DATA SCIENCE PROJECT

Introduction

1.1 Background

Exploring new places in a city can be challenging at times. It can be a daunting task for one to find his way through the convoluted streets of a new city. The city may offer a variety of places to visit but due to time and money restraints for some, it may be a perplexing task to figure which places best suits the traveler. So, to choose the most trendy spots to explore, it only makes sense if someone mines the data of all places and somehow finds the best place to visit in this new city. For this interest, a Location API like Foursquare can come in handy to precisely analyze the data and suggest spots to visit. This is another way where data science comes to rescue.

1.2 Problem

This project aims to guide tourists to select the most trending shopping malls in a predetermined radius (as selected by the user) from the current location of the tourist by clustering the shopping malls of the said region based on the "Likes" as given by other users to those venues. There will be three clusters that will divide the shopping malls into three classes namely – Above Average, Average and Below Average.

Data Processing

2.1 Data Source and Collection

The location data will be collected using the Foursquare API, which is a location service provider. The "Places API" of Foursquare Developer will be used for sourcing the data. In this case, the test data will be selected from Mumbai, India. To further focus on an area, Bandra is selected which is the most popular location for tourists to stay.

2.2 Data Preparation

Using the API of Foursquare, within the radius of 10000 meters, keeping Bandra as the center, a list of nearby shopping malls is generated. The raw data is converted to a JSON data frame for better understanding. A total of 30 shopping malls with 18 columns (features) have populated the data frame.

Then, only columns that include venue name and anything that is associated with location were kept and column names were appropriately edited. A venue ID list is generated to extract "Likes" from each venue. After getting like count of each venue, a list of likes for each venue is produced. This list was appended to the main data frame and then the entire table was sorted in descending order based on "Likes".

2.3 Feature Selection

Finally, only the venue name, the distance from the tourist and the corresponding likes of the venue are kept as the selected features from the aforementioned data frame. This is the final dataset that is used for feeding into the machine learning algorithm. All other columns/features are deemed unnecessary at this point for this project, so they were dropped.

Exploring the Data

3.1 Data Visualization

After normalizing the data and feeding it into the K Means clustering method, labels are assigned to each venue through which the data is visualized by separating it into three bins. The centroids of each cluster are then obtained. Further, the clusters are plotted using the matplotlib library and each cluster can be seen through the different scatter plots.

3.2 Result

The different clusters are displayed with their respective likes and distances from the given location.
Each cluster is displayed using Density-based spatial clustering.