Application of AI and ML Models for a Hotel Recommendation System

Diogo Beltrán Dória, Mariana Lopes Paulino
Universidade de Coimbra
{dorio, marianapaulino}@student.dei.uc.pt

Master's Degree in Engineering and Data Science

Abstract

Nowadays, we have a lot of information, and we need to choose carefully what reaches the end costumer. Big Data is commonly used by artificial intelligence (AI) algorithms like, Recommendation Systems, to suggest or recommend relevant options to consumers. According to Nvidia, these can be done based on various criteria, including past purchases, search history, demographic information, and other factors. [1]

In this project, we aim to create a Recommendation System using Hotel's attributes and reviews by past users, depending on what the user is looking for. Our Recommendation System will find the most suitable places for their stay, either if it's for vacation or business trips.

Keywords: Artificial Intelligence (AI),

Recommendation System, Hotels

1. Problems and Motivations

Since 2022, tourism activity has grown considerably, despite the travelling restrictions and tight rules implemented by countries to address the COVID-19 pandemic. In 2023, we've seen multiple country's breaking tourism revenue records, even considering pre-COVID figures. The National Institute of Statistics of Portugal, INE, revealed that the movement of passengers at national airports continues to reach historical highs in a sustainable manner, with data indicating that the tourism accommodation sector registered 3.5 million guests and 10.1 million overnight stays during August 2023.[2][3][4]

Motivated by the exponential growth of tourism in Portugal, our aim is to create a Hotel Recommendation System that not only allows users to filter by the usual tags most common websites like booking.com, trivago.pt provide. But also adding extra tags which will enrich the user experience, providing additional information that will make their stay memorable.

Quantitative reviews of past users, demographic information, family friendly to adults-only score, nearby points of interest, parking lots availability, public transportation availability and timetables, crime rate, are examples of tags that will contribute to enrich the model.

Safety, comfort, price, location, services available, parking and transportation, nearby restaurants and point of interest, other guest's recommendations are among the criteria used when selecting a hotel. Our enhanced recommend systems will suggest not only the best hotels in town based on reviews of past users but also will recommend those that met the above-mentioned criteria, potentially saving customer's time and money, contributing to the best user experience.

2. Background

Steve Jobs once said: "People don't know what they want until you show it to them". That's one of the reasons Steve Jobs was considered such a visionary. He had an uncanny ability to understand how Apple products would add value to people's lives and ultimately change the world.[5]

Nowadays, recommendation systems are being increasingly used for many applications such as websites, books, e-commerce, tourism, movies, music, news, etc. Behind the algorithm, at the core of any artificial intelligence recommendation system, there is a machine learning model optimized for the key business goals: customer retention, time spent on platform, generated revenue.

Taking Spotify's approach as an example, the platform generates a user profiling by logging all user's listening activity and by saving user's feedback of every track played. The amount of data collected allows the platform to recommend songs that the customer might like and create and suggest playlists according to listening patterns.[6]

The use of content-based filtering employs item features to suggest alternative items that are similar to the user's preferences, based on their pervious actions or feedback.

The aim of this project is to develop a hotel recommendation system based on reviews from past users. Implicit information of hotel attributes will be transformed alongside the features of the items previously rated by the users and the best matching ones that are recommended.

3. Objectives

Our objective is to transform the information of the reviews provided by past users into explicit features such as sentiment, other topics that can provide insights into which aspects are more important to customers, ratings. We will use this information to train a machine learning model to predict user preferences.

Once obtained the features, we'll check similarities between hotels, using a variety of factors such as location (nearby hotels are likely to be similar in terms of nearby attractions and amenities), price (similarly priced hotels are like to be similar in terms of accommodation and services quality) and amenities (hotels with similar amenities are likely to be similar in terms of the overall experience that they offer).

Finally, we will use the geographic coordinates available in the dataset, to query other data sources in order to present information about nearby parking lots, public transportation, nearby monuments, landscapes and other points of interest.

4. Approach

During research, some similar projects were found. The difference between our idea and the projects found is that none of them tries to explain why the recommended hotel is a better option than x or y, it's just a list of recommended hotels based on the score of reviews.[11][12][13][14]

Our goal is to develop an output that specifies what makes the recommend hotel more unique than the others to what the user is looking for. That's where our explicability comes in, not only the specific tags of each hotel are going to be considered, but also whether the hotel is appropriate for holiday or business trips and its demographic information considering the city's landmarks. From other projects, we've learned that our first steps will rely on data cleaning and processing. Making it possible to start the modelling phase of the hotel attributes and reviews. Further ahead, through Feature Engineering, we are going to understand which are the key attributes in our data frame that give us more context. Concluding with the development of the similarity model of the documents, such as reviews and locations, and an output layout built in a way that the user understands why those hotels were recommended.

5. Materials

In this project, we will be using the dataset set "515K Hotel Reviews Data in Europe". The positive and negative columns are already in columns of the CSV file. The data was scraped from Booking.com from 2015 to 2017.[7]

It contains 515738 reviews for 14912 luxury hotels in Europe. The CSV file contains 17 fields as shown in table 1.

Column Name	Description
Hotel Address	Address of the Hotel
Review Date	Date when the reviewer posted the corresponding review
Average Score	Average score of the hotel based on the latest comment in the last year
Hotel Name	Name of the Hotel
Reviewer Nationality	Nationality of the Reviewer
Negative Review	Negative review the reviewer gave to the hotel. If the reviewer does not give a negative review then the value should be "No Negative"
Review Negative Total Word Counts	Total number of words in the negative review
Positive Review	Positive review the reviewer gave to the hotel. If the reviewer does not give a negative review then the value should be "No Positive"
Review Positive Total Word Counts	Total number of words in the positive review
Reviewer Score	Score the reviewer has given to the hotel based on his/her experience
Total Number of Reviews the Reviewer has Given	Number of Reviews the reviewer has given in the past
Total Number of Reviews	Total number of valid reviews the hotel has
Tags	Tags the reviewer gave the hotel
Days Since Review	Duration between the review date and scrape date
Additional Number of Scoring	There are also some guests who just made a scoring on the service rather than a review. This number indicates how many valid scores without review in there
Lat	Latitude of the Hotel
Lng	Longitude of the Hotel

Table 1. Feature information of the dataset

In addition to our data collection, we'll add various coordinates (latitude and longitude) of monuments and important sites of the cities. We'll probably use the Open AI framework, ChatGPT, to obtain the coordinates (latitude and longitude) of the specific landmarks we want to add.[8]

To implement this project, we are going to apply several libraries and methods to test various alternatives. Most of this study will rely on the Natural Language Toolkit (NLTK) that is an open source Python library for Natural Language

Processing, Scikit-Learn for feature extraction and evaluation metrics, Pandas, NumPy and geocoder.[9][10]

5.1 Data Analysis

In this subsection, we are going to display some information that our dataset contains that we consider it's useful to be displayed in a graphic.

In Figure 1, we chose to display a map containing the cities that are present in the dataset, having six countries represented.

Countries contemplated in the dataset:

- Austria
- England, United Kingdom
- France
- Italy
- the Netherlands
- Spain

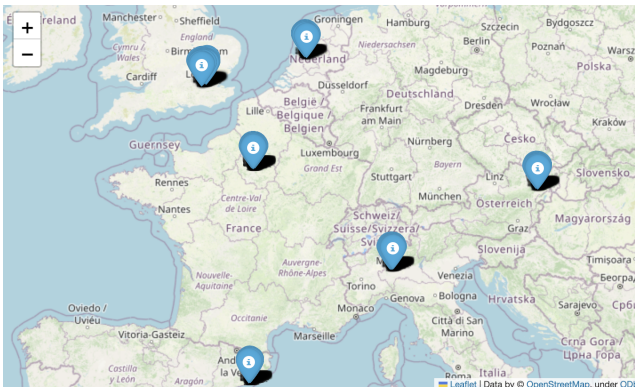

Figure 1 - Europe Map

In this image, we can distinguish the six countries shown above in the list. In the following image, there is a histogram presented to describe even more the distribution of hotels by the countries.

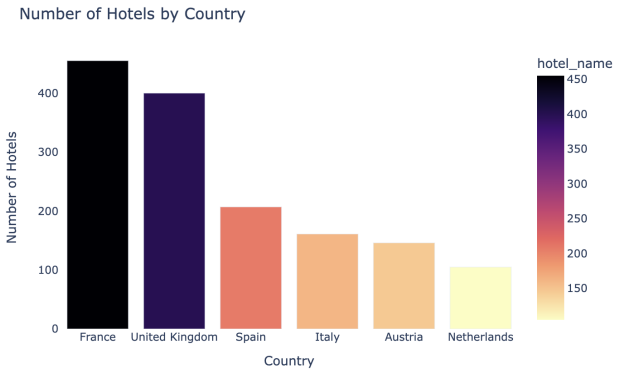

Figure 2 - Count of Hotels per Country Histogram

With the help of these two images, we can clearly check that our dataset has plenty more reviews of hotels in France, meaning there is more information for the French hotels, but not discrediting the other cities that have plenty of data that can give the users very good recommendations as well.

For the recommendations, our system will be based on tags and the comments left by other users. In the next figure, we will present the top 10 tags that are present in the hotel's description whether it is the type of room, type of group, the duration of the stay, and many more.

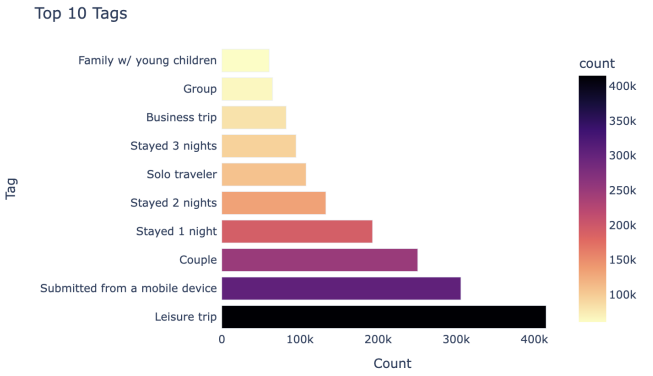

Figure 3 - Top 10 Occurring Tags in Users Review

In the top 10 tags, like demonstrated above, there are a couple of themes that the users mention in their comments that are very useful to the Recommendation System we will develop and its explainability.

Before starting to develop the system, a fault was detected in the names of the hotels, specially in the French ones where the word Hotel is written *Hôtel* and in the data frame the character *ô* wasn't recognized. So, our solution was to replace the *H tel* that was left when removing the character with Hotel. This was done to ensure that the user would get the best experience while using our system.

5.1.1 Data Removal

While analysing the dataset, a few *null* values came to our attention in particular because they were present in the latitude and longitude, meaning that their location was a bit dubious to plot on a map. Considering a deletion threshold of 15% since the absolute value of missing values was 17, it was well under 15% of the entries, since the dataset is composed by over 515 000 reviews. The final decision was to delete the entries without this data, since it doesn't affect our analysis or the recommendation system.

5.1.2 Information Insertion

To make our dataset a bit more detailed and more complete, we used reverse geocode to get the city of the hotels we had in the dataset. This resulted in information that is more easily dealt with and could be helpful when trying to insert the closest landmarks.

5.1.3 Tags

All the hotels that exist in our dataset have tags related to them as mentioned above, and they are going to be essential for our project, being them one of the basis for the recommendation our system will make.

5.1.4 Reviews

Since this dataset is based on reviews of the client of the hotels, when analysing all the information, we could see that there were "*No Positive*" values in the Positive Review column as a sign of the user not saying anything positive of the hotel. The same happened in the Negative Review column but with "*No Negative*". The solution for these values was to replace them by only "*No*" since this value is going to be removed when applying the stopwords.

In the dataset, there are also some duplicated reviews and, we are only interested in getting one set because duplicate information can lead to bias in the recommendation system.
Then, a labelling function was developed to ensure that every duplicate would have the value "*Nothing*" that would also be removed by the stopwords.

5.1.5 Landmarks

We noted that this dataset was composed by hotels in very touristic cities such as Amsterdam, Barcelona, Vienna, London, Milan, or even Paris.
With this conclusion, since these cities are known for their landmarks, a data frame was created in order to plot the landmarks in a map alongside the hotels the system suggests. This way, the user can understand if the hotel is close to the chosen landmark.
Giving one example of landmark, the user can choose for each city:

- Eiffel Tower - Paris
- Sagrada Familia - Barcelona
- The British Museum - London
- Duomo di Milano - Milan
- Anne Frank House - Amsterdam
- Hofburg Palace - Vienna

To choose a landmark, the user will have to choose a city first, and, the system will only be able to calculate the distance to the landmarks present in the dataset and the ones that are in the chosen city.

6. Implementation

In the development of this project, there were different steps that needed to be done in order to achieve our end goal. Starting by applying some knowledge of different areas like Feature Engineering, Natural Language Processing, Data Visualization so we could understand if we were in the right way or not.

6.1 Feature Engineering

In this project, there was a need to rely on Feature Engineering so we could store different attributes in columns containing either 0 or 1 values. These value have the following meanings:

- **0** - The Hotel **doesn't have** the attribute
- **1**- The Hotel **has** the attribute

In order to not keep duplicates and to understand which were the key attributes, a set was used to store the respective attributes for every hotel.

After this step, the attributes were put together in an array so that we could refactor them and refactor the attributes of the hotel to more generic ones.

6.2 Data Visualization

In this section, we present the visualizations that were made to get even more information about the hotels. In this two plots, the goal is to analyse the quantity of room types there are in the dataset and their special characteristics.

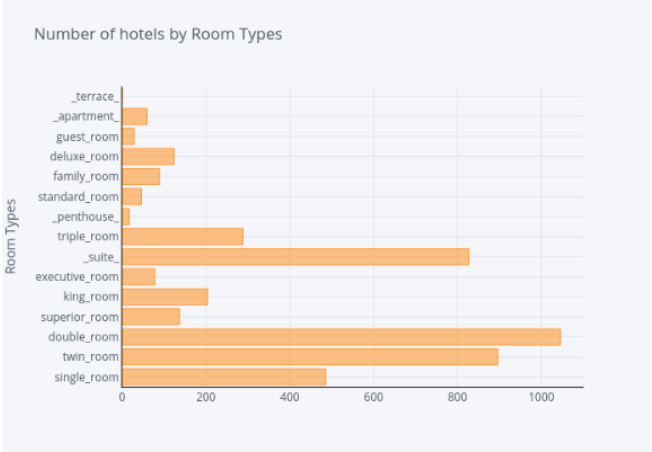

Figure 3 - Number of Hotels by Room Type Bar Chart

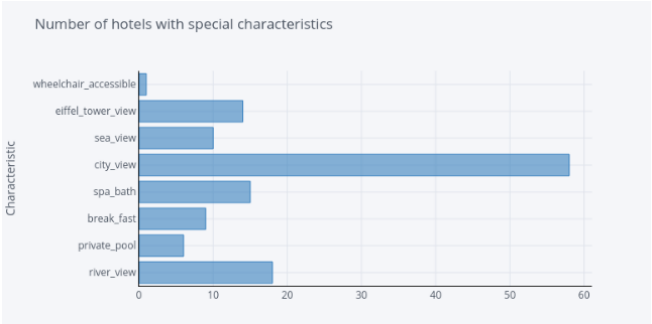

Figure 4 - Number of Hotels with Special Characteristics Bar Chart

6.3 Text Visualization

For a better visualization of the dataset, in this section we are interested in trying to comprehend the frequency of words in the reviews and the occurrence of stop words.

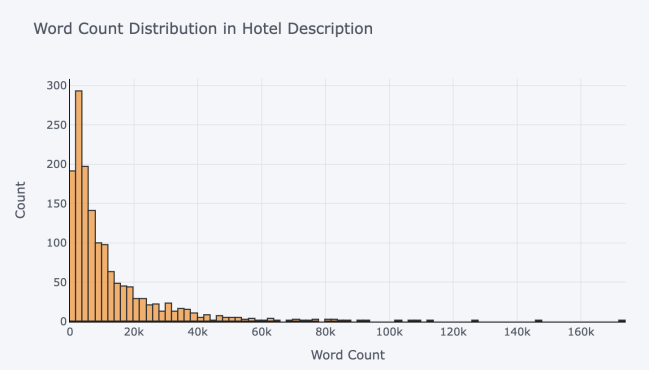

Figure 5 - Word Count in Hotel Descriptions Histogram

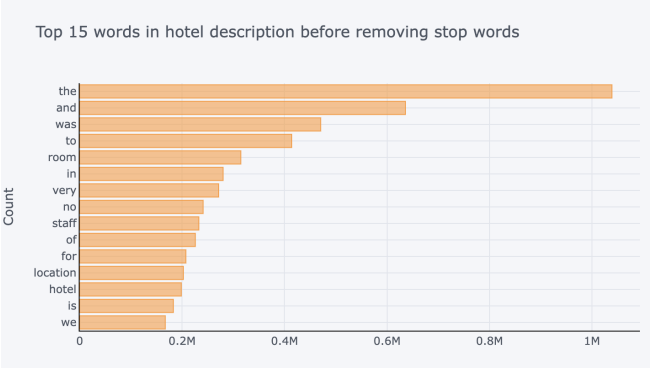

Figure 6 - Bar Graph with the top 15 words in the description without removing stopwords

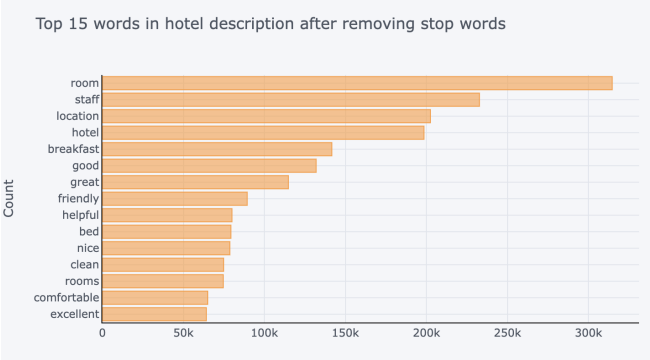

Figure 7 - Bar Graph with the top 15 words in the description after removing stopwords

In the figures presented above (figures 6 and 7) the conclusion we can take is that there is a big difference if we keep the stopwords in the reviews of the hotels.

Furthermore, once the stopwords were taken out, the words became more useful to describe the hotel and provide more details about their facilities and accommodation.

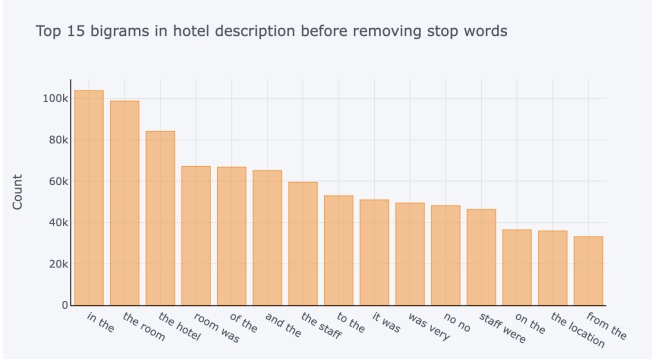

Figure 8 - Bar Graph with the top 15 bi-grams in the description without removing stopwords

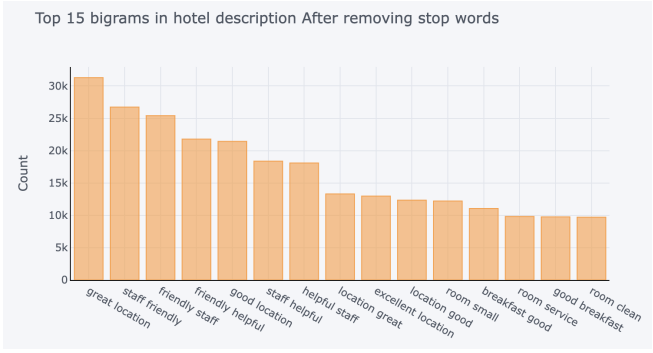

Figure 9 - Bar Graph with the top 15 bi-grams in the description after removing stopwords

Above in figures 8 and 9, instead of plotting graphs regarding words by themselves we combined two words which can be referred as bi-grams. In these figures we can clearly state that the bi-grams are more helpful with getting more information out of the top 15 than with the simple words because it joins even more information and characterization of some features of the hotels.

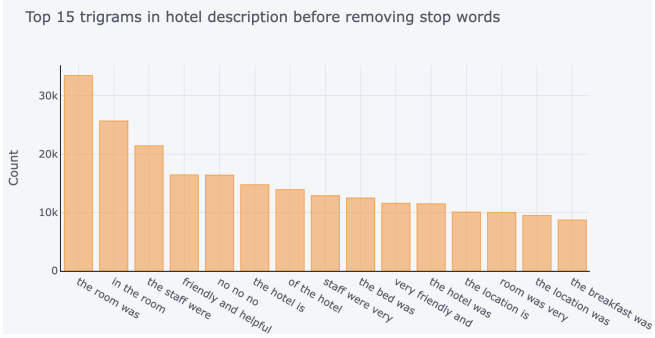

Figure 10 - Bar Graph with the top 15 tri-grams in the description without removing stopwords

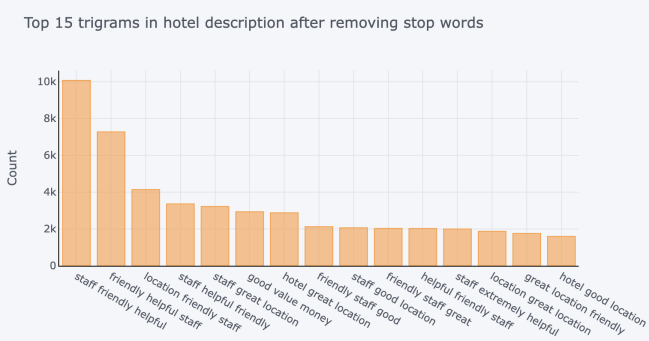

Figure 11 - Bar Graph with the top 15 tri-grams in the description after removing stopwords

Both figures 9 and 10 represent almost the same as the previous four graphs, but instead of presenting single words or just bi-grams this time we analysed tri-grams, these sets are junctions of three words. If two words give more information than just a single word, three words give even more, and this way we can even take more information out of the reviews.

This section's work was developed with the help of iplot and plotly which helps us to create simple but meaningful graphics while keeping the visualization clear.

6.4 Modelling

Based on the steps made beforehand, it is evident that bi-grams and tri-grams have more impact on the modelling process, giving more context of the reviews and making more sense to model the recommender system.

Our model was constructed based on Cosine Similarity as a resource of the sklearn library, we also used TFIDF Vectorizer in order to vectorize words.

7. Results

In this section, we will make a demonstration of our recommendation system.

```
Welcome to MORIANA the Best Hotel Recomendation System
[MORIANA] - The cities we have data available
```

	City
0	Paris
1	London
2	Vienna
3	Barcelona
4	Amsterdam
5	Milan

For the user, the first menu that appears is a list of the cities available in our system.

Afterwards, the user will be asked to choose a city and then the user will be asked to insert a landmark he is interested in knowing the distance to the best hotel in the city. Apart from the best hotel in the city regarding reviews, an additional top five recommendations will appear, and they will be labelled as similar hotels.

```
Which city would you like to search hotels?
PARIS

Great! Searching for hotels in PARIS...
```

For this test, the landmark data frame used was the original one that contemplates sixty-one landmarks spread across the six cities available.

After deciding on a city, the user will have the choice to input a landmark he is interested in or not.

If the user chooses to input a landmark, then the system is going to calculate the distance between the best hotel in the city and the landmark. For this example, the user would like to know the distance between the Hotel the system will recommend him and the Eiffel Tower in Paris.

```
Do you prefer to stay close to any landmark? Type 'no' if you don't have a specific landmark preference.
EIFFEL TOWER

Great! You prefer to stay close to EIFFEL TOWER. Continuing search...
```

Figure 14 - Output of the Landmark Choice

After the user selecting a landmark, the system can give him various answers, such as the example present in figure 14 that is a successful answer and the landmark is presented in the chosen data frame.

If the user answers a landmark, the system doesn't recognize, the answer will be the one demonstrated in figure 15.

```
Which city would you like to search hotels?
PARIS

Great! Searching for hotels in PARIS...

Do you prefer to stay close to any landmark? Type 'no' if you don't have a specific landmark preference.
HYDE PARK

Sorry, we don't have data regarding the landmark HYDE PARK in PARIS. We have limited data about some main landmarks of PARIS.


From the available data, the landmarks in PARIS we have are:
Eiffel Tower
Louvre Museum
Notre-Dame Cathedral
Montmartre
Arc de Triomphe
Musée d'Orsay
Sainte-Chapelle
Panthéon
Musée Picasso
Champs-Élysées
Tuileries Garden
```

Figure 15 - Demonstration of a wrong landmark insertion

For this example we still considered the user wanted to stay in Paris and inserted a landmark that doesn't belong to the city, for example, Hyde Park in London. The system didn't recognize the existence of a Hyde Park in Paris and also gave some example of landmarks that we have data in Paris.

In order to extend the demonstration, Eiffel Tower was chosen again as the preferred landmark. Now, the system will show the user the best hotel he can stay in (demonstrated in figure 16) and then five hotels that are also really great and have similar characteristics to the best hotel in Paris regarding reviews. And, the system also gives the user the possibility to read ten reviews from those hotels. For demonstration purposes, we included one example in figure 17.

Figure 16 - Output of the Best Hotel of the Chosen City and the distance between the landmark and the

Figure 17 - Output of the Similar Hotel of the Chosen City and the distance between the landmark and the hotel

Now, the explanation to why these hotels were selected to be shown to the user relies solely on the comments made by previous users. When comparing hotels the Hotel Regina (the best in Paris) according to our users, the closest one is Holiday Inn Paris Notre Dame as shown above in figure. 17.

To get the best visualization of this, we decided to implement a map where the red marker marks the best hotel in the city, and the blue ones are the hotels classified as similar to the best one. And to complete the map, the green markers with information points are the landmark the system has access to.

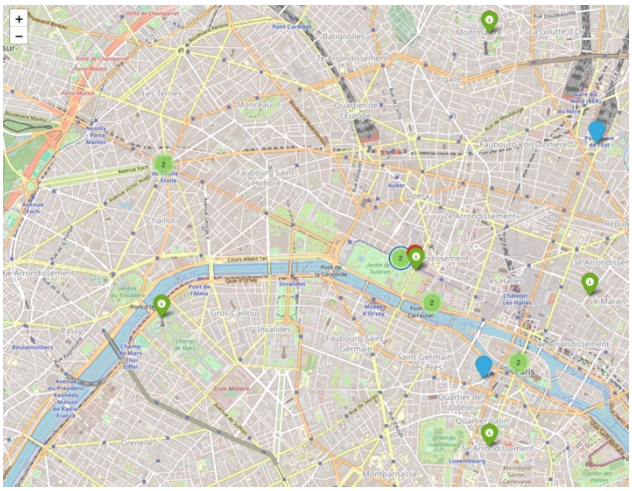

Figure 18 - Map the user sees as soon as the system opens the page

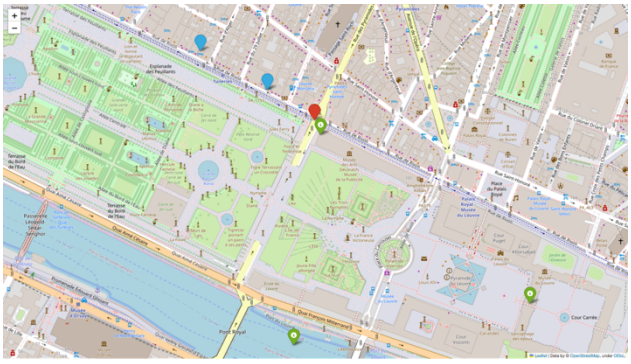

Figure 19 - Close up look for a better view of the best hotel

This map is totally interactive and clickable, where the user can see the name of the markers he is clicking.

8. Validation

To get a better validation of our work, we developed a Google Forms for our users to check the usability, and for them to interact with the system having seven people answering our questionary but with different backgrounds and travel knowledge.

After developing the questionary to share, we asked people that we knew that travel a lot and could use the help of a Recommendation System and people who do not travel so often. [15]

- Do you travel a lot? (1- Never, 2- Rarely, 3- Some times a year, 4- Usually, 5- Very Frequently)

Mean	Standard Deviation	Min Value	Max Value
4	1,069044968	2	5

- Do you know any landmark in each city? (1- Yes for all the cities, 2- Yes, in almost every city, 3- No)

Mean	Standard Deviation	Min Value	Max Value
1	0,5345224838	1	2

- Do you know any of the recommended hotels? (1- Yes or 2- No)

Mean	Standard Deviation	Min Value	Max Value
1	0,4879500365	1	2

- Do you agree with the suggested hotels? (1-Yes or 2-Some of them, 3- No)

Mean	Standard Deviation	Min Value	Max Value
1	0,5345224838	1	2

- Do you consider the system useful? (1 to 5)

Mean	Standard Deviation	Min Value	Max Value
5	0,4879500365	4	5

9. Future Work

In this section, there are some tasks that we took from our validation questionary and some that we think are good additions for our system.

One of the ideas we recommend is the deployment of the system into a platform for it to be more organized and prettier with other elements and a menu. Another suggestion was the separation of the chosen landmark from the others, marking it with a different colour. There are many suggestions we can make for this system to get even better.

10. Conclusion

In conclusion, with the development with this work, we learned a lot about recommendation systems and other topics. As students of the Bachelors Degree in Engineering and Data Science, we have already some previous knowledge in Natural Language Processing which has really helped us to use prior knowledge to develop our recommendation system and use the right techniques available such as the Cosine Similarity and the TFIDF Vectorizer.

References

[1] – Nvidia, Recommendation System, https://www.nvidia.com/en-us/glossary/data-science/recommendation-system/

[2] – Público, 2023, Estrangeiros põem receitas da hotelaria em máximos: mais de 4 mil milhões até Agosto, https://www.publico.pt/2023/10/13/economia/noticia/estrangeiros-poem-receitas-hotelaria-maximos-4-mil-milhoes-ate-agosto-2066643

[3] – The Portugal News, 2023, Record breaking tourism revenue, https://www.theportugalnews.com/news/2023-10-13/record-breaking-tourism-revenue/82284

[4] – INE, 2023, Movement of passengers at national airportscontinues to reach historical highs - August 2023, https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_destaques&DESTAQUESdest_boui=594873783&DESTAQUESmodo=2

[5] – Music Tomorrow, 2022, Inside Spotify's Recommender System: A Complete Guide to Spotify Recommendation Algorithms, https://www.music-tomorrow.com/blog/how-spotify-recommendation-system-works-a-complete-guide-2022

[6] – Inc., 2021, This was Steve Jobs most controversial legacy. It was also his most brilliant, https://www.inc.com/jason-aten/this-was-steve-jobs-most-controversial-legacy-it-was-also-his-most-brilliant.html

[7] - 515K Hotel Reviews Data in Europe, 2017, https://www.kaggle.com/datasets/jiashenliu/515k-hotel-reviews-data-in-europe/data

[8] – OpenAI ChatGPT, https://openai.com/chatgpt

[9] – NLTK library documentation, https://www.nltk.org

[10] – Scikit-learn documentation, https://scikit-learn.org/stable/

[11] – Susan Li, 2018, A Machine Learning Approach – Building a Hotel Recommendation Engine, https://towardsdatascience.com/a-machine-learning-approach-building-a-hotel-recommendation-engine-6812bfd53f50

[12] – Aman Kharwal, 2021, Hotel Recommendation System with Machine Learning, https://thecleverprogrammer.com/2021/02/13/hotel-recommendation-system-with-machine-learning/

[13] – Alantancr, 2020, Hotel Recommender, https://github.com/alantancr/Hotel-Recommender

[14] – Keshav Ramaiah, 2020, Hotel Recommender, https://www.kaggle.com/code/keshavramaiah/hotel-recommender/notebook

[15] - Validation Google Forms, https://forms.gle/KP2riJzw6dM2CbjF9

[16]- GitHub Repository, https://github.com/paulinomary/IACH-Hotel-Recommendation-System