**Analysis of TripAdvisor Hotels’ Customer reviews using Sentiment Analysis**

**BIA 660-B: Web Mining**

**Instructor**: Jingyi Sun

**Group 4**

Ronak Kachalia (10459133)

Devanshi Mehta ()

Krina Shah ()

Sacheth Shetty (10459184)

**Introduction**

Having comments and reviews scrapped as well as analyzed can assist any business in keeping an eye on how their customers are feeling about their product as well as services provided, which in turn provides deep insights to major stakeholders in the business on how to improve their products and what services need to be improved upon.

TripAdvisor is a travel company that assists its customers in finding the best rates for their hotel stay as well as booking tickets for their trip. One of the services it offers is their comprehensive hotel booking suite which enables its users to not only view hotels based on location, cost, cleanliness, and various other factors but also review their stay at each of these hotels. The users are prompted to write a text-based review of more than 200 characters and provide an overall rating as well as a rating for cleanliness, rooms, and location as part of their review. Users can read thousands of reviews left by other users for a specific hotel before making their choice. These reviews are not only useful for other users, but they provided several insights to major stakeholders for the hotels which might help them improve the quality of their services.

TripAdvisor sticks to three main ratings for a specific hotel namely, cleanliness, rooms and location of the hotel. It also prompts the user to provide an overall rating for the hotel but it is not necessary that the guests are always looking for these specific services in the hotel. Adversely the review left by the user might include more details about services which they might be unhappy about, however the overall numerical rating does not provide any information regarding the details of these services.

For example, A guest might be satisfied with the cleanliness of the hotel, their room size as well as the location, but they might extremely be unhappy with another service such as food or value for money. The guests might express these concerns in their text review and change the overall rating for the hotel, but this numerical rating does not provide enough information to the Hotel’s management team to make changes or improve their services.

Our project aims at bridging the gap between these text-based reviews using Sentimental analysis as well as identifying certain other categories from popular words used in the review text which users have left for specific hotels. These new categories not only help the user’s narrow down their search for the perfect hotel room but also help the business to ascertain which services need to be improved in order to increase customer satisfaction and bring in more business into their respective hotels.

**Literature Review**

We reviewed the work detailed in the paper by Hsiu-Yuan Tsao and Ming-Yi Chen in the research paper “The asymmetric effect of review valence on numerical rating”, where the authors have conducted a sentiment analysis via text mining, using self-developed computer programs to retrieve a data set from the TripAdvisor website. This study finds there is an asymmetric relationship between review valence or the verbal review text and numerical rating. The authors further find brand strength to have an important moderating role. For a stronger brand, negative review content will have a greater impact on numerical ratings than positive review content, while for a weaker brand, positive review content will have a greater impact on numerical ratings than negative review content.

Therefore, the overall rating that is provided to a hotel is not a reliable measure of services offered by a specific hotel branch or customer satisfaction. The authors mention that assumption verbal review text is symmetrically related to the numerical rating might be a false one, since Brand image is significant factor that customers consider while writing these reviews on TripAdvisor. Similarly other factors or services offered by a specific hotel might not be considered while providing their independent overall rating to the hotel. The authors further conclude that Marketers could adopt sentiment analysis via text mining of online reviews as a valid measure or predictor of consumer satisfaction or numerical ratings. Strong brands should direct more attention to negative reviews, because in such reviews the negative impact transcends the positive. In contrast, weak brands should aim to exploit as many positive reviews as possible to minimize the impact of any negative reviews.

We noted that part the “Brand Image” of the Hotel is simply just one of the factors that might affect the Review valence and overall rating. Other factors would include the services offered by the specific Hotel Branch, such as the quality of food and dining services, gym and fitness services, staff politeness, etc. All of these keywords can be identified and a sentiment analysis would provide us with more insights as to whether the customer reviewing the hotel had a positive or negative experience on these specific factors. This might in turn help us to bridge the gap between the review valence and the overall rating provided by TripAdvisor.

**Research Area**

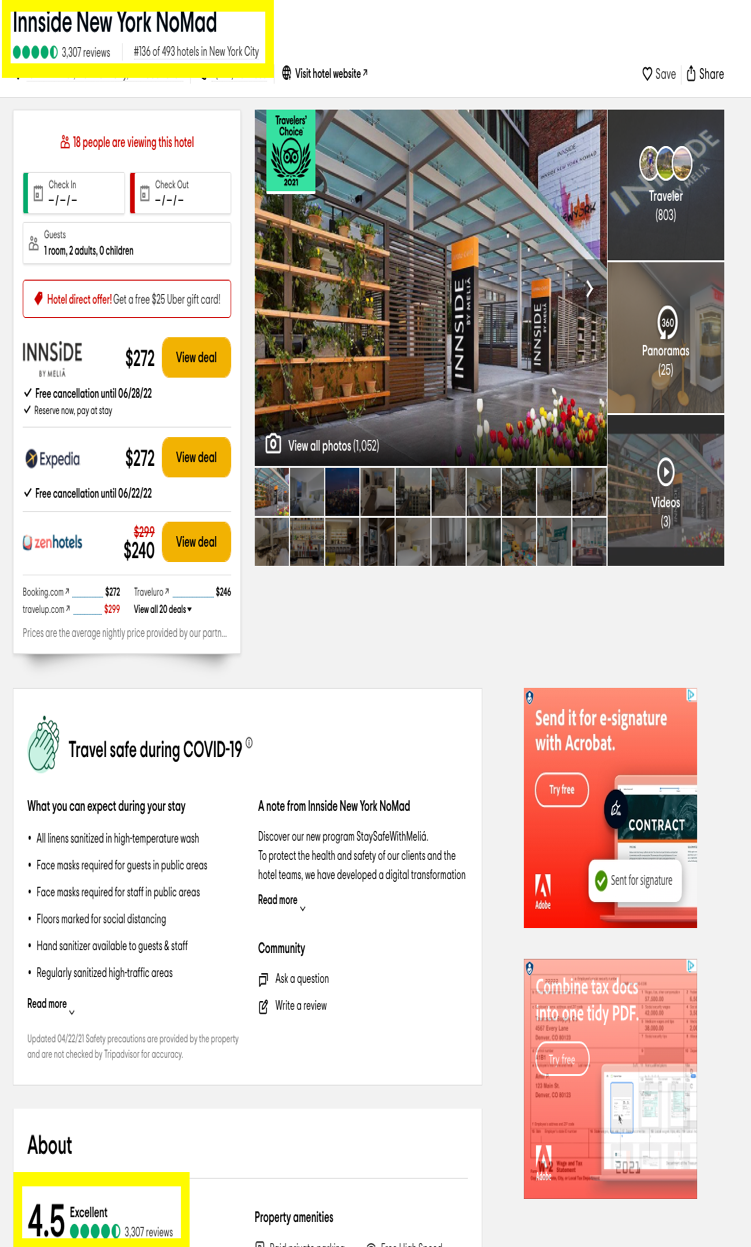
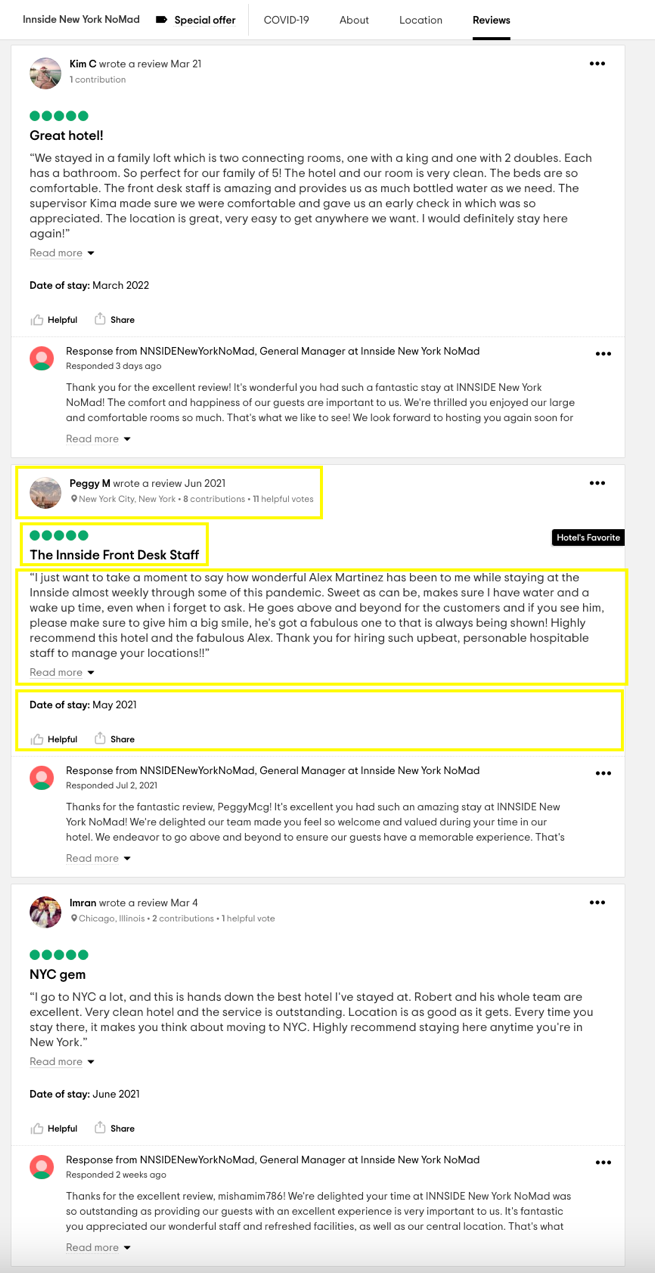
**Objective**

Based on the research question imposed, this experiment will involve various steps before we actually work on building models for review analysis.

In this section, we will focus on extracting data from TripAdvisor, pre-processing the extracted data, perform Exploratory Data Analysis, and draw insightful conclusions.

1. **Data Extraction**

* In order to extract data from TripAdvisor, we will be implementing web scraping using Selenium.
* The bracket for number of hotels to scan is restricted to 6 hotels, and for each hotel we will be scraping 20 review pages where each page constitutes of 10 different reviews.

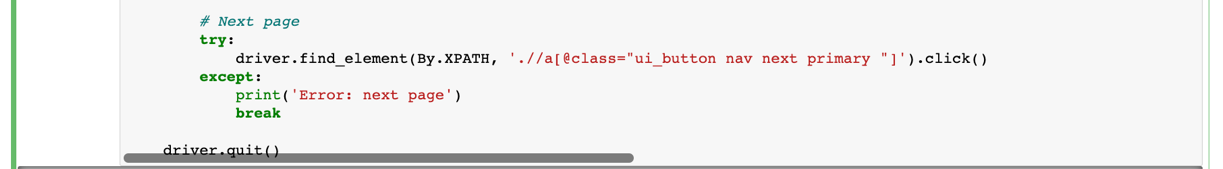
* The above pictures of the TripAdvisor screens illustrate the fields we are scraping using. Selenium. Below are the fields we will be concerned with:
  + Name of the hotel
  + Overall ratings
  + Number of reviews
  + Username of reviewer
  + Review date
  + No. of contributions
  + No. of votes review received
  + Reviewer’s overall ratings
  + Review title
  + Review text
  + Date of stay
  + Individual category ratings (if any)
* Below are the screenshots of the Web Scraping scripts implemented using Selenium in Python:



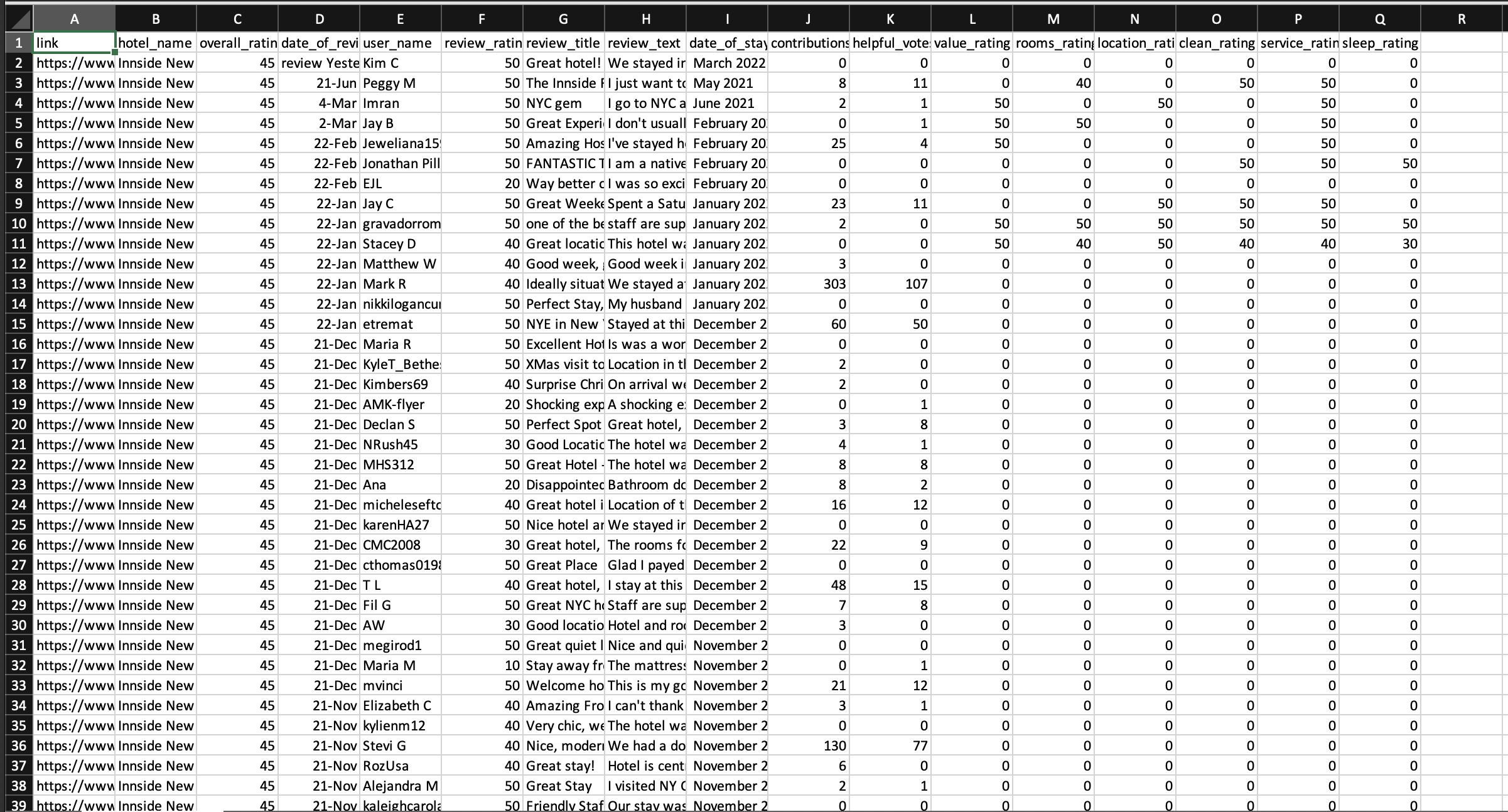








* Scrapped Data in CSV:



1. **Data cleaning & Pre-processing**

The scrapped data from the TripAdvisor website included a number of columns where the data was provided was inconsistent. Many columns in the scrapped data contained both an object of strings, single string values as well as integer values in the same column. Therefore we used certain pre-processing steps in order to clean the data and make it a little more consistent in order to perform out exploratory data analysis. This will provide us with useful insights into the scrapped hotel review data.

* **Dropped column ‘link’ which is not useful for our analysis**

The ‘link’ column contained a URL link which redirected to the actual review in the TripAdvisor website. We concluded that this data was not useful to us in order to perform any analysis.

Text

Description automatically generated with medium confidence

Figure 1 Script for dropping 'link' column

* **Replaced all values in the ‘date\_of\_review’ column which contained string values such as ‘reviewed today’ or ‘reviewed yesterday’ with consistent date value.**

The column ‘date\_of\_review’ contained values such as ‘reviewed today’ & ‘reviewed yesterday’ which was inconsistent with the other values in the column which were in the format of a data value. (21-Mar or YY- shorthand month name). We replaced these values with the current month and year value of ‘22-Mar’.

Text

Description automatically generated

Figure 2 Replacing String values in date columns

* **We noted that the review rating columns had review scores out of a 50. We reduced these scores by a factor of 10 in order to make the review ratings simple.**

We noted that the columns containing integer review ratings for the below given rating columns, contained a integer score out of a 50 which we reduced by a factor of 10 in order to keep it consistent with the reviews present in the Trivago website.

1. Overall rating
2. Value rating
3. Location rating
4. Clean Rating
5. Service Rating
6. Sleep Rating

Text

Description automatically generated

Figure 3 Reducing the factor of review score by 10

* **We noted that columns of ‘date\_of\_stay’ and ’date\_of\_review’ contained different date formats.**

We noted that the above two column contained values in 2 different date formats.

In order to perform our EDA over the months in a year, we extracted only the month parameter in the date values from both the column and stored them in new columns ‘month\_of\_review’ and ‘month\_of\_stay’.

Text

Description automatically generated

Figure 4 Extracting only the month from the date columns

* **Based on the months extracted from the ‘date\_of\_review’ & ‘date\_of\_stay’ column values, we grouped the reviews based on the quarter of the year.**

We used clubbed the reviews based on the financial quarters in a year (Q1,Q2,Q3,Q4) based on the values in the above columns and stored these values in 2 different columns, ‘quarter\_of\_review’ & ‘quarter\_of\_stay’. We will further use these new column to perform EDA in order to understand if there are any insights we can gain based on quarter wise review distribution.

Text

Description automatically generated with medium confidence

Figure 5 Making a new column for quarter of review

Text

Description automatically generated

Figure 6 Making a new column for quarter of stay

1. **Exploratory Data Analysis (EDA)**

**Project Log:**

|  |  |
| --- | --- |
| Tasks | Team Member(s) |
| Scraping data from TripAdvisor | Ronak Kachalia |
| Data cleaning and pre-processing | Sacheth Shetty |
| Exploratory Data Analysis | Krina Shah, Devanshi Mehta |
| Report preparation | Ronak Kachalia, Sacheth Shetty, Krina Shah, Devanshi Mehta |

**References**

* <https://www.researchgate.net/publication/327494800_The_asymmetric_effect_of_review_valence_on_numerical_rating_A_viewpoint_from_a_sentiment_analysis_of_users_of_TripAdvisor>