

An Efficient and Intelligent System for Finding Accommodation

- Husnul Abid

1. Introduction

In recent years, the transportation industry has improved drastically, making it easier for people to travel. Furthermore, websites facilitate the process of booking hotels which make it simpler to make reservations. However, because of growing number of websites, it is a daunting task to find affordable accommodation that suits everyone's preferences. We simplified and improved the accommodation search process for our users. To accomplish this, we scraped accommodation data from several websites and compare among these. Moreover, we have implemented a system that interacts with user and enables to find the best accommodations for the individuals.

We have used python as a programming language for both scraping and data analysis task. More particularly, we used BeautifulSoup and Selenium to scrape data from the websites. Pandas has been used for processing and cleaning the data. To visualize the cleaned data, Matplotlib and Seaborn has been used. To interact with the user, we chose Python ipywidgets library that enables us to use range slider and dropdown box.

2. Data Collection

We have scraped three websites for enrich our data-set. The websites are Booking.com, Agoda and Kayak. From each website, we have looked for eight datapoints such as hotel name, price, number of stars, address of the hotel, distance from city center, review score, summary, link of the hotel and photos of the hotel. We selected a city (Turku in our case), check-in and check-out day as our criteria of scraping.

To scrape the website, we used selenium to navigate to the appropriate page. After loading the website, we fetched the html content and feed it to BeautifulSoup which takes the html content in structured way. By filtering tags using BeautifulSoup's find method we parsed our desire data. Furthermore, to get appropriate data type of the parsed data, we performed some string manipulation and conversion. Initially, all the data was saved in list format. After fetching all the data from a single website, we had eight lists (for eight datapoints) with the same size. For further processing and analysis, we decided to merge these lists into a single pandas dataframe. Pandas offers quick, adaptable, and expressive data structures that are intended to make dealing with "relational" or "labeled" data simple and natural. It aspires to serve as Python's core, high-level building block for performing useful, in-the-real world data analysis.

3. Data Processing

After scraping three websites, we had three pandas dataframe. Pandas strong libraries helped us to merge these into a single large dataframe. However, only larger data-set is not enough for further processing. To increase the quality of the data, we performed several operations in our data-set. First of all, we removed duplicate rows from the data-set. Dealing with missing values in the data-set is always challenging. To overcome this, the missing indexes of hotel stars and review score column was filled with the average value of the column. We have decided to put mean value as these columns had small ranges of value without outliers. However, distance column had outliers which were affecting the average value. We calculated the mean value without the outliers and filled the missing values in distance column. After this step, we had a cleaned data-set without any missing value for further analysis.







	name	price (euro)	stars	address	cc_distance (km)	review_score	summary	photos_link						
0	Hesehotelli Turku Kaskentie	73.0	3.0	Tranbacespajan, 20700 Turku, Finland	1.6	8.4	\nlocated in Turku, within 0.9 miles of Verta...	https://cf.bstatic.com/xdata/images/hotel/m...	https://www.booking.com/hotel/f/hesehotelli-t...					
1	Hesehotelli Turku Linja- autobasema	73.0	3.0	Läntinen pitkälläkatu 1, 20110 Turku, Finland	0.6	8.0	\nConveniently located right next to Turku Bus...	https://cf.bstatic.com/xdata/images/hotel/m...	https://www.booking.com/hotel/f/hesehotelli.e...					
2	Turun Herman	73.0	3.0	33a Eerikinkatu B104, 7. kerros, 20100 Turku, ...	0.8	9.3	\nTurun Herman is set in Turku, 1,650 feet fro...	https://cf.bstatic.com/xdata/images/hotel/m...	https://www.booking.com/hotel/f/turun-herman...					
3	Turun Blanka A30	82.0	3.0	Eerikinkatu 33a A30, 5. kerros, 20100 Turku, F...	0.8	9.3	\nTurun Blanka A30 has a balcony and is set in...	https://cf.bstatic.com/xdata/images/hotel/m...	https://www.booking.com/hotel/f/turun-blanka...					
4	Riverside Lux with 2 bedrooms, car Park garage...	197.0	4.0	Linnankatu 59 F, 20100 Turku, Finland	1.8	9.4	\n\n\nYou're eligible for a Genius discount at R...	https://cf.bstatic.com/xdata/images/hotel/m...	https://www.booking.com/hotel/f/riverside-lux...					

Figure 1: Processed data

4. Data Analysis

It is very important to analyze a data-set carefully before making any constructive decision or training a model. Data visualization is a very effective way to show the glimpse of a whole data-set. We performed several visualizations on different datapoints to get an initial idea of our data-set.

It was interesting when a normal distribution emerged from the rent of the hotels. We used a histogram with ten ranges of rent. Most of the hotel rent lies in the range between 50 to 120 euro.

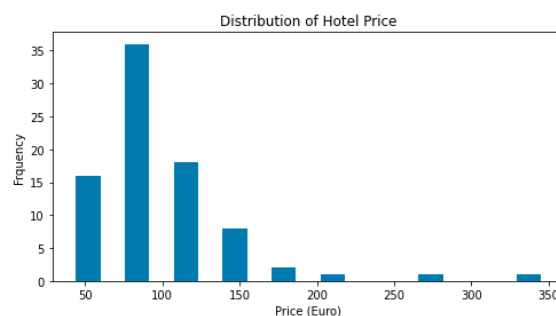


Figure 2: Histogram of hotel price

Another interesting fact was to look into the correlation between number of stars and price of the hotel. It was expected that these two features are strongly correlated. However, in the data-set, we observed a slightly different scenario. Hotel price and number of stars are not so strongly correlated in our data-set. We have used scatter plot and also heatmap to come up with this conclusion.



Figure 3: Price vs Review Score and Stars

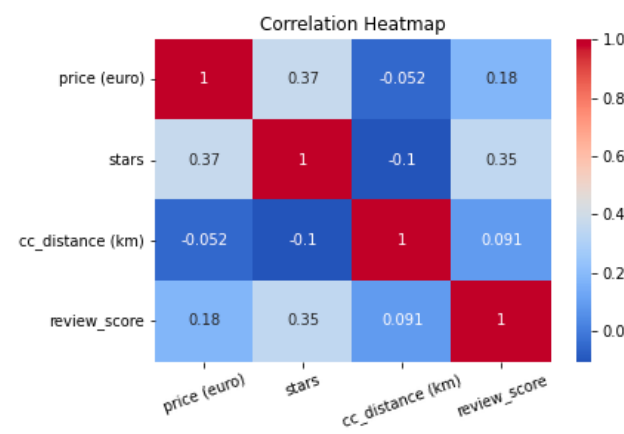


Figure 4: Correlation Heatmap

From figure 3, we can observe price vs review score and also number of stars are in a colored format. In the top left corner, we can see few ash dots which denote five stars hotel with less than 100 euro. It is clearly visible that, there are a lot of three stars hotel than these five stars hotels. From figure 3, we can also come with fact that, in Turku most of hotel belongs to particular group achieving higher review score. Figure 4 confirms that there is no strong correlation between number of stars and price or review score and price.

5. Conclusion

Through this project we have been able to scrape data from website, performed an exploratory data analysis and implemented an efficient and intelligent way to interact with user to find an affordable accommodation. Figure 5 shows the user interface that we build.

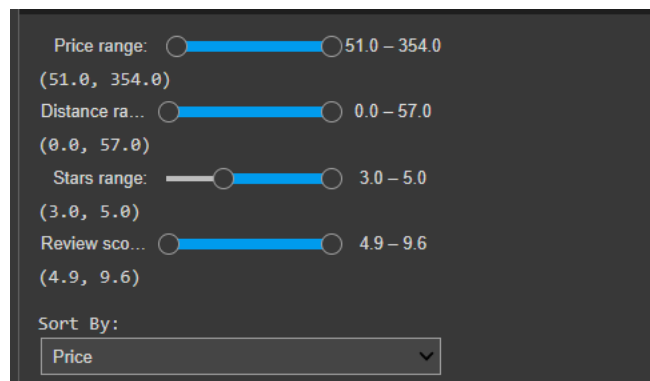


Figure 5: Interaction widgets

It's important to note that there were certain scientific roadblocks. First of all, these websites are highly secure and expect request from browser and not from an automated script. So, we had to use user-agent header which bypass this security issue. Furthermore, these are dynamic websites which means changing of behavior depending on user's location and cookies. For this reason, we were getting different units such as dollar instead of euro. We overcome this challenge by specifying exact value in query parameters. Even if we overcame most the challenges, some future improvements are necessary to increase the performance of the entire system.