonably large dataset,
- Process the dataset using BASH Shell Scripts,
- Conduct aggregation of the dataset content,
- Read data from a file in R, and
- Generate appropriate visualisations in R and output to files

Data

The data provided is a pre-processed data that is derived from corona_tweets_58.zip from https://ieee-dataport.org/open-access/coronavirus-covid-19-tweets-dataset. The data is a twitter dataset for 15th May 2020, filtered by keywords that are related to the COVID-19 situation. Instead of providing the whole twitter data for that day alone, which amounts to more than 6GB of JSON formatted data (*volume*), the dataset provided to you for this assignment has been processed and contains only a portion of the data. There are more than 2 million tweets and it will take more than 24 hours to download all the tweets from twitter (using a standard developer account). Note that to download 1-day data (that is only a small portion of twitter data for that day), it takes more than 1 day to do so (*velocity*). The original data comprises more than 6GB of JSON data which contains a *variety* of information such as data, text, identities, numbers and so on.

The data has been pre-processed into CSV (it's separated using tabs in order to minimize issues with commas (',') as the twitter text may contain commas as well as other fields). It has been converted from JSON and only contains a subset of the data. The file corona_tweets.csv.gz can be downloaded from Moodle and this assignment will be based on this file.

Hand-in Requirements

Please hand in a single PDF file only and a video file (refer to **Part B**).

PDF file should consist of:

1. Answers to the questions. In order to justify your answers to all the questions, make sure to
    a. Include screenshots/images of the graphs or outputs you generate (You will need to use screen-capture functionality to create appropriate images.)
    b. Please be informed that you need to explain what each part of the 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.
    c. Copy and paste of your Unix code from Bash Shell and the R code (Do Not include screenshots of your code).
    d. Kindly Do Not copy the questions, else you might have high Turnitin similarity due to all submissions referring to the same set of questions.

Assignment Tasks:

This assignment is to illustrate working with large data sets (in this case, just more than a million lines of data, not really large but enough for learning) and to also experience the use of shell scripts to process and aggregate data. In the whole exercise, you must NOT uncompress the data and store it. Once the data is aggregated and nicely formatted, you are then to read the data in R where you are to conduct further analysis. In this assignment, you only need to read the data in R and provide some visualisations.

Note, for this assignment you are required to write **shell commands** to answer all questions in Part A unless the instructions specify using **R code**.

Part A:

A1. Inspecting the data

1. Copy the downloaded file to your UNIX (Linux) terminal. State the size (in Bytes or MegaBytes) of the corona_tweets.csv.gz file and provide the shell command that you used to determine the size.

2. The first line of the CSV file contains headers that are “tab” separated. What are the header names and provide the command you used to obtain it. Note that the command provided has to be in one line.

3. How many lines are there in the dataset? Again, provide the single line code on how you obtained it.

A2. Information from Data

1. How many unique twitter users are there in the dataset. Provide a single line code that uses the “awk” and “uniq” commands. You are also required to read the “man” pages of the “uniq” command to figure out if it is sufficient to answer the question. Explain the code you provided.

2. For each of the sub-questions below, provide a single line code (one each) and briefly explain your code.
    a. How many tweets mentioned the word “death” in any combination of uppercase or lowercase letters .
    b. How many of those are not spelt exactly “death” or “Death” but in other combination of uppercase and lowercase (e.g. DEath, deatH), and
    c. Output the lines of **A2.2 (b)** into a file called myText.txt (not the number of lines but the specific lines).

A3. Data aggregation

For this part, let’s assume that you would like to know how many followers each of these twitter users have.

1. Let’s group the twitter user (ID) by the number of followers that they have into the following ranges. Provide the code for them. One line of code for each of them below (11 lines).

    a. Less than or equal to 1000
    b. 1001 to 2000
    c. 2001 to 3000
    d. 3001 to 4000
    e. 4001 to 5000
    f. 5001 to 6000
    g. 6001 to 7000
    h. 7001 to 8000
    i. 8001 to 9000
    j. 9001 to 10000
    k. More than 10000

    *Do note that twitter users may tweet more than once a day and hence they can appear more than once in this dataset.

2. Create a CSV file manually with the output from **Part A3.1** . The CSV file should contain two columns, the first column is the range (e.g. “6001 to 7000” or “6-7k” or other meaningful names) and the second column is the number of twitter users.

3. **[R Code]** Use the output of the above **(A3.2)** and read it using R.

4. **[R code]** Plot the suitable visualisation (Histogram/Bar Chart/ any visualisation that you think is suitable) using the data and output it into a PNG file (.png). For submission, you just show the code, and paste the PNG image in your PDF report.

A4. Small Challenge

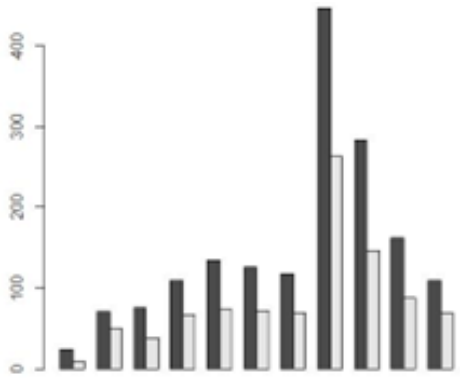
Let's assume that we want to compare against tweets that are not retweets. We can assume that this is indicated at the beginning of the text by the "RT @" (Note that if we are using the raw JSON data, there is a field that indicates it).

1. Provide a single line code that filters out the tweets that contain "RT @" and output the results into a compressed gz file (Note, you need to use the opposite of gunzip).

2. Do the same process as **Part A3.1 to A3.2** with this new file.

3. **[R code]** Copy the output of the above **(A4.2)** and read it using R.

4. **[R code]** Plot a side by side bar chart to visualise the data created from **A3.2** and from **A4.2** .
   (An example of a side by side grayscale bar chart is shown below - the example below is not the answer to this question). The bars should be coloured (any colour/pattern that you like).



Part B: Video Preparation

Presentation is one of the important steps in a data science process. In this task you will need to prepare a short video of yourself (you can share your code on screen) and describe your approach on the above task (**Task A4**).

- Please make sure to keep your camera on (show yourself) during recording.

Good Luck! ☺