Use of AI and ML models for a Hotel Recommendation System

Diogo Beltrán Dória, Mariana Lopes Paulino
University of Coimbra
{dbdorio, 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 Motivation

Since 2022, tourism activity has grown considerably, despite the traveling restrictions and tight rules implemented by countries to address 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, INA, 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 is possible to filter by the usual tags most common websites like booking.com, trivago.pt provide, but also adds 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 abovementioned criteria, potentially saving customer´s time and money, contributing for 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 in 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 recommend.

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, and 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 consider, but also whether the hotel is appropriate for vacation 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 hotel
Review Date	Date when 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 Hotel
Reviewer Nationality	Nationality of Reviewer
Negative Review	Negative Review the reviewer gave to the hotel. If the reviewer does not give the negative review, then it should be: 'No Negative'
Review Total Negative 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 the negative review, then it should be: 'No Positive'
Review Total Positive 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 Reviewer Has Given	Number of Reviews the reviewers has given in the past
Total Number of Reviews	Total number of valid reviews the hotel has
Tags	Tags 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 Open AI framework, chat GPT, 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]

6. Evaluation

Through plots we will understand the frequency of top words and its occurrence with respect to stopwords. After understanding which modelling process has more impact giving us more context of the reviews. Results will be evaluated based on cosine similarities between documents and the use of TfidfVectorizer to vectorize the words and calculate TF_IDF normalization. We will develop more this section of the work, as we progress in the literature review.

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 airports continues 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