# One in a Thousand: Navigating the Noise of Hotel Reviews


| | | | |
|---|---|---|---|
| Amit Halbreich | 20817393 | amithalbreich | Amit.Halbreich@mail.huji.ac.il |
| Omer Mushlion | 208271197 | omer_mushlion | Omer.Mushlion@mail.huji.ac.il |
| Omri Marom | 319133666 | omrimar | Omri.Marom@mail.huji.ac.il |


## Problem description

In this work we plan to explore the *indicativeness* of online hotel reviews across various aspects (for example: breakfast, service quality, cleanliness) to determine the overall usefulness of these platforms.

The project consists of two main parts: first, we plan to explore and process the reviews of each hotel separately to encode them into a compact and meaningful representation, highlighting contradicting reviews and redundant or conflicting information; Second, we plan to to use this summarized representation of hotel reviews to compare between different hotels and try to identify meaningful patterns and insights.

## Data

We implemented a JS script to scrape user reviews from booking.com and packaged it into a Google Chrome extension for easy use. When running the script on the reviews page of a hotel on booking.com, it automatically sifts through all reviews pages and scrapes the reviews into a .csv file. The script's code and instructions for running it will be provided within the final project submission. For each hotel, we get a .csv file with the following scheme:

| **Review Title** | **Negative Reviews** | **Positive Reviews** | **Rating** | **Stay Date** | **Review Date** | **Room Type** | **# of Nights** | **Traveler Type** | **Overall Average Rating** |
|---|---|---|---|---|---|---|---|---|---|
| text | text | text | int {1, 10} | date mm + yy | date dd+mm +yy | categ orical | int | categoric al | float [1, 10] |

For example, for the Radisson Blu Tromsø Hotel, Norway we have a .csv file with 1,935 rows for 1,935 English reviews, and this is one of them:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Excelle nt hotel in | sauna hours could've | great location and | 10 | Nov 22 | 08/11/2 022 | Stand ard Room | 3 | Couple | 8.3 |

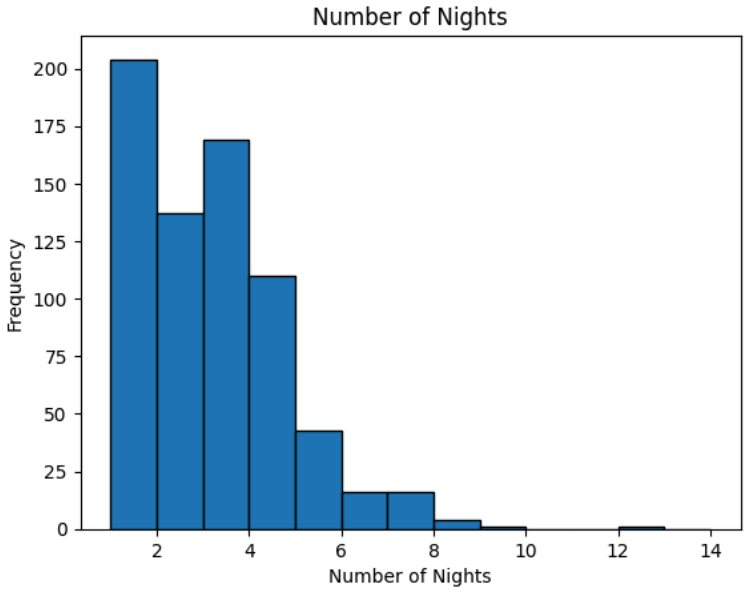
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| tromso! | been more clear | warm staff, comfort able stay | | | | | | | |

For the project we will scrape multiple reviews datasets for different hotels.

## Visualizations

We will use the mentioned Radisson Blu Tromsø Hotel data for the following visualizations.

First, we created a histogram of the number of nights. We expected it to be centered around 3-4 nights, the length of an average city break or a long weekend vacation, with tails for shorter or longer stays. However, we were surprised to find out that the most frequent number of nights is actually only one night; this is unexpected particularly given that the hotel is located in a remote arctic destination where people typically come to travel and enjoy winter safari or northern lights hunting rather than short business trips.




Next, we generated a word cloud from the negative reviews. We expected to see the main topics one might complain about in a hotel: the room, the reception, the cleanliness, etc. Indeed, we observed that the significant words are “room”, “breakfast”, “staff” and “small”. There is, however, one surprising significant word which we didn’t expect, “nothing”, likely indicating that some negative reviews stated that there was "nothing" wrong and that everything was good (we can consider them as noisy negative reviews).