# COMP1204
# Unix Coursework

## 1 OVERVIEW

This coursework will cover two key topics that will have been covered in the first three weeks of the module: Unix and Latex. The coursework is divided into three parts: (i) Unix scripting for basic file processing, (ii) Unix scripting for data analysis and (iii) Report writing using Latex. Each part carries a percentage of the marks for this coursework (out of a 100) as detailed in Table 1. You should be able to complete various parts of this coursework as we go through the lectures, and parts of this assignment will be covered during lab sessions. Help will also be available throughout the labs. The key points are:

- This Coursework counts for 10% of this module.
- The deadline[1] for submission of your report and scripts: **11$^{th}$ March 2019 by 4pm.**
- Feedback will be given within 4 weeks after the deadline.
- You are only allowed to use Bash (Unix) commands and Latex (Overleaf and other editors are acceptable). Use of other scripting languages or text editors will get you zero marks for the relevant sections.

For submission instructions, please see Section 4 at the end of this document.

[1] Failure to submit by the deadline will incur a 10% penalty per working day. Submissions later by more than five working days will not be accepted.

# 2 Learning Outcomes (LOs)

This coursework aims to achieve the following learning outcomes:

- Knowledge of Unix commands including but not limited to sed, awk, grep and pipes.
- Knowledge of Latex commands and document preparation.
- Data cleaning techniques using pattern matching and filtering

# 3 The Assignment

You work as a data scientist for TripAdvisor and your job is to help them make sense of what hotels are performing well. You have been tasked with analysing all the files containing reviews for each hotel as described in the next subsections.
Before you start, create a project on `https://git.soton.ac.uk`. The project should be named comp1204-userid-cwk1. So if your userid is mjk7f49, the project should be named comp1204-mjk7f49-cwk1. **Make sure the visibility level of the project is set to private. Now invite user mjg1e16 as Maintainer on the project.**
Please make sure to regularly commit your work on the Unix scripts to your git project.

## 3.1 Dataset

The dataset to be used for this coursework is the TripAdvisor dataset at: `https://secure.ecs.soton.ac.uk/notes/comp1204/coursework/dataset/reviews_dataset.tar.gz`. Download this file to a folder on your home drive (e.g., myworkspace). Extract the file using standard UNIX file decompression commands.

### 3.1.1 Basic file processing – 40%

You would like to find out what hotels are commented on the most (the frequency of comments may indicate how many guests they actually receive). To answer this question:

1. Copy a hotel data file to your home directory (e.g., hotel_72572.dat).
2. Write a Unix script called *countreviews.sh* that counts the number of reviews in the file that takes input from the command line like this:

   ```
   % ./countreviews.sh  hotel_72572.dat
   15
   ```

   where hotel_72572.dat is an example file name. Note that 15 is just an example and not the actual result for this file. Also note that the argument to the script, hotel_72572.dat, is just an example file PATH; your script must be able to handle both relative and absolute paths correctly. **Do NOT submit this script on the ECS electronic hand-in system – complete the next steps and submit the final script for this section.**

3. Extend *countreviews.sh* to count the number of reviews in each file given the folder name (i.e., where all your files are stored).

   ```
   % ./countreviews.sh  path_to_reviews_folder
   15
   12
   13
   ...
   ```

   Remember again that the argument to the script, path_to_reviews_folder, can be an absolute or a relative path; your script must be able to handle both types of paths correctly.

4. Finally rank all the hotels according to the review count so that the hotel with the most reviews is at the top of your list.

The output of your complete script should be formatted like this:

```
% ./countreviews.sh  path_to_reviews_folder
hotel_1322 50
hotel_21313 49
hotel_31331 45
...
```

The argument to the script, path_to_reviews_folder, can be an absolute or a relative path; your script must be able to handle both types of paths correctly.
Regarding the output of the script, please note the following:

1. The .dat extension is omitted from the filename.
2. The hotel name is separated from the count by a single whitespace.
3. Nothing but the hotel-count pairing should be output by the script.

**Submit only your final script on the ECS electronic hand-in system.**

### 3.1.2 DATA ANALYSIS – 40%

Now that you know the top scoring hotels, you'd like to do some further analysis on the ratings these hotels get.
Write a script that returns the ranked list of hotels according to their **average overall rating** (i.e., the mean rating over all the reviews they've received). Do not use the 'Overall rating' field of the hotel and instead compute your own based on the 'Overall' fields in each review. The command and output (sorted by value) should look as follows:

```
% ./averagereviews.sh  path_to_reviews_folder
hotel_11212 3.51
hotel_2121 2.62
hotel_31212 2.43
...
```

The argument to the script, path_to_reviews_folder, can be an absolute or a relative path; your script must be able to handle both types of paths correctly.
Regarding the output of the script, please note the following:

1. The .dat extension is omitted from the filename.
2. The hotel name is separated from the average by a single whitespace.
3. Nothing but the hotel-average pairing should be output by the script.
4. The average should be rounded to two decimal places. For example, 4.8199 should be rounded to 4.82, 2.0621 should be rounded to 2.06, and 3.7354 should be rounded to 3.74.

#### 3.1.3 REPORT – 20%

Write a report in Latex detailing the following:

1. The cover page of the report have at least your name and ID written on it as well as the title.
2. The report should be divided into two sections:
   - Scripts: the countreviews.sh and averagereviews.sh scripts you wrote. Make sure you clearly explain what the scripts are doing. Your scripts should preferably be written using the 'listings' environment from Latex.
   - Discussion: a discussion of what challenges TripAdvisor faces in collecting reviews in this way (e.g., data storage or query challenges). How would you help them improve the way they collect reviews? How would you help them ensure the reviews are trustworthy?

### 3.2 ASSESSMENT CRITERIA

| | |
|---|---|
| Unix Script for basic file processing | 40% |
| Unix Script for data analysis | 40% |
| Report Writing in Latex | 20% |

Table 1: The weighting given to the different parts of this coursework.

Your code and your report will be evaluated as follows:

**Unix**
For Unix scripts, we will use an automated script checker that will pipe data to your code and check the output. Your scripts will be tested on previously unseen data. In particular, we will check that:

1. Code returns expected output.

2. Code cleanliness and efficiency (slow code will be penalised).

3. Appropriate use of git versioning system in writing your scripts. In particular, we will assess if meaningful commits and commit messages have been provided.

Indicative feedback may include: code does not work, code works partially (i.e., some functions not working), code is inefficient, code not readable, git commit messages not provided, all scripts work.

**Report**
For the report we will check the following in particular:

1. Report written in LATEX and any script listed using standard LATEX environments e.g., verbatim, listings, or algorithmic.

2. Use the 'article' document class to create your report.

3. The title page of the report should contain student name and id information.

4. The report should be appropriately structured into sections and subsections.

5. The script code environment should be appropriately captioned.

6. Single and double quotes should be correctly typeset.

7. For the discussion:
    - Compare the use of unstructured markup vs. structured database for representing the data.
    - Present 2-3 ideas to authenticate the authors of reviews.
    - Present 2-3 ideas on improving the review ranking system.
    - Discuss data storage issues with TripAdvisor's flat file-structure.

Indicative feedback may include: poorly structured document, well structured document, discussion omits key issues, discussion proposes innovative ideas, discussion missing clarity, approach not credible.

## 4 Submission Instructions

While working on your scripts, countreviews.sh and averagereviews.sh, you would have made several commits on git. For your submission generate a git log file – named git.log.
Submit your work using the ECS electronic hand-in system. The submission is to be made by **4pm** on the due date listed above. Please submit a single file to the ECS electronic hand-in system as detailed below:

- Your scripts and report must be in compressed archive named as comp1204.tar.gz. You should submit the scripts (countreviews.sh, averagereviews.sh), the git log file (git.log), your report in LATEX (report.tex and report.pdf), and the LATEX log file (report.log) that is generated when you compile your latex file (on Overleaf you need to navigate the options to download the log file).
- Your report should be in .PDF format and be included in the archive.

Failure to follow these instructions will incur a penalty. In particular, you will lose (possibly all) marks if:

1. You use Word to create your document.
2. You compile to ZIP and change the extension of the file.
3. You submit a word doc and change the extension of the file to PDF.
4. Your code runs for more than 1 min on the test dataset.