Privago: Hotels Search System

André Costa
up201905916@edu.fe.up.pt
Faculty of Engineering - University of Porto
Porto, Portugal

André Ávila
up202006767@edu.fe.up.pt
Faculty of Engineering - University of Porto
Porto, Portugal

Fábio Sá
up202007658@edu.fe.up.pt
Faculty of Engineering - University of Porto
Porto, Portugal

Fábio Morais
up202008052@edu.fe.up.pt
Faculty of Engineering - University of Porto
Porto, Portugal

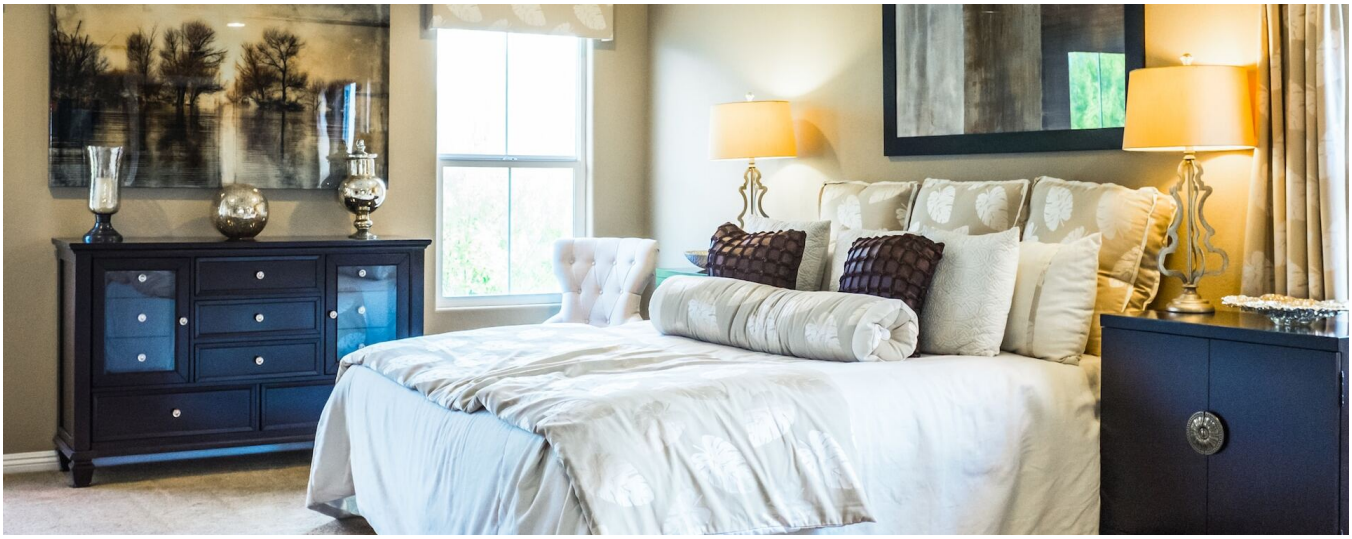

Figure 1: Hotel

ABSTRACT

The internet's exponential growth demands appropriate systems for harnessing and connecting this massive information resource. This project addresses this need by focusing on the hotel industry, where reviews play a crucial role in shaping consumer choices. This article aims to provide a clear and well-documented explanation of the work needed in developing a robust search engine for hotel reviews. To accomplish this, data is collected from multiple sources, cleaned and prepared, and in-depth data analysis is undertaken.

CCS CONCEPTS

• **Information systems** → Information retrieval query processing.

KEYWORDS

Hotels, Reviews, Information, Dataset, Data Retrieval, Data Preparation, Data Analysis, Data Processing, Data Refinement, Pipeline, Data Domain Conceptual Model

ACM Reference Format:
André Costa, André Ávila, Fábio Sá, and Fábio Morais. 2023. Privago: Hotels Search System. In *Proceedings of PRI (G53).* ACM, New York, NY, USA, 6 pages.

G53, October-09, 2023, Porto, Portugal

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00

1 INTRODUCTION

This paper is developed as part of the course "Information Processing and Retrieval" (PRI) within the first year of the Master's in Informatics and Computing Engineering (MEIC) at the Faculty of Engineering of the University of Porto (FEUP).

The choice of the hotel reviews theme is motivated by its significant relevance and the rich diversity of attributes it encompasses. Hotel reviews, as a research focus, hold substantial importance in the modern information landscape. They not only provide valuable insights into the hospitality industry but also serve as a prime example of data diversity, combining numerical rates, submission dates, and personal, subjective narratives. This diversity introduces intricacies in data structuring and presents challenges in contextual

search, making it an ideal choice for aligning the search system with real-world scenarios. Thus, this theme strongly resonates with the course objectives, emphasizing practical applicability and the development of robust information retrieval solutions.

This document is structured into several major sections, each tailored to fulfill the objectives of Milestone 1. We commence with **Data Extraction and Enrichment**, where we introduce the data sources, briefly characterize the datasets, and assess data quality. Subsequently, **Data Preparation** outlines the selection criteria, processing methods, and data storage procedures for hotel-related information and associated reviews, following a clear and reproducible pipeline.

In **Data Characterization** we delve into the evaluation and visualization of the refined data. This involves examining various criteria and relationships, from the Domain Conceptual Model to Word Clouds. Finally, **Possible Search Tasks** and **Conclusions and Future Work** provide an overarching interpretation of the results, guiding the identification of suitable research objectives for the project's next phase.

2 DATA EXTRACTION AND ENRICHMENT

After conducting research for relevant data in terms of variety and quantity, four datasets in CSV format from different regions were selected through the Kaggle [1] platform. Table 1 provides a characterization of the acquired datasets:

Table 1: Initial datasets characterization

Dataset	Features	Hotels	Reviews	Size(MBs)
Datafiniti's Hotel Reviews	26	1400	10000	124.45
Hotel Review Insights	7	570	7000	1.31
London Hotel Reviews	6	20	27329	22.85
Europe Hotel Reviews	17	1493	515000	238.15

The Datafiniti's Hotel Reviews [2] dataset was taken from Datafiniti's Business Database [3] through sampling. Hotel Review Insights [4] is a compilation of hotels around the world through web-scraping of reviews found on Booking.com [5]. London Hotel Reviews [6] is a sample taken and partially refined from a DataStock dataset [7]. Finally, Europe Hotel Reviews [8] also results from web scrapping of hotel reviews across Europe published on Booking.com [5].

All datasets have a public use license and, according to the Kaggle platform, a usability index greater than 8. This index is justified, given that the elimination of null, repeated or non-informative entries practically did not eliminate nothing.

The datasets contain common features, numerical data, such as rate, review date, and textual data, such as review text, hotel location and name. The last dataset contains two additional parameters, positive review and negative review. The features stated were extracted in this step and refined in Data Preparation.

3 DATA PREPARATION

In this section, is presented the structured data preparation pipeline [2] that was developed for the project. This pipeline embrace various data cleaning and restructuring procedures aimed at achieving a clean, uniform, and ready to analyze dataset. The goal of this stage is to build a strong basis for insightful analysis.

The data preparation process began with a comprehensive cleaning phase, where the primarly focus was the removal of records containing **empty or null values** in any attribute. Simultaneously, was identified and eliminated incomplete or uninformative data, including strings with **uninformative text**, such as "no comments available for this review.", using Python. This combined cleaning step ensured that the dataset was cleansed in detail for the normalization phase.

With the data cleaned, attention was turned to **attribute normalization**. Given the presence of diverse datasets with varying formats, comprehensive normalization process was needed. This included standardizing attributes such as **"positive_reviews"** and **"negative_reviews"** into a unified **"review_text"** attribute for the 4th dataset as is demonstrated in the Pipeline Diagram [Figure 2]. Additionally, **date formats** were normalized to ensure uniformity and suitability for analysis. The date format was set as "year-month" due to the absence of day-specific review information in the second dataset and its irrelevance to the research targets. In fact, people search for seasons of the year, months, and not for a specific day. **Rate scales** were also normalized to a common range and converted to floating-point values, facilitating comparative analysis ([0.0, 5.0]).

In addiction, was established a **standardized naming** convention to address variations in **hotel names**, such as from *"45 Park Lane - Dorchester Collection"* to *"45 Park Lane Dorchester Collection"*. This step was necessary to facilitate the aggregation step and the addition of the feature "average_rate" to each hotel entity, referenced below. **Location standardization** involved reducing location names to their last two words, preserving only the capital and country names.

To gain insights into the textual content, we calculated the temporary column **word count** for each review across all datasets. This analysis was facilitated using the Pandas [9] Python tool, allowing us to extract valuable information such as quartile ranges and make informed decisions during the review deletion phase. This process enabled us to identify and manage reviews with either an insufficient word count or an excessively high word count. We achieved this by removing reviews falling below the 25% threshold (first quartile) and those exceeding the 75% threshold (fourth quartile). This step was done separately for each dataset, due to the discrepation of each average word counting [Figure 9] [Figure 10].

At this stage, all the datasets were successfully merged into a single, consolidated dataset, which streamlined the remaining preparation tasks. These tasks commenced with the computation of the temporary column **"average_rate"** for each unique hotel. This information may prove valuable for defining search criteria in future milestones.

After completing the aforementioned steps, the next phase involved determining the minimum and maximum number of **reviews per hotel** to be retained. To accomplish this, the same approach used for analyzing the number of words per review was employed, utilizing the Pandas [9] .describe() function. This statistical analysis provided essential insights into the distribution of reviews across hotels. This step was important due to the discrepation of the number of reviews per hotel [Figure 11].

First, hotels with fewer reviews, falling below the established minimum threshold (first quartile), were addressed, and they were subsequently removed from the dataset. This step was crucial in ensuring that the dataset focused exclusively on hotels with a substantial volume of reviews, thus providing more meaningful insights.

Subsequently, hotels with an excessive number of reviews were taken into account. To handle this situation, a strategy was implemented, allowing the selection and retention of reviews while preserving the proportion of reviews per rate category for each specific hotel, i.e., its global rate. This approach guaranteed a balance between the "average_rate" and the rate of the selected reviews.

The final step in the data preparation phase involved organizing the data into the **desired JSON file format**, which was designed based on our UML diagram [Figure 8] and with a focus on the primary objective of the search tool. This format consisted of a collection of JSON objects, each representing a "Hotel" entity. Within each "Hotel" object, were included not only its associated attributes (name, location, average_rate) but also the related reviews, presented as JSON objects themselves. Each review has a corresponding text, rate and submission date.

4 DATA CHARACTERIZATION

In this section there is a characterization of the documents that are produced by the pipeline. The graphs and tables were obtained using the Matplotlib [10], Numpy [11] and NLTK [12] libraries.

4.1 Reviews Word Cloud

Figure 3: Reviews Word Cloud

The word cloud based on the texts of the reviews supports the search tasks of the next milestone as well as the contextual search to be implemented. As expected, the words that stand out the most are those intrinsically related to the hotel industry, such as "staff", "room", "location", "breakfast". Generally, the highlighted words have a neutral tone, although several with a very positive connotation are also detectable, such as "clean", "great", "lovely", "excellent", which can also be proven by the average rate [Figure 6] given to hotels.

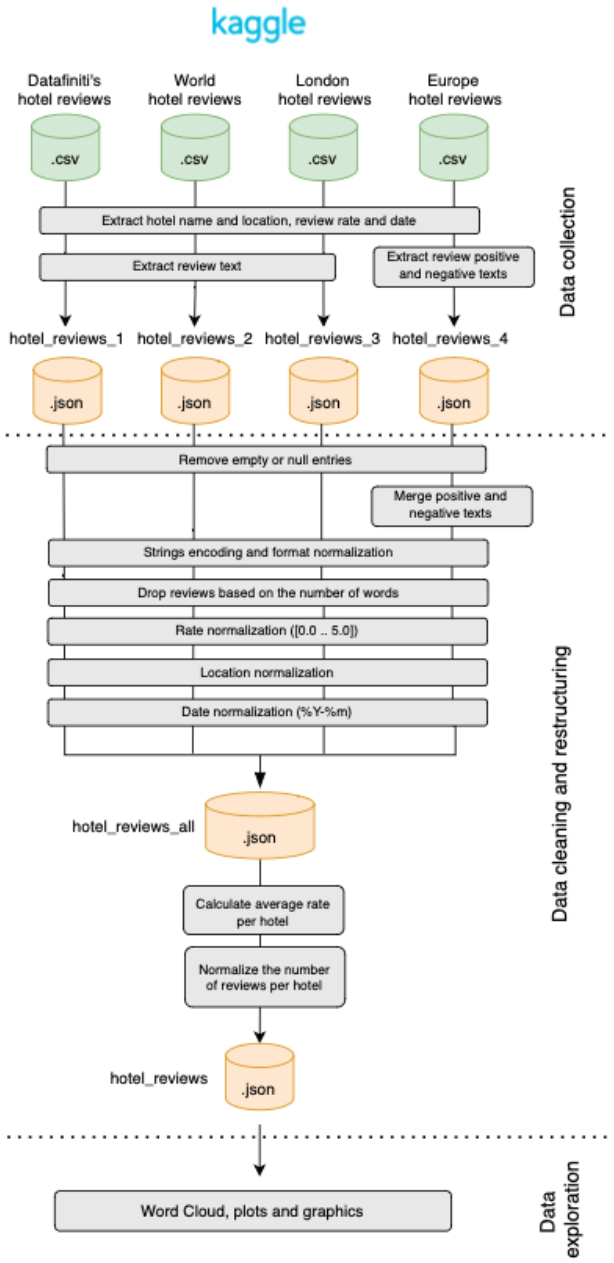

Figure 2: Data preparation pipeline

4.2 Hotel location distribution

Figure [5] shows the distribution of the 10 most frequent hotel locations in the system. As expected, the capitals and large cities of the main tourism countries concentrate the largest number of hotels and their reviews.

4.3 Average rate distribution

As we can see in figure [6] Hotels tend to have very good reviews. The result is encouraged by their locations, as the majority are in cities with major tourist attractions and therefore with a greater cadence of reviews.

4.4 Reviews per year

From the analysis of the figure [7], it can be concluded that the system includes hotels with reviews from 2010 to 2023. However, the choice of initial datasets with information relating mainly to the period 2015 and 2017 influenced their representativeness.

Let's look, for example, at the distribution of reviews by month in 2016 in the figure [8].

As we can see, it is July and August when the number of reviews is higher. This period corresponds to the Summer time when people normally take vacations and therefore choose to travel. This pattern is repeated for the other years under analysis.

4.5 Data Conceptual Model

After the Data preparation phase, our documents contain the following relationships:

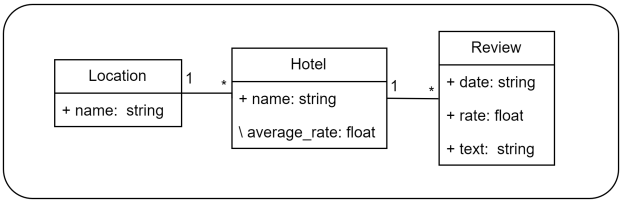

Figure 4: Data Conceptual Model

A hotel is made up of the attributes name, location and average_rate. Note that average_rate is a derived attribute that was calculated in the process based on the selected reviews. Each review has the corresponding text, rate and submission date.

Location was also considered a class because, as expected in the hotel scene, there are several hotels per region, mainly in large cities, as shown in figure [5].

5 POSSIBLE SEARCH TASKS

In the course of the data analysis journey, were unveiled valuable insights through the use of a Word Cloud Diagram [Figure 3]. This visual representation highlighted the most frequently occurring words in hotel reviews, shedding light on what matters most to travelers. Among these words, some stood out as pivotal in understanding the key factors that influence hotel choice and guest satisfaction.

"Location" emerged as one of the top considerations in travelers' decision-making processes. Whether it's proximity to local attractions, accessibility to transportation hubs, or the overall neighborhood ambiance, the location of a hotel can greatly influence the overall travel experience. Queries like **"best hotels in [City/Region/Country]"** and **"hotels near the airport"** can help travelers selecting accommodations that align with their preferred locations and provide convenient access to their destinations.

For many tourists, a good breakfast is an essential element of their stay. The word **"breakfast"** featured prominently in our Word Cloud [Figure 3], suggesting a keen interest in this aspect. Whether it's a hearty breakfast buffet or specialty morning treats, travelers seek accommodations that cater to their breakfast preferences. Queries like **"Hotels with breakfast/good breakfast"** can assist those who prioritize morning meals.

The words "staff" and "service" were significant contributors. This emphasizes the critical role that hotel staff play in the overall guest experience. From warm welcomes at the reception desk to prompt and efficient room service, exceptional staff service can elevate a stay. Queries such as **"hotel with a helpful staff"** and **"affordable room service"** could be instrumental in identifying hotels that excel in providing exceptional service to their guests.

Another one of the foremost considerations in hotel selection is the quality of the room and its amenities. Words like "room," "bed," and "bathroom" featured prominently in our Word Cloud [Figure 3], underscoring the importance of these aspects to travelers. Queries related to "room"/"bed" quality or "bathroom" sanitation can guide travelers to accommodations that prioritize comfort and cleanliness.

6 CONCLUSIONS AND FUTURE WORK

In conclusion of this milestone, all the planned tasks within the data preparation phase of the project have been successfully completed. This accomplishment marks a crucial turning point in the process of creating a useful hotel search engine that will give tourists useful information and help them make informed choices.

One of the most challenging aspects of the work was developing effective strategies to address the issue of an excessive number of reviews. Substantial effort was invested in determining the best approach to manage and utilize this abundance of information. Through meticulous analysis and innovative methods, a balance was struck between data volume and relevance, ensuring that the dataset remained rich with insights while maintaining a manageable size.

As the project progresses, there are always opportunities for further enhancements and refinements. With the cleansed and consolidated dataset, the next phase of the project will focus on the development of a robust hotel search engine. This engine will allow travelers to explore and filter accommodations according to their preferences, whether related to location, room quality, staff service, or other factors identified during the analysis phase.

A ANNEXES

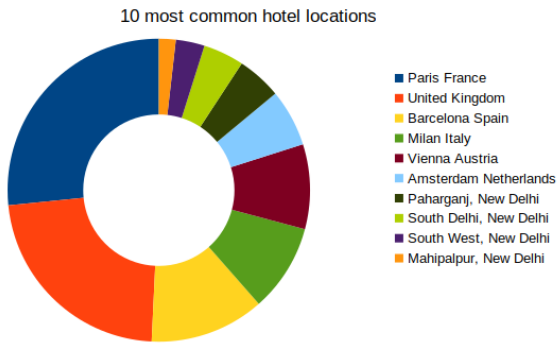

Figure 5: Hotel location distribution

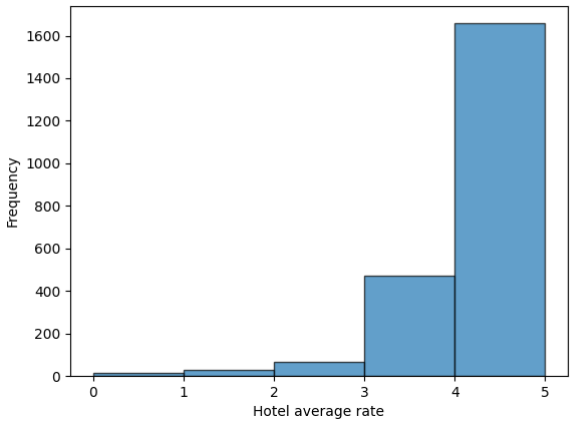

Figure 6: Average rate distribution

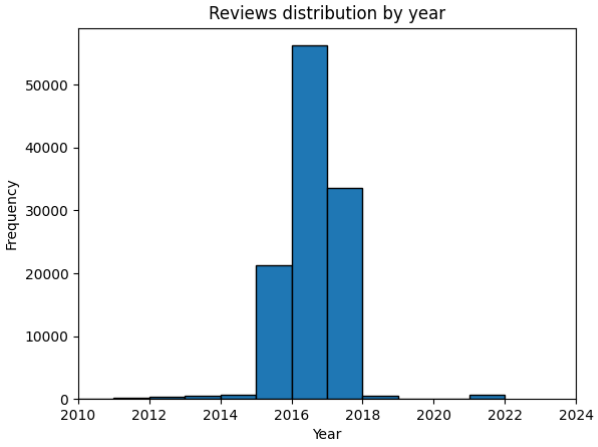

Figure 7: Reviews distribution per year

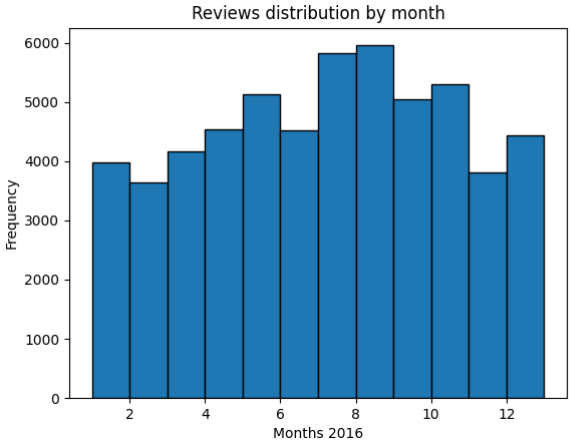

Figure 8: Month reviews distribution in year 2016

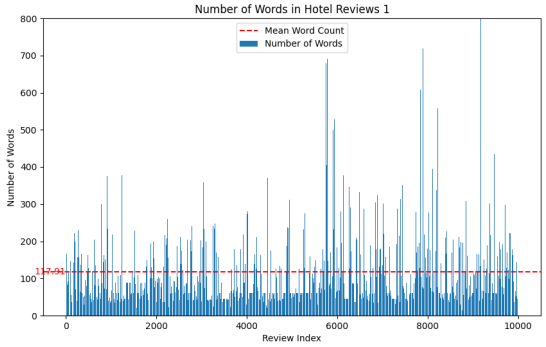

Figure 9: Average words per review in Datafiniti's Hotel Reviews dataset

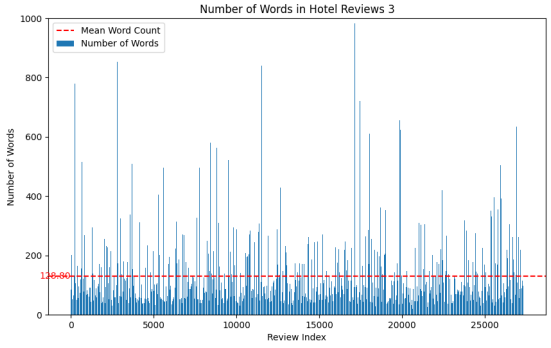

Figure 10: Average words per review in London Hotel Reviews dataset

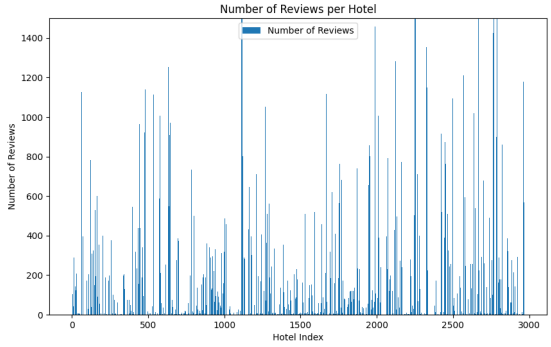

Figure 11: Number of Reviews per Hotel

REFERENCES

[1] Kaggle, 2022. https://www.kaggle.com.
[2] Datafiniti, 2017. https://www.datafiniti.co.
[3] Datafiniti, 2017. http://www.ctan.org/pkg/acmart.
[4] Hotel reviews, 2017. https://www.kaggle.com/datasets/juhibhojani/hotel-reviews.
[5] Booking, 1996. https://www.booking.com.
[6] Reviews of london-based hotels, 2018. https://www.kaggle.com/datasets/PromptCloudHQ/reviews-of-londonbased-hotels.
[7] Datastock, 2018. https://datastock.shop.
[8] 515k hotel reviews data in europe, 2017. https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe.
[9] Pandas, 2023. https://pandas.pydata.org.
[10] Matplotlib, 2023. https://matplotlib.org.
[11] Numpy, 2023. https://numpy.org.
[12] Nltk, 2023. https://www.nltk.org.