COMP1204
Unix Coursework

Alberto Tamajo

Student ID: 30696844

March 2020

Electronics and Computer Science Department
University of Southampton

Introduction

This report has been written in order to explain how the script **countreviews.sh** that I have **personally** written for the COMP1204 Unix coursework achieves the objectives stated by the following assignment sheet: `https://secure.ecs.soton.ac.uk/notes/comp1204/coursework/UNIX_Coursework.pdf`

How the script works

The aim of the bash script below is to count the number of reviews of each TripAdvisor hotel contained in the following archive: `https://secure.ecs.soton.ac.uk/notes/comp1204/coursework/dataset/reviews_dataset.tar.gz` (or in an archive that stores information about hotels with the same structure) and rank the hotels according to the their review count so that the hotel with the most reviews is at the top of the list.

This is achieved by using a for-loop that iterates over all files inside a directory whose names start with the string "hotel". The directory path is provided as a command-line argument and it can be both an absolute or relative path. The directory path is obtained by using the argument variable $1 used in line 4 of the script.

For each loop, the name of the hotel separated from its review count by a single white space are piped to the command located in line 10 of the script. This command *sort -nr -k 2* sorts all the hotels iterated by the for-loop in reverse order *"option -r"* by comparing the numeric value *"option -n"* contained in the field number 2 *"option -k 2"* which represents the hotel review count.

In order to understand how the name of the hotel and its review count are computed we need to look at line 8 of the script. The file path is extracted by using the following command *echo $file* which pipes the path of the file to the *sed 's/.dat//'* command that removes the ".dat" extension of the file. In order to retain just the name of the file (name of the hotel) the *awk -F/ '{print $NF}'* command is used which separates the file path by using the delimiter '/' *"option -F/"* and prints the last field *'{print $NF}'* (name of the hotel).

Similarly, the review count of each hotel is computed by using the command *grep -c "<Overall>" $file* which counts the number of lines inside each hotel file "option *-c"* that contain the following string *"<Overall>"*. Since, each review has an overall rating then counting the number of lines that contain *"<Overall>"* is equal to counting the number of reviews of each hotel.

```
1  #!/bin/bash
2
3  #Loops through all the files inside a directory  whose names start with "hotel"
4  for file in $1/hotel*
5  do
6
7    #The name of the hotel and its review count are printed out
8    echo $( echo $file | sed 's/.dat//'| awk -F/ '{print $NF}') $(grep -c "<Overall>" $file)
9
10 done | sort -nr -k 2   #Ranks the hotels into descending order based on their review count
```

Listing 1: countreviews.sh