

# Data Management Course Work

Anish Katariya

February 2016

Student ID:27561879

Email ID: ak7n14@soton.ac.uk

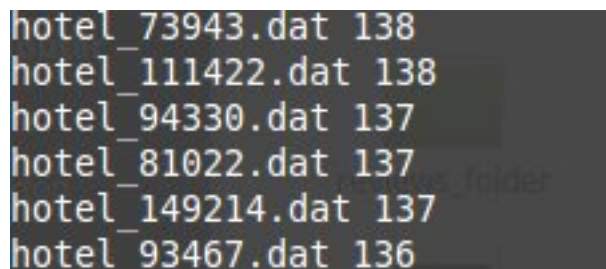
# 1 Scripts:

## 1.1 countreviews.sh :

The 1st script is countreviews.sh. The script takes in the reviews folder as an argument, counts the number of reviews in each file(hotel) in the reviews folder directory, ranks them according to the number of reviews in each hotel and prints the number of reviews with the hotel name. The script is given below:

Listing 1: countreviews.sh

```
#!/bin/bash
for f in $1/*.dat;
do
    echo -n $(basename $f)" "
    grep '<Author>' $f | wc -l
done | sort -nrk2
```



Hotel File	Number of Reviews
hotel_73943.dat	138
hotel_111422.dat	138
hotel_94330.dat	137
hotel_81022.dat	137
hotel_149214.dat	137
hotel_93467.dat	136

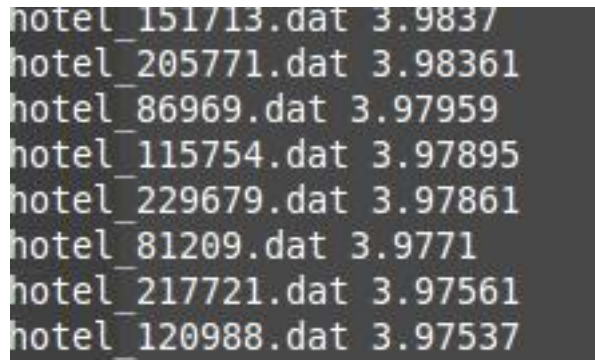
Figure 1: countreview output example

## 1.2 averagereview.sh :

The 2nd script is averagereviews.sh . The script takes in the reviews folder as an argument, goes through all the overall ratings given to each file(hotel) in directory and takes their average overall rating and displays them according to their ranking in decreasing order of overall ranking. The code is given in the next page

Listing 2: avreagereview.sh

```
#!/bin/bash
for f in $1/*.dat;
do
    echo -n $(basename $f) " "
    grep "<Overall>" $f | egrep -o "[0-9]+" | awk '{
        count+=1;
        SUM+= $1;
    } END{
        print SUM/count}'
done | sort -nrk2
```



The screenshot shows the output of the script, which lists hotels and their average review scores. The output is as follows:

Hotel	Average Review Score
hotel_151713.dat	3.9837
hotel_205771.dat	3.98361
hotel_86969.dat	3.97959
hotel_115754.dat	3.97895
hotel_229679.dat	3.97861
hotel_81209.dat	3.9771
hotel_217721.dat	3.97561
hotel_120988.dat	3.97537

Figure 2: averagereview output example

### 1.3 statistical\_sig.sh:

The 3rd script is statisticalsig.sh. The script takes in two hotels as arguments, calculates their average, standard deviation, t test statistics, t value and prints 1 if the t value is significant and zero if the t value is not significant i.e. 0 if the t value is less than the critical value and 1 if the t value is more than the critical value. The code for the script has been given on the next page.

Listing 3: Statistical<sub>s</sub>*ig.sh*

```
#!/bin/bash
for f in $@;
do
    echo -n $(basename $f) " "
    grep "<Overall>" "reviews_folder/$f.dat" | egrep -o "[0-9]+" | awk '{
        count+=1;
        SUM+= $1;
        Sum_square+= $1^2;
    } END{
printf "%d %d %d ",count,SUM,Sum_square;

    },
done | awk '{

    split($1 " $5, hotelName, " ");
    split($2 " $6,count, " ");
    split($3 " $7,sum, " ");
    split($4 " $8,Sum_square, " ");
    for (i=1;i<=2;i++){
        sd[i] = sqrt( ( Sum_square[i] - ( sum[i]^2 / count[i] )
                                                                / ( count[i] - 1) ));
    }
    df=count [1]+count [2]-2;
    sx1x2 = sqrt( ( ( count [1]-1 ) *sd [1]^2 +
                    ( count [2]-1 ) *sd [2]^2 ) / df );
    t= ( (sum [1]/count [1] - sum [2]/count [2]) / (sx1x2 *
                                                sqrt( 1/count [1]+1/count [2])));

    printf "t : %.2f \n",t;
    for(i =1; i<=2;i++){
    printf " Mean %s : %.2f SD :
%.2f\n",hotelName [i],sum [i]/count [i],sd [i];
    }
    if(t>1.965261468090270)
        print "1"; #Significant
    else
        print "0"; #insignificant

    },
```

```

l_203921
t : 0.03
Mean hotel_188937 : 4.78 SD : 0.63
Mean hotel_203921 : 4.78 SD : 0.53
0

```

Figure 3: Statistical\_sig output example

## 2 Hypothesis Testing :

We were given row data containing reviews given to hotels on TripAdvisor with a file provided for each hotel. The file contained all reviews that the hotel had got including its rating in individual departments, its overall rating and comments given by the visitors of that hotel.

The hypothesis tests we ran was t-tests and found out the t-statistics as the type of data given to us had its in dependant variable as categorical and its dependant variable as continuous and taking averages of the overall ratings of the hotel would not be accurate as the number of reviews in each hotel were different and each hotel was tried by different groups of people and as a result the average would have no standard benchmark.

The t-test that was run on the top two hotels showed that the t value was less than the critical value as a result the difference in the ratings of the hotels was not significant. These tests showed us that the rankings of the hotels purely on the basis of average overall rating as done by TripAdvisor was not accurate and as a result it could be misleading to the users of Trip Adviser

## 3 Discussion:

The current method by which TripAdvisor ranks its data is inaccurate. Ranking hotels by average overall rating is unpredictable as the number of reviews on each hotel are not the same and may lead to misleading results. Different hotels are visited by different groups of people who may prefer other services as compared to another group of people as a result classifying the data according to average overall rating might not help the user to choose the best hotel for themselves.

To overcome these problems TripAdvisor should rank their hotels in accordance with their t-statistic values for a better and accurate rankings. TripAdvisor should also take regular surveys from their users about the services in the hotel they care about the most and should provide them with personalized rankings with hotels having good ratings in those results having a higher ranking.

To make sure that the users leaving comments are trustworthy TripAdvisor should crosscheck bookings to see if that user actually visited that hotel or not.

TripAdviser should also advertise and use reviews left by regular users in order to get more trustworthy and accurate reviews.