Mining Massive Datasets: Hotel Recommendation System

Project Proposal

Group Members:

Anzla Aslam 20L-2109

Shahbano Waqar 20L-2162

National University of Computer and Emerging Sciences
Department of Computer Science
Lahore, Pakistan

1. Mining of Massive Data Techniques

In our data mining project, we are aiming to analyze hotel reviews for personalized recommendations and insights. Here is the tentative plan that how we can apply the mining techniques:

- **Clustering:**
  We will use clustering algorithms like K-means to group similar reviews and user behavior. This will help in identifying patterns in hotel satisfaction and user preferences. Clustering will be applied to features such as review ratings, review text sentiments, and possibly other extracted features.

- **Sentiment Analysis:**
  Sentiment analysis techniques will be employed to classify reviews into positive, neutral, and negative categories. This will help in understanding user opinions and sentiments towards hotels, which is important for providing personalized recommendations.

- **Recommendation Systems:**
  We will develop a recommendation system for hotels based on user clusters and similarities. This could involve collaborative filtering or content-based recommendation algorithms. The recommendation system will utilize the results from clustering and sentiment analysis to suggest hotels tailored to individual user preferences.

- **Data Preprocessing Techniques:**
  Techniques for data cleaning and preprocessing will be applied to handle missing values, convert data types, and process the reviews.date field to understand the timeline of reviews. Additionally, feature extraction techniques will be used to derive meaningful features such as review text length, positive/negative sentiment scores.

- **Visualization Techniques:**
  Visualization techniques using libraries like Matplotlib or Seaborn will be employed to present insights and patterns in the dataset. This will include visualizing clusters, timelines of reviews, user demographics, common themes in reviews, popular hotel features, and user satisfaction trends.

- **PySpark for Big Data Handling:**
  PySpark will be utilized for handling and processing the data efficiently due to its capability to handle large datasets. PySpark will provide clustering algorithms for grouping similar reviews and users, and sentiment analysis tools for classifying review texts.

2. Project Objective

The actual problem is the lack of personalized recommendations for hotel selection based on individual preferences and sentiments expressed in reviews. Users often face difficulty

in sifting through large volumes of reviews to find accommodations that match their specific needs and preferences. This project aims to address this problem by employing data mining techniques to analyze reviews and offer tailored recommendations, thus improving the overall user experience in hotel selection. The objective of this project is to analyze a dataset containing hotel reviews to extract valuable insights and provide personalized recommendations to users based on their preferences and similarities with other reviewers.
Here is the expected outcomes of our project:

- **Clustered Reviews:**
  The project expects to cluster hotel reviews based on similarities in user preferences and sentiments. These clusters will help identify different types of reviews and user behavior patterns.

- **Sentiment Analysis:**
  Sentiment analysis will classify reviews into positive, neutral, and negative categories, providing insights into overall user opinions towards hotels.

- **Personalized Recommendations:**
  By combining clustering results with user behavior analysis, the project aims to develop a personalized recommendation system for hotel choices. This system will recommend hotels based on user preferences and similarities to other users, thereby improving the relevance and usefulness of recommendations.

- **Insights and Visualizations:**
  The project will present insights such as common themes in reviews, popular hotel features, and patterns in user satisfaction. Visualizations will help depict clusters, timelines of reviews, and user demographics, facilitating a better understanding of the dataset and its implications.

3. Related Research Paper:

We studied the related research paper [1] and here is the key findings:

- Migration Trends: The paper highlights the prevalence of migration, particularly among students and employees, driven by factors such as work opportunities and higher education. It emphasizes the challenges faced by migrants, particularly students, who often relocate to unfamiliar locations alone, lacking support networks and facing difficulties in finding suitable accommodations.

- Challenges in Accommodation Search: One of the primary challenges identified is the difficulty in finding suitable accommodations, exacerbated by factors such as language barriers and cultural differences. The absence of local knowledge makes it challenging for migrants to locate subsidized housing that meets their needs, potentially leading to their return to their home location.

- Recommendation System: The paper proposes the development of a recommendation system to address the challenges in finding accommodations for migrants. This system aims to assist users in identifying suitable accommodations based on their preferences, budget, amenities, and proximity to desired locations.

- Machine Learning Approach: The research paper presents an efficient approach utilizing machine learning methodologies, specifically K-Means clustering algorithm and content-based filtering. These techniques are employed to recommend paying guest rooms and hostels based on user preferences and desired locations.

- Proposed Solution: The research paper aims to address the problem statement by proposing an Immigration Geolocation Analysis and Recommendation System. This system aims to recommend the best accommodation facilities based on users' preferences and provide information on nearby transportation options from the desired destination.

- Data Collection and Cleaning: The project utilizes data collection techniques to gather information about the user's city and hostel/PG rooms in the locality. The collected data undergoes cleaning to eliminate noisy and outlier data, ensuring its relevance to the objective of the recommendation system.

- Clustering and Geolocation Analysis: Through the use of the K-means clustering algorithm and geolocation data, the project clusters hostel and paying guest room data based on the user's location. This clustering helps identify accommodations nearest to the user's desired location, facilitating personalized recommendations

- Data Visualization: The project employs data visualization techniques to present the gathered data on maps, providing users with a visual representation of available accommodations in their desired location. Visualization tools such as Folium or Seaborn are utilized to achieve this.

- Content-Based Filtering: The recommendation system utilizes content-based filtering, a natural language processing approach, to recommend accommodations based on user preferences. By analyzing descriptions and tags of accommodations, the system suggests options that closely match the user's requirements.

- Price Prediction Model: Additionally, the project includes a price prediction model using linear regression. This model predicts the price of hostels based on features such as the number of beds, amenities, and ratings, providing users with estimated pricing information for accommodations.

- Results: The result and analysis demonstrate the effectiveness of the Hostel-Finder recommendation system in providing personalized accommodation recommendations based on user preferences and location.

4. Dataset Details:

The dataset utilized for this project was sourced from Kaggle, comprising over 35,000 entries of hotel reviews. It encompasses various features including address, categories, city, country, latitude, longitude, name, postal code, province, review date, review addition date, review recommendation status, review ID, review rating, review text, review title, reviewer's city, reviewer's username, and reviewer's province. References to previous work or similar datasets could involve research articles or datasets related to hotel reviews, recommendation systems, and geolocation analysis.
https://www.kaggle.com/datasets/datafiniti/hotel-reviews?rvi=1

5. Bibliography

[1] "Hostel Finder: Location-Based Recommendation System for Hostels and PGS with Transit Information, 28-May-2023. [Online]. Available: https://ijrpr.com/uploads/V4ISSUE5/IJRPR13360.pdf [Accessed: 17-April-2024].