Tanuj Sood, Neeraj Pandey and Suvam Hota
Prof. Anirban Mondal
22 April 2020

Project Report

Data Mining Lab Assignment

Introduction

For this lab assignment, our task was to mine data on twenty hotels from three countries with a total of a hundred comments for each hotel. The objective of this lab assignment was to analyse this data and conduct sentimental analysis to attain the top ten hotels from each country based on our mined data. For our project, we took several steps which helped us acquire our final results from our preexisting dataset. These parts can be divided into the following components:

1. Mining Data
2. Clustering Dataset
3. Performing Sentimental Analysis
4. Visualisation of Results
5. Using Mapbox to create comprehensible visual data

Following our desired plan, we hope to acquire conclusive results and gain interesting insights from our analysis of the given comments. Our code has been posted on the Github link and contains all parts of our planned process as well as the images we created using softwares such as WEKA and Mapbox.

Mining Data

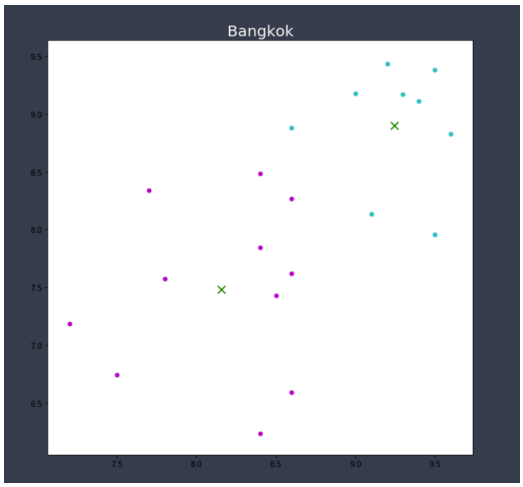
In order to acquire a big dataset where we could acquire conclusive results, we chose to use the famous travel website booking.com to look for twenty hotels each in three different countries. The countries we chose were Singapore, Bangkok and Kuala Lumpur as these had a sufficient amount of rated hotels with hundreds of comments we could mine.

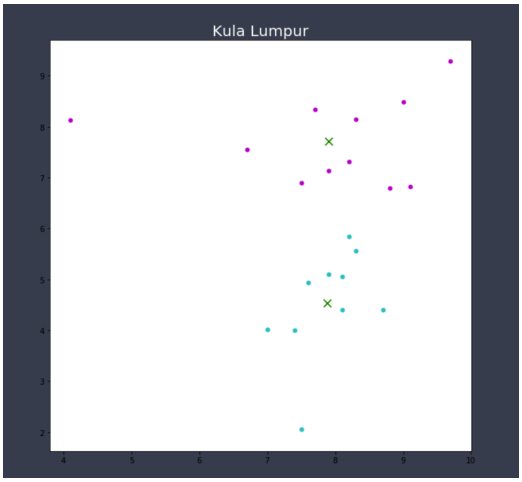
Our mining process included the use of the online tool selenium through which we copied the comments and added them to our hotel for all of our chosen list of sixty hotels gathering a

total of around six thousand comments as required. To store this data, we created a JSON database as well as a CSV file which included the name of the hotel, its address and all the reviews we received. Thus, through this we acquired our apt dataset and proceeded to its analysis.

Clustering Dataset

We clustered the data for our hotel list and through softwares such as WEKA, plotted them on to scatter graphs to get a better understanding of the overall rating for the hotels in our list. We repeated this for each chosen country and acquired the scatter plots shown below:



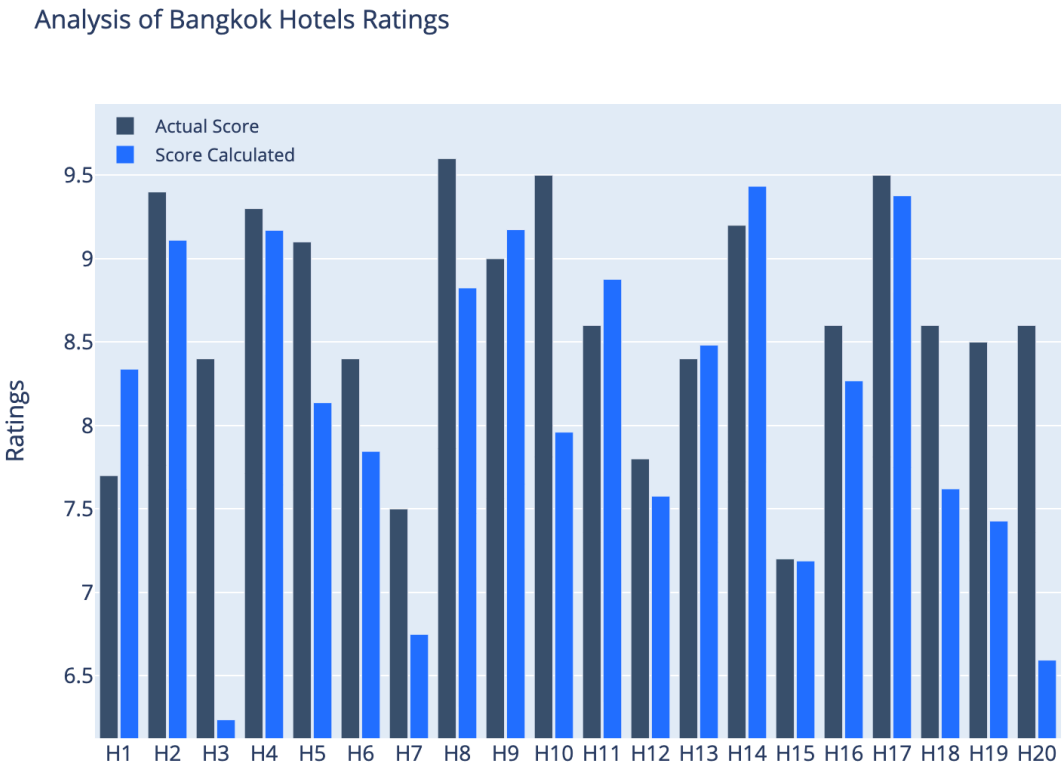


The results gave us a rough idea of the overall ratings of the twenty chosen hotels in each country and we moved on to our main process which is to conduct the sentimental analysis on our dataset.

Performing Sentimental Analysis

Sentiment analysis, also called opinion mining, is a text mining technique that could extract emotions of a given text — whether it is positive, negative or neutral, and return a sentiment score. Once we have successfully extracted all reviews from Booking.com, we are ready to get the sentiment score for each review using Python. First, we'd import the libraries. We will use a library in Python called NLTK. We apply this technique to acquire a sentiment score for each of our hotels based on its reviews. There are four types of scores: negative, neutral, positive and compound. Then we have the sentiment score for each review. Each review has a negative, neutral, positive and compound score. The compound score is a comprehensive assessment of the first three scores. This score ranges from -1 to 1. We will set a threshold of the compound score to identify the sentiment. Here we could set the threshold as ±0.2. If the compound score of a review is greater than 0.2, then the review is positive. If the compound score of a review is less than 0.2, then the review is negative. If the compound score is

between -0.2 and 0.2, then the review is neural. Through our process, we visualised the data we acquired using bar graphs given below. The y-axis represents the given and actual scores whereas the x-axis shows the rating for each hotel.



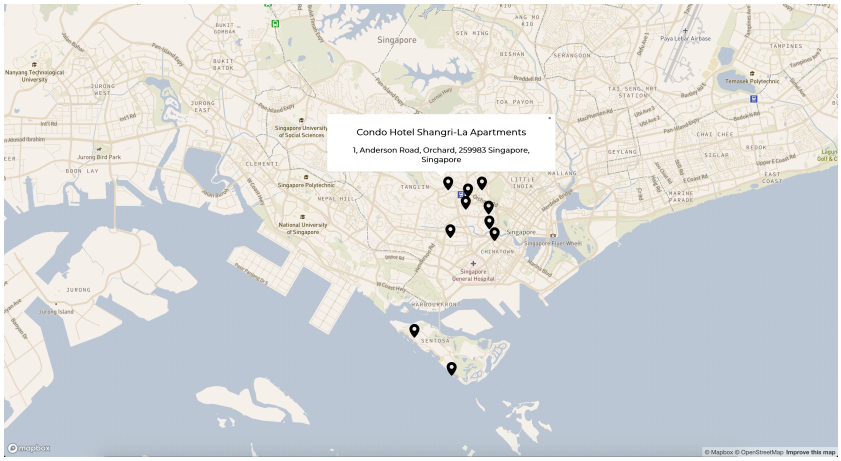
After assigning the polarity scores to each hotel, we realised that given scores for some hotels vastly differed from our acquired scores and could be taken as outliers. For other valid scores, we went on to agree that at most points, the given score was often higher than the calculated score meaning people often rate leniently compared to their sentiment evoked by their reviews. Visualising this data, the data was then plotted on to maps according to the ten best hotels in each country according to our results.

Visualisation of Results

As mentioned above, using softwares such as WEKA, scatter plots were created to give us a better understanding of the ratings we mined in our dataset. Through bar diagrams, we were able to compare the given scores versus the polarity scores we received from our sentimental analysis. Visual data in this exercise was extremely important as it aided in the comprehension of results and to determine any present outliers in our process. This allowed us to move on to the final step of picking the top ten hotels for each country and create visualisations on the map to interactively compare our results with the rating on the website.

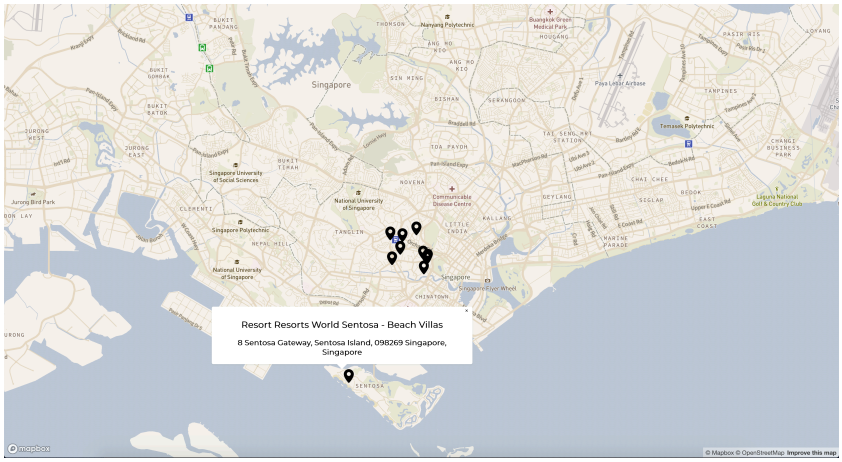
Creating comprehensible visual data

Using Mapbox, we plotted each countries top ten hotels on to a map based on their actual scores given on websites in comparison to our calculated scores through our analysis. The map below is an example of our findings. This allowed us to clearly compare the top ten hotels acquired through two different processes.



Singapore Map based on Initial scores.

Singapore Map based on our calculated Polarity scores.



Conclusion

Through this exercise we looked at the data given to us by a popular website and tried to question and verify this information by conducting our own sentimental analysis of the consumer reviews of thirty different hotels in three different cities. Though we found out that the top ten hotels we found in each country were similar, there were points of difference where the given scores were extremely different from what we acquired in our analysis. Thus, we can conclusively say that the data online is not always trustworthy and thus to gain a true insight on user reviews, sentimental analysis is a tool through which we can attain trustworthy data by analysing the sentiments from a large number of consumer reviews.