IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
ISSN: 2457-0362
www.ijarst.in

DATA MINING AND ML MODELS FOR AIRBNB DATA ANALYSIS WITH COSINE SIMILARITY

A. Lavanya[1], Korupathi Sarika[2], Kalvakuntla Sadweep[2], Thotala anil Yadav[2], Gummadi Avinash[2]

[1]Assistant Professor, [2]UG Scholar, [1,2]Department of CSE (Data Science)
[1,2]Malla Reddy Engineering College and Management Sciences, Kistapur, Medchal, 501401, Hyderabad, Telangana

ABSTRACT

In 2019, PricewaterhouseCoopers conducted a study highlighting five sharing sectors: travel, car sharing, finance, recruiting, and streaming music and videos. According to their findings, these sectors have the potential to significantly boost global revenue from $158 billion (estimated in 2019) to approximately $735 billion by the year 2033. Among these sectors, the travel industry has been significantly impacted by the emergence of new peer-to-peer (P2P) models like Airbnb. These platforms have disrupted the conventional reservation systems by offering a unique and personalized experience to consumers. As of now, Airbnb stands as the largest P2P hosting platform, boasting around 4 million ads in 2017 and having a valuation of $25 billion in 2015. One of the major reasons for the popularity of Airbnb is its ability to make rental sharing more cost-effective and convenient for both hosts and customers. However, as the customer experience plays a vital role in shaping their opinions and recommendations, online user reviews have a significant impact on the consumer interest in P2P platforms like Airbnb. Potential guests heavily rely on the feedback and ratings provided by previous users to make informed decisions about their bookings. Since trying out properties before making a reservation is not feasible, these reviews become crucial in influencing the customers' choices. With Airbnb emerging as the leading platform for short-term rental accommodations, it has become essential to understand the factors that contribute to customer satisfaction in the P2P hosting landscape. Various studies have been conducted in this direction, but there are still certain gaps that need to be addressed, especially concerning how different categories of customers perceive and approach their rental experiences. This understanding will be vital for the continued success and growth of P2P hosting platforms like Airbnb in the future. Therefore, this work proposes a data mining and machine learning models for the analysis of Airbnb data with cosine similarity.

Keywords: Data mining, Recommendation system, Airbnb analysis, Cosine similarity.

1. INTRODUCTION

AIRBNB and the accommodation business have been growing significantly in the world. In recent years, Airbnb hosts have provided services for nearly 6.5 million customers for over 2.1million nights with the total revenue of $540 million. This booming business shows growth where the listings of hotel in Airbnb continue to grow from 15,000 in 2016 to 30,000 in 2019. As the number of hotels increase, the market has become more competitive. With the prevailing cutthroat competition, for a listing to find the maximum occupancy in a year, they must be appealing for the customer on the Airbnb platform. In 2015, PricewaterhouseCoopers [1] pointed out that five sharing sectors (travel, car sharing, finance, recruiting, and streaming music and videos) have the potential to increase global revenue from $15 billion (estimated in 2015) to around $ 335 billion in 2025. By focusing on the travel sector, the appearance of new peer-to-peer (P2P) models such as Airbnb has the effect of disrupting the classic reservation system [2] by offering an experience different to consumers. Airbnb is currently the largest P2P hosting platform. It contained around 4 million ads in 2017 and was valued at $25 billion in 2015 [3]. Taking into account the popularity of Airbnb, customers and hosts are predisposed to view rental sharing as cheaper as the platform allows each of them to rent the property with global visibility instead of using traditional intermediaries [4] while offering an unprecedented experience given the particularity of certain properties on offer (such as igloos or castles) compared to traditional accommodation, in particular that offered by hotels [5]. In addition, this different experience stimulates the hedonic value of the customers [6]. However, since the customer experience is a determining factor that can have an impact on the recommendations of the latter and taking into account that products related to hosting cannot be tried before purchase [7], consumer interest in listings displayed on P2P platforms (of which Airbnb is one) is influenced by online user reviews [8]. It even happens that the latter have an impact on the purchasing decision since they constitute a major source of information [9] and thus contribute to electronic word of mouth (eWOM) which has radically restructured the relationship between the business and the consumer [10].

2. LITERATURE SURVEY

The sharing/collaborative consumption economy has made great strides in recent years. It is a form of consumption where people share goods or services online. It represents collaborative activities to benefit, provide, or share access to goods or services, coordinated by online services that are based on a community of users [12]. Interactions between users of this form of consumption are often provided by P2P sites or platforms that facilitate contact and coordinate the exchange [14]. Moreover, the platforms of this new type of economy have acquired a significant market share in several segments such as transport (Uber and Cabify) and accommodation (Airbnb and CouchSurfing) [13]. Although the sharing economy has long existed in tightly knit communities, its shift to a much larger scale was the result of certain conditions including the rapid adoption of new technologies and low entry requirements for startups [13].

Collaborative consumption allows people to perceive the benefits of ownership while at a reduced cost, ensuring it becomes an alternative to traditional home ownership [14]. Among the services that are part of collaborative consumption is P2P tourism. P2P tourism brings together activities carried out by tourists who interact with the attributes of the destination (including gastronomy, entertainment, and visits to natural and cultural heritage) made available by peers [15]. P2P tourism ensures a direct relationship between the host and the customer, which has the effect of promoting the authenticity perceived by the latter vis-à-vis their tourist experience [65]. Additionally, hosts can offer local experiences to their guests who, by engaging with the community, discover how the city of their stay lives [17]. Residents themselves can, through P2P tourism, contribute to tourism-related business activities. Indeed,

Hamari et al. [18] identified four main factors that arouse the willingness to participate in the activities of the sharing economy as a service provider, namely, sustainability, pleasure, reputation, and economic benefits, especially with the development of political reforms in favor of the collaborative economy by certain large cities including San Francisco, Paris, London, and Singapore [19].

Since 2007, the first P2P tourism platforms have appeared, except they were not very popular. However, some of them have had great success over time, in this case, Airbnb whose activity is linked to P2P hosting which, in March 2018, had more than 150 million users and 640,000 hosts [20].

Airbnb is a leading platform for short-term accommodation and a pioneer in the sharing economy. It is a service that connects people who have a space to share with people who are looking for accommodation. Airbnb describes itself as a trusted community marketplace for people to list, discover, and book unique accommodations around the world [21]. Since its launch in 2008, Airbnb has grown very rapidly with more than 2 million ownerships worldwide and over 50 million customers who used their services in 2015 [22].

As with the rest of the collaborative consumption platforms, technological innovations have simplified the process of entering the market and allowed it to facilitate the list of searchable for consumers and reduce transaction costs. Airbnb provides better reach by reducing consumer search costs as electronic marketplaces reduce inefficiencies caused by buyer search costs [23]. This significant advantage has placed it at the forefront of competition with traditional providers of accommodation services (such as hotels and guesthouses). Indeed, certain stays with Airbnb can replace certain hotel stays, which affects the turnover of the latter. This impact can be differentiated by geographic area, by hotel market segment, or by season [21]. For example, Credit Suisse analysts estimated that Airbnb led to an 18.6% drop in revenue per room in January 2015 in New York [22].

Faced with this situation, the managers of hotel chains sometimes make contemptuous statements concerning competitors such as Airbnb, arguing their remarks by the fact that these platforms are a niche market, or target market segments complementary to those targeted by hotels. In fact, Airbnb for its part announced that 70% of the properties offered on its platform are located outside hotel zones [23]. What is certain is that P2P hosting platforms (in this case, Airbnb) have changed customers' perceptions of their trepidation. Many of the latter are looking for low-cost housing and direct interaction with the local community. This interaction was preceded by direct interaction with the host through the P2P platform. This has helped transform the market and attracted mainstream consumers by giving them the opportunity to rent properties as tourist residences [23].

3. PROPOSED METHODOLOGY

3.1 Overview

The primary objective of this project is to analyze and gain insights from a dataset containing information about Airbnb listings. Additionally, the project aims to build a recommendation system to suggest similar listings to users based on textual descriptions. Overall, this project combines data analysis, data visualization, and natural language processing (NLP) techniques to extract valuable insights from Airbnb listing data and provide users with recommendations based on listing descriptions. It demonstrates a comprehensive approach to exploring and utilizing data for practical applications in the hospitality industry.

Step 1: Data Loading and Exploration

- The project begins by loading the dataset from a CSV file ("train.csv") using the pandas library.
- Initial data exploration is conducted to understand the dataset's structure, including data types, missing values, and a preview of the data.

Step 2: Data Visualization

- Various data visualization techniques are employed to better understand the dataset:
- A pie chart is used to visualize the distribution of room types in Airbnb listings.
- Price data is cleaned and plotted to show price fluctuations over a three-month period.
- The top neighbourhoods with the most reviews are identified and displayed in a bar chart.
- The neighborhood with the most real estate listings is determined.
- The number of entries per neighborhood and per month is presented.
- Histograms are used to visualize the distribution of neighbourhoods.

Step 3: Map Visualization – It includes an interactive map visualization, displaying Airbnb listings for the month of February. Each listing is marked on the map with additional transit information available in pop-up form.

Step 4: Neighborhood Review Analysis – It identifies and analyzes neighbourhoods with the highest review scores. This information is presented in a horizontal bar chart.

Step 5: Room Type Analysis – It calculates and displays the average number of people that each room type can accommodate.

Step 6: Text Data Preprocessing – The textual data from the "name" and "description" columns is preprocessed for further analysis. This includes steps such as removing punctuation, converting to lowercase, removing stop words, and performing lemmatization.

Step 7: Recommendation System

- It builds a recommendation system for Airbnb listings based on the textual descriptions of each listing.
- It uses the TF-IDF (Term Frequency-Inverse Document Frequency) technique to vectorize the textual data.
- Cosine similarity is computed between listings to identify similar listings.
- A function called recommend() is provided to recommend similar listings based on a given listing ID.

Step 8: Top Co-Occurring Words – It identifies the top 10 pairs of words that commonly co-occur in the descriptions of listings, providing insights into frequently associated terms.

3.2 Data Preprocessing

Data pre-processing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. When creating a machine learning project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. So, for this, we use data pre-processing task. A real-world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for machine learning models. Data pre-processing is required tasks for cleaning the data and making it suitable for a machine learning model which also increases the accuracy and efficiency of a machine learning model.

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Encoding Categorical Data
- Splitting dataset into training and test set

Importing Libraries: To perform data preprocessing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data preprocessing, which are:

Numpy: Numpy Python library is used for including any type of mathematical operation in the code. It is the fundamental package for scientific calculation in Python. It also supports to add large, multidimensional arrays and matrices. So, in Python, we can import it as:

import numpy as nm

Here we have used nm, which is a short name for Numpy, and it will be used in the whole program.

Matplotlib: The second library is matplotlib, which is a Python 2D plotting library, and with this library, we need to import a sub-library pyplot. This library is used to plot any type of charts in Python for the code. It will be imported as below:

import matplotlib.pyplot as mpt

Here we have used mpt as a short name for this library.

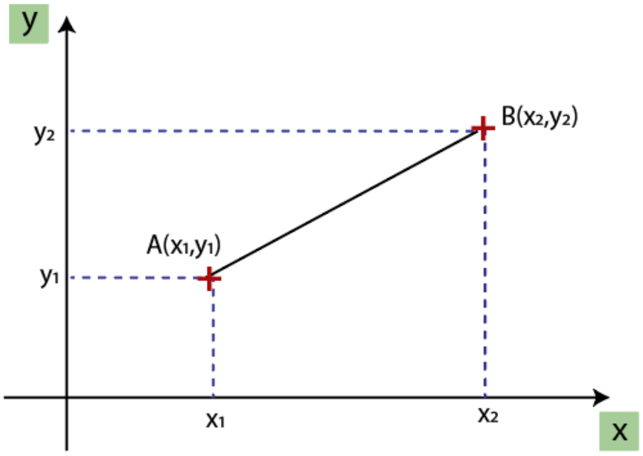
Pandas: The last library is the Pandas library, which is one of the most famous Python libraries and used for importing and managing the datasets. It is an open-source data manipulation and analysis library.

Handling Missing data: The next step of data preprocessing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset. There are mainly two ways to handle missing data, which are:

- By deleting the particular row: The first way is used to commonly deal with null values. In this way, we just delete the specific row or column which consists of null values. But this way is not so efficient and removing data may lead to loss of information which will not give the accurate output.
- By calculating the mean: In this way, we will calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age, salary, year, etc.

Encoding Categorical data: Categorical data is data which has some categories such as, in our dataset; there are two categorical variables, Country, and Purchased. Since machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So, it is necessary to encode these categorical variables into numbers.

Feature Scaling: Feature scaling is the final step of data preprocessing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no variable dominates the other variable. A machine learning model is based on Euclidean distance, and if we do not scale the variable, then it will cause some issue in our machine learning model. Euclidean distance is given as:



Euclidean Distance Between A and B=$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

Figure 1: Feature scaling

3.3 Splitting the Dataset

In machine learning data preprocessing, we divide our dataset into a training set and test set. This is one of the crucial steps of data preprocessing as by doing this, we can enhance the performance of our machine learning model. Suppose if we have given training to our machine learning model by a dataset and we test it by a completely different dataset. Then, it will create difficulties for our model to understand the correlations between the models. If we train our model very well and its training accuracy is also very high, but we provide a new dataset to it, then it will decrease the performance. So we always try to make a machine learning model which performs well with the training set and also with the test dataset. Here, we can define these datasets as:

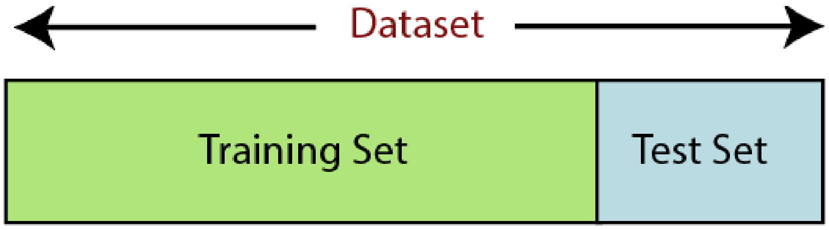


Figure 2: Splitting the dataset.

Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

For splitting the dataset, we will use the below lines of code:

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test= train_test_split(x, y, test_size= 0.2, random_state=0)
```

Explanation

- In the above code, the first line is used for splitting arrays of the dataset into random train and test subsets.
- In the second line, we have used four variables for our output that are
- x_train: features for the training data
- x_test: features for testing data
- y_train: Dependent variables for training data
- y_test: Independent variable for testing data
- In train_test_split() function, we have passed four parameters in which first two are for arrays of data, and test_size is for specifying the size of the test set. The test_size maybe .5, .3, or .2, which tells the dividing ratio of training and testing sets.
- The last parameter random_state is used to set a seed for a random generator so that you always get the same result, and the most used value for this is 42.

3.4 TF-IDF Feature Extraction

TF-IDF, short for Term Frequency-Inverse Document Frequency, is a commonly used technique in NLP to determine the significance of words in a document or corpus. To give some background context, a survey conducted in 2015 showed that 83% of text-based recommender systems in digital libraries use

TF-IDF for extracting textual features. That's how popular the technique is. Essentially, it measures the importance of a word by comparing its frequency within a specific document with the frequency to its frequency in the entire corpus. The underlying assumption is that a word that occurs more frequently within a document but rarely in the corpus is particularly important in that document.

3.4.1 Mathematical formula for calculating TF-IDF

TF (Term Frequency) is determined by calculating the frequency of a word in a document and dividing it by the total number of words in the document.

- TF = (Number of times the word appears in the document) / (Total number of words in the document)
- IDF (Inverse Document Frequency), on the other hand, measures the importance of a word within the corpus as a whole. It is calculated as:
- IDF = log((Total number of documents in the corpus) / (Number of documents containing the word))

The tf-idf value increases in proportion to the number of times a word appears in the document but is often offset by the frequency of the word in the corpus, which helps to adjust with respect to the fact that some words appear more frequently in general. TF-IDF use two statistical methods, first is Term Frequency and the other is Inverse Document Frequency. Term frequency refers to the total number of times a given term t appears in the document doc against (per) the total number of all words in the document and The inverse document frequency measure of how much information the word provides. It measures the weight of a given word in the entire document. IDF show how common or rare a given word is across all documents. TF-IDF can be computed as tf * idf

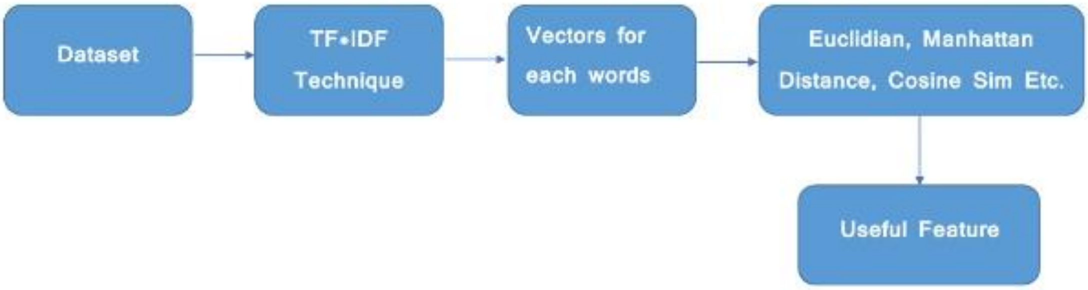


Fig. 3: TF-IDF block diagram.

TF-IDF do not convert directly raw data into useful features. Firstly, it converts raw strings or dataset into vectors and each word has its own vector. Then we'll use a particular technique for retrieving the feature like Cosine Similarity which works on vectors, etc.

Terminology

t — term (word)

d — document (set of words)

N — count of corpus

corpus — the total document set

Step 1: Term Frequency (TF): Suppose we have a set of English text documents and wish to rank which document is most relevant to the query, "Data Science is awesome!" A simple way to start out is by eliminating documents that do not contain all three words "Data" is", "Science", and "awesome", but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency. The weight of a term that occurs in a document is simply proportional to the term frequency.

$$tf(t,d) = count\ of\ t\ in\ d\ /\ number\ of\ words\ in\ d$$

Step 2: Document Frequency: This measures the importance of document in whole set of corpora, this is very similar to TF. The only difference is that TF is frequency counter for a term t in document d, whereas DF is the count of occurrences of term t in the document set N. In other words, DF is the number of documents in which the word is present. We consider one occurrence if the term consists in the document at least once, we do not need to know the number of times the term is present.

$$df(t) = occurrence\ of\ t\ in\ documents$$

Step 3: Inverse Document Frequency (IDF): While computing TF, all terms are considered equally important. However, it is known that certain terms, such as “is”, “of”, and “that”, may appear a lot of times but have little importance. Thus, we need to weigh down the frequent terms while scale up the rare ones, by computing IDF, an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely. The IDF is the inverse of the document frequency which measures the informativeness of term t. When we calculate IDF, it will be very low for the most occurring words such as stop words (because stop words such as “is” is present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.

$$idf(t) = N/df$$

Now there are few other problems with the IDF, in case of a large corpus, say 100,000,000, the IDF value explodes, to avoid the effect we take the log of idf . During the query time, when a word which is not in vocab occurs, the df will be 0. As we cannot divide by 0, we smoothen the value by adding 1 to the denominator.

$$idf(t) = log(N/(df + 1))$$

The TF-IDF now is at the right measure to evaluate how important a word is to a document in a collection or corpus. Here are many different variations of TF-IDF but for now let us concentrate on this basic version.

$$tf - idf(t,d) = tf(t,d) * log(N/(df + 1))$$

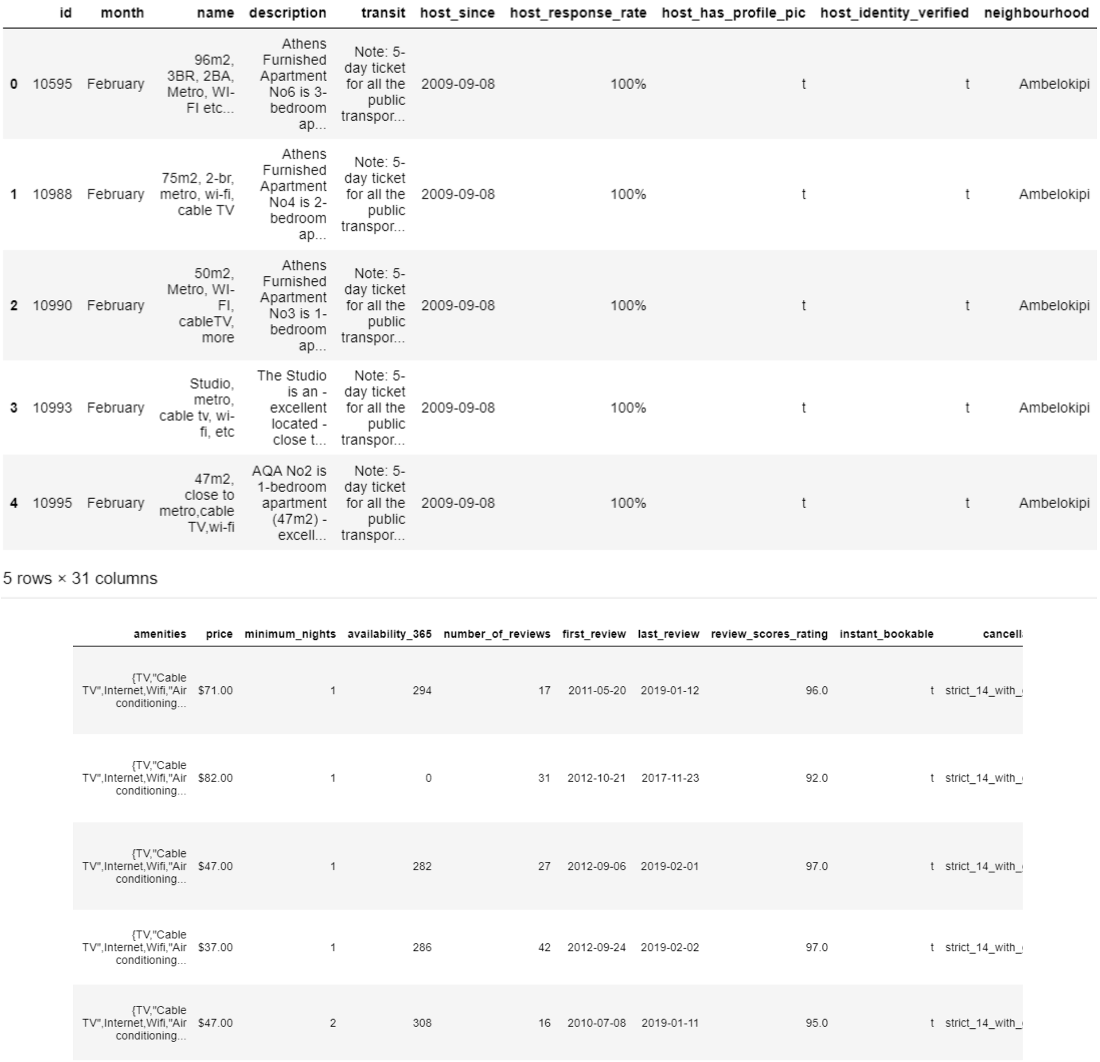
Step 4: Implementing TF-IDF: To make TF-IDF from scratch in python, let’s imagine those two sentences from different document:

first sentence: “Data Science is the sexiest job of the 21st century”.

second sentence: “machine learning is the key for data science”.

4. RESULTS AND DISCUSSION

Figure 4 shows the top rows of the Airbnb dataset, providing a glimpse of the structure and content of the dataset. It includes columns like listing ID, host ID, name, neighbourhood, room type, price, and other relevant information.



	id	month	name	description	transit	host_since	host_response_rate	host_has_profile_pic	host_identity_verified	neighbourhood
0	10595	February	96m2, 3BR, 2BA, Metro, WI-FI etc...	Athens Furnished Apartment No6 is 3-bedroom ap...	Note: 5-day ticket for all the public transpor...	2009-09-08	100%	t	t	Ambelokipi
1	10988	February	75m2, 2-br, metro, wi-fi, cable TV	Athens Furnished Apartment No4 is 2-bedroom ap...	Note: 5-day ticket for all the public transpor...	2009-09-08	100%	t	t	Ambelokipi
2	10990	February	50m2, Metro, WI-FI, cableTV, more	Athens Furnished Apartment No3 is 1-bedroom ap...	Note: 5-day ticket for all the public transpor...	2009-09-08	100%	t	t	Ambelokipi
3	10993	February	Studio, metro, cable tv, wi-fi, etc	The Studio is an - excellent located - close t...	Note: 5-day ticket for all the public transpor...	2009-09-08	100%	t	t	Ambelokipi
4	10995	February	47m2, close to metro,cable TV,wi-fi	AQA No2 is 1-bedroom apartment (47m2) - excell...	Note: 5-day ticket for all the public transpor...	2009-09-08	100%	t	t	Ambelokipi

5 rows × 31 columns

amenities	price	minimum_nights	availability_365	number_of_reviews	first_review	last_review	review_scores_rating	instant_bookable	cancell
{TV,"Cable TV",Internet,Wifi,"Air conditioning...	$71.00	1	294	17	2011-05-20	2019-01-12	96.0	t	strict_14_with_
{TV,"Cable TV",Internet,Wifi,"Air conditioning...	$82.00	1	0	31	2012-10-21	2017-11-23	92.0	t	strict_14_with_
{TV,"Cable TV",Internet,Wifi,"Air conditioning...	$47.00	1	282	27	2012-09-06	2019-02-01	97.0	t	strict_14_with_
{TV,"Cable TV",Internet,Wifi,"Air conditioning...	$37.00	1	286	42	2012-09-24	2019-02-02	97.0	t	strict_14_with_
{TV,"Cable TV",Internet,Wifi,"Air conditioning...	$47.00	2	308	16	2010-07-08	2019-01-11	95.0	t	strict_14_with_

Figure 4: Displaying the header of sample dataset from Airbnb.

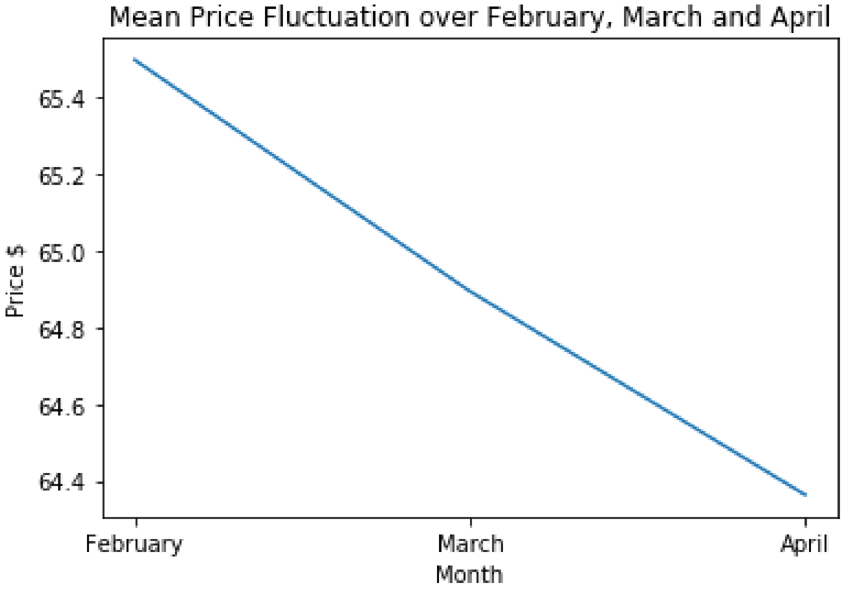


Figure 5: Plot of price fluctuation over 3 months.

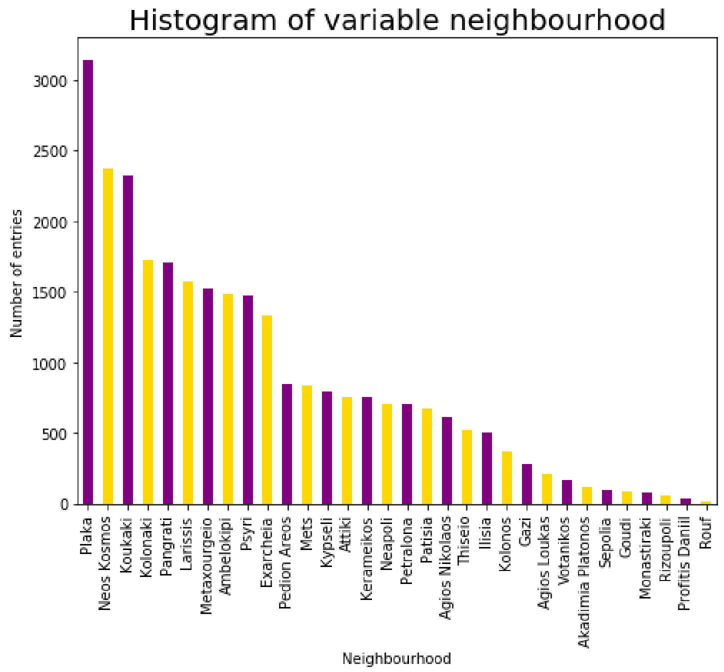


Figure 6: Histogram of variable neighbourhood.

Figure 6 shows a histogram depicting the distribution of listings across different neighbourhoods. It gives an overview of how listings are distributed among various areas. Figure 7 displays a plot comparing the prices of different room types to identify the most expensive one. It' s bar chart highlighting the variation in prices.

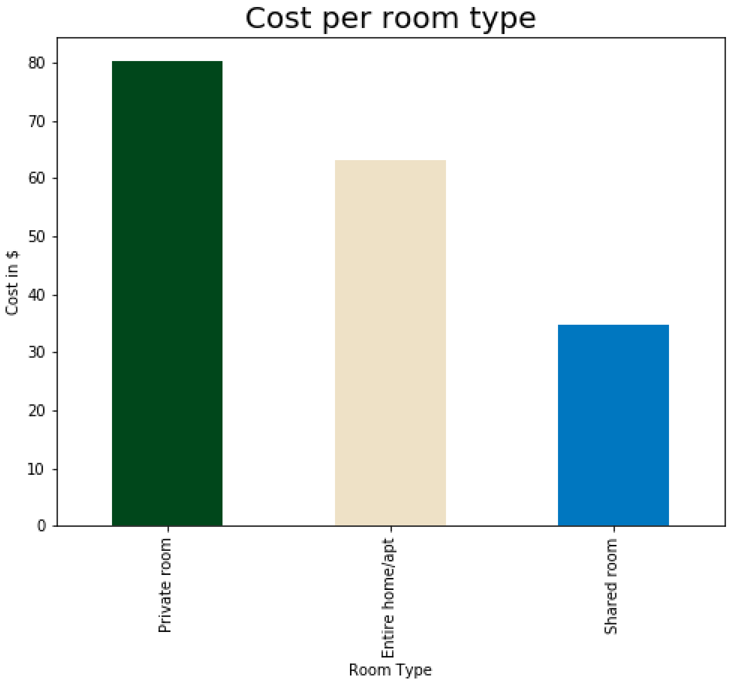


Figure 7: Plot of the most expensive room type.

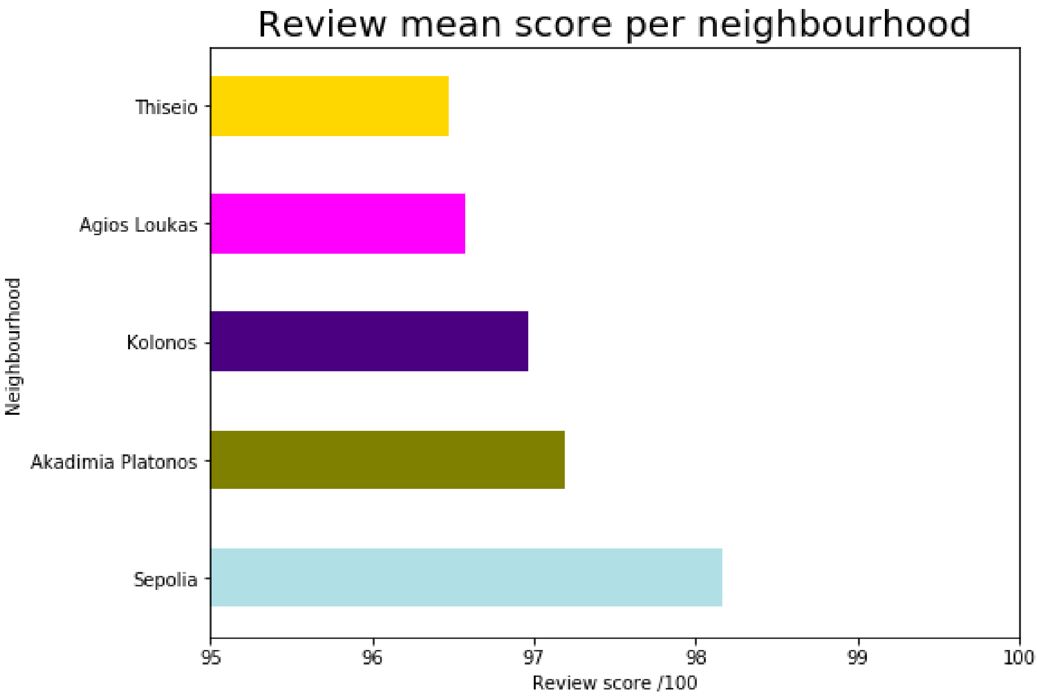
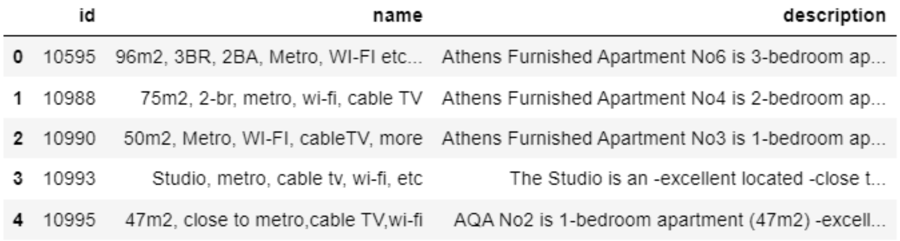


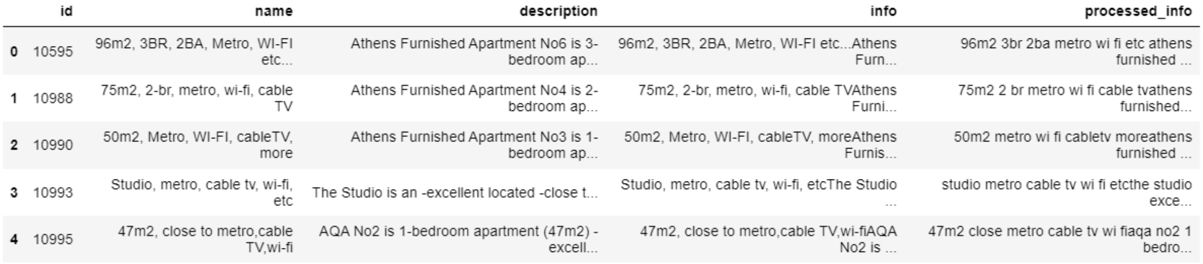
Figure 8: Neighbourhoods with the best review score.

In Figure 8, it reveals the neighbourhoods that receive the highest review scores from guests. It’s simple visualization showing the average review scores for different areas. Similarly, Figure 9 illustrates the process of creating a new data frame with only the columns related to listing ID, name, and description. This subset of data could be used for generating recommendations. In Figure 10, it shown a snippet of preprocessed data where multiple columns or datasets have been combined (concatenated) to create a comprehensive dataset for analysis. This includes columns related to listing details, review scores, and more.



	id	name	description
0	10595	96m2, 3BR, 2BA, Metro, WI-FI etc...	Athens Furnished Apartment No6 is 3-bedroom ap...
1	10988	75m2, 2-br, metro, wi-fi, cable TV	Athens Furnished Apartment No4 is 2-bedroom ap...
2	10990	50m2, Metro, WI-FI, cableTV, more	Athens Furnished Apartment No3 is 1-bedroom ap...
3	10993	Studio, metro, cable tv, wi-fi, etc	The Studio is an -excellent located -close t...
4	10995	47m2, close to metro,cable TV,wi-fi	AQA No2 is 1-bedroom apartment (47m2) -excell...

Figure 9: New data frame creation containing only id, name, and description for recommendations.



	id	name	description	info	processed_info
0	10595	96m2, 3BR, 2BA, Metro, WI-FI etc...	Athens Furnished Apartment No6 is 3-bedroom ap...	96m2, 3BR, 2BA, Metro, WI-FI etc...Athens Furn...	96m2 3br 2ba metro wi fi etc athens furnished ...
1	10988	75m2, 2-br, metro, wi-fi, cable TV	Athens Furnished Apartment No4 is 2-bedroom ap...	75m2, 2-br, metro, wi-fi, cable TVAthens Furni...	75m2 2 br metro wi fi cable tvathens furnished...
2	10990	50m2, Metro, WI-FI, cableTV, more	Athens Furnished Apartment No3 is 1-bedroom ap...	50m2, Metro, WI-FI, cableTV, moreAthens Furnis...	50m2 metro wi fi cabletv moreathens furnished ...
3	10993	Studio, metro, cable tv, wi-fi, etc	The Studio is an -excellent located -close t...	Studio, metro, cable tv, wi-fi, etcThe Studio ...	studio metro cable tv wi fi etcthe studio exce...
4	10995	47m2, close to metro,cable TV,wi-fi	AQA No2 is 1-bedroom apartment (47m2) -excell...	47m2, close to metro,cable TV,wi-fiAQA No2 is ...	47m2 close metro cable tv wi fiaqa no2 1 bedro...

Figure 10: Sample preprocessed data with concatenation.

5. CONCLUSION

In our analysis of Airbnb data, several valuable insights have been uncovered in short-term accommodations. One of the key findings was the prevalence of "Entire home/apartment" as the most common room type among Airbnb listings. This suggests that many hosts offer entire properties for rent, catering to guests looking for private and self-contained accommodations. It also delved into the fluctuation of prices over a three-month period, spanning February, March, and April. This analysis revealed that the mean prices of Airbnb listings vary throughout the year. Such insights can be incredibly useful for travellers planning their trips, helping them choose the most budget-friendly times to book their stays. This analysis identified the top five neighbourhoods with the highest number of reviews, shedding light on the most popular areas among Airbnb guests. Additionally, the neighborhood with the most real estate listings, providing information about areas with a high density of available properties also highlighted. These findings are essential for both travellers and property owners, as they offer valuable location-based insights. This research analysis also examined the number of entries per neighborhood and per month, offering a detailed breakdown of listing activity. This information can be beneficial for hosts looking to optimize their pricing and availability strategies based on seasonal demand trends. Furthermore, it also explored the most common room types in each neighborhood, helping travellers understand the diversity of accommodations in different areas. For travellers looking for luxury or unique stays, our system identified the most expensive room type based on mean prices. This insight can guide travellers in making informed choices that align with their budget and preferences. To provide a visual representation of Airbnb listings, the proposed system used interactive maps that displayed property locations along with transit information, enhancing the travel planning experience.

REFERENCES

[1] https://www.pwc.com/us/en/technology/publications/assets/pwc-consumer-intelligence-series-the-sharing-economy.pdf, 2015.

[2] G. Zhang, R. Cui, M. Cheng, Q. Zhang, and Z. Li, “A comparison of key attributes between peer-to-peer accommodations and hotels using online reviews,” Current Issues in Tourism, vol. 23, no. 5, pp. 530–537, 2019.

[3] L. Zhu, M. Cheng, and A. Wong, “Determinants of peer-to-peer rental rating scores: the case of Airbnb,” International Journal of Contemporary Hospitality Management, vol. 31, no. 9, 2019.

[4] S. Moro, P. Rita, J. Esmerado, and C. Oliveira, “Unfolding the drivers for sentiments generated by Airbnb Experiences,” International Journal of Culture, Tourism and Hospitality Research, vol. 13, no. 4, 2019.

[5] M. Chattopadhyay and S. K. Mitra, “Do airbnb host listing attributes influence room pricing homogenously?” International Journal of Hospitality Management, vol. 81, pp. 54–64, 2019.

[6] S. Rosengren, “Experience value as a function of hedonic and utilitarian dominant services,” International Journal of Contemporary Hospitality Management, vol. 28, no. 1, 2014.

[7] G. Cetin and A. Walls, “Understanding the customer experiences from the perspective of guests and hotel managers: empirical findings from luxury hotels in istanbul, Turkey,” in Proceedings of the 17th Annual Graduate Student Research Conference in Hospitality and Tourism, Washington, DC, 2012.

[8] D. E. Boyd, T. B. Clarke, and R. E. Spekman, “The emergence and impact of consumer brand empowerment in online social networks: a proposed ontology,” Journal of Brand Management, vol. 21, no. 6, pp. 516–531, 2014.

[9] E. Martin-Fuentes, C. Mateu, and C. Fernandez, “The more the merrier? number of reviews versus score on TripAdvisor and booking.com,” International Journal of Hospitality & Tourism Administration, vol. 21, no. 1, pp. 1–14, 2018.

[10] J. Mellinas, “Average scores integration in official star rating scheme,” Journal of Hospitality and Tourism Technology, vol. 10, no. 3, 2019.

[11] T. Radojevic, N. Stanisic, and N. Stanic, “Inside the rating scores: a multilevel analysis of the factors influencing customer satisfaction in the hotel industry,” Cornell Hospitality Quarterly, vol. 58, no. 2, pp. 134–164, 2017.

[12] http://insideairbnb.com/get-the-data.html, 2020.

[13] G. Santos, V. F. S. Mota, F. Benevenuto, and T. H. Silva, “Neutrality may matter: sentiment analysis in reviews of Airbnb, booking, and Couchsurfing in Brazil and USA,” Social Network Analysis and Mining, Issue, vol. 1, 2020.

[14] J. Hamari, M. Sjöklint, and A. Ukkonen, “The sharing economy: why people participate in collaborative consumption,” Journal of the Association for Information Science and Technology, vol. 67, no. 9, 2015.

[15] Dianne Dredge and S. Gyimóthy, “The collaborative economy and tourism: critical perspectives, questionable claims and silenced voices,” Tourism Recreation Research, vol. 40, no. 3, 2015.

[16] R. Botsman and R. Rogers, Beyond Zipcar: Collaborative Consumption, Harvard Business Publishing, Boston, MA, USA, 2010.

[17] J. Batle and B. Joan, “Are locals ready to cross a new frontier in tourism? factors of experiential P2P orientation in tourism,” Current Issues in Tourism, vol. 23, no. 10, 2020.

[18] B. Hasan, K. Berezina, and C. Cobanoglu, “Comparing customer perceptions of hotel and peer-to-peer accommodation advantages and disadvantages,” International Journal of Contemporary Hospitality Management, vol. 30, no. 2, 2018.

[19] G. Zervas, D. Proserpio, and J. W. Byers, “The rise of the sharing economy: estimating the impact of airbnb on the hotel industry,” Journal of Marketing Research, vol. 54, no. 5, pp. 687–705, 2017.

[20] https://www.Airbnb.com.

[21] J. Bakos, “Reducing buyer search costs: implications for electronic marketplaces,” Management Science, vol. 43, no. 12, 1997.

[22] New York City hotel rooms are getting cheaper thanks to Airbnb, Quartz, New York, NY, USA, 2015.

[23] D. Guttentag, “Airbnb: disruptive innovation and the rise of an informal tourism accommodation sector,” Current Issues in Tourism, vol. 18, no. 12, pp. 1192–1217, 2013.