Recommend System for opening Asian Restaurant in Toronto

Yu Wang
June 12, 2020

1. Introduction

1.1 Background

Toronto is a city consist of people from all over the world. This is a open, fair and multi-cultures city. It also means that they are many different cultures' special food in Toronto. The idea of this project is that you want to open a featured and special food restaurant in Toronto. For me, I will open a Sichuan food restaurant. Sichuan food is one piece of Chinese food and also very similar to Asian food. As a result, if I want to open a successful Sichuan restaurant, I firstly need to find a suitable location. So, I decide to use this project to help us find suitable locations for opening Sichuan restaurant.

1.2 Business Problem

The objective of this capstone project is to find the most suitable location for opening a new Sichuan food restaurant in Toronto, Canada. By using data science methods and machine learning methods, such as clustering, K-means, Numpy, Pandas and etc. As a result, the business question: In Toronto, if an entrepreneur want to open a Sichuan restaurant, where should they consider opening it ?

1.3 Interest

The people who want to open a local special restaurant in Toronto, Canada.

2. Data Acquisition and Cleaning

2.1 Data Sources

- Getting Toronto neighborhoods and boroughs data via Wikipedia
- Getting Latitude and Longitude data of these neighborhoods and boroughs via Geocoder package
- Using Foursquare API to get venue data related to these neighborhoods.

3. Analyze and Methodology

3.1 Processing Data

We download and scrap data from multiple sources, such as Wikipedia, and using pandas methods to remove empty values of neighborhoods and boroughs of data, aggregating all neighborhoods of same borough. Then getting coordinates data from Geocoder package, and merging coordinates data with neighborhoods data into a new data which contain all data of these two tables.

3.2 Analyzing data

- **Simply analyze**: First, we want to find the most suitable location in a borough with a target feature. Therefore, we calculate average latitude and longitude of borough which we want to

analyze, then we use folium package to draw map of this borough and add markers of all neighborhoods in this map. Second, we used Foursquare API to pull list of top 100 venues within 500 meters. I have created a developer account and API key to pull the data. From Foursquare I am able to pull the names, categories, latitude and longitude of the venues. Then I get the target borough (Downtown Toronto) data, I can check how many unique categories and analyze each neighborhood by grouping the rows by neighborhood and taking the mean on the frequency of occurrence of each venue category. This is to prepare clustering to be done later. Third, we use index of Chinese Restaurant as our factor to cluster data, we firstly convert all venues categories into one hot code and get mean result. Lastly, I cluster data method by using k-means clustering, K-means clustering algorithm identifies K number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible, K-means is a simple and popular unsupervised machine learning algorithms. In this lab, we set K equal to 3, it means we clustered the neighborhoods in Toronto into 3 clusters based on their frequency of occurrence for Chinese Food. Based on clustering result, we can recommend the ideal place to our customer

- **Further Analyze:** In the further study, we make some improvements in this lab, firstly, we enlarge our search area from a borough to the whole Toronto, secondly, we use multiple factors, not only Chinese food, but many other Asian food, like Korean, and Japan food. Just like Simply Analyze, we also use Foursquare API and folium API key to analyze data, in this lab, we set cluster equal to 5, and we set key words including: Chinese food, Korean food, Japanese food, Thai food, Taiwanese food. then we merge these indexes into one index by adding these index's value together as total index. Then we cluster all data into 5 clusters again, and analyze final result to recommend to our customers who want to open a Sichuan restaurant in Toronto.

4. Detail Process

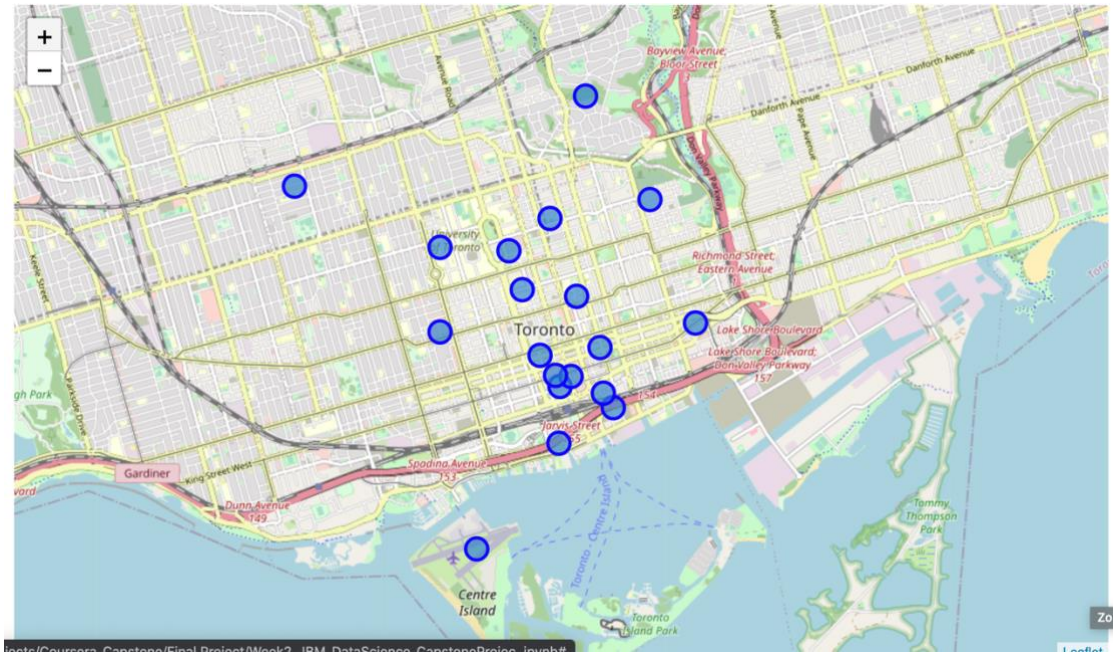
- I get data from Wikipedia and aggregate all neighborhoods of one borough.

	Postal Code	Borough	Neighborhood
0	M1B	Scarborough	Malvern, Rouge
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek
2	M1E	Scarborough	Guildwood, Morningside, West Hill
3	M1G	Scarborough	Woburn
4	M1H	Scarborough	Cedarbrae

- Merge coordinate table with Neighborhoods table into a new table.

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

- Using a folium to draw a map of Toronto Downtown borough's neighborhoods.



- Using Foursquare API to pull top 100 venues within 500 meters.

```
  # Use Forsquare API to get top 100 venues, radius = 500 meters.
radius = 500
LIMIT = 100

venues = []

for lat, lng, post, neighbourhood in zip(df_Downtown['Latitude'], df_Downtown['Longitude'], df_Downtown['Postal Code'],
    url = 'https://api.foursquare.com/v2/venues/explore?&client_id={}&client_secret={}&v={}&ll={},{}&radius={}&limit={}
        CLIENT_ID,
        CLIENT_SECRET,
        VERSION,
        lat,
        lng,
        radius,
        LIMIT)

    result = requests.get(url).json()['response']['groups'][0]['items']

    for venue in result:
        venues.append((
            post,
            neighbourhood,
            lat,
            lng,
            venue['venue']['name'],
            venue['venue']['location']['lat'],
            venue['venue']['location']['lng'],
            venue['venue']['categories'][0]['name']))
```

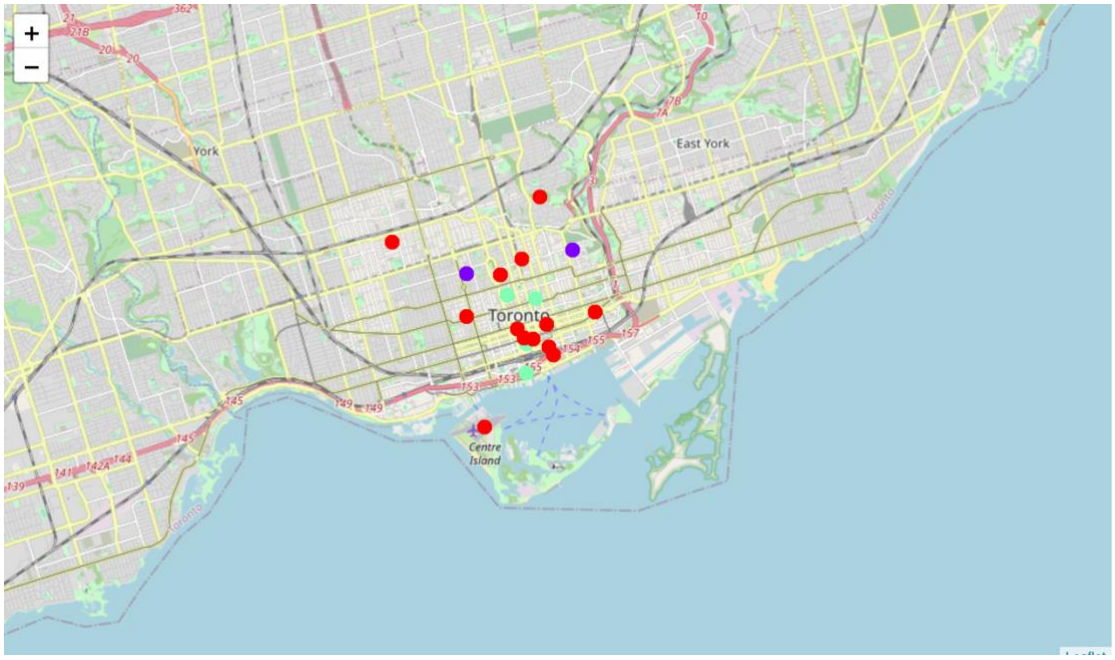
- Convert all venues into one hot table.

	Neighborhoods	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	Art Gallery	Art Museum	Arts & Crafts Store
0	Berczy Park	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.017857	0.000000	0.000000
1	CN Tower, King and Spadina, Railway Lands, Har...	0.000000	0.055556	0.055556	0.055556	0.111111	0.166667	0.111111	0.000000	0.000000	0.00	0.000000	0.000000	0.000000
2	Central Bay Street	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.015385	0.000000
3	Christie	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	0.000000	0.000000	0.000000
4	Church and Wellesley	0.012821	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.012821	0.000000	0.00	0.000000	0.000000	0.012821
5	Commerce Court, Victoria Hotel	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.040000	0.000000	0.00	0.010000	0.000000	0.000000

- Cluster Data and add cluster column into original Data Frame

	Neighborhood	Chinese Restaurant	Clusters Label	Postal Code	NeighborhoodLatitude	NeighborhoodLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	St. James Town, Cabbagetown	0.041667	1	M4X	43.667967	-79.367675	Cranberries	43.667843	-79.369407	Diner
1	St. James Town, Cabbagetown	0.041667	1	M4X	43.667967	-79.367675	F'Amelia	43.667536	-79.368613	Italian Restaurant
2	St. James Town, Cabbagetown	0.041667	1	M4X	43.667967	-79.367675	Butter Chicken Factory	43.667072	-79.369184	Indian Restaurant
3	St. James Town, Cabbagetown	0.041667	1	M4X	43.667967	-79.367675	Kingyo Toronto	43.665895	-79.368415	Japanese Restaurant
4	St. James Town, Cabbagetown	0.041667	1	M4X	43.667967	-79.367675	Murgatroid	43.667381	-79.369311	Restaurant

- Then map again with clusters which has different colors represent different clusters.



- Getting detail data of one cluster.

Out[35]:

	Neighborhood	Chinese Restaurant	Clusters Label
10	Queen's Park, Ontario Provincial Government	0.0	0
3	Christie	0.0	0
4	Church and Wellesley	0.0	0
5	Commerce Court, Victoria Hotel	0.0	0
6	First Canadian Place, Underground city	0.0	0
1	CN Tower, King and Spadina, Railway Lands, Har...	0.0	0
13	Rosedale	0.0	0
0	Berczy Park	0.0	0
12	Richmond, Adelaide, King	0.0	0
14	St. James Town	0.0	0
16	Stn A PO Boxes	0.0	0
11	Regent Park, Harbourfront	0.0	0
9	Kensington Market, Chinatown, Grange Park	0.0	0

5. Conclusion

In this lab, I have done two analyze, one is trying to find the best suitable location in Toronto Downtown Borough to open a Chinese Restaurant. Another is trying to find the best suitable location in the whole Toronto area to open a Asian food restaurant, Sichuan food restaurant.

In the first topic, we find these conclusions and predictions from K-means clusters data. We can see that Cluster 1 has most Chinese restaurant which values is highest, and cluster 0 has least Chinese Restaurant which value is equal to 0. So, from these cluster's value, we prefer to open a Chinese restaurant in neighborhood which are belonged to Cluster 1. because these place (like CN Tower, King and Spadina, Railway Lands, etc.) value is equal to 0, it means there rarely have Chinese restaurants, so less same type restaurants less competition, meanwhile, it also means more opportunities to earn more money. As a result, it your restaurant's food is good taste, localization and enough features, I am confident that you will get success. There are clusters pictures as follow.

In [35]:
```
# get detail data of every cluster
to_merge.loc[to_merge['Clusters Label'] == 0]
```

Out[35]:

	Neighborhood	Chinese Restaurant	Clusters Label
10	Queen's Park, Ontario Provincial Government	0.0	0
3	Christie	0.0	0
4	Church and Wellesley	0.0	0
5	Commerce Court, Victoria Hotel	0.0	0
6	First Canadian Place, Underground city	0.0	0
1	CN Tower, King and Spadina, Railway Lands, Har...	0.0	0
13	Rosedale	0.0	0
0	Berczy Park	0.0	0
12	Richmond, Adelaide, King	0.0	0
14	St. James Town	0.0	0
16	Stn A PO Boxes	0.0	0
11	Regent Park, Harbourfront	0.0	0
9	Kensington Market, Chinatown, Grange Park	0.0	0

```
In [36]: to_merge.loc[to_merge['Clusters Label'] == 1]
```

Out[36]:

	Neighborhood	Chinese Restaurant	Clusters Label
15	St. James Town, Cabbagetown	0.041667	1
18	University of Toronto, Harbord	0.028571	1

```
In [37]: to_merge.loc[to_merge['Clusters Label'] == 2]
```

Out[37]:

	Neighborhood	Chinese Restaurant	Clusters Label
2	Central Bay Street	0.015385	2
17	Toronto Dominion Centre, Design Exchange	0.010000	2
7	Garden District, Ryerson	0.010000	2
8	Harbourfront East, Union Station, Toronto Islands	0.010000	2

In the second topic. We can see from value of Cluster Label, we find that Cluster 2 has the most Asian restaurant, because cluster 2's Asian restaurant index is maximum. On the contrary, cluster 0 has lowest value which are equal to 0. So, from business view, we better to open a restaurant in a area where has least same type restaurants. As a result, from above data, we really recommend you open a Sichuan restaurant in these areas in Cluster 0. Clusters pictures as follow.

But good location is only one factor of getting successful, you need to have good taste, better service and good localized, etc. we can't make everything perfect, so choose a suitable location is very important.

```
In [96]: final_toronto_mergerd.loc[final_toronto_mergerd['Cluster Label'] == 0]
```

]:

	Neighborhood	Asia restaurant index	Cluster Label	PostalCode	Borough	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Agincourt	0.000000	0	Scarborough	M1S	43.794200	-79.262029	Panagio's Breakfast & Lunch	43.792370	-79.260203	Breakfast Spot
1	Agincourt	0.000000	0	Scarborough	M1S	43.794200	-79.262029	El Pulgarcito	43.792648	-79.259208	Latin American Restaurant
2	Agincourt	0.000000	0	Scarborough	M1S	43.794200	-79.262029	Twilight	43.791999	-79.258584	Lounge
3	Agincourt	0.000000	0	Scarborough	M1S	43.794200	-79.262029	Mark's	43.791179	-79.259714	Clothing Store
4	Agincourt	0.000000	0	Scarborough	M1S	43.794200	-79.262029	Commander Arena	43.794867	-79.267989	Skating Rink
11	Alderwood, Long Branch	0.000000	0	Etobicoke	M8W	43.602414	-79.543484	Alderwood Pool	43.601802	-79.547247	Pool
	Alderwood, Long										

```
In [97]: final_toronto_mergerd.loc[final_toronto_mergerd['Cluster Label'] == 1]
```

	O'Shanter, Sullivan										
359	Clarks Corners, Tam O'Shanter, Sullivan	0.166667	1	Scarborough	M1T	43.781638	-79.304302	Remezzo Italian Bistro	43.778649	-79.308264	Italian Restaurant
360	Clarks Corners, Tam O'Shanter, Sullivan	0.166667	1	Scarborough	M1T	43.781638	-79.304302	The Royal Chinese Restaurant 避風塘小炒	43.780505	-79.298844	Chinese Restaurant
362	Clarks Corners, Tam O'Shanter, Sullivan	0.166667	1	Scarborough	M1T	43.781638	-79.304302	TD Canada Trust	43.779169	-79.303617	Bank
363	Clarks Corners, Tam O'Shanter, Sullivan	0.166667	1	Scarborough	M1T	43.781638	-79.304302	Kub Khao	43.780438	-79.299837	Thai Restaurant
	Clarks										

```
In [98]: final_toronto_mergerd.loc[final_toronto_mergerd['Cluster Label'] == 2]
```

Out[98]:

	Neighborhood	Asia restaurant index	Cluster Label	PostalCode	Borough	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
35	Bayview Village	0.5	2	North York	M2K	43.786947	-79.385975	Kaga Sushi	43.787758	-79.381090	Japanese Restaurant
34	Bayview Village	0.5	2	North York	M2K	43.786947	-79.385975	Maxim's Cafe and Patisserie	43.787863	-79.380751	Café
32	Bayview Village	0.5	2	North York	M2K	43.786947	-79.385975	Sun Star Chinese Cuisine 翠景小炒	43.787914	-79.381234	Chinese Restaurant
33	Bayview Village	0.5	2	North York	M2K	43.786947	-79.385975	TD Canada Trust	43.788074	-79.380367	Bank

```
In [100]: final_toronto_mergerd.loc[final_toronto_mergerd['Cluster Label'] == 4]
```

0]:

	Neighborhood	Asia restaurant index	Cluster Label	PostalCode	Borough	BoroughLatitude	BoroughLongitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
49	Bedford Park, Lawrence Manor East	0.041667	4	North York	M5M	43.733283	-79.419750	Bruno's Fine Food	43.736642	-79.419870	Butcher
59	Bedford Park, Lawrence Manor East	0.041667	4	North York	M5M	43.733283	-79.419750	Harvey's	43.730256	-79.418589	Restaurant
58	Bedford Park, Lawrence Manor East	0.041667	4	North York	M5M	43.733283	-79.419750	Baskin-Robbins	43.733314	-79.419452	Ice Cream Shop
57	Bedford Park, Lawrence Manor East	0.041667	4	North York	M5M	43.733283	-79.419750	Jolie Beauty Bar	43.732757	-79.419552	Cosmetics Shop
56	Bedford Park, Lawrence	0.041667	4	North York	M5M	43.733283	-79.419750	Pizza Pizza	43.731952	-79.415019	Pizza Place

6. Future directions

Everyone should know a suitable location from our analyze doesn't mean that you must get success in this place and your restaurant. Because we know that many factors can cause whether a restaurant can get popular in city, just like, a good taste is very important, and a good service also very crucial. What's more, I think localized is also necessary, because if local people don't like eating your food although your feature food is very authorized, it also makes you failed. And still have many factors that I don't mention above. Consequence, A good location doesn't means success. But what I really want to say is that if we could make most factors better than others, the highly percentage that we get success. So do not ignore every factor that you can change better, just like choose a better location to open a restaurant. Because a good location could make you have less competition and make you have more choice to show you food to people, if you food is good enough, people will eat again and again, and tell others , and then others will eat and tell their friends. This process will happen again and again, finally, you will find that your restaurant is famous in local area.

As a result, Do more thing better, you will more likely get successful!