TripAdvisor European Restaurant's Information
Information Processing and Retrieval

Maria Francisca Almeida
up201806398@up.pt
FEUP

Mafalda Magalhães
up201707066@up.pt
FEUP

Tomás Torres
up201800700@up.pt
FEUP

Abstract

In the current days, we come across big amounts of data and so an increasing concern to index and search efficiently appears. In this paper, one can see the process of dataset preparation with the goal of creating a restaurant search system. To obtain a dataset with relevant and suitable information for the theme, data refinement and enrichment were performed. Furthermore, the dataset was analyzed for a better understanding of the available data, with some statistics being made for that same purpose.

Keywords

Dataset, restaurant, review, pipeline, data, preparation, analysis, retrieval, rating, ranking, cuisine style.

1. Introduction

Restaurants have always played an essential role in business, social, intellectual and artistic life of a thriving society. Nowadays, it's still a growing sector and clients have a bigger need to filter their never-ending choices.

The current panorama for restaurant's search systems is pretty decent in regards of the information that is able to retrieve, letting users search for names, type or location. The main goal of this project is to complement this type of search systems with a search engine that allows users to search for restaurants based on ratings, reviews, cuisine styles, etc., in order to provide an easier and better experience when trying to find a restaurant that fits their preferences.

That being said, this paper aims to explain our process to extract and process information about restaurant's reviews in order to assemble an engine capable of filtering them according to the user preferences.

2. Dataset

The main dataset chosen contains the general information needed to describe restaurants, and it's reviews, from thirty one cities in Europe, gathered by TripAdvisor (TA). It was obtained by scraping TA for information about restaurants for a given city. The scraper goes through the restaurants listing pages and fulfills a raw dataset. The raw datasets for the main cities in Europe have been then curated for further analysis purposes, and aggregated to obtain this dataset.

2.1. Data Source

As for the authority of the data source, we considered the author, Damien Beneschi, to be experienced in the area, already having other projects similar to this one. There is also a good feedback on this specific dataset.

This dataset was a personal project of his to learn to scrape and was published in a very well-known website, Kaggle, and he also shared the code of the program. However, since the dataset was published in 2018, the code is no longer available.

Therefore, it is concluded that it is a good data source.

2.2. Data Preparation

Initially, it was observed that the dataset had 125527 rows which was a good number to work with (table 1). However, with a closer inspection, we found some irregularities that needed to be fixed. To initiate the preparation and cleaning process, we started by converting the "Ranking" column to a categorical datatype and the "Number of Reviews" column from a float to an int. Some duplicated values were also found and we kept only the first entry of each restaurant. After this, we renamed all columns by removing blank spaces and capital letters, since it would help us later to have these names normalized. Additionally we discovered that some ratings had negative values (-1), which is clearly impossible. Therefore, we replaced these values by zero. Finally, it was observed that there were several cases of missing values and, after analyzing some of these rows we decided that it did not make sense to include them in the dataset. In other words, all the lines with missing values were discarded. After this process, we ended up with approximately seventy five thousand rows.

Besides this, every restaurant only had two reviews, which was too little information. A new table was created for the reviews, and through python scripts and the pandas library, we tried to scrap the TripAdvisor website in order to obtain more reviews. For the python script, the links of

the reviews in the column URL_TA were used, meaning a normalization was made to add 'www.tripadvisor.com' to all rows. However, we did not manage to finish this scraping, since the only package that the website allowed was the selenium and we had to use a web driver in order to get the wanted information (title of the review, content of the review, date and rating) and it would take several days to scrap all the information related to the seventy five thousand restaurants.

Consequently, we had to create a new .csv file for the reviews using only the reviews of the original dataset. In order to do this, we had to separate the values of the column "Reviews" so that we would have one row for each different review. This row contained the restaurant id, the content of the review and its date. During this process, we found a couple of restaurants that had no reviews and we eliminated them. A new .csv file was also created for the cuisine style, using the same process.

In the end, we copied the cleaned dataset to a new .csv file. Therefore, we finished this procedure with three .csv files that contained all the information that we need to our project.

TABLE 1. NUMBER OF NON-NULL VALUES OF EACH COLUMN OF THE ORIGINAL DATASET

Name	125 527 non-null
City	125 527 non-null
Cuisine Style	94 176 non-null
Ranking	115 876 non-null
Rating	115 897 non-null
Price Range	77 672 non-null
Number of Reviews	108 183 non-null
Reviews	115 911 non-null
URL_TA	125 527 non-null
ID_TA	125 527 non-null

2.3. Data Collection

Using the link to web page of each review in the column URL_TA, after the normalization, we tried to perform web scraping. Since this task was more expensive than we anticipated, we will do this as a future work in order to complete our datasets. This will be done with the main purpose of increasing the number of reviews available so that users can better access their options when searching and choosing a restaurant. This will also help with the search engine in the future to show what restaurants are considered 'good' or 'bad'.

2.4. Data Enrichment

The final step of the developed pipeline consisted of combining the datasets that resulted from the cleaning stage in a SQL database.

2.5. Data Characterization

Throughout the analysis of the dataset, we gathered information regarding the mean value (table 2), minimum and maximum values for each of its numerical properties (ranking, rating and number of reviews).

TABLE 2. MEAN VALUE OF RATING IN EACH CITY

City	Mean Rating
Amsterdam	4.130654
Athens	4.233831
Barcelona	4.023047
Berlin	4.150000
Bratislava	4.087699
Brussels	3.899121
Budapest	4.098214
Copenhagen	4.006950
Dublin	4.084636
Edinburgh	4.095541
Geneva	3.979210
Hamburg	4.085597
Helsinki	3.950197
Krakow	4.201651
Lisbon	4.070515
Ljubljana	4.102167
London	3.977194
Luxembourg	3.945578
Lyon	3.993521
Madrid	3.895141
Milan	3.877356
Munich	4.036678
Oporto	4.168436
Oslo	3.912037
Paris	3.981964
Prague	4.063735
Rome	4.170070
Stockholm	3.900000
Vienna	4.067200
Warsaw	4.085290
Zurich	4.036463

We also calculated the number of restaurants per city (figure 1), per inhabitant and per square kilometer (figure 2). In order to be able to do this, we had to search for some specific information, such as the population of each city and its area.

Furthermore, we also analysed the cuisine style column. We found out that there were one hundred and twenty-six different cuisine styles and we got the global (figure 3) and local (figure 4) occurrence of one of each styles.

Finally, we thought that it would be interesting to find out which cities where the best for special diets, such as vegetarianism, veganism and a gluten free diet, among others. In order to discover this, we plotted some graphics that would be useful in this analysis (figure 5).

3. Pipeline

In order to achieve a greater quality of the chosen data, various processing tasks were executed. These steps are represented in the data pipeline (figure 6).

4. Conceptual Model

The conceptual model consists of three main classes: Restaurant, Review and CuisineStyle. Each restaurant is characterized by its id, name, link, ranking, rating, price range and city. On the other hand, each review consists of a commentary and a date and each cuisine style is represented by its name.

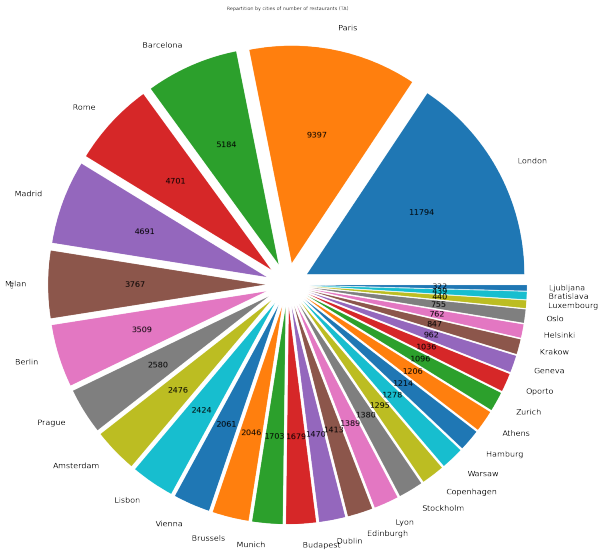

Figure 1. Number of restaurants per city

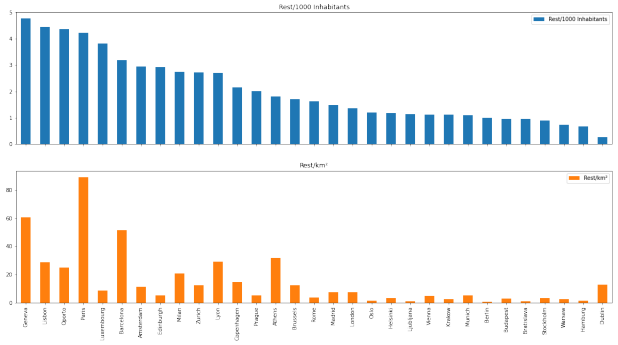

Figure 2. Number of restaurants per inhabitant and per square kilometer

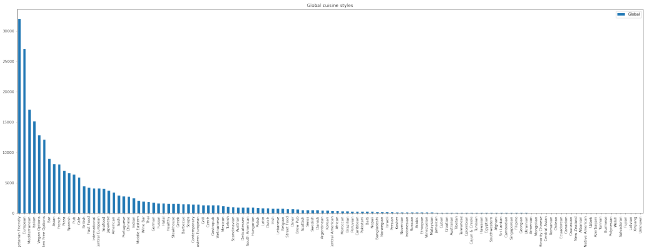

Figure 3. Global cuisine styles in all cities

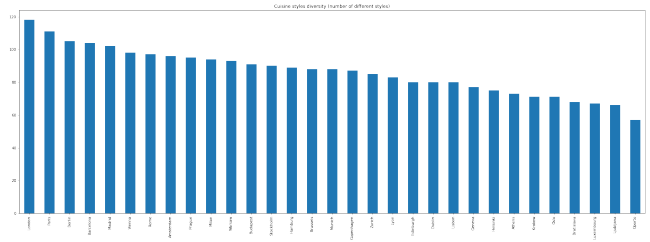

Figure 4. Cuisine styles diversity in each city

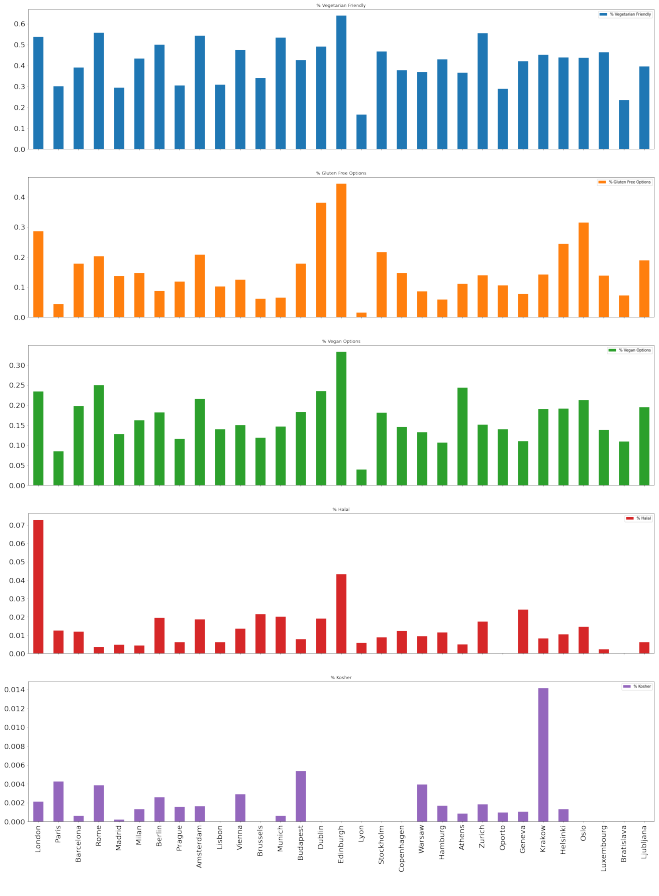

Figure 5. Special diets ratio for each city

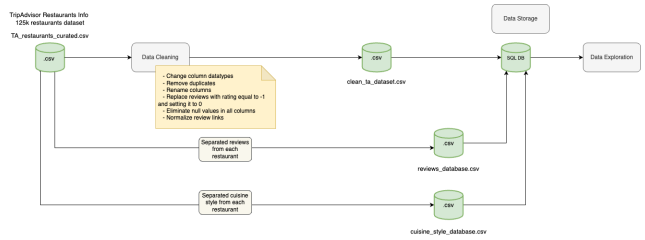

Figure 6. Data processing pipeline

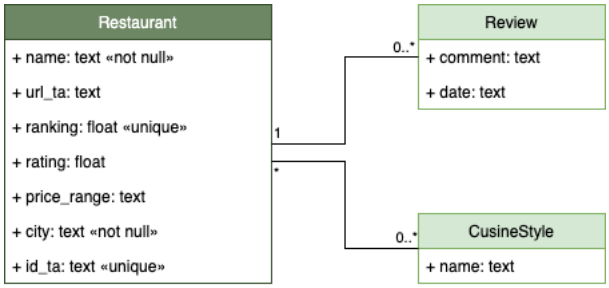

Figure 7. Conceptual model

5. Conclusion

In this paper is presented the process the datasets went through to reach its final state, ready to use.

Throughout this part of the project, the dataset was well analysed and studied in order to conclude which data cleaning and preparation tasks were necessary for it to help us accomplish the project goal.

5.1. Future Work

As it was previously stated, we would like to complete the web scraping of reviews in order to enrich our datasets and to improve the user experience by having more information for retrieval. In addition, we would like to explore some scenarios that we thought that would be useful in order to simulate the client experience.

5.1.1. User Stories.

- As a client I want to search for the best restaurant in the city, so that I can visit it
- As a client I want to search for cuisine styles by their rating value so that I can choose the type of restaurant that I will visit
- As a client I want to search for the cuisine style so that I can choose a restaurant with my preferences
- As a client I want to search for the best reviews so that I can choose a good restaurant

References

[1] H. Kopka and P. W. Daly, *A Guide to LaTeX*, 3rd ed. Harlow, England: Addison-Wesley, 1999.