CSE4077- Recommender Systems

J Component – Project Report

Airbnb Recommender System

By

19MIA1047	Mohammed Imran Z
19MIA1061	Aravindan TR
19MIA1082	Alagarsamy N
19MIA1062	Abinandhan Kumar T S S

M.Tech Computer Science Engineering with Specialization in Business Analytics

Submitted to

Dr.A.Bhuvaneswari,
Assistant Professor Senior,
SCOPE, VIT, Chennai

School of Computer Science and Engineering

VIT®
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

November 2022

BONAFIDE CERTIFICATE

Certified that this project report entitled “Airbnb Recommender System” is a bonafide work of Mohammed Imran 19MIA1047, TR Aravindan 19MIA1061, Alagarsamy N 19MIA1082, Abhinandan Kumar T S S 19MIA1062 who carried out the J-component under my supervision and guidance. The contents of this Project work, in fullor in parts, have neither been taken from any other source nor have been submitted to any other Institute or University for award of any degree or diploma and the same is certified

Dr.A.Bhuvaneswari,

Assistant Professor Senior,

SCOPE, VIT, Chennai

ABSTRACT

Airbnb is an online marketplace and hospitality service where people can rent short-term lodgingsuch as apartments, hostel beds, hotel rooms and cottages. People can also organize or participatein holiday activities and experiences such as walking tours, concerts, workshops and restaurant dining. There are more than 4 million accommodation listings on Airbnb in 191 countries and 65000 cities, with over 260 million check-ins facilitated

Airbnb can be accessed via its website or mobile apps. Accommodation listings are generated when users search by destination and use filters such as Dates, No. of Guests, Home Type, Priceand Trip Type. Airbnb is popular among travelers, especially those who are budget-conscious, because of its various advantages. For example, travelers have the option to stay in an entire apartment which offers greater flexibility compared to a hotel room. They are also able to have amore authentic travel experience by staying in a local’s home, and prices are generally lower thanhotels .

One downside of relying on Airbnb for travel planning, however, is that listings can be fully booked very quickly, especially those in desirable locations and during peak travel periods. In addition, travelers would have to look through reviews of listings carefully to ensure safety andsecurity as well as to be better informed about the amenities provided by the host.

This project aims to build a recommendation system based on the polarity scores of the user review using sentiment analysis. It makes use of datasets provided by Inside Airbnb, which are sourced from publicly available information from the Airbnb site. These datasets include detailedinformation on listings and reviews by Airbnb users, for a number of cities and countries. The project focuses on accommodation listings in London.

ACKNOWLEDGEMENT

We wish to express our sincere thanks and deep sense of gratitude to our project guide, **Dr. A. Bhuvaneswari Assistant** Professor, School of Computer Science Engineering, for his consistent encouragement and valuable guidance offered to us in a pleasant manner throughout the course of the project work.

We express our thanks to our HOD **Dr Sivabalakrishnan** for his support throughout the course of this project.

We also take this opportunity to thank all the faculty of the School for their support and their wisdom imparted to us throughout the course.

We thank our parents, family, and friends for bearing with us throughout the course of our project and for the opportunity they provided us in undergoing this course in such a prestigious institution.

TR Aravindan 19MIA1061

Mohammed Imran 19MIA1047

Alagarsamy N 19MIA1082

Abinandan Kumar T S S 19MIA1062

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computing Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai -
600 127
FALL SEM 22-23

Worklet details

Programme	Computer Science and engineering with specialization in business analytics	
Course Name / Code	Recommender Systems	
Slot	E1+TE1	
Faculty Name	BHUVANESWARI	
Component	J – Component	
J Component Title	Airbnb recommendation system	
Team Members Name | Reg. No	ARAVINDAN T R	19MIA1061
	MOHAMMED IMRAN Z	19MIA1047
	ALAGARSAMY N	19MIA1082
	ABINANDHAN KUMAR T S S	19MIA1062

Team Members(s) Contributions :

Worklet Tasks	*Contributor’s Names*
Preprocessing	Mohammed Imran, Abinandhan
Clustering and Recommender system	TR Aravindan, Mohammed Imran
NLP	Alagarsamy
Technical Report writing	TR Aravindan, Alagarsamy
Presentation preparation	TR Aravindan, Abinandhan

TABLE OF CONTENTS

INTRODUCTION

Travel industries are having important reflection of the economy from past few decades, and Airbnb housing price ranges are of great interest for both Hosts and Traveller. Airbnb is a $75 Billion online marketplace for renting out homes/villas/ private rooms. The website charges a commission (3 to 20 percent) for every booking. Even though the prospects are sound, but there are critics who argue that this has driven up rent, and caused damage to the local communities living in the vicinity.

With several use cases that cover a wide range of characteristics of Airbnb listings, we are evaluating various components in this project. It aids in not only comprehending the significant connections between features but also enables us to do independent study and present our own conclusions. With the help of incredible Python libraries like Plotly and Word Cloud, this project tries to improve our visual comprehension of the data. Also the aim of the project is to build a recommender system using Clustering and Natural Language Processing.

Literature Review

1. Airbnb Price Prediction Using Machine Learning and Sentiment Analysis, by Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, Hoormazd Rezaei – The Author reflected the need to develop a reliable price prediction model using deep learning, machine learning and natural language techniques to aid both the property owners and the customers with price evaluation given minimal available information about the property.

2. A Hotel Recommendation System Based on Reviews - by Koji Takuma, Junya Yamamoto, Sayaka Kamei, Satoshi Fujita – This paper is focused on evaluation values (ratings) given by the contributors whose references are similar to user's preference. The Authors proposed a method to extract the user preferences from a collection of reviews and perform analytics.

3. A theme of extraction transformation and loading was imminent in most of the papers that we surveyed which brings us to the conclusion that the basics of data analytics were utilized to interpret important results which are crucial for decision making.

4. A major drawback which was discovered was the ever so slight change in accuracy scores over the course of our survey. The models are volatile with either high bias and low variance or vice-versa.

S.No	Title	Author / Journal name / Year	Technique	Result
1	Airbnb Price Prediction Using Machine Learning and Sentiment Analysis	Pouya Rezazadeh Kalehbasti, Liubov Nikolenko, Hoormazd Rezaei	K-means Clustering, Support Vector Regression, Neural Network, Gradient Boosting.	Support Vector Regression (SVR) performed the best and produced an R2 score of 69% and a MSE of 0.147 (defined on ln(price)) on the test set.
2.	Airbnb Price Prediction using sentimental analysis	Peilu Liu	Linear Regression, Grdient boost, Neural network, SVR.	Among the models tested, Support Vector Regression (SVR) performed the best and produced an R2 score of 69% and a MSE

				of 0.147 (defined on ln(price)) on the test set.
3.	Applying Deep Learning To Airbnb Search.	Malay Haldar & Moose Abdool	CNN, LTSM, DNN	Overall, we found this led to an increase in the diversity of our search results, along with a +0.4% global booking gain in an online A/B test.
4.	Realtime personalization using embeddings for search ranking at5. airbnb	Kamelia aryafar, Devin Guillory, and Liangjie Hong	Embedded models	From the results of SLR, we analyze that various MBSE activities are simultaneously researched to provide a complete development solution for embedded systems.
5.	Self-Supervised learning on graph	Kadhar Moidheen	GCN	The model which they used got the accuracy of 81.32%, which is cora, then 71.43% which is citeseer, then they got 71.28% for pumped.
6.	A Hotel Recommendation System Based on Reviews, 2016 Fourth International Symposium on Computing and Networking (CANDAR)	Koji Takuma, Junya Yamamoto, Sayaka Kamei, Satoshi Fujita	method to extract the preference of review contributors from a collection of reviews.	result of questionnaire based evaluations indicates that our proposed method can recommend hotels that matches the user preference.
7.	Integrating contextual sentiment analysis in collaborative recommender systems	PLOS ONE 2021, Nurul Aida Osman, Shahrul Azman	a sentiment based model with contextual information for	Results showed that the proposed contextual information

		Mohd Noah, Mohammad Darwich, Masnizah Mohd	recommender system was proposed.	sentiment-based model illustrates better performance as compared to the traditional collaborative filtering approach.
8.	Reviews Sentiment analysis for collaborative recommender system	Alia Karim Ahmed Bahaa, 1st International Conference on Engineering and Computing, 2017 (ICEC2017) 2017	Sentiment analysis system implemented using NLP techniques with machine learning to predict user rating form his review	Sentiment analysis success in predicting user satisfaction or dissatisfaction by classifying reviews into either positive or negative. This approach could compensate the deficiency in user rating about an item in recommender system (data sparsity).
9.	Collaborative Filtering Recommender System: Overview and Challenges	Hael Al-bashiri, Mansoor Abdullateef Abdulgabber, Awanis Romli, Fadhl Hujainah, Journal of Computational and Theoretical Nanoscience 2017	defined the main challenges which have clearly impact on the performance and accuracy of CF recommender system.	This paper summarizes the limitations of the existing methods and recommendations.

Dataset and Tools

We will be using the **Boston Airbnb open data** dataset from Kaggle

Listings.csv: It contains full descriptions and average review score

Reviews.csv: It contains unique id for each reviewer and detailed comments

Calendar.csv: It contains listing id and the price and availability for that day

Tools: Python, Google Colab

In the datasets we are provided with 26 columns (Features) of data, We have chosen only the following important attributes

- listing_id : Unique for each Airbnb listings
- name : Name of the Airbnb
- neighbourhood : Name of the Neighbourhood
- city : Name of the City
- property_type : Property Type of Airbnb
- room_type : Airbnb Room Type
- amenities : Name of the Amenities provided by Airbnb
- price : Airbnb Price per night
- monthly_price : Monthly Price of the Airbnb
- comments : Traveller Comments/Reviews

Proposed Methodology

Our project is a full functioning aggregation of recommender systems algorithms like collaborative filtering, content-based filtering along with sentiment analysis.

We have split the project into 4 distinct modules

- Data cleaning
- Natural Language Processing
- Clustering
- Recommender System

- **Data Cleaning**

  In general, the following steps below describe the major data cleaning and preprocessing performed before conducting analysis. For more detailed steps, please refer to the 'Data Cleaning' and 'Preprocessing, Data Visualization, Clustering' notebooks found in this repository.

  Handling Missing Values: There were many missing values discovered in dataset. For example, host_response_time contained over 2,100 rows of missing entries. Since this was a categorical ordinal column, these missing values were imputed with an 'N/A' value to represent Airbnb hosts who have not responded back to hostees. Other numerical missing values such as security_deposits were imputed with the value 0 (assuming that a security deposit was not needed for the listings).

  Encoding Categorical Features and Values: Categorical features were split into ordinal and nominal features. Ordinal features (columns where the values have a structured order) consisted of host_response_time and cancellation_policy and were encoded using an OrdinalEncoder. Nominal features (columns where values have no order of precedence) consisted of all other categorical features (ie. property_type) and were one hot encoded.

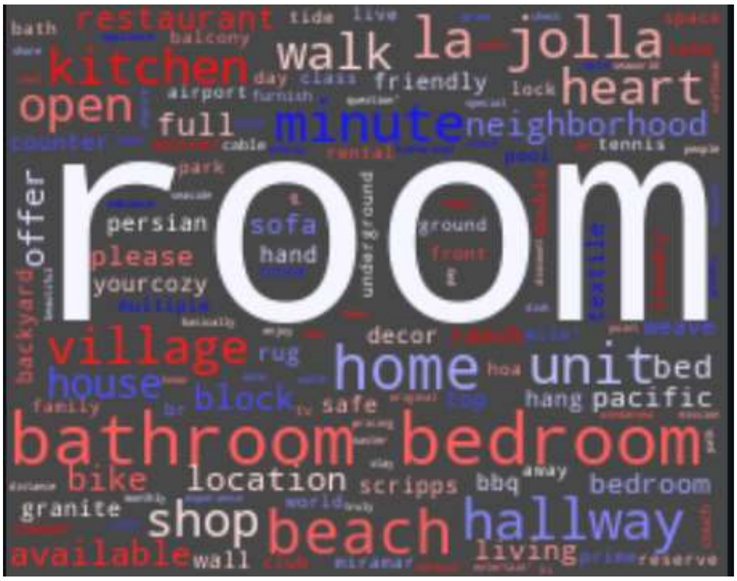
  Standard Scaling: All other numerical columns consisting of integer and float values were subsequently scaled using a StandardScaler.

  In total, the preprocessed dataset consisted of 13,039 listings and 240 features.

- **Natural Language Processing**

  Wordcloud visualizations were constructed for each text column in the Airbnb Listings dataset. The text columns were preprocessed and normalized as follows:

Missing values are imputed with 'blank' text
Text in each column are tokenized. Tokenization is a process by which the text is broken down into smaller units and subwords. Stopwords and other text without much value were also removed during this step.



Findings

- Most listings tend to be described as scenic and picturesque by the beach (or some variation of paradise).
- A lot of listings are ironically "hidden".
- Must have wifi, tv, parking, and large beds!
- The hosts must have a lot of spare time to rent out Airbnbs as their side jobs are also involved in entertainment.

- A sentiment analysis was also performed and to gather further insights about how Airbnb listings are generally described. The process of understanding sentiment scores is described as follows:

  TextBlob Module: Allows for the ability to place a score on sentiment of words based on where it is in a sentence.

  Sentiment Labels: Each word in a corpus is labeled in terms of polarity and subjectivity.

  Polarity: How positive or negative a word is; -1 is most negative, +1 is most positive

  Subjectivity: How subjective, or opinionated a word is; 0 is fact, +1 is an opinion

- **Clustering**

A Uniform Manifold Approximation and Project (UMAP) dimensionality reduction technique was leveraged to create a clustering visualization of all data points following preprocessing.

Clustering labels were constructed using a MiniBatch KMeans iterating through the preprocessed dataset to determine optimum cluster size. A total of 5 unique cluster groups were generated with labels assigned to each individual listing.

Below is a snippet image of San Diego Airbnb Listings Embedding via UMAP whereby users can hover over each individual data point to obtain a better understanding of the features associated within each cluster. Visualization is generated using Bokeh.

- **Recommender Engine**

In order to construct a recommendation for an individual Airbnb listing, the mathematical concept of cosine similarity was leveraged.

In short, cosine similarity measures the similarity between two vector points in a defined space using the cosine angle between these two vectors. For two items, cosine similarity measures how far apart (or similar) each item are away from each other.

In order to calculate cosine similarity, the preprocessed dataset and user selected listing need to be converted to 2-dimensional arrays. In this context, these are individual arrays of Number of Rows x Number of Columns in each respective dataset.

Cosine similarity values for all listings consisting of all features in the dataset are calculated and sorted by top 5 most similar listings to generate recommendations.

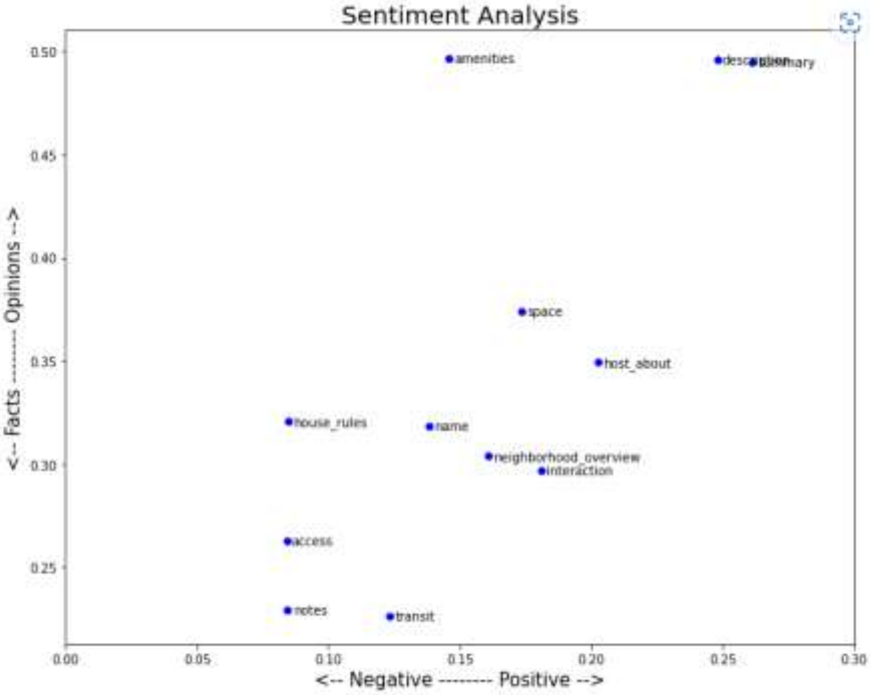
Algorithms Used

- **Kmeans Clustering:**
  k-means clustering tries to group similar kinds of items in form of clusters. It finds the similarity between the items and groups them into the clusters. K-means clustering algorithm works in three steps. In our project Kmeans is used to assign cluster labels to Airbnb listings, to recommend similar Airbnb listings later.

- **Cosine Similarity:**
  Cosine similarity measures the similarity between two vectors of an inner product space. It is measured by the cosine of the angle between two vectors and determines whether two vectors are pointing in roughly the same direction. In our project cosine similarity is the similarity metric, based on which the recommender system gives similar items.

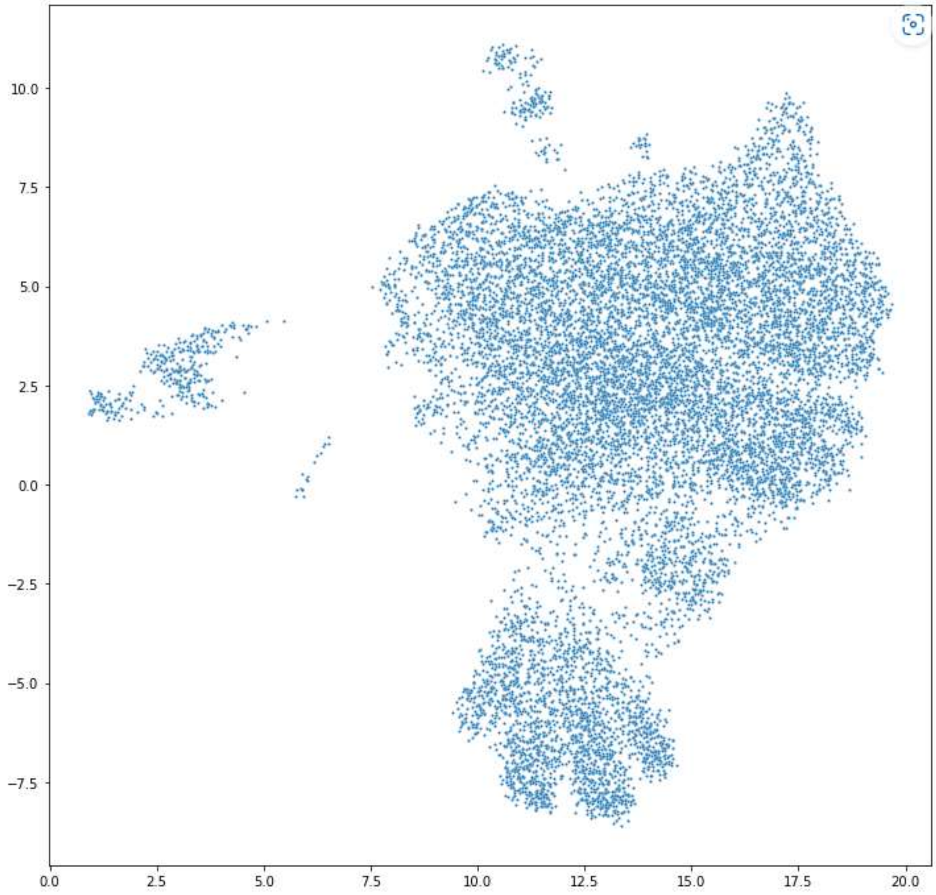
- **Natural Language processing**
  Natural language processing (NLP) algorithms support computers by simulating the human ability to understand language data, including unstructured text data. Here TextBlob library is used for sentiment Analysis for finding subjectivity and Polarity of text fetures.

Experimental Analysis

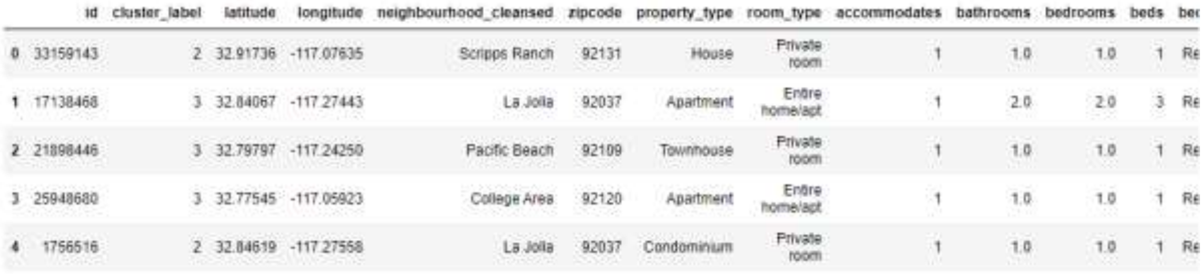


- Airbnb listings tend to be positive when it comes to descriptions and summaries.
- This makes sense, hosts want to encourage people to stay at their Airbnb and having a positive description is beneficial.
- However, these descriptions tend to be grounded in opinion
- Interesting to note that amenities are considered very opinionated.

- One would expect that amenities would be more grounded in facts.



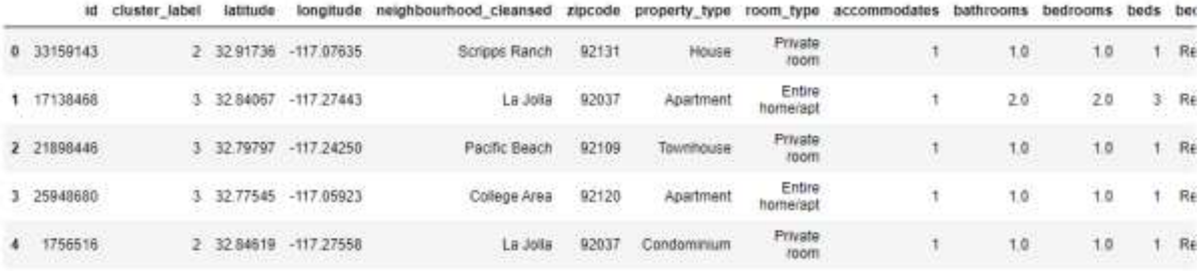
- From the embedding plot, we can see that the dataset has different clusters
- We performed clustering to find the cluster labels for each and every review
- We will use the cluster labels and the calculated features to recommend similar Airbnb listing using Cosine Similarity



	id	cluster_label	latitude	longitude	neighbourhood_cleansed	zipcode	property_type	room_type	accommodates	bathrooms	bedrooms	beds	bec
0	33159143	2	32.91736	-117.07635	Scripps Ranch	92131	House	Private room	1	1.0	1.0	1	Re
1	17138468	3	32.84067	-117.27443	La Jolla	92037	Apartment	Entire home/apt	1	2.0	2.0	3	Re
2	21898446	3	32.79797	-117.24250	Pacific Beach	92109	Townhouse	Private room	1	1.0	1.0	1	Re
3	25948680	3	32.77545	-117.05923	College Area	92120	Apartment	Entire home/apt	1	1.0	1.0	1	Re
4	1756516	2	32.84619	-117.27558	La Jolla	92037	Condominium	Private room	1	1.0	1.0	1	Re

Discussion on Results

From the cluster plot , we can see that the dataset has different clusters
• We performed clustering to find the cluster labels for each and every review
• We then used the cluster labels and the calculated features to recommend similar Airbnb listing using Cosine Similarity



	id	cluster_label	latitude	longitude	neighbourhood_cleansed	zipcode	property_type	room_type	accommodates	bathrooms	bedrooms	beds	be
0	33159143	2	32.91736	-117.07635	Scripps Ranch	92131	House	Private room	1	1.0	1.0	1	Re
1	17138468	3	32.84067	-117.27443	La Jolla	92037	Apartment	Entire home/apt	1	2.0	2.0	3	Re
2	21898446	3	32.79797	-117.24250	Pacific Beach	92109	Townhouse	Private room	1	1.0	1.0	1	Re
3	25948680	3	32.77545	-117.05923	College Area	92120	Apartment	Entire home/apt	1	1.0	1.0	1	Re
4	1756516	2	32.84619	-117.27558	La Jolla	92037	Condominium	Private room	1	1.0	1.0	1	Re

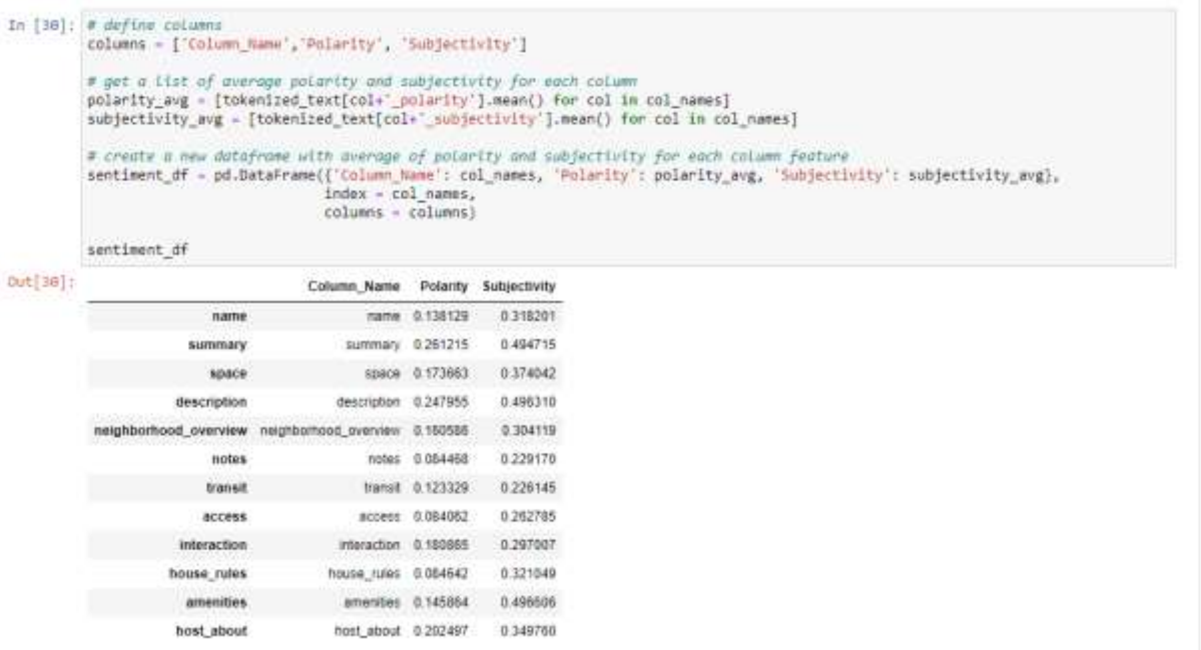
Conclusion and Future Works

In our project, We have implemented various imputations and pre-processing techniques for to clean the data for extracting meaningful insights from the data. We have done Clustering which is a un-supervised machine learning to assign cluster labels to Airbnb listings. Also we have used Natural Language Processing for sentiment analysis of user reviews. Finally we created a Utility matrix containing all the pre-processed features and we have implemented Recommendation system to recommend Top 6 similar Airbnb Listings for any Airbnb Listing.

For improved performance of recommendation system, we can implement a hybrid recommendation system which is a combination of both content and collaborative based in the future.

Screenshots

Sentiment Analysis

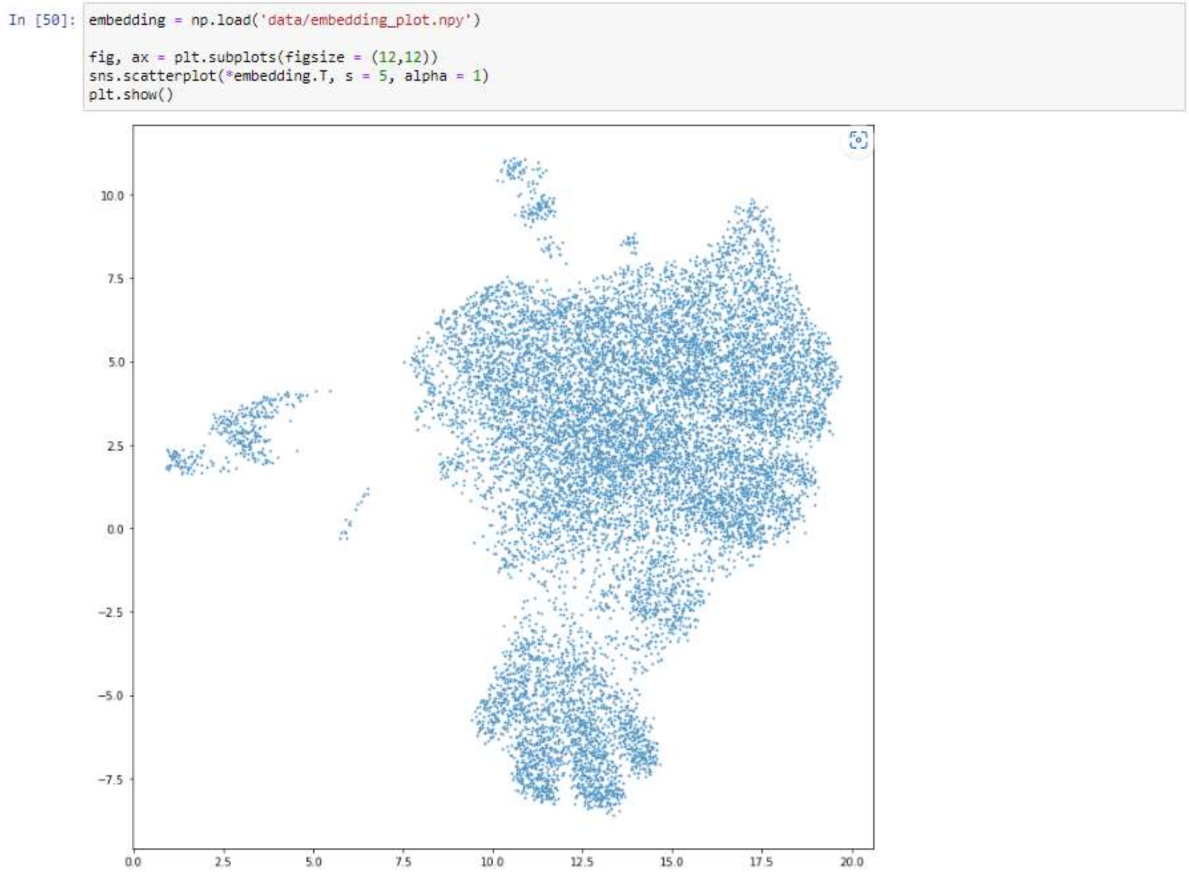


In [30]:
```
# define columns
columns = ['Column_Name','Polarity', 'Subjectivity']

# get a list of average polarity and subjectivity for each column
polarity_avg = [tokenized_text[col+'_polarity'].mean() for col in col_names]
subjectivity_avg = [tokenized_text[col+'_subjectivity'].mean() for col in col_names]

# create a new dataframe with average of polarity and subjectivity for each column feature
sentiment_df = pd.DataFrame({'Column_Name': col_names, 'Polarity': polarity_avg, 'Subjectivity': subjectivity_avg},
                            index = col_names,
                            columns = columns)

sentiment_df
```

Out[30]:

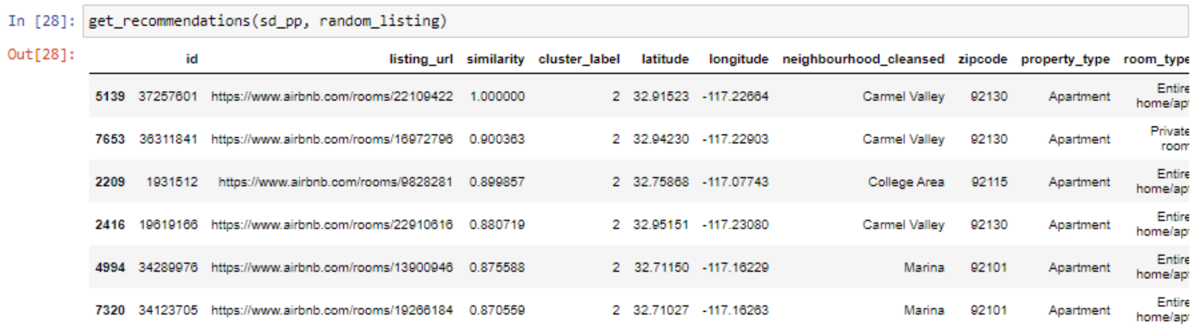
	Column_Name	Polarity	Subjectivity
name	name	0.138129	0.318201
summary	summary	0.261215	0.494715
space	space	0.173663	0.374042
description	description	0.247955	0.496310
neighborhood_overview	neighborhood_overview	0.160586	0.304119
notes	notes	0.084468	0.229170
transit	transit	0.123329	0.226145
access	access	0.084062	0.262785
interaction	interaction	0.180865	0.297007
house_rules	house_rules	0.084642	0.321049
amenities	amenities	0.145864	0.496606
host_about	host_about	0.202497	0.349760

Clustering



In [50]:
```
embedding = np.load('data/embedding_plot.npy')

fig, ax = plt.subplots(figsize = (12,12))
sns.scatterplot(*embedding.T, s = 5, alpha = 1)
plt.show()
```

Recommendations



```
In [28]: get_recommendations(sd_pp, random_listing)
Out[28]:
```

	id	listing_url	similarity	cluster_label	latitude	longitude	neighbourhood_cleansed	zipcode	property_type	room_type
5139	37257601	https://www.airbnb.com/rooms/22109422	1.000000	2	32.91523	-117.22664	Carmel Valley	92130	Apartment	Entire home/ap
7653	36311841	https://www.airbnb.com/rooms/16972796	0.900363	2	32.94230	-117.22903	Carmel Valley	92130	Apartment	Private room
2209	1931512	https://www.airbnb.com/rooms/9828281	0.899857	2	32.75868	-117.07743	College Area	92115	Apartment	Entire home/ap
2416	19619166	https://www.airbnb.com/rooms/22910616	0.880719	2	32.95151	-117.23080	Carmel Valley	92130	Apartment	Entire home/ap
4994	34289976	https://www.airbnb.com/rooms/13900946	0.875588	2	32.71150	-117.16229	Marina	92101	Apartment	Entire home/ap
7320	34123705	https://www.airbnb.com/rooms/19266184	0.870559	2	32.71027	-117.16263	Marina	92101	Apartment	Entire home/ap

Github Repository:
https://github.com/aravindsriraj/Airbnb-Recommendations-system.git

References

1. Lovedeep Singh (2020). Fake News Detection: a comparison between available Deep Learning techniques in vector space, 4th Conference on Information & Communication Technology (CICT), Chennai, India, 2020, pp. 1-4, doi: 10.1109/CICT51604.2020.9312099.

2. Al-Ghuribi S. M. & Mohd Noah S. A., "Multi-criteria review-based recommendersystem—the state of the art", in IEEE Access, 7, pp. 169446–169468, 2019.

3. Osman N. A. & Mohd Noah S. A., "Sentiment-based model for recommender systems" in Proceedings of the Fourth International Conference on Information Retrieval and Knowledge Management (CAMP), 2018.

4. Ghabayen A. S., Mohd Noah S. A., "Using tags for measuring the semantic similarity ofusers to enhance collaborative filtering recommender systems," in International Journal on Advanced Science, Engineering and Information Technology, vol., no. 6, 2063–2070, 2017.

5. Darwich M., Mohd Noah S. A. and Omar N., "Automatically generating a sentiment lexiconfor the Malay language," in Asia-Pacific Journal of Information Technology and Multimedia, vol. 5., no. 1, 2016.

6. Darwich M., Noah S. A. M., and Omar N., "Minimally-Supervised Sentiment Lexicon Induction Model: A Case Study of Malay Sentiment Analysis," Multi-disciplinary Trends inArtificial Intelligence—11th International Workshop, MIWAI 2017, Proceedings. Springer Verlag, Vol. 10607 LNAI, pp. 225–237, 2017.