|  |
| --- |
| NEW Restaurant recommender system |
|  |

|  |  |
| --- | --- |
| 05/01/2020 | IBM Capstone project |

Restaurant recommender system is a machine learning model, developed to demonstrate as a capstone project to IBM through coursera. It recommends restaurants based on available restaurants around. **Table of contents**

|  |  |
| --- | --- |
| **Topic** | **Page number** |
| Table of contents | 2 |
| Introduction | 3 |
| Data section | 4 |
| Methodology section | 7 |
| Result section | 8 |
| Discussion section | 11 |
| Conclusion section | 12 |

**Introduction:**

**Background:**

Running a restaurant is more than just offering food in exchange for money. You will be offering an experience to customers based on items such as the decor, food, and service. As the owner, it is up to you to determine the type of experience you want to give customers. Even after you've decided on the type of restaurant you want to open, you need to make sure there is a market for it's **cuisine**, and you can find **the right location**.

Recommendation systems are now very popular in many aspects such as e-commerce services, movie rating (Netflix), social networking and a lot of services helping users finding preferred items, in this work i will focus on (LBSN) location-based social networks.

Location-based social networks, e.g., **Foursquare**, Gowalla and Whrrl4, have seen soaring popularity, attracting millions of users. People are increasingly using these location-based social networking services to connect with friends, explore places (e.g., restaurants, stores, cinema theaters, etc.) and share their locations.

**Business Problem:**

Choosing a Location for Your Restaurant Location is vital to the success of any restaurant. There are several factors to consider when searching for that perfect restaurant location, including population base, local employment figures, and accessibility and a critically important factor is to **select an area where the competition is not intense for the food you are offering**.

**Interest:**

Our goal is to find a solution made for who is looking to open a restaurant in Toronto, by asking him what kind of food (cuisine) he will offer, i will recommend him which neighborhoods will not be a good choice for him. In this work, I will design a restaurant recommender system based on k-means an unsupervised clustering model, which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean of the frequency of occurrence of each category (Restaurant Cuisines).

# **DATA ACQUISITION AND CLEANING**

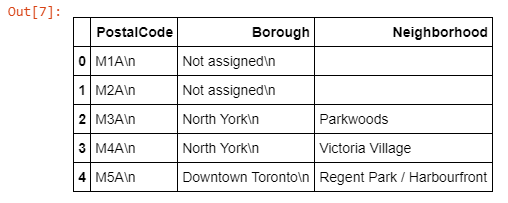
**Data sources:**

I will use a combination of data types:

* List of postal codes of Toronto from wikipeda.
* Geospatial Coordinates CSV file for Toronto postal codes from <http://cocl.us/Geospatial_data>.
* Foursquare API

**Data Cleaning:**

I will read first the HTML text using BeautifulSoup library of the List of postal codes of Toronto to group neighborhoods for each zip code.

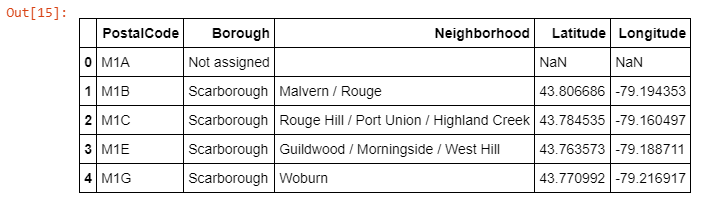


Group each zip code with all neighborhoods in:

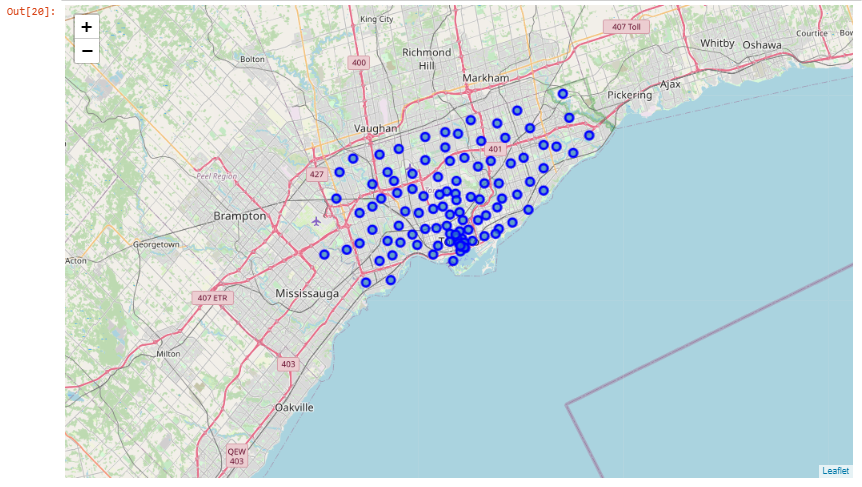


**Insert Geo-spatial Coordinates:**

I will insert Geospatial Coordinates csv file to the code from my IBM Cloud Object Storage, to add Latitude and Longitude for each zip code area and store it into pandas dataframe and I will perform a join with the previous dataframe to get a new dataframe with each zip code and neighborhood and the latitude and longitude coordinates.



Then we can plot a map of the neighborhoods using Folium Library:



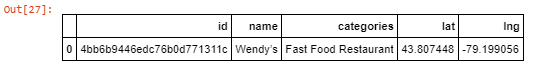
**Foursquare API utilizing:**

Finally I will utilize Foursquare API to explore neighborhoods and segment them.

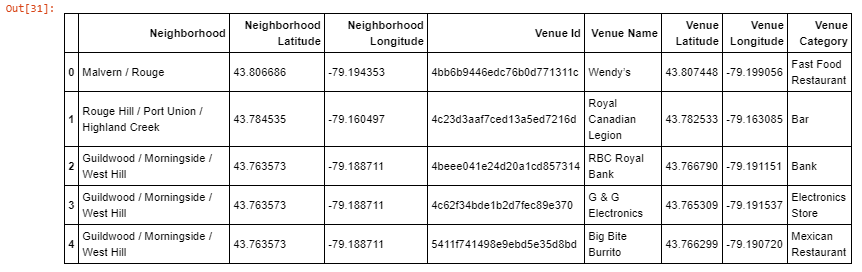
We can get the top 100 venues that are within a radius of 500 meters for a zip code area, the result is a json object. For example the result of Rouge, Malvern area will be.



We need to clean the json to extract venues categories and structure it into pandas dataframe.



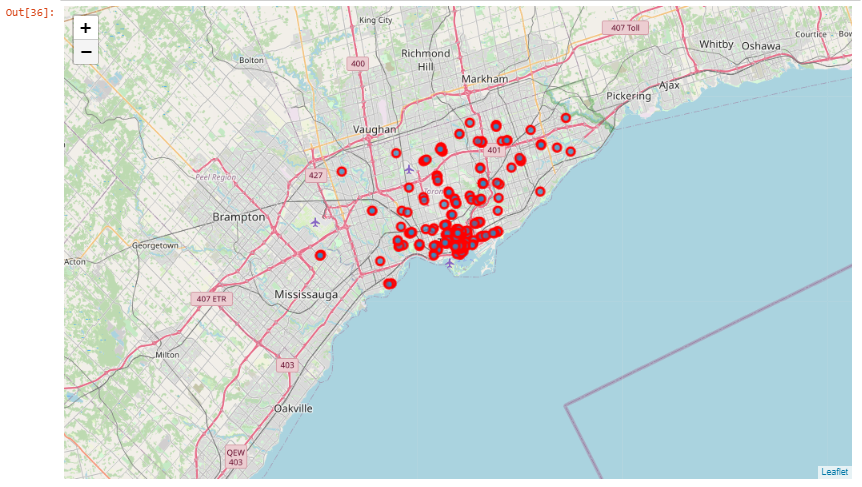
Then I will perform top venue exploration for all neighborhoods and merge it to one data frame, that will include neighborhoods and their top rated venues, and because our main goal is to recommend a neighborhood for a new restaurant, we need to filter our data, so we will have only venues categories that contains the word 'Restaurant'



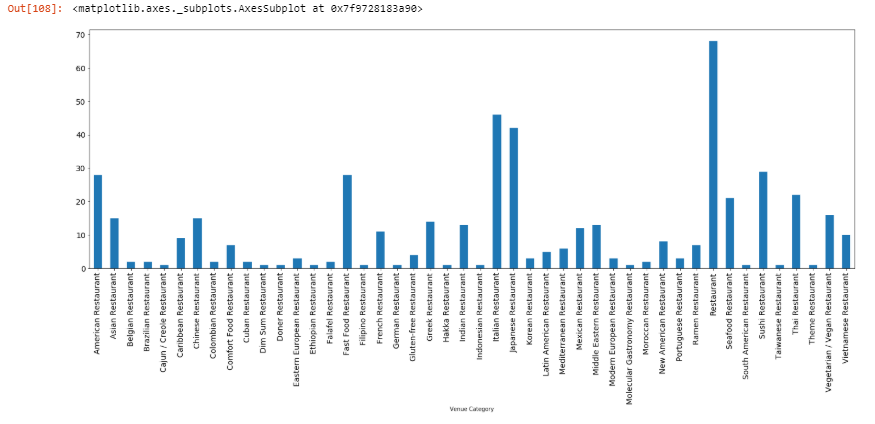
# **Methodology section (EDA - Exploratory Data Analysis)**

From the dataset, we filtered out the neighborhood which have venue contains the word 'Restaurant'. After filtering out the data, our dataset shape is (514, 9) that means that the data contains 514 rows (Restaurant venues) and 9 relevant columns.

Here is a map of all the top rated restaurant in Toronto:



Here is the distribution of the restaurant cuisines:

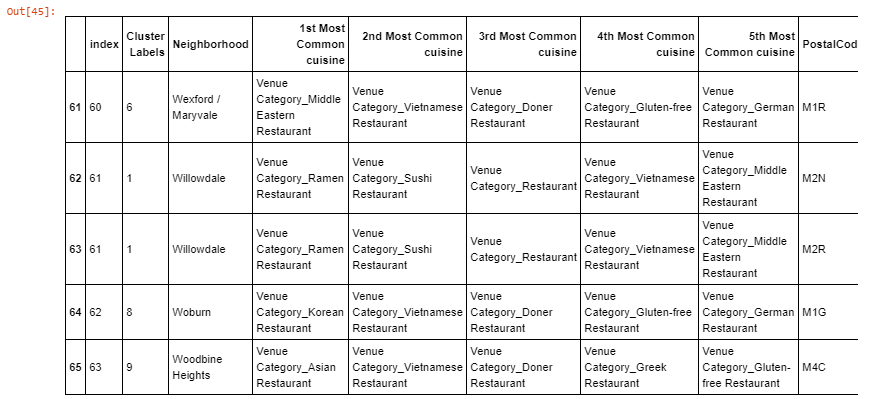


# **Results section**

We can see the most frequent restaurant type is a ‘General Restaurant’ show in light blue. Then for each one of the neighborhood we can extract the top frequent restaurant categories:



We can see for ‘Adelaide, King’ Neighborhoods Thai Restaurant is the most frequent category with 0.15 frequency mean of all the top rated restaurants, and then in the second place is the American Restaurant with 0.11 frequency mean.



# Clustering Modeling

K-means clustering is a type of unsupervised learning, which is used when you have unlabeled data (i.e., data without defined categories or groups). The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable K. The algorithm works iteratively to assign each data point to one of K groups based on the features that are provided. Data points are clustered based on feature similarity. The results of the K-means clustering algorithm are:

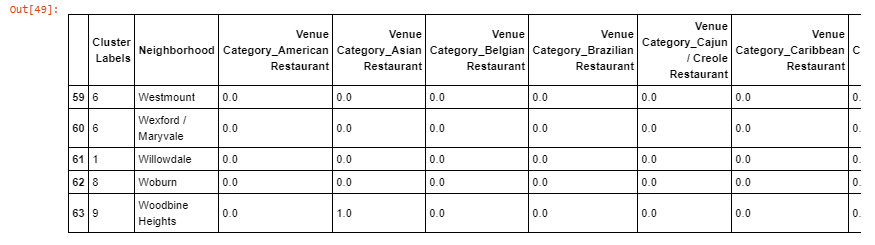
* The centroids of the K clusters, which can be used to label new data.
* Labels for the training data (each data point is assigned to a single cluster)

Rather than defining groups before looking at the data, clustering allows you to find and analyze the groups that have formed organically. The "Choosing K" section below describes how the number of groups can be determined .I will use popularity recommendation filtering approach in order to make recommendations.

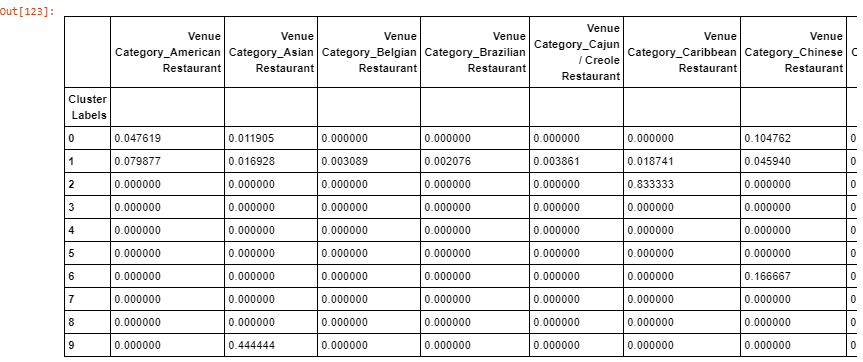
**Popularity Recommendation:** Given the number of play counts (occurrences) for items songs, movies, whatever, just sort by descending ordering and recommend everyone what’s popular. This actually makes a good shelf for any content platform, such as Netflix, Spotify or Amazon.

I performed a qualitative comparison of the existing techniques used in the Data I got from Foursquare, followed by a clustering on the type of the services and the location features the Foursquare utilizes to perform the recommendations.

First of all I performed one hot encoding on the data to get a binary represented data set for each neighborhood venue (Restaurant)



and then grouped rows by neighborhood and by taking the mean of the frequency of occurrence of each category I got :



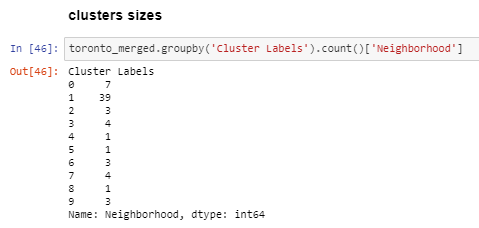
S dataframe representing each neighborhood and the mean frequency of occurrence for each restaurant cuisine in it. In my solution I chooses K = 10, so I want to split neighborhoods to 10 clusters:

For the first 10 neighborhoods, the model result is :

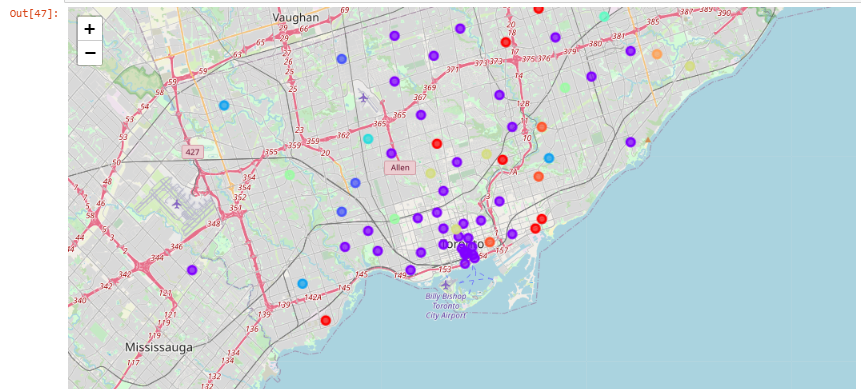
First neighborhood is in cluster number 2, second neighborhood in cluster number 3, third neighborhood in cluster number 8 and so on…

# **Discussion section**

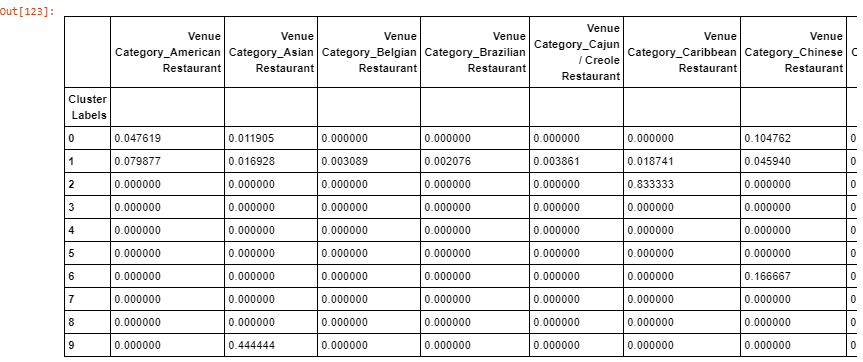
For simplicity we can see clusters size:



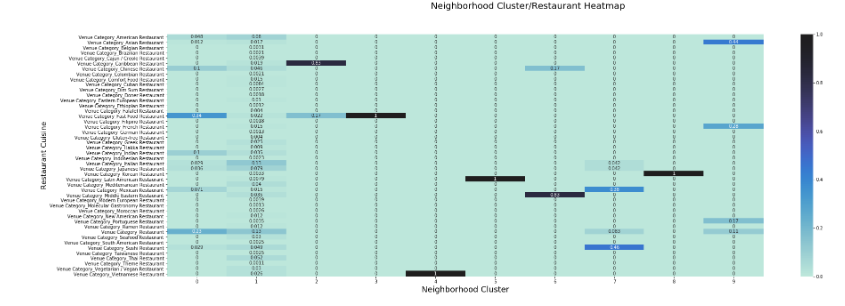
Easily we can see that cluster 2 is the biggest one, that means that all neighborhoods in cluster 2 are similar with top rated restaurants cuisines, now, let’s plot a map of neighborhood and colored clusters:



In this stage, I merged cluster labels to the dataset, so the new dataset will be:

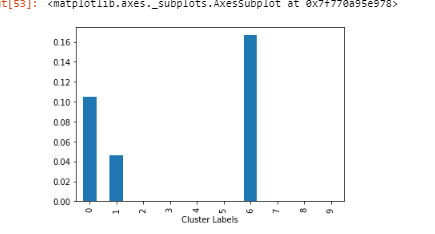


Then I grouped the dataset by cluster to get the mean of the frequency occurrence of occurrence of each cluster, we then I examined each cluster by a heatmap:



# **Conclusion section**

Conclusion can be taken from the previous heatmap, Dark color means that the restaurant cuisine is very common in the corresponding neighborhood cluster, for example if someone is willing to open a new Chinese Restaurant, we can recommend him where not to open his new restaurant.



Clusters 6,1 & 0 has a high frequency of Chinese restaurants, so try not to open a new Chinese restaurant in clusters 6,1 and 0 neighborhoods.