CptS 315

Restaurant Recommender System

12/12/2021

Introduction

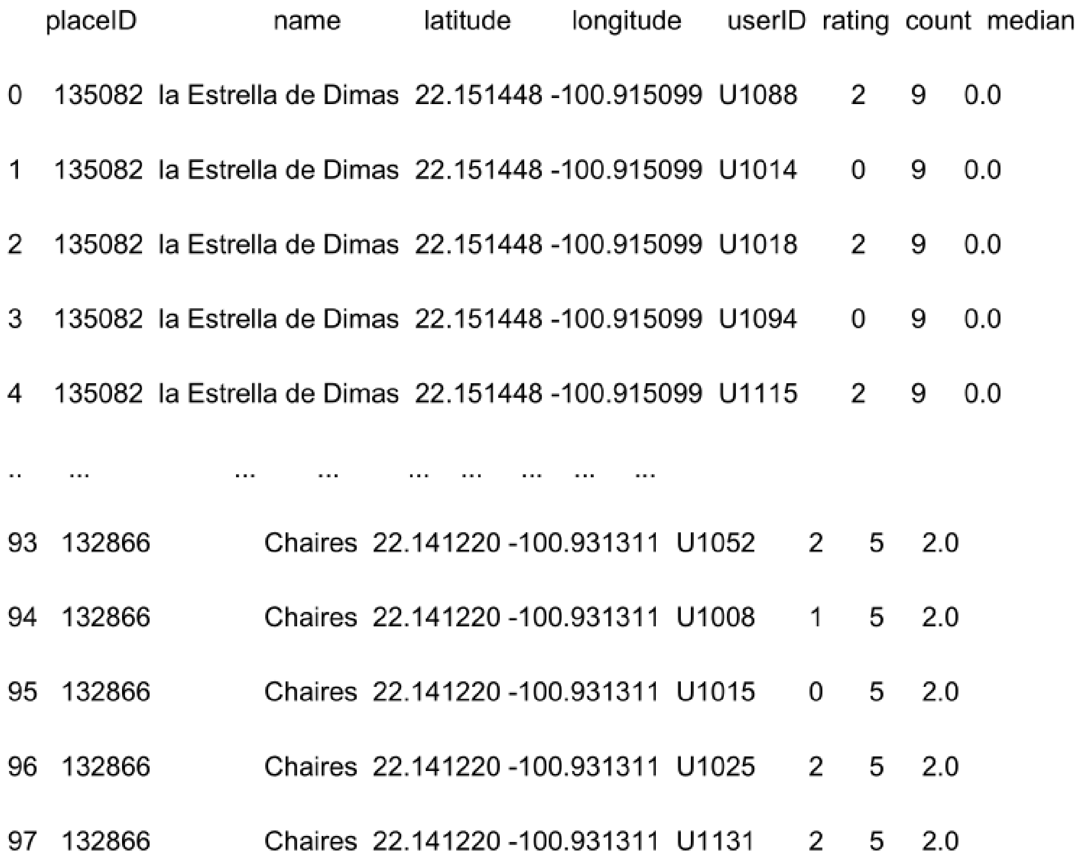
We have many choices when choosing a restaurant, and it is difficult to decide which restaurant to go to. There are currently many different applications that can recommend restaurants based on their location, ratings, user preferences, etc. Recommendation systems are widely used in different fields to predict users' preferences or ratings for products or services. What I want to create is also a restaurant recommendation system, which considers the user's location, and then recommends restaurants based on the user's preferences.

There are two main methods for the implementation of the recommendation system, namely content-based filtering, and collaborative filtering. Content-based filtering makes recommendations based on a comparison between item descriptors and user profiles. Collaborative filtering is a technology that automatically predicts user interests by collecting the preference or rating information of many users.

The challenge of this project is to pre-filter the context and then apply collaborative filtering. Focus on geographic location and use the user's location to provide relevant recommendations. The K-Means clustering algorithm will be used to define a restaurant cluster by its location. The K-nearest neighbor algorithm will be used to implement item-based collaborative filtering. K-nearest neighbor will calculate the distance between a given restaurant and all other restaurants in the cluster, and then return the first k nearest neighbor restaurants as suggestions. The program can output recommendations based on user location and item-item similarity.

Data Mining Task

The input data is obtained as a .csv file from the UCI machine learning repository. The input data contains various restaurant information and user information, including geographic data. For the first part of the filtering, the pandas data frame is used to merge the geoplaces2.csv and rating_final.csvl files together after removing unnecessary attributes. Mapbox from the plotly express module is used to visualize the location of the restaurant based on the given input data. Select the city with the most data for Kmeans clustering. The city name consists of all capital letters. The determination of the optimal k value is obtained from the evaluation of the elbow method. The following data shows restaurants in the same cluster based on the user's location.



```
    placeID            name      latitude    longitude   userID  rating  count  median
0   135082  la Estrella de Dimas  22.151448 -100.915099  U1088      2      9     0.0
1   135082  la Estrella de Dimas  22.151448 -100.915099  U1014      0      9     0.0
2   135082  la Estrella de Dimas  22.151448 -100.915099  U1018      2      9     0.0
3   135082  la Estrella de Dimas  22.151448 -100.915099  U1094      0      9     0.0
4   135082  la Estrella de Dimas  22.151448 -100.915099  U1115      2      9     0.0
..     ...                  ...        ...         ...    ...    ...    ...     ...
93  132866              Chaires  22.141220 -100.931311  U1052      2      5     2.0
94  132866              Chaires  22.141220 -100.931311  U1008      1      5     2.0
95  132866              Chaires  22.141220 -100.931311  U1015      0      5     2.0
96  132866              Chaires  22.141220 -100.931311  U1025      2      5     2.0
97  132866              Chaires  22.141220 -100.931311  U1131      2      5     2.0
```

For the second part of the filter, the Item-item similarity will be calculated, and then the K nearest neighbor method will be used to generate recommendations to the user based on the restaurant that the user likes. In this project, we will randomly select restaurants that users like.

The data mining problem investigated in this project.

1. Scoring basis
2. According to the given data, the location of the restaurant
3. Cluster data by location
4. Recommend restaurants based on user location
5. Recommend restaurants using collaborative filtering technology.

Learn how to use the sklearn, scipy and plotly modules. There are some tutorials on the Internet, but there are many different problems that affect the desired output.

Technical Approach

The first part uses the Kmeans clustering method of the sklearn module. Kmeans is a very simple algorithm that can cluster data into K clusters.

Step1-Select K random points as cluster centers, called centroids, assuming random centroids = c1, c2, ... ck

Step2-Assign each input to the nearest cluster by calculating the distance from each input to each centroid.

Euclidean distance: $d(\mathrm{C},\mathrm{X}) = \sqrt{(C1-X1)^2+(C2-X2)^2+\cdots+(Cn-Xn)^2} =$

$\sqrt{\sum_{i=1}^{n}(Ci - Xi)}$

Step3- Find a new cluster center by taking the average of the distribution points. Si = the set of all points assigned to the i-th cluster: $Ci = \frac{1}{|Si|}\sum_{xi\varepsilon Si} Xi$

Step4- Repeat steps 2 and 3 until there is no change in the cluster allocation

For collaborative filtering implementation, use the sparse matrix in the SciPy module and the K nearest neighbors in the sklearn module.

Step1- Create a pivot table and fill empty cells with the value 0.
Step2- Calculate the cosine similarity between each restaurant. Cosine similarity is a metric used to measure the similarity of documents regardless of their size.:$Cos\theta = \frac{\vec{a}*\vec{b}}{||\vec{a}||\,||\vec{b}||} = \frac{\sum_{i}^{n} aibi}{\sqrt{\sum_{i}^{n} ai^{2}}\sqrt{\sum_{i}^{n} bi^{2}}}$

Step3- Perform k-nearest neighbor classification and fit the given data.

```
k-nearest neighbor
Classify(X,Y,x)//X: training data, Y: class labels of X, x: unknown sample
For i=1 to m do
    Compute distance d(Xi,x)
End for
Compute set I containing indices for the k smallest distances d(Xi,x)
Return majority label for {Yi where i ∈I}
```

Step 4- Predict the recommended restaurant for a given restaurant.

Evaluation Methodology

The research data set comes from the National Research and Technology Development Center of CENIDET, Mexico. There are 130 restaurants, 138 users, and 1,161 restaurant ratings. The attributes used in this study are location ID, name, user ID, rating, latitude, and longitude. The accuracy of the k nearest neighbor classifiers cannot be tested. It is difficult to find any evaluation indicators.

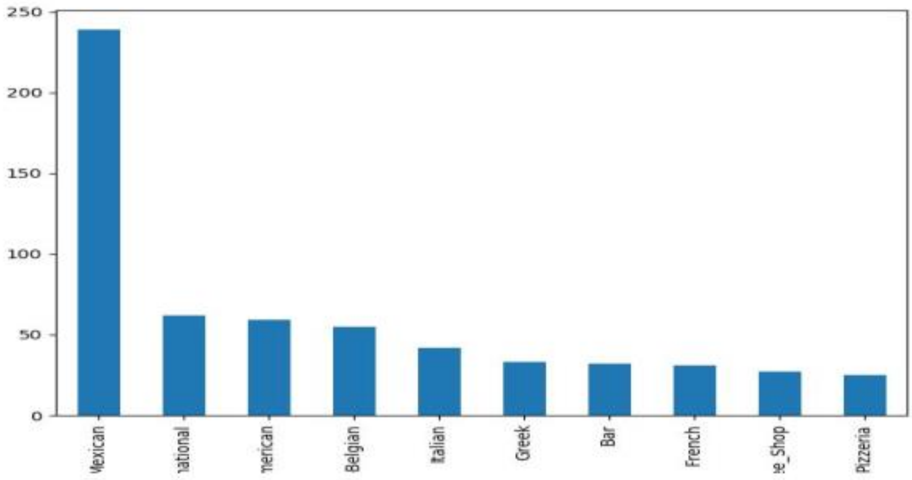
Contrary to the supervised learning that evaluates the basic facts of model performance, cluster analysis has no reliable evaluation indicators, and it can be used to evaluate the results of different clustering algorithms. Since Kmeans requires k as input and not from the data, there is no correct answer in terms of the number of clusters you should have in any question. Sometimes domain knowledge and intuition may be helpful, but usually this is not the case. In the cluster prediction method, the performance of the model can be evaluated based on different K clusters. The Elbow method is based on the sum of the squared distances between the data points and the centroids of their assigned clusters. Choose k where the SSE starts to flatten and form the elbow. You will use the geyser data set and evaluate the SSE for different values of k and see where the curve might form an elbow and flatten.

Results and Discussion

Visualization of restaurant locations and user locations

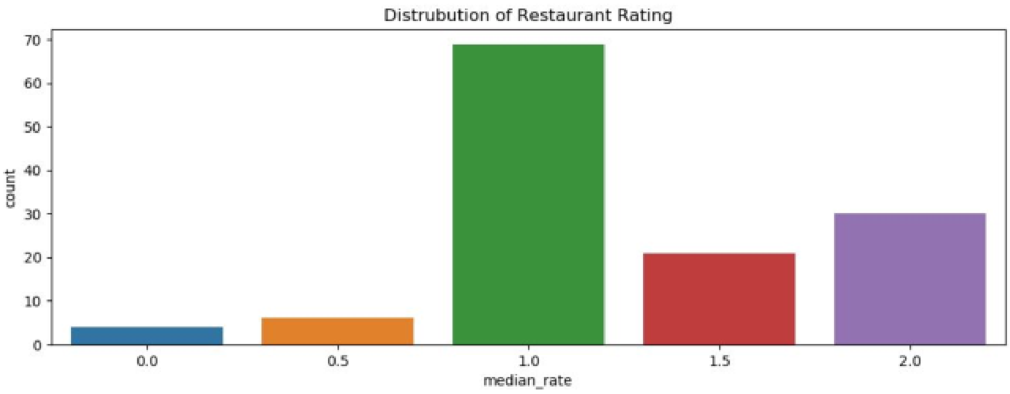
There are 130 restaurants in database There are 138 users in database

Find the top 10 foods with the highest ratings from a given data set. The Matplotlib module is used for visualization. The code counts the number of food ratings for each unique time, and then generates a bar graph. Mexican cuisine has the highest score, followed by national cuisine, and third is American cuisine.



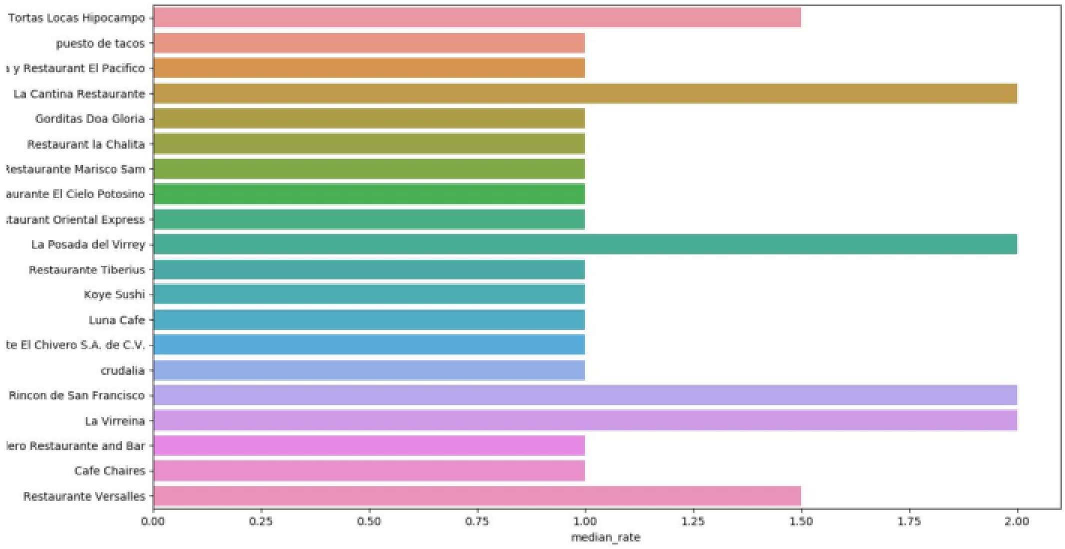
Based on the input data, Mexico’s top 10 cuisines

The median rating approximately has a normal distribution centered on 1. The median rating is calculated based on the sum of the ratings of each restaurant divided by the rating count.



Restaurant median score distribution

To calculate the popular restaurants, the values are sorted according to the number of restaurant ratings and the median of their ratings. The most popular restaurant is Tortas Locas Hipocampo, and the second one is Puesto de Tacos.

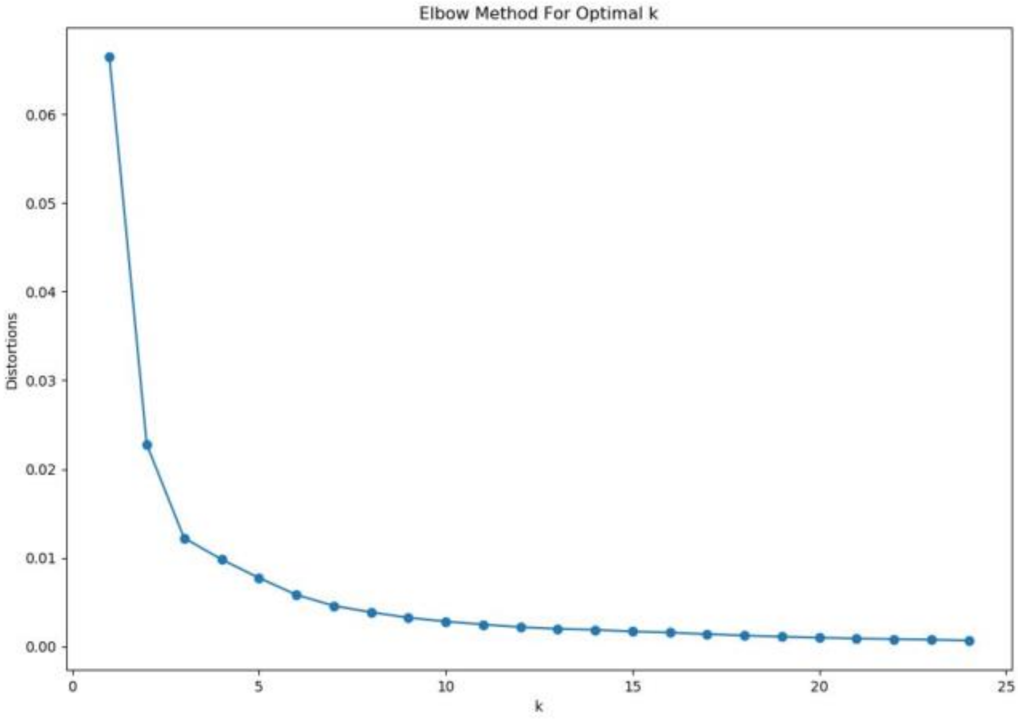


Enter the top 20 restaurants in the data.

Plotly express provides a scatter plot frame in which I can illustrate geographic data. The data sets are distributed in 3 different states of Mexico, namely San Luis Potosí, Tamaulipas, and Morelos.

Expand the Mapbox to show the location of the restaurant in the input data set, and Kmeans clusters by its front and back locations.

Reduce the scope to a certain city as a data set to be able to perform K-means clustering. Users will most likely need recommendations from their city. Five clusters can be selected using sklearn.After filtering out recommendations based on the user's location, the k-nn algorithm provided by sklearn is implemented to apply item-item collaborative filtering. If the user likes a restaurant named a certain restaurant, the system will use the cosine similarity index to recommend similar restaurants that the user may also like.



Using K-nearest neighbor algorithm to make recommendation
The fundamental assumption of item-based collaborative filtering is that a user gives similar ratings for similar restaurants. Also, we can assume, that similar users would give similar ratings to similar restaurants. Therefore, if we have similar users, then we can recommend restaurants to certain user at the ground of another users ratings for these restaurants. We need to create a database of users to differentiate them by certain groups. Only ratings of users inside similar group will be counted when making recommendations. Three files contain information about users: userprofile.csv, userpayment.csv, usercuisine.csv. According to paper by VargasGovea, Gonzales-Serna, Ponce-Medellin, 2011, usually most of information about objects is excessive because some of them are interdependent, and some have no influence on possible classification of users. We select following features from file userprofile.csv to be enough to characterize users: latitude, longitude, budget Reading data from rating_final.csv with medium budget

Conclusion:

I learned that having more data helps collaborative filtering technology to provide better accuracy. Having appropriate data suitable for research purposes is important to achieve better performance. It is more challenging than usual, and there are many solutions. There is also learning about K-means. The K-means algorithm iteratively improves cluster quality by assigning and reallocating data to different groups until equilibrium is reached. Change at any time without obvious errors, just like in k-means. I think it is more flexible. If I have the opportunity, I would like to further optimize the project, combining other aspects of user preferences, such as price, fashion preferences, restaurant atmosphere, etc. And have a deeper understanding of evaluation indicators and machine learning algorithms. All in all, I have gained a lot of knowledge about AI machine learning and data mining through this course, and I also hope that I can follow the understanding and in-depth things in the future.