# UNIX Coursework (COMP1204)

Georgios Alexiou
30097762

March 11, 2019

# 1 Scripts

## 1.1 Basic File Processing

For the Basic File Processing (3.1.1) Section of the Unix Coursework the task was to create a script that would count the number of reviews in each file given the directory that they are stored in. The script was compiled using *chmod 755 countreviews.sh* and ran using *./countreviews.sh [directory path].*

The script can be seen below:

```
1   #!/bin/bash
2   cd "$1" || exit 1
3   grep -c Author hotel_*.dat | sort -t : -k2nr | sed 's/\.dat:/ /'
```

As mentioned above the user passes the path of the desired directory through as an argument when running the program. In line 2 we navigate to the indicated directory. If the directory does not exist the program will exit.

In line 3 we use the grep -c command to count how many times the string Author appears in each *hotel_*.dat* file signifying the beginning of a new review. We then sort the output in decreasing order according to how many reviews each file contains. Finally using sed we modify the output to match the specification. Below is a sample output of the script when run in the terminal.

```
hotel_218524 2686
hotel_149399 1552
hotel_208454 1225
hotel_150841 1213
hotel_93569 1211
hotel_87016 1196
hotel_149397 1119
hotel_218492 1093
hotel_115644 917
hotel_218486 763
hotel_93507 746
hotel_113317 721
hotel_149395 718
hotel_150849 710
hotel_86978 667
hotel_148598 610
hotel_149921 587
hotel_93593 577
```

The output shows the filename followed by the number of reviews it contains.

## 1.2 Data Analysis

For the Data Analysis (3.1.2) Section of the Unix Coursework the task was to create a script that would rank the hotels based on their average overall reviews. Similarly to Section 1.1 Basic File Processing the script was compiled using *chmod 755 averagereviews.sh* and ran using *./averagereviews,sh [directory path].*

The scroipt can be seen below:

```
#!/bin/bash
1    #!/bin/bash
2    cd "$1" || exit 1
3    grep "<Overall>" hotel_*.dat | sed 's/\.dat:<Overall>/ /' |
     awk '{sum[$1] += $2; counts[$1]++;} END {for (i in sum)
     printf "%s %.3g\n",  i, sum[i]/counts[i];}' | sort -nrk2
```

Similarly to 1.1 the user passes the path of the desired directory as an argument when running the program. In line 2 we navigate to the indicated directory. If the directory does not exist the program will exit.

In line 3 we use the grep -c command to count how many times the string <Overall>appears in each file. We then use sed to convert the output from:

```
hotel_99762.dat:<Overall>5
```

to:

```
hotel_99762 5
```

As seen above, the output now consists of two columns, the filename and the overall score. Using the awk command we find the average overall score for each filename. This means that the script reads through the filenames and for all the rows that have the same filename it takes the average of their overall score. We then print the result and what we can observe that for each filename there is only one value corresponding to it being the average. As indicated in the coursework specification the average values should be output using 2 decimal places which is done with the parameters next to the printf command. We then sort the output of awk in a decreasing order and are thus left with the final output. This can be seen below:

```
hotel_203921 4.78
hotel_188937 4.78
hotel_230572 4.75
hotel_185406 4.74
hotel_190664 4.73
hotel_188961 4.73
hotel_193121 4.72
hotel_224953 4.71
hotel_194233 4.7
hotel_147790 4.7
hotel_187737 4.69
hotel_195006 4.67
hotel_224221 4.66
hotel_197794 4.66
hotel_190694 4.66
hotel_194308 4.65
hotel_190614 4.65
hotel_197666 4.64
hotel_195703 4.64
```

The output is the filename followed by the average overall score for each hotel.

# 2 Discussion

## 2.1 Unstructured vs. Structured Data

Structured data refers to the data that follows a specific pattern thus making them readable and easily searchable. On the other hand, unstructured data is the data that is not searchable usually because it does not follow a clear structure or because it consists of different types of file formats such as video or audio. The choice between unstructured and structured data is done depending on the use of the database and what it contains.

More recently there has been a growing trend in the adoption of structured data in databases as a way of making them more organised and efficient. Structured data follows specific formats that aid in their ease of search and accessibility by the user or a computer system.

Unstructured data is everything that might have an internal structure but us not organised using specific standards like the structured data. It usually consists of user-generated files, such as emails, websites, text files, media or scientific data.

## 2.2 Authentication of review authors

We live at a time where travel-related user generated content is increasingly popular. One of the websites that provides that provides such a service is TripAdvisor and a common problem that the website constantly is whether the review is credible enough. There is a current system in place where specific reviewers that have been using the platform for a long time receive virtual badges that show that their review is verified by TripAdvisor themselves.

A way this could be further improved is using a downvote - upvote system where users will be able to rate other people's reviews based on their credibility and accuracy. This will result in users receiving a score according to the nature of their activity on the website that may or may not make them seem more credible.

Another way to authenticate users on the platform is making them link other social media accounts with their TripAdvisor account which will in turn force review authors to write more credible reviews as their friends might be able to see the review.

## 2.3 Improving the ranking system

To improve the ranking system, the users should be able to assess their experience in the hotel they stayed using more detailed 'grades'. That means that more areas of the hotel need to be assessed in order for the reader of the review to gain an idea as to what the hotel actually have to offer. Doing this will help get an overall score that is more accurate according to the experience the respective reviewers had in that hotel.

There should also be a level of content curation as far as reviews are concerned where reviewers that have been using TripAdvisor for a long time and are trusted by the company, should have their reviews appear on top, as well as be able to select other good reviews and rate them based on their quality. TripAdvisor is community driven and such initiatives would help the platform grow and develop in terms of its ranking system, as people will have influence over

## 2.4 Data Storage Issues with flat file structure

There is a variety of issues involving the storing data using flat file databases such as the one presented in the coursework. A main problem is that the computer has more data to read through making features such as searching, accessing or editing the data a difficult and slow process. This is because the computer has a larger amount of data to access in order to reach the desired file that contains the information that is requested.

Another error in flat file databases is the fact that data has to be repeated which usually leads to inconsistencies and inputting errors. That is very common as a lot of the data needs to be input by the user and the file format requires the rewriting of specific parts of the document several times. This was observed in the hotel files where they all have the same structure. This repetition also leads to larger file sizes which in turn take up more space on TripAdvisor's servers.