RouteQuest: The Road Trip Planner

Krishna Sameera Surapaneni, Maharsh Soni, Siddharth Solanki

Motivation

Embarking on a journey in a completely new country brings forth a myriad of challenges, with navigating unfamiliar landscapes and discovering must-see attractions among the most complex. Concurrently, the current spike in interest in transforming long road journeys into immersive full-day events heightens the demand for a solution to improve the road trip experience for tourists**.** Recognizing this trend, the motivation for establishing a travel suggestion system derives from a general desire to improve tourists' road trip experiences. This approach primarily depends on tourist departure location and arrival destination information to help people choose the best sights along their chosen routes.

Background

The basis of this program is based on the concept that individuals migrating to a new nation face unique challenges in trip planning, notably in terms of navigation complications and unfamiliarity with notable sites.The recommendation system incorporates the Open Route Service API, by recommending locations that are strategically located between the trip's start and end points. The smooth incorporation of navigation features and real-time location data highlights the commitment to offering comprehensive recommendations. This comprehensive method ensures that travelers not only obtain directions to intriguing locations but also receive suggestions based on distance, prospective costs, popularity, and rating, eventually enriching their whole road trip experience.

Literature Review

Kokate et al. (2018) developed a tourism recommender system using data mining techniques, offering personalized recommendations for destinations, routes, accommodations, and attractions. Researchers used travelers' preferences, suggestions, popular places info, and reviews for informed decision-making. The proposed system used algorithms like Euclidean Distance, KNN, and Apriori to recommend and classify based on user data, location, and preferences. Nitu et al. (2021b) built a personalized travel recommendation system using social media data, analyzing Twitter activity to offer tailored travel suggestions. The system prioritized recent user interests, enhancing accuracy. Various packages like Tweepy, BotoMeter, arules, and GoogleAPI are used for data collection, sentiment analysis, and recommendation. The research categorized Points of Interest (POIs) and leverages machine learning for tweet classification. Despite achieving 75.23% accuracy, the system faced challenges in classifying certain travel tweets, limited training data, and the need for real-time updates. In the study by Bigdeli et al. (2008), research focused on comparing correlation-based and cosine-based similarity algorithms.These algorithms are fundamental in collaborative filtering, a technique used to predict user preferences.The research revealed that the correlation-based algorithm outperformed the cosine-based similarity algorithm in the context of tourism recommendations. User profiling, which plays a vital role in addressing challenges faced by recommender systems, involves the collection of user data. This study underscores the significance of user profiling and the superior performance of correlation-based algorithms in tourism recommendations.

In comparison to existing models like Kokate et al.'s (2018) tourism recommender system, which might lack real-time adaptability, and Nitu et al.'s (2021b) approach facing challenges in real-time updates and limited training data, the travel suggestion system stands out for its superior features. Prioritizing real-time location data and navigation features ensures users receive up-to-the-minute recommendations, addressing potential drawbacks in existing systems. The RouteQuest project goes beyond conventional approaches by considering factors such as distance, value at each price point, popularity, and rating. Leveraging user reviews to capture actual sentiments results in more accurate recommendations, offering a comprehensive approach to road trip planning. Incorporating the Open Route Service API strategically recommends locations between trip start and end points, enhancing the user experience by

displaying only the types of places users want to see. In contrast to previous research works excelling in specific aspects, the method seeks to offer a complete road trip experience, making it an excellent option in the field of travel recommendation systems and encouraging special trips.

Methodology

Data Collection

The data for this project was obtained from GitHub. In total, there were four datasets. In total, there were four datasets. Critical information included in these databases included prices, names, reviews, and ratings of attractions, as well as geographic coordinates, state and country, and pricing information.

Of the four datasets, two in particular had detailed information about the attractions, such as their latitude and longitude coordinates, the accompanying information about the state and country, and their prices. At the same time, the remaining two datasets contained information critical to user experiences, such as attraction names, reviews, and ratings. These four datasets were combined into two separate datasets during a deliberate preprocessing step. Attraction details were smoothly merged into one dataset to create a comprehensive library of contextual, price, and geographic data. The other dataset balanced opinions and reviews related to various attractions, creating a unified and consolidated dataset.

Note. Although the datasets were originally obtained via GitHub, it is important to emphasize that the present project is substantially different in how it is implemented than the GitHub project where the data was originally obtained. The main focus of the GitHub project is to provide users who are traveling to a new place with recommendations for hotels and surrounding attractions. The present study, on the other hand, adopts a unique strategy, concentrating on recommending points of interest that coincide with the path between two user-selected destinations.

Exploratory Data Analysis

Upon initial analysis of the dataset including attraction details, a consistent observation emerged: the majority of attractions were located in Canada, with a smaller representation from the United States and a scarce presence from other nations. An intriguing data structure was discovered in which latitude and longitude values were combined into a single column designated "location." Furthermore, it was noted that this "location" column had null values. One other significant discovery concerned the price values, which showed several instances of irregularities, particularly when the values were listed as negative. Subsequent exploration of the dataset containing attraction reviews revealed a distinct inconsistency in the rating values—some ratings were assigned negative values.

A closer look at the dataset containing attraction reviews revealed a key inconsistency: negative ratings in the rating column. Furthermore, a structural observation revealed that the "attraction_id" column contained numerous duplicate values, which might be attributed to each attraction having multiple reviews listed, resulting in the dataset having more than 33000 instances. The dataset also had the date on which the reviews were given.

Feature Extraction

In the attraction dataset, which includes multiple reviews and ratings for each attraction, two new features, namely 'sentiment' and 'popularity,' were derived based on the reviews and ratings, respectively. The sentiment feature captured the overall sentiment conveyed in the reviews, while the popularity feature reflected the collective rating scores. Following the creation of these new features, an analysis was conducted to observe how the sentiment evolved over time.

Data Cleaning

The initial step involved the separation of the "location" column in the attraction details dataset into two distinct columns, namely "latitude" and "longitude." Following this separation, the dataset was scrutinized, revealing the presence of null values in the newly created latitude

and longitude columns. The handling of these null values was approached with caution, recognizing that imputation with mean, mode, or nearest values might introduce inconsistencies, given the nature of geographical data. Consequently, the decision was made to remove instances with null values in the latitude and longitude columns. This approach aimed to preserve the accuracy and integrity of the geographical information, ensuring that any imputation did not compromise the precision of location coordinates for attractions in the dataset. Secondly, the price column presented challenges in the form of both null values and negative values. To address the issue of negative values, they were converted to their absolute values. Subsequently, the null values within the column were addressed through imputation. In this particular instance, the missing values were filled by employing the mean value of the column.

Then, outliers in the price column were discovered. To overcome this issue, an imputed outlier value technique was used. The outlier values were imputationally replaced with the largest value within the interquartile range (IQR) that did not qualify as an outlier. This method was designed to reduce the influence of outliers on the dataset by ensuring that the imputed values were within a fair and representative range. A critical step in data processing in the attraction reviews dataset was the removal of duplicate values based on the "attraction_id" to ensure the uniqueness of each attraction's representation. Furthermore, several columns, such as "date," "user_id," and "reviews," were deemed unnecessary for the project's specific aims, as important information such as sentiment and popularity had already been obtained. As a result, the dataset's unnecessary columns were removed.

Then, when the ratings in the attraction reviews dataset were examined, it was discovered that some of the numbers were negative. Given that online ratings cannot normally be negative because the lowest rating is zero stars, the absolute values of the negative ratings were used to ensure that all ratings in the dataset were genuine and aligned with the inherent constraints of online rating systems. In the final step, both datasets were merged using an inner join. This choice aimed to avoid potential complications associated with other join types, such as outer joins, which might introduce more null values. An inner join ensured that only the overlapping records between the two datasets were retained in the merged dataset, minimizing the occurrence of null values and maintaining data consistency.

Data Transformation

To standardize the sentiment and popularity scores in the dataset, Z-score normalization was used. A crucial element is the sentiment score, which captures the overall sentiment expressed in user reviews. Meanwhile, the popularity score is an average of the rating scores and serves as a measure of public favor.

Z-score normalization, also known as standard score normalization, is a statistical technique that translates data points in terms of their standard deviation from the mean. This procedure yields a distribution with a mean of 0 and a standard deviation of 1. This normalization method properly scaled both the sentiment and popularity scores in the application to fall inside a range of -1 to 1.

The sentiment and popularity scores contributed equally to the analysis by normalizing them in this way, regardless of their original magnitude or distribution. This methodological decision contributed significantly to the robustness and dependability of a data-driven conclusions.

Similarly, with the price column, it starst by extracting values from the CSV file and converting them into a list of floating-point numbers. This step is critical for later data processing. After creating this list, calculate the minimum (min_price) and maximum (max_price) values among these prices, which sets the stage for the normalization step.

The code then starts the normalization step. It processes each row of data sequentially, applying a normalization method to the 'price' value. The price is standardized to a scale of 1 to 5 using the following formula: **1 + (price - min_price) / (max_price - min_price) * 4**. This

formula is intended to linearly adjust each price so that the least price corresponds to 1 and the greatest price corresponds to 5. As a result, a consistent price scale is created, which improves the data's consistency and comparability, which is an important step in preparing the dataset for further analysis or application within the project. Figure 1 shows the final dataset after data preprocessing.

Figure 1

attraction_id	name	country	province	city	price	rating	latitude	longitude	sentiment_y	popularity	normalized_popularity
0	vancouver_city_sightseeing_tour	canada	british_columbia	vancouver	1	4.5	49.1978322	-123.0649959	0.4413471554668297	57	0.583789370060629
1	vancouver_to_victoria_and_butchart_gardens_tour_by_bus	canada	british_columbia	vancouver	2	5.0	49.1978322	-123.0649959	0.4103325643502209	89	1.1588650497649926
2	quebec_city_and_montmorency_falls_day_trip_from_montreal	canada	quebec	montreal	2	4.5	45.5001458	-73.5720264	0.32393798459003686	67	0.7635005199682428
3	niagara_falls_day_trip_from_toronto	canada	ontario	toronto	2	5.0	43.6561507	-79.3842642	0.3610255959379636 3	27	0.044655920337787966
4	best_of_niagara_falls_tour_from_niagara_falls_ontario	canada	ontario	niagara_falls	2	5.0	43.0857136	-79.0824311	0.389974211495044 85	111	1.554229579561743
5	niagara_falls_in_one_day:_deluxe_sightseeing_tour_of_american_and_canadian_sides	canada	ontario	niagara_falls	2	5.0	43.102436	-78.961638	0.44663402711607264	367	6.154835017196654
6	whistler_small-group_day_trip_from_vancouver	canada	british_columbia	vancouver	2	5.0	49.1978322	-123.0649959	0.3687468390133831	34	0.17045372527311756
7	ultimate_niagara_falls_tour_plus_helicopter_ride_and_skylon_tower_lunch	canada	ontario	niagara_falls	3	5.0	43.0857136	-79.0824311	0.3484333148663506	28	0.06262703532854934
11	vancouver_seaplane_tour	canada	british_columbia	vancouver	2	5.0	48.4241277	-123.3707833	0.3157020170375434	57	0.583789370060629
12	niagara_falls_grand_helicopter_tour	canada	ontario	niagara_falls	2	5.0	43.1888394	-79.1714688	0.3103026581002772	21	-0.06317076960678025
13	whistler_helicopter_tour	canada	british_columbia	whistler	2	5.0	49.1830074	-123.176599	0.41477732832643543	26	0.026684805347026597
14	niagara_falls_helicopter_tour	canada	ontario	niagara_falls	2	4.5	43.1186039	-79.0743842	0.3569561153575744 3	246	3.9803301033145275
15	7-minute_helicopter_tour_over_toronto	canada	ontario	toronto	2	4.5	43.632485	-79.3956627	0.339708978191121	13	-0.2069396895328712

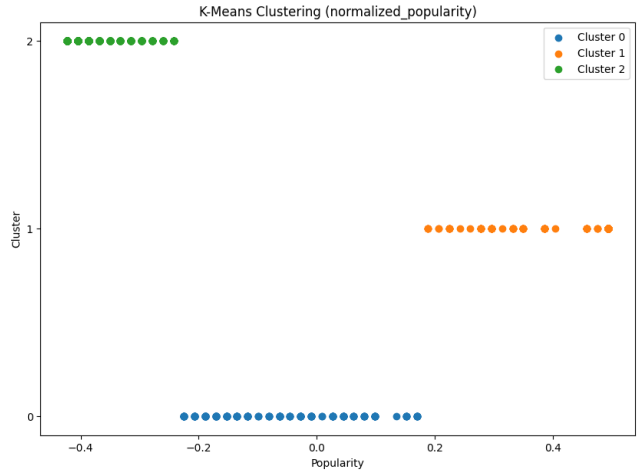
Experiments and Results

K-Means Clustering and Cluster Validation

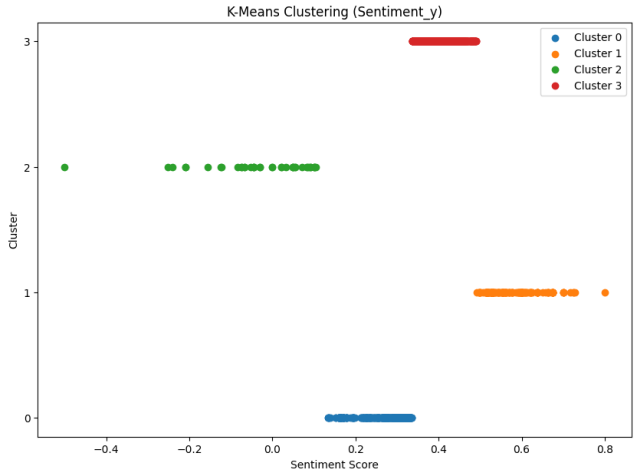
K-means clustering was used to cluster the normalized sentiment and popularity columns, with several values of k being tested. The Silhouette score was used to assess the efficacy of the clusters. The elbow approach was then used to calculate the ideal value of k. The silhouette score was generated for each candidate k, assisting in determining the point at which additional increases in k did not improve clustering quality appreciably.

Following the determination of the optimal value of k, sentiments were grouped into categories based on the clusters and the values within those clusters. This process entailed assigning descriptive labels to the clusters, such as "Bad," "OK," "Good," and "Very Good," based on the observed sentiment scores and their association with specific clusters. For instance, cluster 2 exhibited sentiment scores greater than 0.4, all data points within this cluster were categorized as "Very Good." The same was implemented for normalization, where the groupings were ‘Not Popular’, ‘Popular’, and ‘Very Popular’. The figures below show the three different clusters and their scores. Figure 2 shows the clusters for normalization and sentiment respectively. Figure 3 depicts the cluster considering both sentiment and normalization.

Figure 2

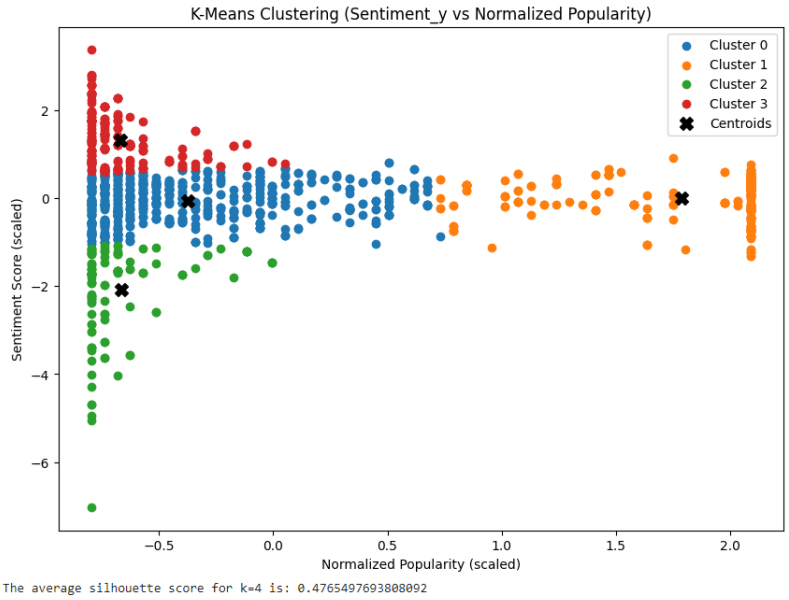


The average silhouette score for k=3 is: 0.7342778075415882



The average silhouette score for k=4 is: 0.5636775248216486

Figure 3



Feature	Optimal k	Silhouette Score
Sentiment_y	4	0.57
Normalized_popularity	3	0.73
Sentiment_y Vs Normalized_popularity	4	0.47

Recommendation System

The dataset, which has been optimized through extensive pre-processing, serves as the cornerstone for the recommendation engine. The project begins by compiling a list of coordinates for various cities across Canada. This list is used to acquire the corresponding coordinates after receiving the user's input for starting and finishing locations. The OpenRouteService API then uses these coordinates to generate precise route information. This provides street names, starting and ending coordinates (latitude and longitude), distance in meters, and estimated travel time in seconds for each street segment. The resulting data, which is initially in JSON format, is saved carefully into a JSON file. This file is then converted to CSV format for improved usage.

The final dataset includes the precise coordinates of all Points of Interest (POIs), as well as various other features. Calculation of the distance to the POIs, is next, using these coordinates and the coordinate range of every particular street on the suggested route. In addition, if the user indicates a preference for a specific type of POI, the system generates a curated list of relevant POIs. This output list includes not only the name of each POI but also useful information such as its sentiment type, popularity, and price level. The findings part of the report includes a visual representation of this output, providing a clear and succinct summary of the system's capabilities.

Results

The system utilizes input parameters such as the starting city, ending city, and category filter. Based on these inputs, it generates a comprehensive list of Points of Interest in Canada. The

list includes how people feel about each place (sentiment), how popular they are, and their price level. The resulting output comprises four columns:

- **Name**: Represents the name of the tour.
- **Sentiment**: Reflects the average rating of the POI, ranging from "Very Good" to "Bad."
- **Popularity**: Indicates the relative popularity of the POI, spanning from "Not Popular" to "Very Popular."
- **Price Level**: Represents the price level of the POI, categorized from "$" to "$$$$$." This structured output format provides users with clear and organized information, aiding in their decision-making process.

A sample output is shown in Figure 4.

Figure 4

Discussion & Future Improvement

Dataset Augmentation for Enhanced Robustness

To improve the stability and usefulness of the dataset, there is a plan to add a larger number of Points of Interest (POIs) and the corresponding geographic coordinates. The aim of this expansion is to broaden the demographic reach of the target audience and diversify the recommendations provided.

Integration of User-Specific Data for Enhanced Personalization

In order to enhance recommendation algorithms further, there is a plan to investigate integrating user-specific data in future updates of the project. This initiative aims to improve the level of customization offered by the service by more precisely tailoring recommendations to each user's preferences. The goal is to provide each user with a more customized and fulfilling experience by integrating personalized data.

Development of an Intuitive User Interface for Enhanced Accessibility

In order to improve the application's accessibility, there is a planned initiative to design an intuitive user interface. The goal of this strategic improvement is to provide end users with a more seamless and engaging experience, encouraging greater user participation and overall happiness. The objective of the user interface development is to make the program easy to use so that users may interact with the trip recommendation system in a pleasant and positive way.

References

Bigdeli, E., & Bahmani, Z. (2008). Comparing accuracy of cosine-based similarity and correlation-based similarity algorithms in tourism recommender systems. IEEE. https://doi.org/10.1109/icmit.2008.4654410

Kokate, S., Gaikwad, A., Patil, P., Gutte, M., & Shinde, K. (2018). Traveler’s Recommendation System Using Data Mining Techniques. IEEE. https://doi.org/10.1109/iccubea.2018.8697862

Nitu, P., Coelho, J., & Madiraju, P. (2021). Improvising personalized travel recommendation system with recency effects. Big Data Mining and Analytics, 4(3), 139–154. https://doi.org/10.26599/bdma.2020.9020026