FIT1043 Assignment 3 Specifications

23th September 2021 – Version 1.0

Due date: Thursday 14th October 2021

Objective

Assignment 1 & 2 walked you through what you have learnt in Lectures 1 to 7 and also the *Collection*, *Wrangling*, *Analyse* and *Present* phases of our Standard Value Chain. It provides you an introduction to the data analytics lifecycle. This assignment relates to the latter part of this unit, in the use of the BASH shell and the R programming language to work on larger datasets. It will test your ability to:

- Navigate the BASH shell
- Process large file using BASH shell commands
    - Use online resources or the "`man`" pages to assist in the commands
- Output a processed file to CSV format using BASH shell
- Read a processed file in R
- Conduct visualisation using R
- Challenge: Conduct classification using R (sentiment analysis)

Note that unlike the previous assignments, you will notice that there are open ended tasks, especially for the challenges, with less detailed information pertaining to the data and the expected process. In other words, there will be less guidance and you are expected to be able to understand the requirements, provide suitable answers and explanation for the tasks.

Data

The dataset for this assignment is available on Moodle. It is a compressed file that contains pre-processed twitter content sourced from Sentiment140 Dataset on Kaggle. The original source contained 1.6 million tweets, extracted using the Twitter API and they have been labelled as negative (0), neutral (2), or positive (4). The data on Moodle for this assignment is a subset of that data.

The columns are the same, as follows:

- target: the polarity of the tweet (e.g., 0 = negative, 2 = neutral, 4 = positive)
- ids: The id of the tweet (e.g., 2087)
- date: the date of the tweet (e.g., Sat May 16 23:58:44 UTC 2009)
- flag: The query. If there is no query, then this value is NO_QUERY.
- user: the user that tweeted (e.g., robotickilldozr)
- text: the text of the tweet (e.g., Lyx is cool)

Note: You will need to use either a Linux machine, a Mac terminal or Cygwin on a Windows machine for this purpose.

For those who are more curious, the paper describing the dataset and how it was labelled is here (this link does expires, but you can search and access it (by searching on scholar.google.com) using the following citation:

- Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N project report, Stanford, 1(12), 2009.

Submissions

For this assignment, you are to hand-in on Moodle, only 1 well formatted PDF file. Details for the submission:

1) Hand in a PDF file containing your answers to all the questions and, numbered correspondingly.
2) Your report should include the following cases:
   a) The screenshots/images of the outputs/graphs you generate in order to justify your answers to all the questions. Ensure that they are legible, such as making sure that the image resolution is sufficient.
   b) Copies of all the BASH command lines and R scripts you use. Ensure that the grader can copy and paste your code from the PDF document (not an image).
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 decompress the zip file.
4) Please don't include the questions into the assignment, this will cause a high TurnItIn percentage and we will impose a penalty if you include the questions in the PDF.

Moodle Submission

You are to hand in only one file:

1. A **PDF** file that is generated from your word processor, e.g., from Microsoft Word.

Clarifications

This assignment has less specific guidance and has open ended tasks. Do use the ED Discussions so that other students can participate and contribute in any discussions. For postings on the forum, do post as though you are asking others (instead of your lecturer or tutors) for their opinions or interpretation. Just note that you are not to post answers directly.

Assignment

There are two tasks that you need to complete for this assignment. Students that complete only Tasks A1 – A12 and B1 can only get a maximum of D (Distinction). Students that attempt tasks A13 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 (HD). You need to use the BASH shell and R to complete the tasks.

Tasks

Part A: Investigating Sentiment140 Data using shell commands

Download the file `FIT1043_Dataset.gz` from Moodle. Use the BASH shell to manipulate the file and answer the following questions. Show the BASH shell command you used and also the displayed output where appropriate.

1. Decompress the file (if it isn't already) What is the file size in megabytes? Show the shell command and state the answer.

2. What delimiter is used to separate the columns in the file? Do illustrate how you deduced this.

3. How many rows are there in the file?

4. How many columns are there? Do not do a manual count, it has to be done using BASH shell commands. This wasn't directly "taught" in this Unit and you are expected to explore how you may want to do this task.

5. How many unique users are there? (You will need to understand how to interpret "unique" on your own)

6. What is the date range for the Twitter posts in this file? (Assume that the data is ordered by date in chronological order)

7. When was the first mention of a person with the name "Ian" that is not a retweet or a reply-to? For this you will have to search on what a retweet or a reply-to is on twitter looks like and also you can manually look at the output and state the answers for the date and time of the tweet, the full message, and which user mentioned it.

8. How many tweets have the word "Australia" in the message column of the file? How did you find this? (*Ignore the case*, i.e., lower/upper case).

9. Similar question to (8), how many times the word "Australia" appeared in the message column. Note that one tweet can contain more than one occurrence of the word.

10. What about if I only want the exact word “Australia” (ignoring cases), not “Australian” or other combinations of the word?

11. What about “`India`” (mentioned in the message column of the file)? Which country is more popular, Australia or India? (*Do not ignore the case,* and you are to define what is meant by popular, and also ensure that you keep both comparisons the same. You can use any of the methods you have used for answering questions 8, 9, or 10.)

12. Select the posts where country “India” (*Ignore the case*) is mentioned in the twitter message and determine how many unique users are there that mentions the country “India” (not nationality “Indian”, nor any other area such as “Indiana”). To get full marks for this, it is expected that the BASH Shell is done in one single line.

13. *Challenge*: Find the total number of positive, neutral, and negative tweets for “Australia” and “India” and store them in respective files named `sentiment-australia.csv` and `sentiment-india.csv`, using the following format:

    ```
    Negative, 99
    Neutral, 99
    Positive, 99
    ```

    Which country seem to have more positive tweets related to it? Do you think your answer is justified? (Do not need to ask for clarification for this, you are to justify your interpretation of the question and your approach).

    You can do this challenge using multiple lines of commands.

(Hint: You may (optional) use BASH commands such as echo and also the use of backticks (or also called backquotes) to figure this out. Note that there are many ways to do this and this is just an optional suggestion)

Part B: R, Graphing and Machine Learning

1. We want to consider tweets that contain the word “Australia” on a weekly basis from the data provided.
   i. To answer this question, you will need to extract the timestamps of the tweets referring to “Australia” (ignore case) using the BASH Shell.

   ii. You will then need to read them in R and R will not recognise the strings as timestamps automatically, and for this task you are to convert them from text values using the `strptime()` function. Instructions on how to use the function is available here. You will need to write a format string, e.g., starting with “`%a %b`” to tell the function how to parse the particular date/time format in your file. Explain this in your answer.

   iii. Plot a line graph, with the dates on the x-axis, and the number of tweets that was tweeted on that particular day (using the data that you have processed for Part B Question 1 (i) and (ii))

2. *Challenge*: This dataset is for sentiment analysis, where the tweets are labelled as “negative”, “neutral”, or “positive”. The challenge is to use the dataset to build a model using the R programming language. You will need to (but not limited to the following):
   a) conduct any data pre-processing (if necessary),
   b) determine the training and testing datasets,
   c) use the Naïve Bayes algorithm to train and build the model,
   d) conduct model testing, and
   e) interpret the output confusion matrix of the model.

   This is similar to your last assignment but it is to be done in R. It is a challenge but the accuracy is of no importance for this task. This challenge is for you to be able to do machine learning in R, with minimal prior sessions of R. Have fun and best of luck!

   Hints:
   - You are allowed to install and use other libraries that are useful, for example the library `e1071`, `caret` or even `RTextTools`.
   - You can read the whole `FIT1043_Dataset`, but some of you may encounter issues with memory. If you do, you are allowed to use a subset of it.
   - There are many resources online on how to split the dataset for training and testing in R, please ensure that you provide proper referencing to the source.

   For text processing, you are encouraged to study basic text processing and how the text are manipulated to be used as features for the Naïve Bayes algorithm. This is an optional exercise. However, if you do provide the code (can be from external sources) and explain what you understand from it, bonus marks may be awarded. In addition, if you are able to explain your (probably poor) confusion matrix results, bonus marks may also be awarded.

Marking Rubrics

There are two tasks that you need to complete for this assignment. Students that complete **only** Tasks **A1 – A12** and **B1** can only get a **maximum of Distinction**. Students that attempt tasks **A13 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 (HD)**. You need to use the Unix shell and R to complete the tasks.

General points:

- **Late submission:** Late submissions will have a **penalty of 10% per day**, including weekends and public holidays **for up to 7 days.** Assessment items handed in after 7 days will not be considered.
- **Zip (compressed) file submission:** Zip file (or any compressed file) submission will have a **penalty of 10%.**
- **Drafts (not submitted):** There have been many of you who left your submission in Draft mode. Please make sure to submit your assignments that are in draft mode. For this assignment, we reserve the right **not to accept** the assignments that are **not yet submitted**.
- **Screen shots of codes:** are not acceptable (the grader will need to be able to copy and paste your code) and there will be no marks awarded for the code portion.
- **Acknowledgement of sources**: Plagiarism or unauthorised collaboration will result in an **automatic failure**

Task A – Is worth slightly more than 63% (28 out of 44) of total mark
(Challenge Question is worth 4 mark out of the 28 marks)

	You are currently working in this approximate range:	
Coding 70%	Some errors	Error free
Answers or justify answers (if applicable) 30%	Poor/no justification (if applicable)	Correct answer, strong justification (if applicable)

Task B – Is worth slightly less than 37% (16 out of 44) of the total mark
(Challenge Question is worth 7 mark out of the 16 marks)

	You are currently working in this approximate range:	
Coding and Visualisation 80%	Some errors or misleading visuals	Error free or clear visuals
Answers or justify answers (if applicable) 20%	Poor/no justification (if applicable)	Correct answer, strong justification (if applicable)

Congratulations!

You have completed your in-semester assessments for FIT1043. I hope that you have enjoyed the course assignments, starting from a very guided assignment 1, to something with a little bit of flexibility for you to try out new stuff and complete in assignment 2, and finally assignment 3 with less guidance and requires you to state your interpretation on the tasks.