UNIX Coursework

Data Management (1204)

Denitsa Radeva
Student ID: 31693849

1 My Script

Listing 1: Extracting information from files.

```
#!/bin/bash

for file in $1/*;
do
filename=$(basename $file);
printf $filename" "|sed 's/.dat//g';
cat $file |grep 'Content'|sed 's/>/ /g'
|awk 'BEGIN{count=0}($1=="<Content"){count+=1}
END{print count}';done|sort -k2,2gr
```

2 Description

The above script is extracting the number of reviews that are mentioned in all data files from a specified directory, which is given as a user input. The name of the script is 'countreviews.sh' and it expects a name or an absolute path of a directory as an argument. The following methodology is used:

1. **For cycle** to loop through all files within a given directory.
2. The **basename** utility to extract the current file's name.
3. Usage of **sed** to remove the suffix '.dat' from the current file's name.
4. The utility **grep** to find all occurrences of the word "Content", because each time it is found, that's exactly +1 review more.
5. The utility **sed** to replace the symbol '>' with a white space in order to separate "<Content" from the text of the review and to spot it easily in the next step.
6. The utility **awk** to count how many times "<Content" is found throughout the document.
7. Usage of **sort** in the end to sort in a descending order by the second column of the output, which contains the number of reviews found.