**Peek-A-Business**

Titas Sarkar

May 4, 2020

**1. Introduction**

**1.1 Background**

The type pf business one wants to set up is chosen from a list of options. The user gives the location and choice of business as input. Major venue categories in the location along with their percentages are displayed after categorizing them into different clusters. The input business category is predicted. User is also required to provide some venue category names around which the user feels comfortable opening his business. Based on this information, new location coordinates are provided to the user.

**1.2 Problem**

The user wants to set up a business in a town/city but is unable to decide any particular location. A specific latitude and longitude is calculated and provided to the user.

**1.3 Interest**

Anyone who wants to open a business or a recreational venue and is unable to find a location to set up his/her business would be benefited from the project.

**2. Data acquisition and cleaning**

**2.1 Data Sources**

The only data source used is Foursquare.com. Data is retrieved by making the Foursquare API calls. No other data sources have been used in this project.

**2.2 Data Cleaning**

The API call request returns a json file which is normalized and converted into a dataframe. A new dataframe is created by retrieving only columns named 'name', 'categories', 'lat'(latitude) and 'lng'(longitude). Data in the 'categories' column is refined to make the machine learning algorithms more effective. Categories separated by a comma are split and only the string before comma is accepted. Unncessary white spaces are removed from the strings in 'categories' column. The input data is also refined to make easier prediction.

**2.3 Feature Selection**

The columns mentioned in the data cleaning sub-section are initially retrieved. And then for one hot encoding, the required columns are again retrieved from the dataframe. The only required column in this case is the 'categories' column.

**3. Exploratory Data Analysis**

After selecting the essential features or columns, in this case a single column 'categories', one hot encoding is performed and the result is stored in a variable.

**4. Clustering and Predicting**

**4.1 Clustering**

The clustering algorithm is divided into two parts due to the fact that the number of rows cannot be less than the number of k's. The 'sse' or the 'sum of the squared distance' is calculated to find out the optimum value of k. The algorithm is run 20 times to find the maximum optimum value of k. After finding the value of k, we fit the one hot encoded data with the number of clusters or the maximum optimum k value. The cluster label is inserted into the main dataframe.

**4.2 Predicting**

The major categories are formed into clusters and are portrayed on a folium map. Similar venue categories have the same color. One hot encoding is again performed on the existing 'categories' column. The one-hot encoded 'categories' column becomes the 'x' data while the 'cluster' column which was later added becomes the 'y' data. Training is performed using train\_test\_split with 0% test size. The refined input data is made into a row and is appended into the dataframe only if the input category matches with any of the categories in the dataframe. The category in which the input data shoud fall is predicted by using the support vector machine algorithm.

**5. Finding Location**

The user is asked to enter a number of venue categories according to the user's choice. Based on this input, the average latitude and longitude of each venue category is calculated. The final coordinates are given by the mean of the calculatd average latitudes and longitudes.

**6. Results**

The major venue categories along with their percentages are displayed. The business or the recreational venue which the user wanted to set up is shown as in which category it falls. A tip is also given based on what type of business the user wants to open. Also, coordinates are provided to the user where the business could be set up. A popop appears at the coordinates which is diplayed on a folium map with all the other venues surrounding the popup.

**7. Conclusion**

These predictions are typically more accurate for bigger towns and cities where data is abundant. It might not be very accurate in case of small towns or might even fail in case of villages. But this prediction gets better as an as Foursquares updates it's database. The categorization of venues are made as best as possible. The best clustering occurs when the selected feature data is cleaned first. Also, the project is highly case-sensitive.