An Integrated Approach to Immigrant Housing Search using KMeans Clustering and NLP

Prof. Ranjeetsingh Suryawanshi
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
ranjeetsingh.suryawanshi@vit.edu

Tanish Modase
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
tanish.modase20@vit.edu

Gautam Mudawadkar
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
gautam.mudawadkar20@vit.edu

Prathmesh Nagpure
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
prathmesh.nagpure20@vit.edu

Komalesh Patil
Department of Computer Engineering
Vishwakarma Institute of Technology
Pune, India
komalesh.patil20@vit.edu

***Abstract*: The increasing global migration trends have highlighted the critical need for efficient and personalized immigrant housing solutions. This report presents an innovative approach, combining K Means clustering and Natural Language Processing (NLP), to bridge the gap between immigrants' preferences and available accommodations. The primary goal of this project is to streamline the decision-making process by offering personalized recommendations and providing insightful data to both immigrants and accommodation providers.**

Keywords: K Means Clustering, Natural language processing, accommodations

I. Introduction

In today's fast-paced lifestyle, individuals often seek convenient dining options. Whether opting for home-cooked meals or dining out, food is a significant aspect of one's daily routine. When relocating to a new area, aligning with preferred dining establishments can enhance convenience and satisfaction for individuals. This project explores the concept of guiding students to live near their favored food outlets, benefiting both immigrants and food providers. Moreover, this information can aid restaurant and hotel managers in making strategic location decisions. By analyzing demographics and preferences, managers can select prime locations with a high concentration of their target audience.

The aim of this project is to address the gap between immigrants' preferences and available accommodations, streamline decision-making through personalized recommendations, and empower both immigrants and food providers with insightful data for informed choices. By integrating data scraping from Google for amenities and Natural Language Processing for sentiment analysis, the project seeks to offer a holistic solution that optimizes the accommodation search process, enriches the overall immigrant experience, and assists businesses in strategic location decisions.

II. Literature review

An Investigation into Geographical Data to Provide Direction for Accommodation and Food Resources (Turkish Journal of Computer and Mathematics Education 2022):

Proposed by Kolla Alekhya, Mallavarapu Harika, et al, this study investigates how migrants' preferences for facilities, cost, location, and other factors can be taken into account when classifying available housing options using K-Means Clustering. It uses the geolocational data to segregate groups with similar traits and assign them into clusters. It uses maps to provide visual representation of the locations.[1]

A systematic review and research perspective on recommender systems (Springer 2022):

This paper presents a thorough and insightful review of recommender systems, addressing their increasing importance in filtering online information amid evolving user habits and the widespread accessibility of the internet.he systematic review delves into diverse applications such as books, movies, and products, offering a nuanced algorithmic analysis and establishing a taxonomy that comprehensively covers the components essential for developing effective recommender systems. The inclusion of details on datasets, simulation platforms, and performance metrics across various contributions enhances the paper's value.[2]

Geolocation Analysis using Machine learning (IJERCSE 2022):
The project addresses a significant challenge faced by students and young adults relocating for education or work—finding suitable accommodation. The objective of creating a system that classifies users based on preferences, such as budget and proximity to daily necessities, is commendable. The expansion of the system's applicability to various purposes, such as identifying optimal locations for businesses or agriculture, adds versatility. The literature survey's focus on machine learning in geostatistics and spatial applications provides a solid foundation.[3]

An Effective Hotel Recommendation System through Processing Heterogeneous Data (MDPI 2021):
The authors introduce a novel framework for ranking hotels by incorporating consumer reviews and nearby amenities with a focus on hotel selection. Using information from well-known online booking platforms, the framework incorporates an algorithm to assess review keywords and generates numerical scores. The strategy combines these review scores with evaluations of nearby facilities in various categories. The study runs experiments using trip organizer websites and Booking datasets to demonstrate the effectiveness of the suggested system in producing relevant and personalized hotel recommendations. The framework contributes to improving user choices in the context of hotel bookings by utilizing textual reviews, Google Place API data, and a unique RSG algorithm, setting itself apart from current recommendation systems.[4]

NLP-Based Movie Recommendation System (IEEE 2020):
In this paper, Nimish Kapoor, Saurav Vishal, et al. propose a sentiment analysis system based on the TMDB dataset to predict positive or negative sentiment from user reviews. After that, the sentiment analysis and review score are used to rate the films. Sentiment analysis is used on reviews and comments to help improve movie ratings and, ultimately, provide more recommendations. It detects positive reviews by using text mining to calculate polarity scores.[5]

A Multi-Element Hybrid Location Recommendation Algorithm for Location Based Social Networks (IEEE Access 2019):
The paper discusses the difficulty of personalized POI recommendation in location-based social networks (LBSNs). The proposed GFP-LORE system integrates social, popularity, geographic, and sequential influences to provide a novel hybrid recommendation algorithm. The study effectively directs new POI recommendations by modeling user social correlations and point of interest popularity as power-law distributions. The system determines the likelihood of user arrival at new locations using Kernel Density Estimation (KDE) based on individual check-in behaviors. The GFP-LORE algorithm outperforms competing approaches by fusing these influences into a single framework, resulting in increased recommendation accuracy. The system's thorough methodology highlights its potential to improve personalized POI recommendations within LBSNs.[6]

Using Location-Based Context to Enhance Point-of-Interest Recommendations in New Areas (IEEE Access):
The paper discusses the difficulty of offering users visiting new areas personalized POI recommendations. Using the integration of location data from the Internet-of-things (IoT) with social media network data,the study introduces the Latent Factor model-based New Place Recommendation Algorithm (N-PRA).. This algorithm takes into account a variety of context features, including changes in user interest, POI popularity, geographic factors, and social network connections. Through testing on actual Yelp urban data, N-PRA is shown to outperform baseline algorithms, demonstrating its effectiveness in providing precise personalized recommendations. In the changing environment of contemporary sensor networks and wireless internet, the suggested method has application potential in a number of scenarios.[7]

A Literature review on spatial location analysis for retail site selection

The paper effectively underscores the strategic importance of optimal site selection for businesses and retailers, offering a nuanced analysis of various approaches ranging from multi-criteria decision-making models to GIS-based and deductive/inductive reasoning methods. The delineation of factors influencing site selection, encompassing locational and demographical attributes, reflects a comprehensive understanding of the intricacies involved. Notably, the integration of social media as a potential factor in decision-making, while still underexplored, adds a contemporary dimension to the discussion. The paper's broad search strategy, encompassing diverse electronic databases, contributes to the richness and reliability of the literature review..[8]

Reviews on Machine learning for spatial analyses in urban areas

The paper acknowledges certain limitations, including its exclusive focus on spatial urban data and its scoping review methodology, which only provides an initial mapping of this evolving field. However, the paper offers valuable insights to the scientific community, serving as a guide for understanding the applicability of various approaches and datasets for specific urban challenges. It identifies promising areas for future research, emphasizing the need for more comparative studies, a better understanding of the impact of data and algorithm choices, and the promotion of explainable ML methods for urban applications. Ultimately, the paper highlights the importance of knowledge transfer to equip cities with the tools needed to address their complex challenges.[9]

ZenDen - A Personalized House Searching Application

Paper highlights the limitations of current real estate mobile applications, which heavily rely on filter-based searches and complex map interfaces, often neglecting the specific needs of the student and lower-income demographics. The paper introduces a novel mobile application that promises to transform the rental market, especially for university students, by employing a user-friendly swipe-based interface and an efficient user-house recommender system. Notably, the application leverages deep learning techniques, particularly convolutional neural networks (CNNs), for image analysis, which underpins its recommendation system. The ultimate aim of this innovative approach is to offer a personalized, user-friendly solution that significantly reduces the time and effort required for individuals in the rental market to secure suitable housing options.[10]

Developing Session-based Personalized Accommodation Recommender System by Using LSTM

The tourism sector, greatly influenced by Internet technology, has seen a transformation. Travelers now have the capability to independently search for information and select destinations, prompting the demand for personalized recommender systems. The development of such systems is intricate and involves analyzing demographic data, user interactions, clicks, and hotel features to provide tailored hotel recommendations. Given that user interactions and hotel history constitute time series data, Long Short-Term Memory (LSTM) models prove to be an ideal choice for hotel recommendations. In this study, we introduce a session-based accommodation recommender system utilizing LSTM, which has yielded promising results. The research spans areas like time series analysis, signal processing, data modeling, Internet technology, and the travel and hotel industry.[11]

Identifying Real Estate Opportunities Using Machine Learning

The research paper addresses the challenges of the real estate market, where fluctuating prices can be influenced by various uncontrollable factors. The inconsistency in online listings can further complicate the situation, as outdated or deliberately mispriced properties are common. To tackle this, the study focuses on developing a real-time machine learning application that detects properties listed significantly below their market value. Specifically, the application was designed for the Salamanca district in Madrid, Spain, utilizing data from a prominent Spanish online real estate platform. Implementing a regression-based approach, the study undertook comprehensive feature engineering to enhance predictive accuracy. Various machine learning algorithms, such as regression trees, k-nearest neighbors, support vector machines, and neural networks, were evaluated for their efficacy in identifying undervalued properties, highlighting the strengths and limitations of each approach.[12]

Research on House Rental Recommendation Algorithm Based on Deep Learning

The research introduces an innovative house rental recommendation algorithm, employing a fusion of Deep Learning and a text convolutional Neural Network with a content-based recommendation approach. It aims to address the limitations of conventional recommendation systems in capturing user and housing source attributes. By proficiently categorizing housing options, the algorithm adeptly captures user preferences based on behavioral data, leading to a more refined personalized recommendation of housing sources.[13]

Habitation recommendation using machine learning

The research paper highlights the significance of renting as a prevalent habitation method. It emphasizes the challenge individuals face in finding suitable living spaces. In response, the paper proposes a novel approach utilizing user interactions, artificial intelligence, and machine learning. This method aims to comprehend user requirements, gathered via a chatbot, and subsequently employs a machine learning model. The identified recommendations are then seamlessly integrated onto the Google Maps interface, facilitating a user-friendly experience. Ultimately, the proposed methodology streamlines the process of locating an ideal living space, ensuring convenience and effectiveness for users.[14]

Recommendation system using machine learning techniques

This paper provides a comprehensive exploration of recommendation systems, with a specific focus on movies, utilizing machine learning techniques. The author effectively communicates the primary goal of prediction of user interests and the inference of their mental processes, highlighting the relevance of recommendation systems Overall, the paper contributes valuable insights into the current landscape of recommendation systems, specifically within the context of movies, and suggests promising directions for future research and development..[15]

A Recommendation Engine to Estimate Housing Values in Real Estate Property Market

The paper outlines a compelling research endeavor focused on addressing the challenges within the housing market through the development of an efficient recommendation engine. The identified issue of information overload in the existing methods is a pertinent problem, and the proposed collaborative technique aims to provide an optimal solution.The emphasis on the increasing demand and limited supply of housing properties sets the context for the research, highlighting the need for automated machine learning techniques to filter and recommend properties based on user preferences. The incorporation of logistic regression and K-nearest neighbor techniques in the implementation showcases a thoughtful approach to model selection.The K-nearest neighbor's impressive 100% prediction accuracy stands out, and the contrast with logistic regression's 54.6% accuracy rate adds depth to the evaluation[16]

A Development of Accommodation Facility Selection Recommendation System

This paper proposes a collaborative filtering based recommendation system for hotel selection using a virtual user concept. Collaborative filtering faces challenges in accommodation recommendation due to sparsity of user ratings. The authors introduce a virtual user assumed to have rated all accommodation options based on average ratings of actual users. By analyzing user rating patterns on a hotel dataset with quantification theory, they find users form a single rating cluster. Hence a single virtual user with average ratings suffices. In experiments, adding this virtual user increased recommendable user-item pairs by 2% without changing rankings of pairs recommended without it. The paper provides a novel way to address data sparsity for accommodation recommendation. Future work involves evaluating accuracy and precision gains across different datasets. In summary, this paper puts forth an effective collaborative filtering approach tailored to hotel recommendation by leveraging a virtual user.[17]

A Development of Accommodation Facility Selection Recommendation System

This paper proposes an accommodation recommendation system called ACRoSS that summarizes online reviews to help users make hotel selection decisions. It collects accommodation reviews from multiple websites, calculates average scores, and selects representative comments for each hotelemploying a weighted non-negative matrix factorization (FNMF) technique that is feature-based. Features include the following five important aspects of lodging: staff, value for money, cleanliness, comfort, and location.. Experiments on Japanese hotel reviews show the system achieved 61% precision in classifying representative sentences. User studies indicate 55.48% high satisfaction with the representative summaries. While limited to hotels and a small dataset, the paper demonstrates an effective application of FNMF based multi-document summarization for travel recommendations. Key challenges identified are normalizing hotel names across sites and improving review classification accuracy. In summary, this is a novel accommodation recommendation approach using FNMF based summarization of online reviews.[18]

A Location-based personalized recommendation systems for the tourists in India

This paper proposes a location-based personalized recommendation system for tourists visiting India using collaborative filtering techniques. It collects the opinions of local users regarding sites, food items, and products in different locations through a survey of 200 individuals. User preferences for place type, food type, and product type are captured. Cosine similarity is employed to find users similar to a given tourist based on these preferences. The system then recommends the top sites, along with highly rated food and products available there, based on the opinions of those similar local users. To provide the most recent recommendations, the authors implement a time decay function that assigns greater importance to recent user opinions. Evaluation using precision, recall, and F-measure metrics on sample data shows good recommendation accuracy. Comparisons to a model without time weighting demonstrates improved performance in suggesting latest items. While promising, limitations include the small single-city dataset and simplified representation of user preferences. Proposed future work involves expanding the dataset across India and adding travel time minimization for recommended itineraries. Overall, this paper puts forth a collaborative filtering approach customized to travel recommendations in India by leveraging local user knowledge and time-aware preferences. Preliminary results are positive but more extensive evaluations are needed.[19]

Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language

This paper provides a valuable contribution to the field of Natural Language Processing (NLP) by addressing challenges in multi-modal environments and proposing a novel approach to feature extraction using deep learning techniques. The introduction of a multi-modal neural network, with individual sub-neural networks for each mode, demonstrates a thoughtful strategy to handle structural differences in diverse data modalities. The paper's emphasis on word segmentation processing adds to its significance, especially considering the present problems with training prediction time and the long-term dependence of text semantics. For English word segmentation, combining the Conditional Random Field (CRF) model with the Bidirectional Gated Recurrent Unit (BI-GRU) offers a practical solution to increase processing speed and lessen long-distance dependency issues.[20]

Normalization of Microarray Data: Single-labeled and Dual-labeled Arrays

The reliability of high-throughput data analysis is greatly enhanced by the use of normalization techniques, which are discussed in the paper "Normalisation of microarray data: single-labeled and multiple-labeled arrays," especially when it comes to DNA microarrays. The principles of normalization, which address issues such as handling different labeling procedures and differences in experimental circumstances, have broader applicability even though the research focuses on biological systems. The review emphasizes the significance of robust normalization methods in handling large datasets, emphasizing their impact on downstream analyses and the need for careful consideration of assumptions made during the normalization process across diverse domains. This is relevant to your project, which aims to assist managers in making strategic location decisions and to guide students to live near their preferred food outlets.[21]

An Effective Data Normalization Strategy for Academic Datasets using Log Values

In many fields, data normalization is essential for guaranteeing the accuracy and readability of datasets for further examination. Conventional methods of normalization, such Z-score and Min-Max normalization, are well-known. On the other hand, the study under review presents a two-phase normalization technique, starting with a logarithmic transformation and ending with Cube Root normalization. Multi-step normalization methods are a notion that has been studied in the literature to improve the quality of data for better analytical results. In particular, the incorporation of logarithmic transformations into data normalization is acknowledged for its capacity to tackle issues related to data distribution, like skewness. The literature regularly assesses the effectiveness of normalization techniques using common performance evaluation metrics, such as accuracy, precision, recall, and F1 score. Although the study concentrates on academic datasets, a more thorough investigation of educational data normalization procedures provides insights into the difficulties and uses of these methods. The backdrop for comprehending the suggested two-phase normalization technique and its possible ramifications for data analysis and decision-making projects in a variety of fields is established by this literature review.[22]

On the Generalization of Max-Min Normalization for Dimensionality Reduction and Data Representation

The paper "On the Generalisation of Max-Min Normalisation for Dimensionality Reduction and Data Representation" offers an interesting viewpoint in the literature that is relevant to our project, which focuses on helping restaurant managers make strategic location decisions and advising students on where to live close to their favorite food outlets. Beyond its conventional use, this work investigates the wider applications of Max-Min normalization, with a focus on improving data representation and dimensionality reduction. Although dimensionality reduction may not be directly related to our research, the idea of generalizing normalization approaches is relevant to our objective of optimizing a variety of variables that contribute to the recommendation system. Through the application of Max-Min normalization to our dataset, which includes user ratings, distances, and preference scores, among other criteria, we hope to take advantage of this generalization to standardize data representation and enhance the performance of our recommendation algorithm. The goals of the project are connected to the wider normalization tactics investigated in pertinent research by means of this literature review.[23]

Feature Scaling and Selection in Anomaly Detection

The paper "Feature Scaling and Selection in Anomaly Detection" is pertinent to our project because it helps restaurant management make strategic location selections and helps students select lodging close to their favorite food outlets. The methods for choosing and scaling features inside the anomaly detection framework are examined in this study. The idea of feature scaling and selection is relevant to our objective of optimizing the recommendation system, even though anomaly detection is not our main focus. Through the analysis and application of the approaches described in this paper, we hope to improve the robustness and accuracy of our recommendation algorithm by implementing efficient feature scaling strategies to guarantee consistency across a range of parameters, including user ratings, distances, and preference scores. This relationship between feature selection and scaling in the context of the literature on anomaly detection enhances our method for customizing the recommendation system to the unique requirements of food providers and students.[24]

Min Max Normalization Based Data Perturbation Method for Privacy Protection

The focus is on data perturbation as an effective method for privacy preservation while maintaining accuracy. To safeguard sensitive data, a new technique called min-max normalization transformation-based data perturbation is presented. The outcomes of the experiment show how well the strategy works to hide sensitive information and maintain data mining techniques' functionality following data distortion.. The introduction further emphasizes the privacy challenges in data mining, discussing various approaches such as randomization, secure multiparty computations, and anonymization, with a specific focus on data perturbation as a robust technique for privacy preservation.[25]

K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data

Despite its low computational complexity and ease of use, the K-means algorithm is still the most widely used clustering algorithm in the literature.The K-means clustering algorithm and its variations are thoroughly reviewed and classified in the current work. The literature review explores the history of K-means, present trends, unresolved problems, and difficulties related to the algorithm. Cluster analysis is essential for organizing unlabeled data, and the main objective of data mining is to extract meaningful information from collected data. Data clustering issues have been effectively resolved by cluster analysis in a variety of fields, including artificial intelligence, robotics, manufacturing, medical science, and finance..[26]

III. Datasets

1. Web Scraping for Location and Amenities Data:

The data for locations and nearby amenities were collected through web scraping techniques. The process involved extracting information from online sources including Google Maps and online portals. Specific steps for data collection included:

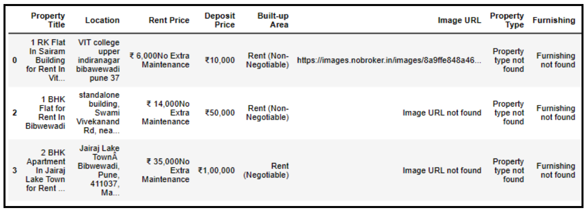
- Selection of Data Sources: Identified and selected reputable online sources known for providing accommodation information about the target locations and amenities.
- Web Scraping Methodology: Utilized web scraping techniques to extract relevant data. This involved parsing the HTML structure of the chosen rental accommodation website and retrieving details such as location coordinates, names of amenities, types of amenities, and other pertinent information.
- Data Normalization: Normalized the collected data to ensure uniformity and consistency across different data points. This step involves standardizing the format of location coordinates, amenity names, and other relevant attributes to facilitate easy analysis and comparison.
- Data Clustering: Applied K-Means Clustering, to group the collected data points based on similarities in location and types of amenities. This step aimed to identify distinct clusters of amenities in proximity to specific locations, providing valuable insights for accommodation recommendations.

2. Google Reviews Data Collection and Processing:

To obtain Google reviews for accommodations, the Google Maps API was utilized. The procedure involved the following steps:

- Google Maps API Integration: Integrated the Google Maps API to retrieve detailed information about accommodations, including user reviews, ratings, and other relevant data points.
- Data Preprocessing: Processed the collected reviews using natural language processing techniques to extract sentiments and key features. This step involved tasks such as text tokenization, sentiment analysis, and feature extraction.
- Data Normalization: Normalized the processed review data to ensure consistency in the sentiment analysis and feature extraction. This step included standardizing the format of sentiment scores, review text, and other relevant attributes.
- Data Integration: Integrated the processed review data with the previously collected location and amenities data to analyze the dataset for accommodation recommendations. This allowed for a comprehensive analysis of user sentiments, preferences, and location-based amenities, facilitating the generation of personalized recommendations for users.

By following these data collection, normalization, and processing procedures, a robust dataset was compiled, enabling the development of an individualized system of recommendations for lodging based on the tastes and emotions of the user.



	Property Title	Location	Rent Price	Deposit Price	Built-up Area	Image URL	Property Type	Furnishing
0	1 RK Flat In Sairam Building for Rent In Vit...	VIT college upper indiranagar bibawewadi pune 37	₹ 6,000No Extra Maintenance	₹10,000	Rent (Non-Negotiable)	https://images.nobroker.in/images/8a9ffe848a46...	Property type not found	Furnishing not found
2	1 BHK Flat for Rent In Bibwewadi	standalone building, Swami Vivekanand Rd, nea...	₹ 14,000No Extra Maintenance	₹50,000	Rent (Non-Negotiable)	Image URL not found	Property type not found	Furnishing not found
3	2 BHK Apartment In Jairaj Lake Town for Rent ...	Jairaj Lake TownÂ Bibwewadi, Pune, 411037, Ma...	₹ 35,000No Extra Maintenance	₹1,00,000	Rent (Negotiable)	Image URL not found	Property type not found	Furnishing not found

Fig. 1. Depicts the apartments data scraped through NoBroker website

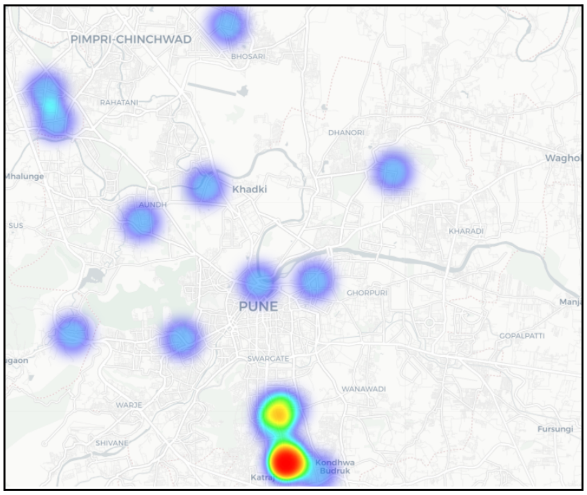


Fig. 2. Preview of Listings of Apartments on a map

IV. Requirements

A. Functional Requirements

1. Data Gathering and Preprocessing
   - Collect and clean housing data and immigrant preferences.
   - Convert text data into numerical features for NLP.
2. NLP Analysis and Clustering
   - Apply NLP to analyze housing descriptions.
   - Group similar listings using K-Means.
3. Recommendation Engine
   - Build a recommendation system based on clustering results and user preferences.
4. User Interface
   - Develop a user-friendly interface for input and housing recommendations.

V. Proposed Design

The Proposed design can be divided into various layers right from data collection layer to mapping and visualization layer . Given below is the system architecture.

- Data Collection Layer

Fetch geolocation data from through web scraping for nearby food outlets and amenities.

Collect Google Maps reviews data for accommodations

- Data Preprocessing and Cleaning Layer

Apply Natural Language Processing (NLP) techniques to preprocess Google Maps reviews, extracting relevant features and sentiments.

Process and clean the collected datasets using Pandas.

- Data Visualization and Analysis Layer

Perform exploratory data analysis to gain insights into student accommodation preferences.

Utilize Matplotlib, Seaborn, and Pandas for data visualization, including box plots to showcase attributes.

- K-Means Clustering Layer

Apply the K-Means Clustering algorithm from Scikit-Learn to cluster accommodation options based on preferences, budget, and proximity.

- NLP Analysis and Accommodation Recommendation Layer:

Build a recommendation engine that suggests accommodations based on user preferences derived from the analysis of reviews.

Rank accommodations based on proximity to preferred amenities and positive sentiments in reviews.

- Mapping and Visualization Layer

Utilize Folium or Seaborn to present clustered accommodations on a map.

Display recommended accommodations along with nearby food outlets and amenities.

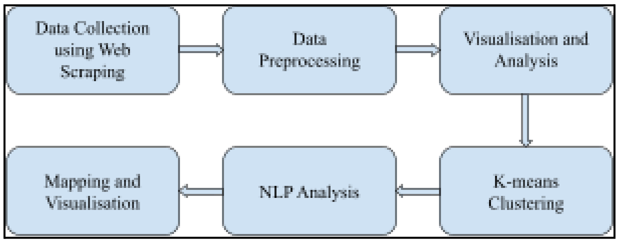

Data Collection using Web Scraping
Data Preprocessing
Visualisation and Analysis
K-means Clustering
NLP Analysis
Mapping and Visualisation

Fig. 3. Depicts the system Block Diagram

VI. Methodology

Step 1: Data Preparation

1. Apartment Data:

Organize the apartment data you scraped from NoBroker. This data should include information about each apartment's location, rent, number of bedrooms, etc.

Ensure that each apartment listing has a unique identifier, such as a property ID.

2. Foursquare Data:

Make API requests to Foursquare for each apartment's location to retrieve information about nearby amenities (e.g., restaurants, schools, parks).

Collect relevant data from Foursquare, including the type of amenity, distance, and any other attributes you want to consider.

Step 2: Feature Engineering

1. Create a Feature Matrix:

Create an empty feature matrix or DataFrame where each row corresponds to an apartment listing, and each column represents a feature (e.g., apartment characteristics or nearby amenities).

2. Apartment Characteristics:

Populate the feature matrix with apartment characteristics, such as rent, number of bedrooms, and any other relevant details. Each of these characteristics becomes a feature column.

3. Amenity Features:

For each apartment, populate the feature matrix with binary columns representing the presence or absence of specific amenities (e.g., Asian restaurants, schools, parks).

Each amenity type from Foursquare should have its column in the feature matrix, and each cell will contain a 1 if the amenity is nearby and a 0 if it's not.

Step 3: Normalization

1. Normalize Data:

Depending on the algorithms you plan to use, you may need to normalize the data to guarantee that the scale of each feature is the same. Common normalization techniques include Min-Max scaling.

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Where:

- X is the variable's initial value.
- The variable's minimum value in the dataset is denoted by Xmin.
- The variable's maximum value in the dataset is denoted by Xmax.
- $X_{Normalized}$ is the normalized value.

This formula ensures that the variable is scaled to a range between 0 and 1.

Eg. Let's say you have a dataset of ages, and you want to normalize the ages using the Min-Max normalization algorithm. Suppose the minimum age is 20, the maximum age is 60, and you want to normalize the age 30:

$$X_{normalized} = \frac{30 - 20}{60 - 20} = \frac{10}{40} = 0.25$$

So, the normalized value for age 30 using Min-Max normalization is 0.25. This process ensures that all age values in the dataset will be scaled proportionally between 0 and 1 based on the minimum and maximum values in the dataset.

Step 4: Feature Vector for Recommendations

1. Create Feature Vectors:

Each apartment listing should now have a feature vector that represents its characteristics and nearby amenities.

The feature vector is essentially a row in your feature matrix, including apartment characteristics and binary values for nearby amenities.

Step 5: Recommendation Algorithm

1. K-means Clustering:

Apply K-Means algorithm to cluster apartments based on user preferences (as converted into a preference vector).

Here's a simplified representation of the K-means algorithm equation:

a. Randomly select k initial centroids: $\mu_{1,} \mu_{2,} \mu_{3, \ldots,} \mu_k$
b. Assign each data point x_i to the cluster with the nearest centroid: $\text{argmin}_j \| x_i - \mu_j \|^2$
c. Recalculate the centroids based on the assigned data points in each cluster:

$$\mu_j = \frac{1}{|C_j|} \Sigma_{x_i \epsilon C_j} x_i$$

Where C_j is the set of data points assigned to cluster j.

d. Repeat Steps 2 and 3 until Convergence.

Suppose we have a dataset with two features, 'Distance to Work' and 'Monthly Rent', and we want to cluster apartments into two groups (k=2).

$$X = \{ (x_1, x_2), (x_3, x_4), \ldots, (x_n, x_{n+1}) \}$$

a. Randomly select two initial centroids:

$\mu_1 = (d_1, r_1)$ and $\mu_2 = (d_2, r_2)$

b. Calculate the distance to both centroids:

- If $\| x_i - \mu_1 \|^2 < \| x_i - \mu_j \|^2$, assign x_i to cluster 1.
- If $\| x_i - \mu_2 \|^2 < \| x_i - \mu_2 \|^2$, assign x_i to cluster 1.

c. Recalculate centroids based on the assigned data points in each cluster:

- $\mu_1 = \frac{1}{|C_1|} \Sigma_{x_i \epsilon C_1} x_i$
- $\mu_2 = \frac{1}{|C_2|} \Sigma_{x_i \epsilon C_2} x_i$

d. Repeat Assignment and Update Steps until Convergence.

The algorithm continues iteratively until the centroids and assignments no longer change significantly. The final result is two clusters of apartments based on their distance to work and monthly rent features.

2. Rank Apartments:

Rank apartments based on their recommendation scores. Higher scores indicate a better match with the user's preferences.

Step 6: Presenting Recommendations

1. Display Recommendations:

Present the top-ranked apartments to the user, along with details like location, rent, and nearby amenities. You can provide an interactive interface for users to explore and filter recommendations based on their preferences further.

Step 7: Evaluation and Optimization

1. User Feedback:

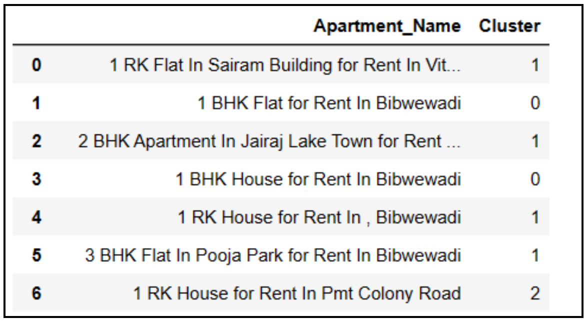
Collect feedback from users about the recommended apartments to continuously improve the recommendation system.

2. Iterate and Optimize:

Based on user feedback, iterate on the recommendation system, adjust the weighting of features, and optimize the algorithms to improve recommendation quality.

VII. Results

The system effectively provided personalized accommodation recommendations based on user preferences, budget, and proximity to essential amenities, thereby addressing a critical need in the immigrant community. Incorporating feedback and user-centric design principles enhanced the overall usability and effectiveness of the system.



	Apartment_Name	Cluster
0	1 RK Flat In Sairam Building for Rent In Vit...	1
1	1 BHK Flat for Rent In Bibwewadi	0
2	2 BHK Apartment In Jairaj Lake Town for Rent ...	1
3	1 BHK House for Rent In Bibwewadi	0
4	1 RK House for Rent In , Bibwewadi	1
5	3 BHK Flat In Pooja Park for Rent In Bibwewadi	1
6	1 RK House for Rent In Pmt Colony Road	2

Fig. 4. Apartments clustered in three classes based on user preferences

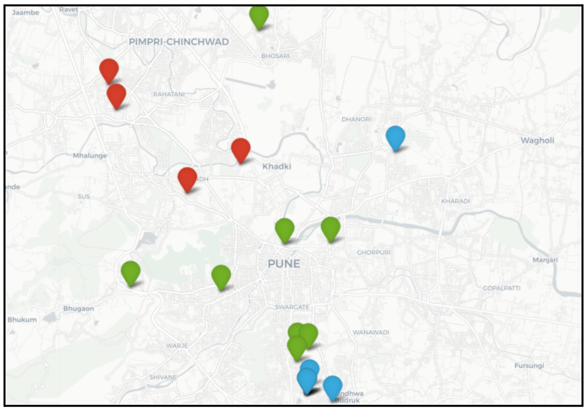


Fig. 5. Visualization of Apartments on map based on Scores generated

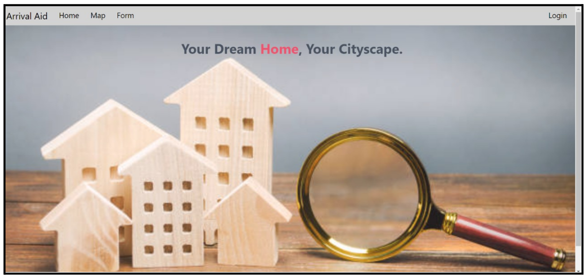
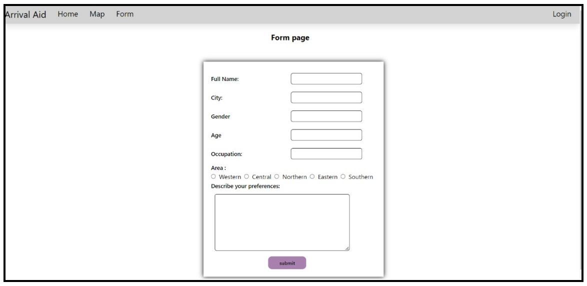


Fig. 6. Home Page



Form page

Full Name:

City:

Gender

Age

Occupation:

Area :
○ Western ○ Central ○ Northern ○ Eastern ○ Southern

Describe your preferences:

Fig. 7. User Form Page

VIII. Conclusion

Our project presents a comprehensive and data-driven solution to the challenges of guiding students and immigrants to find accommodations near their preferred food outlets. By leveraging web scraping, NLP, and K-Means clustering, we can provide highly personalized recommendations based on user preferences, budget constraints, and proximity to amenities. This approach enhances the user experience, streamlines decision-making, and empowers businesses with valuable location insights. The incorporation of mapping and visualization layers adds a unique dimension, presenting clustered accommodations and nearby amenities on interactive maps for a holistic view. This integrated approach promises to simplify the search process and significantly improve user satisfaction while aiding businesses in strategic decision-making.

IX. Future Enhancements

Integrating with the real time data incorporating real-time data sources, such as current local events, traffic conditions, or weather data. This can help immigrants make more informed decisions about their accommodations based on dynamic factors. Also a mobile application that allows users to access the service on the go. Mobile apps can provide location-based services, making it easier for immigrants to find accommodations and amenities when they are out and about.

X. Acknowledgement

We would like to show our gratitude to Prof. RanjeetSingh Suryawanshi, who served as our project advisor, for providing us with the wonderful opportunity to complete the project on the topic of "An Integrated Approach to Immigrant Housing Search using KMeans Clustering and NLP" and for his assistance in doing extensive research.

XI. References

1. Dr.M.Narendra et al, "An Investigation into Geographical Data to Provide Direction for Accommodation and Food Resources" in Turkish Journal of Computer and Mathematics Education, 2022
2. Deepjyoti Roy and Mala Dutta, "A systematic review and research perspective on recommender systems" in Journal of Big Data, 2022
3. Sakshi Rajesh Sinha and Prof. Sumedh Pundkar, "Geolocation Analysis Using Machine Learning" in International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), 2022
4. Md. Shafiul Alam Forhad et al, "An Effective Hotel Recommendation System through Processing Heterogeneous Data" in MDPI, 2021
5. Nimish Kapoor et al, "Movie Recommendation System Using NLP Tools" in the Fifth International Conference on Communication and Electronics Systems (ICCES), 2020
6. Ren Yue-Qiang et al, "A Multi-Element Hybrid Location Recommendation Algorithm for Location Based Social Networks" in IEEE Access, 2019
7. Keyan Gao et al, "Exploiting Location-Based Context for POI Recommendation When Traveling to a New Region" in IEEE Access, 2020
8. Omar Ibrahim Aboulola, "A Literature Review of Spatial Location Analysis for Retail Site Selection" in Data Science and Analytics for Decision Support (SIGDSA), 2017
9. Ylenia Casali, "Machine learning for spatial analyses in urban areas: a scoping review" in Sustainable Cities and Society (Elsevier), 2022
10. Kristina Milkovich et al, "ZenDen - A Personalized House Searching Application" in IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService), 2020
11. Yekta Said Can et al, "Developing Session-based Personalized Accommodation Recommender System by Using LSTM" in 30th Signal

Processing and Communications Applications Conference (SIU), 2022

12. Alejandro Baldominos et al, “Identifying Real Estate Opportunities Using Machine Learning” in Applied Sciences, 2018
13. Xian Shi et al, “Research on House Rental Recommendation Algorithm Based on Deep Learning” in Big Data Economy and Information Management (BDEIM), 2022
14. Mahesh Yadav et al, “Habitation recommendation using machine learning” in International Journal of Creative Research Thoughts (IJCRT), 2023
15. Pramila M. Chawan, “Recommendation System using Machine Learning Techniques” in International Research Journal of Engineering and Technology (IRJET), 2022
16. Stanley Ziweritin1 et al, “A Recommendation Engine to Estimate Housing Values in Real Estate” in International Journal of Scientific Research (IJSR), 2021
17. Keitaro Naruse et al, “Development of accommodation facility selection recommendation system” in 4th International Conference on Awareness Science and Technology (iCAST), 2021
18. Thanatcha Lerttripinyo et al, “Accommodation Recommendation System from User Reviews based on Feature-based Weighted Non-negative Matrix Factorization Method” in 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2015
19. Madhusree Kuanr et al, “Location-based personalized recommendation systems for the tourists in India” in International Journal of Business Intelligence and Data Mining, 2020
20. Dongyang Wang et al , “ Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language” in IEEE, 2020
21. Normalization of Microarray Data: Single-labeled and Dual-labeled Arrays, Jin Hwan Do, Dong-kug PubMed2007
22. An Effective Data Normalization Strategy for Academic Datasets using Log Values, V.SathyaDurga, ThangaKumar JeyaPrakash
23. On the Generalization of Max-Min Normalization for Dimensionality Reduction and Data Representation, Alfredo Cuesta-Infante, Sebastián VenturaPublished in: Information Sciences, 2015.
24. Feature Scaling and Selection in Anomaly Detection ,Authors: Varun Chandola, Arindam Banerjee, and Vipin Kumar
25. Yogendra Jain et al, “Min Max Normalization Based Data Perturbation Method for Privacy Protection” in International Journal of Computer and Communication Technology
26. Abiodun M. Ikotun, et al, “K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data” in Elsevier Information Sciences, April 2023