Capstone Project – The Battle of Neighborhoods Opening Chinese Restaurant in Toronto

Carol Wang

1. Introduction :

Problem background:

Toronto is Canada's Largest city and a world leader in such areas as business, finance, technology, entertainment and culture. With a recoded population of 2,731,571 in 2016, it is the most populous city in Canada and the fourth most populous city in North America. Toronto's large population of immigrants form all over the globe has also made Toronto one of the most multicultural cities in the world.

One of the best things about living in the world's most multicultural city is all the tasty and diverse food options. The diversity of the cuisine available is reflective of the social and economic diversity of Toronto. Whether you are in the mood for Italian, Mexican, Chinese or Ethiopian, you won't have a problem finding a delicious meal to satisfy your craving in Toronto.

Business Understanding/Problem Description:

In this project, we will find the best neighborhood in Toronto for opening a new Chinese restaurant.

There are some questions we need to address:

1. People live in which neighborhood likely eat in this kind of restaurant?
2. How many "similar" restaurants are available nearby?
3. Do the "similar" restaurants cost more? If so, what specialty do that have?

Target Audience:

Target audiences for this project does not limit to a person who keeps travelling but everyone. People could simply decide to look for a similar restaurant all the time because they are addicted to a specific category of food. People who rarely use restaurants would prefer to have the most rated

restaurants nearby them and all this could be easily handled by our recommender system. So target for this project is basically everyone who is exploring different places or similar places.

2. Data :

To find a solution to the questions and build a recommender model, we need data and lots of data.

Here are some questions we need to consider but not limited to when gathering data to suggest a location for new restaurant:
1. Geographical coordinates (latitude and longitude) to find out where exactly it is located.
2. Population of the neighborhood where the restaurant is located.
3. Average income of neighborhood to know how much is the restaurant worth.

Data Source:

1. List of postal code for communities in Toronto from Wikipedia: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
2. Geospatial Coordinates CSV file for Toronto postal codes from http://cocl/us/Geospatial_data
3. Foursquare API venues explore methods to get venues of given neighborhoods
4. Foursquare API venues methods to get ranks and likes of restaurants by given venue id

Methodology:

1. I will use of demographic data by neighborhoods, which allows me to gauge people's taste in each neighborhood. For example, I assume areas with a majority of Chinese people would be prefer to go to Chinse restaurant.

2. I will use the Foursquare API to explore neighborhoods in Toronto and apply the explore function to get the most common venue categories in each of the neighborhoods.

3. I will then use this feature to group the neighborhoods into clusters and use the k-means clustering algorithm to refine the groupings.

4. Finally, I will use the Folium library to visualize the neighborhoods in Toronto and their emerging clusters as well as push the right neighborhood for setting up the restaurant business.

3. Exploratory Data Analysis:

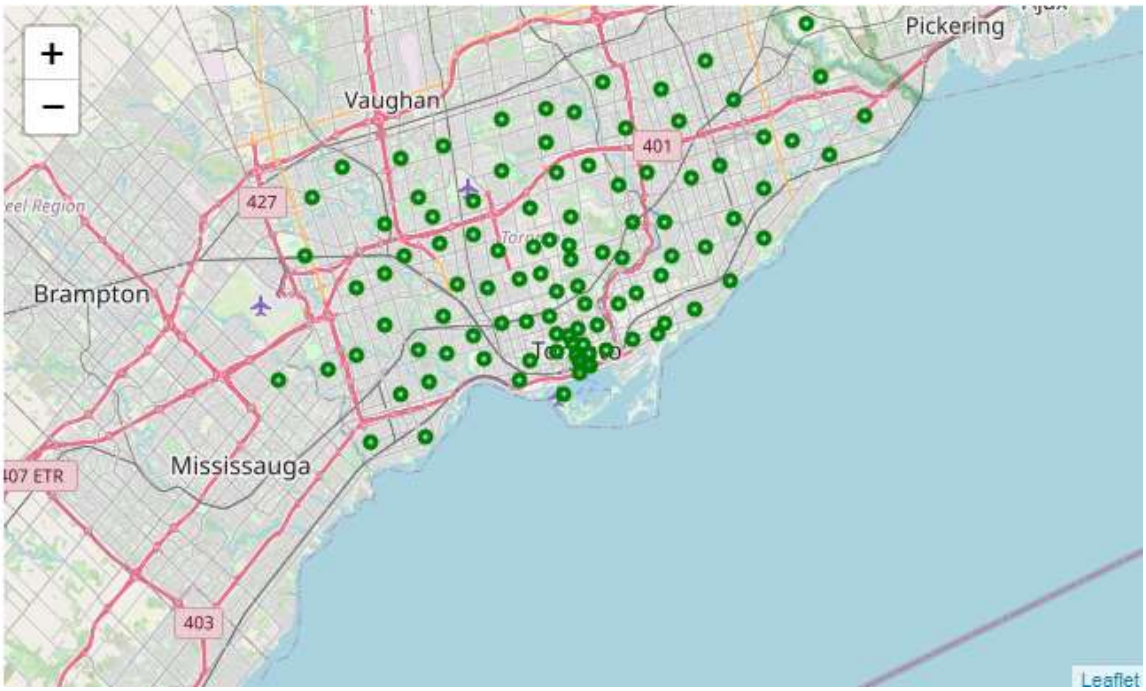
Data Cleaning:
I will use BeautifulSoup library to scrape Wikipedia to list of postal codes for each community. Remove the empty data in each row and group each zip code with all neighborhoods:

	Postcode	Borough	Neighbourhood
0	M3A	North York	Parkwoods
1	M4A	North York	Victoria Village
2	M5A	Downtown Toronto	Regent Park, Harbourfront
3	M6A	North York	Lawrence Manor, Lawrence Heights
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Then geospatial data (coordinates) is read into pandas dataframe. Then joined with the previous dataframe to get a new dataframe with zip code, neighborhood, latitude, and longitude coordinates.

	Borough	Neighbourhood	Postcode	Latitude	Longitude
0	Scarborough	Malvern, Rouge	M1B	43.806686	-79.194353
1	Scarborough	Rouge Hill, Port Union, Highland Creek	M1C	43.784535	-79.160497
2	Scarborough	Guildwood, Morningside, West Hill	M1E	43.763573	-79.188711
3	Scarborough	Woburn	M1G	43.770992	-79.216917
4	Scarborough	Cedarbrae	M1H	43.773136	-79.239476

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



Foursquare API

I will use Foursquare API to explore neighborhoods and merge it to dataframe including neighborhoods and top-rated venues.

	neighborhood	neighborhood latitude	neighborhood longitude	venue	venue latitude	venue longitude	venue category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	SEBS Engineering Inc. (Sustainable Energy and ...	43.782371	-79.156820	Construction & Landscaping
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store

Features

When extracting features from a dataset, it is often useful to transform categorical features into vectors. Then group them by neighborhood by taking the mean of the frequency of occurrence.

	neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Ac
0	Agincourt	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	
1	Alderwood, Long Branch	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	
2	Bathurst Manor, Wilson Heights, Downsview North	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	
3	Bayview Village	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	
4	Bedford Park, Lawrence Manor East	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.038462	0.0	

Then a dataframe is created with top types of venues around each neighborhoods in Toronto:

	neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Mc Comn Ver
0	Agincourt	Lounge	Latin American Restaurant	Skating Rink	Clothing Store	Breakfast Spot	Doner Restaurant	Dim Sum Restaurant	Diner	Discount Store	Distribut Cen
1	Alderwood, Long Branch	Pizza Place	Coffee Shop	Gym	Skating Rink	Pharmacy	Pub	Sandwich Place	Pool	Yoga Studio	Dess Sh
2	Bathurst Manor, Wilson Heights, Downsview North	Bank	Coffee Shop	Pharmacy	Pizza Place	Restaurant	Deli / Bodega	Ice Cream Shop	Fried Chicken Joint	Frozen Yogurt Shop	Mid East Restaur
3	Bayview Village	Café	Bank	Japanese Restaurant	Chinese Restaurant	Deli / Bodega	Dessert Shop	Dim Sum Restaurant	Diner	Discount Store	Distribut Cen
4	Bedford Park, Lawrence Manor East	Sandwich Place	Coffee Shop	Italian Restaurant	Restaurant	Thai Restaurant	Pharmacy	Pizza Place	Indian Restaurant	Pub	C

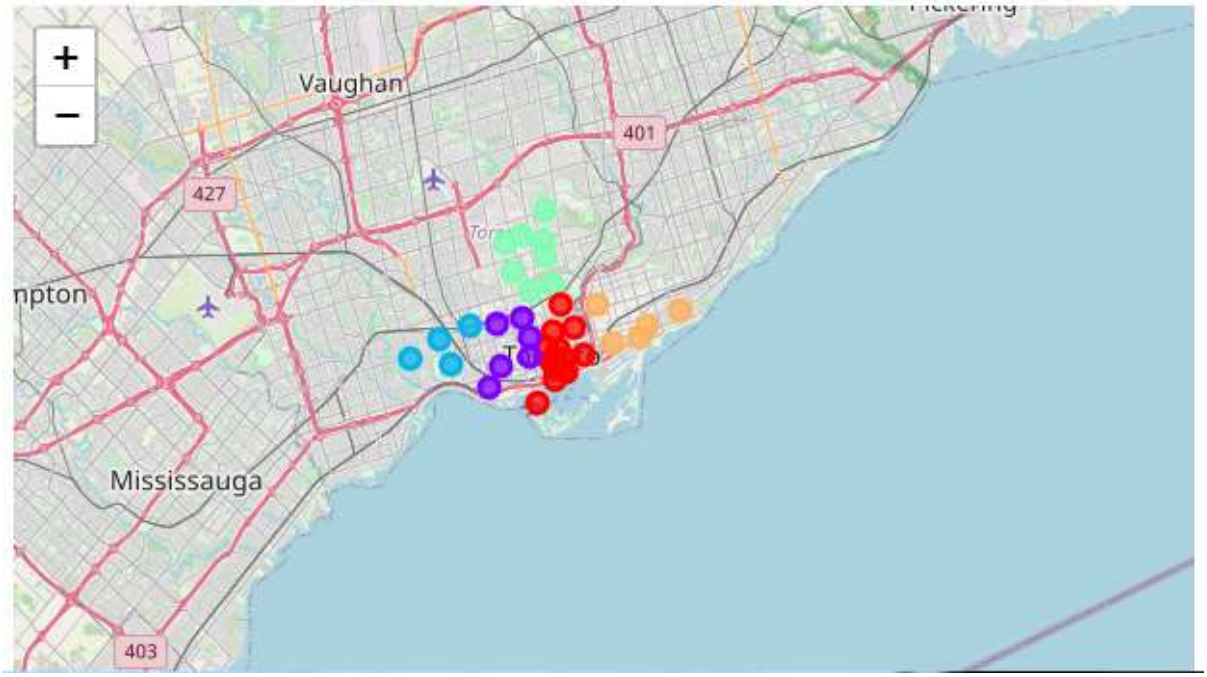
4. Clustering Modeling and Recommendations:

In clustering, I try to find homogeneous subgroup within our data such that data points in each cluster are as similar as possible according to a similarity measure such as Euclidean-based distance or correlation-based distance.

Clustering is considered an unsupervised learning method since we don't have the ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance. In this project, I use K-means clustering, which is a type of unsupervised learning, which is used when have unlabeled data. The K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups where each data point belongs to only one group. In this case, the clustering is based on types of nearby venues.

Then data is grouped into 5 clusters. The clusters and the top 10 venues for each neighborhood are shown below:

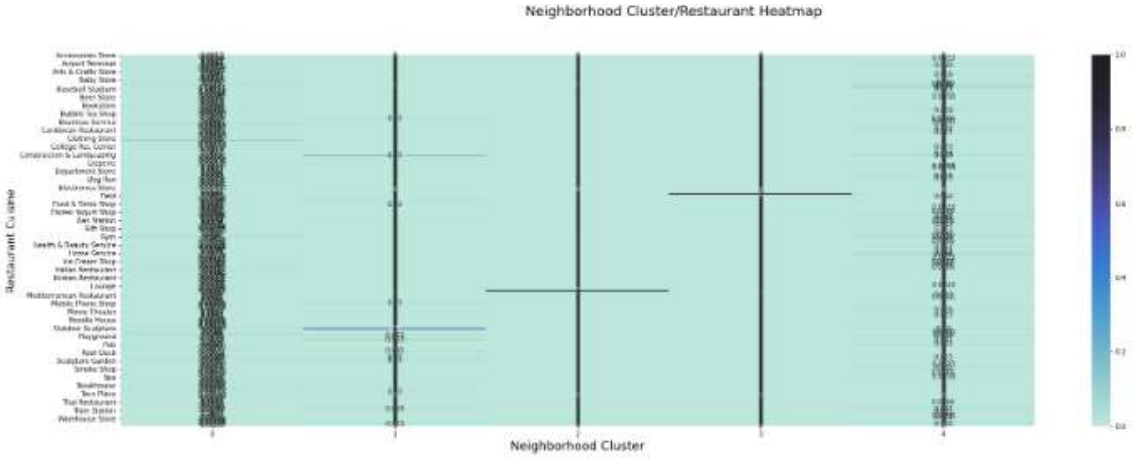
	Borough	Neighbourhood	Postcode	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5
0	Scarborough	Malvern, Rouge	M1B	43.806686	-79.194353	4.0	Fast Food Restaurant	Yoga Studio	Dog Run	Department Store	
1	Scarborough	Rouge Hill, Port Union, Highland Creek	M1C	43.784535	-79.160497	4.0	Bar	Construction & Landscaping	Yoga Studio	Doner Restaurant	D Re
2	Scarborough	Guildwood, Morningside, West Hill	M1E	43.763573	-79.188711	4.0	Electronics Store	Mexican Restaurant	Restaurant	Rental Car Location	E
3	Scarborough	Woburn	M1G	43.770992	-79.216917	1.0	Coffee Shop	Korean Restaurant	Yoga Studio	Dog Run	
4	Scarborough	Cedarbrae	M1H	43.773136	-79.239476	4.0	Fried Chicken Joint	Gas Station	Hakka Restaurant	Bakery	

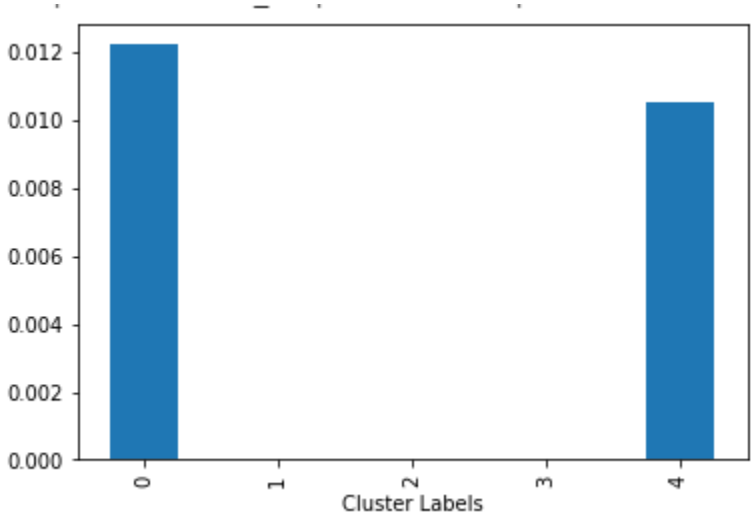


Number of restaurant in each of the 5 clusters are listed below:

```
Cluster Labels
0.0    65
1.0    11
2.0     1
3.0     1
4.0    22
```

Then dataset is grouped by clusters to get the means of the frequency occurrence of each cluster, and examined by a heatmap.





5. Conclusions:

Conclusions can be taken from the previous maps, tables, and exploratory data analysis. For example, cluster 1 and 4 has high frequency of Chinese restaurants, so try not to open a new Chinese restaurant in cluster 1 or 4 neighborhoods.