UNIX Coursework Report
Data Management COMP1204

University of Southampton

Suyash Datt Dubey

sdd1n17

ID: 29533414

1 Scripts

1.1 Counting Reviews Script

The countreviews script counts the number of reviews for each hotel database file in the folder. It prints the file name and number of reviews, sorted in descending of the number of reviews.

```
#!/bin/bash

for i in $1/*
do
    grep -Hc \<Location\> $i | sed 's/.*\///g' |
    sed 's/.dat:/ /g'
done | sort -k2 -nr
```

The for loop iterates over each file in the folder. The grep -Hc command prints the hotel name and counts the occurrences of "Location" in each file in order to count the number of reviews. It then prints the name of the hotel followed by the number of reviews for that particular hotel. The first sed command it removes the entire path of the file. The second sed command replaces the extension of the file with a space. The sort command sorts the files in descending order of the number of reviews.

1.2 Calculating Average Reviews Script

The averagereviews script calculates the average overall rating of each hotel database in the folder. It prints the file name and average overall rating, sorted in descending order of the average overall rating.

```
#!/bin/bash

for i in $1/*
do
    grep -H \<Overall\> $i | sed 's/.*\///g' |
    sed -e 's/.dat:<Overall>/ /g' |
    awk '{sum += $2} END {print $1; printf "%1.2f\n", sum/NR}' |
    xargs
done | sort -k2 -nr
```

The for loop iterates over each file in the folder. The grep -H gets the value of "Overall" for the hotel file. The first sed command removes the entire path of the file. The second sed command replaces the extension of the file with a space. The awk command adds the value from the second column to a variable sum. It then prints the first column and average overall rating of the hotel up to 2 decimal places. The xargs ensures that both print commands are executed and hence prints out the hotel name as well as the average overall review. The sort command sorts the files in descending order of the average overall rating.

2 Discussion

2.1 Challenges Faced

One of the biggest challenges faced by websites such as TripAdvisor is the legitimacy of reviews. Anyone can log on to the website, create an account, and a leave a review for a property. There is no form of verification if the user writing the review has actually visited the property. A person can use multiple email addresses and devices to write several reviews. Property owners themselves can manipulate reviews in order to generate more publicity and attract more customers to their properties. These reviews can be extremely misleading for other users.

The falsified reviews tamper with the legitimate ones and make the collected data unreliable. A user may accidentally post the same review multiple times, thus making it harder to use scripts and queries to access the data. The data collected is vast and multiplies over the years. The website needs to devise innovative techniques to store this data as traditional storage methods are not going to be effective.

2.2 Improving Review Collection Methods

TripAdvisor collects vast quantities of data every day. Thousands of reviews of properties are written on the website by users every day. Not all of these reviews are relevant for data analysis. It is imperative to establish what data is essential to the website. The forms and questionnaires to be filled out by users must be simple to comprehend and short. Lengthy ones take a long time to fill out as well as to process. Confidentiality of users is also essential and should be kept in mind when collecting personal data and reviews from them.

2.3 Ensuring Reviews are Trustworthy

One method of ensuring reviews are legitimate is the website checking with the property for the proof of user's visit as a form of verification. Other feasible methods include uploading an image of the receipt or verifying phone number with the property and ensuring the phone number is valid and unique to the user account. The user community should also take responsibility to flag and report reviews that they think might be misleading. These reviews should then be investigated by a guidelines compliance team. Detection software should be used to detect fraudulent reviews and suspicious activities.