IFN509 Portfolio Item 1 - Requirements

Semester 1, 2018

This portfolio item is making you apply all the practical skills in data manipulation you have learnt in IFN509 so far.

This assessment requires you to answer 3 questions about some data contained in a file. Each question should be answered in two ways: (1) creating bash scripts using `bash` commands (cannot use SQL commands for this part), and (2) creating bash scripts using SQL commands (i.e. writing an SQL command within a bash script like `sqlite3 database.sqlite 'SELECT * FROM mytable;'`).

Marking Guide:

The portfolio is worth 15 points in total (15%). Question 1 (Q1) is the easiest question in the assessment, but it requires to do some initial legwork (for example, explore the data). Question 2 (Q2) is moderately difficult. Question 3 (Q3) is the most difficult in the assessment item.

	Bash	**SQL**
Q1	5 marks	5 marks
Q2	1.5 marks	1.5 marks
Q3	1 mark	1 mark

For Q1 we will also award part marks, i.e. if your answer is incorrect, you will not get full mark, but we will weight the error/s you did and assign you a portion of the 5 marks. Points for question Q2 and Q3 are only awarded in their entirety (or nothing).

REMEMBER: this is an individual assessment, so **you should not work in group** for this assessment, or pass your scripts to a friend.

Along with marks, we also will award the following Honorable Mentions (these are not worth marks - just honour and pride):

1. *most creative solution*: this will be awarded to the student(s) that have answered the problem in the most creative way, i.e. in a way we have not thought about.
2. *most elegant solution*: this will be awarded to the student(s) that have answered the problem in an elegant way - this often coincides with writing the minimal amount of commands required to answer the question.
3. *best commented solution*: this will be awarded to the student(s) that have answered the problem by documenting their scripts in the clearest way (note however that you are actually not required to

document, i.e. comment, your scripts to be awarded full marks).

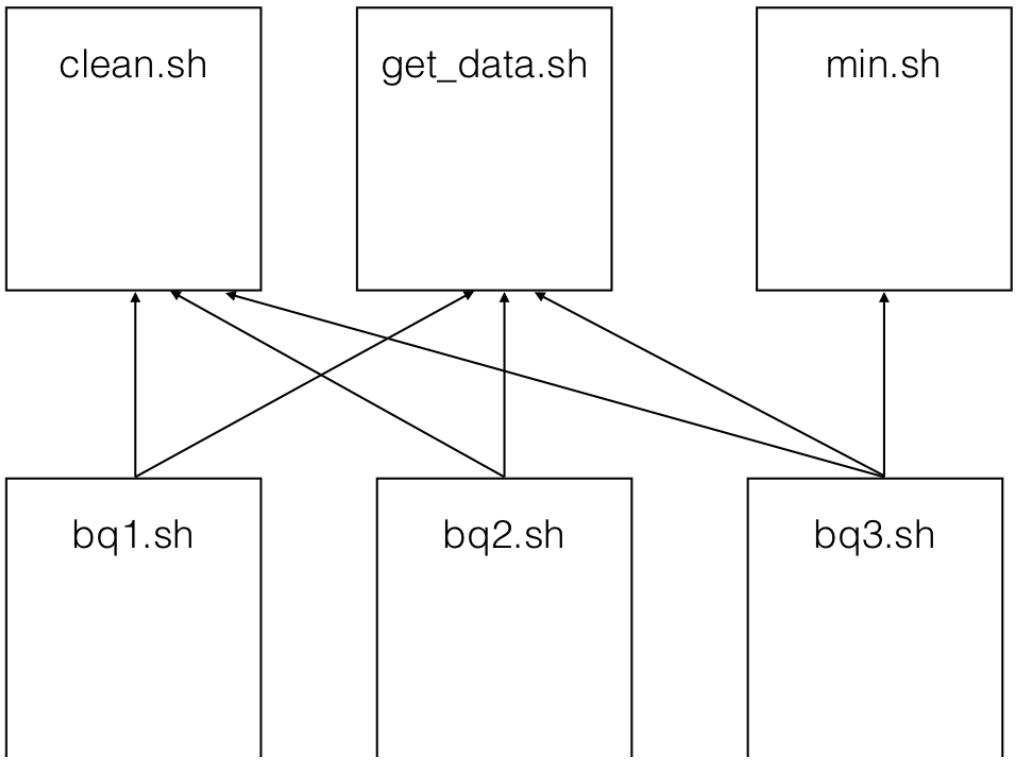
Templates for structuring the assignment:

`portfolio 1.zip` contains the templates. Along with the data to work with, the file contains a `.sh` file for each question. You must write your answer to each question into that file. Your answer must be an executable script that produces an answer to the question the script refers to. You may use auxiliary `.sh` scripts to break up your question scripts to make them easier to work with if you wish: in this way you can also use common scripts across questions, for example to clean the data and make sure you "link" your scripts in the answer script where relevant. We should be able to take your scripts for an answer, run them on the VM that has been distributed to you and obtained the correct output.

The template files where you have to store your solution scripts for the questions are:

- `bq1.sh`, `bq2.sh`, and `bq3.sh` for your solutions to the `bash` portion of the assessment (Q1, Q2, Q3 respectively).
- `sq1.sh`, `sq2.sh`, `sq3.sh` for your solutions to the SQL portion of the assessment (Q1, Q2, Q3 respectively).

Example script files you may produce for the assignment (bash part only shown), and references between them (arrows) are portrait in the image below:



Finally, you have hung on very well during these first 4 weeks of IFN509 and positively committed in learning to use the command line, an environment unfamiliar to most of you. For this you deserve to be rewarded. Therefore, for working on portfolio item 1 we want to make you a gift. We can created a new virtual machine. The instructions for a new VM is in IFN509 Blackboard. Download the VM and "fire it up". As you will see once the VM is running, you will presented with a graphical interface (Ubuntu Desktop), rather than the command line. Although the system is still Linux based, you now have a more familiar interface, that lets you open a browser, navigate the internet, etc. Importantly for the assessment, it lets you open a CLI (called `Terminal`): this is exactly the CLI interface we have been working with up until this point.

However, this time you will be able to resize the window with the console, scroll up and down the screen, and also copy and paste on the command line. We hope you will like our gift: you still have to work in the CLI and write scripts for the CLI, but now you can do so in a comfortable environment. Indeed, you can even use a graphical text editor to write your scripts! The important note is that you make sure they execute on CLI and get the right answer. You may even decide to reboot the VM in command line mode, rather than graphical, if you want.

Please, make sure you use this virtual machine (and not the old one). You are requested to submit your scripts as a zip file in Blackboard (see the end of this portfolio item assignment): to do so you need that the VM in which you write your scripts has internet access and a browser you can use.

Next, we describe the 3 questions you need to solve. Remember, for each question we expect you to produce one `bash` script that uses `bash` instructions only, and one `bash` script that uses SQL instructions.

Preamble: what the data is about

First though, let us let you what the data is about.

This dataset contains details about biopics: a specific genre of movies. Unlike documentaries, which typically include raw footage and interviews, biopics are dramatizations, loosely based on the real-life events of actual people. Biopics offer an interpretation of lives deemed important (and profitable) by Hollywood, and they often try to make a statement about their subjects' historical or cultural significance. So which figures filmmakers spotlight matters, as does whom they ignore (or can't get the funding to feature).

(text and data have been taken from the article at https://fivethirtyeight.com/features/straight-outta-compton-is-the-rare-biopic-not-about-white-dudes/. Note, the original data has been manipulated to ease the tasks you are faced with)

Q1: How many movies has a director made?

Firstly, we would like you to spend some time exploring with the data, and cleaning it up. Then, you should

code up scripts that answer the question: How many movies has a director produced?

The `bash` solution is to be written in **`bq1.sh`** and the SQL solution is to be written into **`sq1.sh`**.

Requirements

Your **`bq1.sh`** script must do the following:

1. download the data from https://goo.gl/BhphrS and store it in a file called `biopics.csv`
2. clean the data in an appropriate way to answer this question (and possibly the next ones)
3. produce a text file called **`ba1.txt`** that contains a list of directors along with how many movies they have produced. Note, each director should be named only once. The first column of your answer must be a list of directors sorted alphabetically and the second column must be the total number of movies they have made. You can **only** use `bash` commands for your answer.

Your **`sq1.sh`** script must do the following:

1. download the data from https://goo.gl/BhphrS and store it in a file called `biopics.csv`
2. clean the data
3. import the data into SQLite using (all in one line):

   ```
   python3 csv2sqlite.py --table-name biopics --input biopics.csv
   --output biopics.sqlite
   ```

   `csv2sqlite.py` is a *Python script* that converts any delimited-type file (for instance csv). This file is included in the archive distributed with the assignment. Just like `sqlite3` and `bash` scripts, we need to invoke it with the `python3` command. It takes several arguments, but the required ones have been filled out for your in the command above. If you would like an explanation of the arguments and the command, run `python3 csv2sqlite.py --help`. If you are feeling brave, you may wish to read the *source code* to understand what the script is doing (using more or less, etc.).

4. produce a text file called **`sa1.txt`** that contains a list of directors along with how many movies they have produced. Note, each director should be named only once. The first column of your answer must be a list of directors sorted alphabetically and the second column must be must be the total number of movies they have made. You **must** use SQL queries to manipulate data for your answer as part of a `bash` script pipeline.

Example Output

```
A.J. Edwards,1
A.W. Vidmer,1
Adam Green,1
Adrian Shergold,1
Alan Rudolph,1
Alexandre Moors,1
Alfred Hitchcock,1
Alfred L. Werker,1
Allen Coulter,1
Anatole Litvak,1
Andrew Bergman,1
Andrew Dominik,1
Andrew V. McLaglen,1
```

Pay particular attention that your output does not have: string that are not names, for example `-` or `"`. Note there are some names that have a strange encoding and thus may contain weird characters, e.g. `Roland JoffÌÎå©`. This is fine. Also, for the purpose of this exercise, consider strings like `Roland JoffÌÎÌãÌ´å©` and `Roland JoffÌÎå©` as referring to two different directors.

Q2: Does gender influence how much a movie earns at the box office?

For the second question, we would like you to process the data to extract aggregate information. The `bash` solution is to be written in **`bq2.sh`** and the SQL solution is to be written into **`sq2.sh`**.

Requirements

Your **`bq2.sh`** script must produce a HTML file called **`ba2.html`** that contains a HTML table that matches the example output (but with the actual correct amount of money). You do not need to style it, basic is better. Your table should summarise the total amount of money earned by each gender. Note that a movie may have more than one subject, and they may be of different genders. For example, a movie that erned $10M at the box office, may have a female and a male subject. In this case, count $10M as the contribution of the femal subject, as well as the one of the male subject (i.e. no need to average or divide the contribution). You can **only** use `bash` commands for your answer.

Your **`sq2.sh`** script must produce a HTML file called **`sa2.html`** that contains a HTML table that matches the example output (but with the actual correct amount of money). You do not need to style it, basic is better. Your table should summarise the total amount of money earned by each gender. Note that a movie may have more than one subject, and they may be of different genders. For example, a movie that erned $10M at the box office, may have a female and a male subject. In this case, count $10M as the contribution of the femal subject, as well as the one of the male subject (i.e. no need to average or divide the contribution). You **must** use SQL queries to manipulate data for your answer as part of a `bash` script

pipeline.

Note, a movie for which box office income is not reported, should not be calculated in the total amount.

Example Output

Gender	Total Amount
Female	$
Male	$

Q3: How much box office earnings do biopsies generate in each year?

For the final question, we would like you to again process the data to extract some aggregate information. The `bash` solution is to be written in **`bq3.sh`** and the SQL solution is to be written into **`sq3.sh`**.

Requirements

Your **`bq3.sh`** script must produce a HTML file called **`ba3.html`** that contains a HTML table that matches the example output. You do not need to style it, basic is better. Your table should summarise the total amount of money earned each year by a biopic movie (each movie in the table is a biopic). You can **only** use `bash` commands for your answer.

Your **`sq3.sh`** script must produce a HTML file called **`sa3.html`** that contains a HTML table that matches the example output. You do not need to style it, basic is better. Your table should summarise the total amount of money earned each year by a biopic movie (each movie in the table is a biopic. You **must** use SQL queries to manipulate data for your answer as part of a `bash` script pipeline.

Example Output

Year	Average Gross
2010	$
2011	$
2012	$

Note, you need to include all the years in the final table.

Submission of the assessment

When working on this portfolio item, make sure you work in a new empty folder in your virtual machine. Once you have finished working on the assessment, you have to upload a zip file with all the content of your directory (and importantly the answer scripts) to blackboard (submission link in the BB folder for Portfolio Item 1, under the Assessment section in Blackboard). The zip file should be called `yourStudentNumber-pi1.zip`, where `yourStudentNumber` is your student number without the leading `n`.

Remember, this assignment is to be performed **individually**. Attempt to answer all questions.

The deadline for submitting your work for this portfolio item in BB is 11.59pm on Sunday 1st April 2018. No late submissions will be allowed. You can submit unlimited attempts, so once you have something that you believe is correct, submit it through BB: you can always go back and change it until the deadline. Start the assignment soon, do not leave it to the last minute: the assignment takes time to resolve, especially to get setup for the first answer.

If you are sick or require special consideration/deadline extension, you need to apply through the official process at QUT. Please also inform the teaching team of this so that we are aware: however this is not a requirement if you do not wish to do so.

Questions and doubts

Please, if you are unsure about what the instructions are or believe there is a mistake or problem in the assignment, inform as *ASAP* through Emails or Slack. Everyone has been invited to the channel. Please `do not` share solutions in the channel, but only doubts, problems, possible mistakes in the requirements, questions, etc.