IBM Data Science Professional Certificate: Capstone Project Report

Where is your familiar working and living place?

A recommendation system for finding neighborhoods with similar venues in a new city

Author: Zicao Fu
Email: zicao.fu@gmail.com

May 2020

This project aims to build a **recommendation system** which helps people to find neighborhoods in a new city with venues similar to those of their old living places.

Here I use myself as an example and try to look for neighborhoods in **New York** and **Toronto** that are most similar to my current working and living place, **Sha Tin district of Hong Kong**. Venues around a neighborhood are obtained from **Foursquare** and are then used to construct a feature set which characterizes that particular neighborhood. Afterwards, different neighborhoods in the new cities are compared to my current place based on their respective feature sets, and those that share most similarities with my current place are recommended and get marked on an interactive **Folium map**.

From the total number of recommended neighborhoods in a particular city, one can also see which city on the whole most resembles my current place, at least as far as living is concerned. In the case studied here, we will see whether it is New York or Toronto that is most similar to the Sha Tin district of Hong Kong.

This article is a project report. For the implementation details, please refer to my Jupiter Notebook at `https://tinyurl.com/fu-notebook` (GitHub) or `https://tinyurl.com/fu-notebook-WS` (IBM Watson Studio).

1 Introduction

1.1 Business Background

We currently live in an age of globalization. People travel around the world more frequently than in the past. More and more people face the problem of finding a neighborhood in a foreign city to live and work. Moving to a new neighborhood which contains many venues similar to his/her old working and living place can save the time and cost needed to adapt to the new environment. Thus, the issue of finding neighborhoods with similar venues in a new city becomes more and more relevant to many people. However, it is not a easy task,

especially when people need to make their decisions remotely and even before they first arrive the new city.

This is the place where data science can help. Utilizing the vast amount of location and venue information available online, if we can build a feature set for every neighborhood in the cities concerned, we can then use these feature sets to compare different neighborhoods with the customer's current working and living place (or any other places he/she picks) and come up with the top neighborhoods in those cities that most resemble this place. The result can aid customers to make a well-informed choice in deciding where to work and live.

As an analogy, the recommendation system which will be built here is similar in essence to that used by Amazon which suggests the next commodity for you to buy.

1.2 Business Problem

In this article, I use myself as an example for this project. Currently, I work and live in **Sha Tin district of Hong Kong**, which is the most populous district in Hong Kong and is away from the city center of Hong Kong. I will use the subdistricts in this district as a benchmark in my search for similar neighborhoods in two cities on the other side of the globe.

In Hong Kong, the words "borough" and "neighborhood" are not used to describe regions in the city. Instead, the words "district" and "subdistrict" are used. In this article, "district" and "borough" are used interchangeably. The two words "subdistrict" and "neighborhood" are used interchangeably as well.

Since my career goal is to apply data science to financial service industry, the next city which I might move into probably will be a financial center. I pick a major financial center in the US and a major financial center in Canada to do the battle (comparison) of neighborhoods. The two cities chosen in this study are **New York** and **Toronto**. The question which this project answers is stated as: "**Which 20 neighborhoods in New York and Toronto do have sets of venues most similar to those in Sha Tin district of Hong Kong?**"

The main objective is to recommend a total number of 20 neighborhoods in these two cities. From the recommendation, we can also learn which city has the largest number of recommended neighborhoods, thus most similar to my current working and living place, Sha Tin district of Hong Kong. Choosing a right neighborhood is a daunting task. Therefore, the analysis done here would be interesting to **anyone who plans to move into a new city**.

2 Data

Based on the business problem described above, there are two data sets needed. The first data set contains data of neighborhoods with latitudes and longitudes in concerned cities. The second data set contains data of the information of venues in concerned neighborhoods.

2.1 Neighborhood Location Data

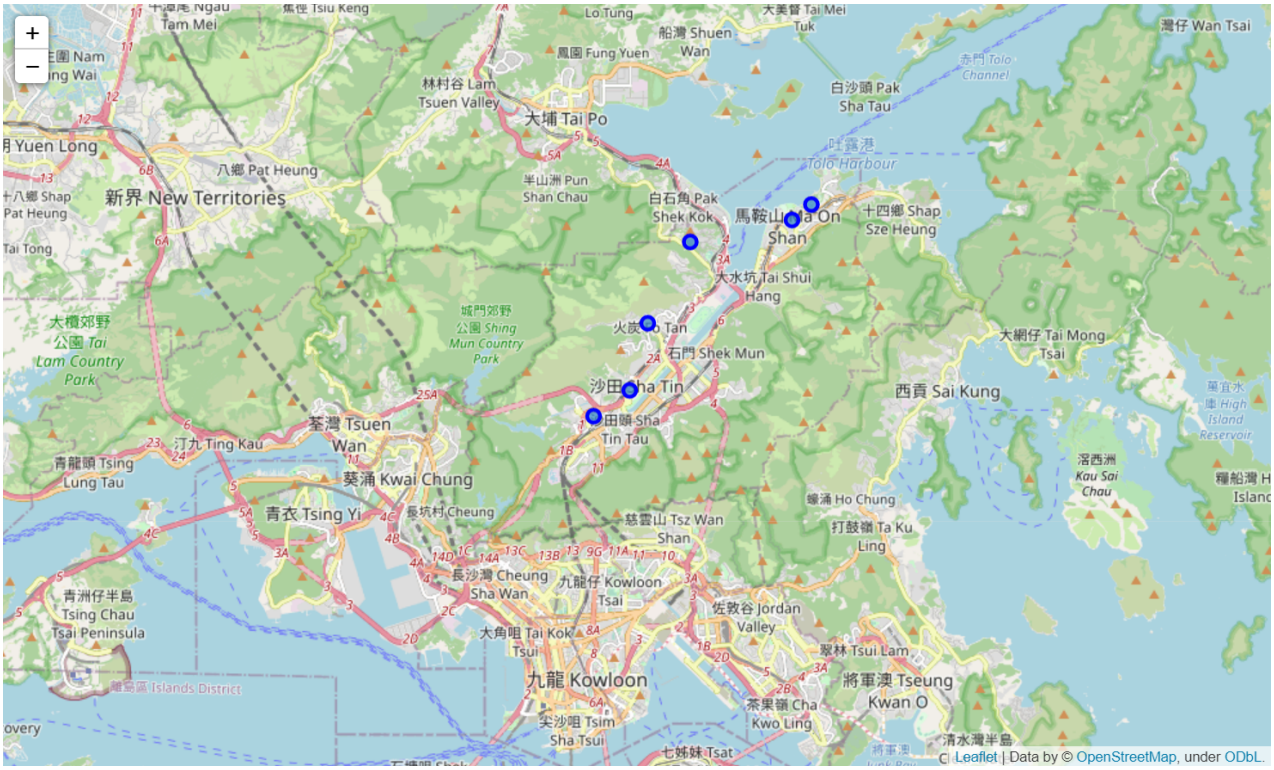
For the first data set, we need data in the form of tables with five columns: city, borough, neighborhood, latitude, longitude. We need three such tables, respectively for Sha Tin district of Hong Kong, New York, and Toronto. In addition, in order to make data processing easy,

we will create a master data frame `TwoCities` for the two tables for New York and Toronto combined and a grand master data frame `ThreeCities` for all three tables combined.

2.1.1 Sha Tin, Hong Kong

The list of subdistricts of Sha Tin can be found at the document "Areas and Districts" published by Hong Kong government at `https://www.rvd.gov.hk/doc/tc/hkpr13/06.pdf`. Since there are only six neighborhoods, the data frame can be created by inputting the values by hand. Below is the full resulting data frame, `Sha_Tin_Data`, where all the latitude and longitude data are obtained from Google map.

	City	Borough	Neighborhood	Latitude	Longitude
0	Hong Kong	Sha Tin	Tai Wai	22.375944	114.178626
1	Hong Kong	Sha Tin	Sha Tin	22.382165	114.188160
2	Hong Kong	Sha Tin	Fo Tan	22.398566	114.192940
3	Hong Kong	Sha Tin	Ma Liu Shui	22.418777	114.204337
4	Hong Kong	Sha Tin	Wu Kai Sha	22.427887	114.236608
5	Hong Kong	Sha Tin	Ma On Shan	22.424187	114.231480

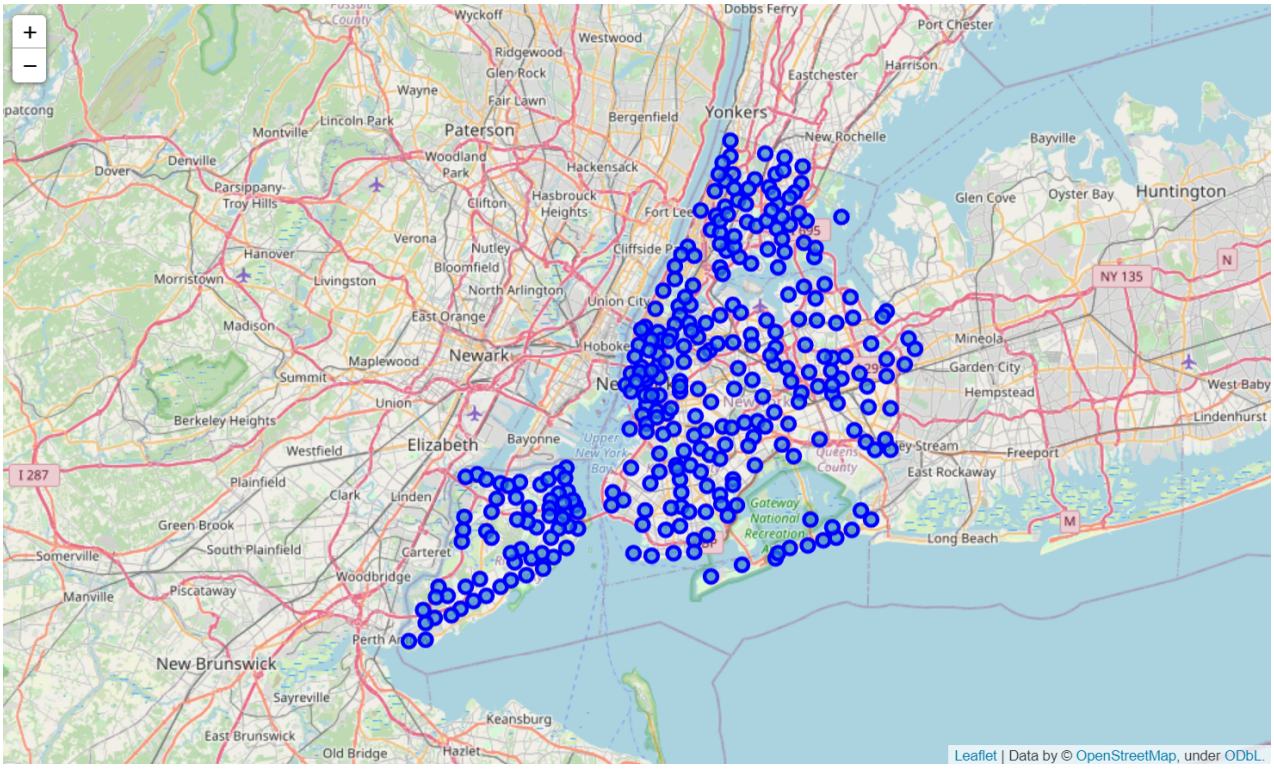


Above is a Folium map centered at Sha Tin, Hong Kong, with blue dots denoting the six neighborhoods.

2.1.2 New York

We have already created a location data frame for New York in our example project of exploring neighborhoods in New York. There, the raw data was fetched from the json file at `https://cocl.us/new_york_dataset`. I simply repeat the procedures in that example project. The first five rows in the location data frame for New York, `NYLocation`, is shown below. The data frame has five boroughs and 306 neighborhoods.

	City	Borough	Neighborhood	Latitude	Longitude
0	New York	Bronx	Wakefield	40.894705	-73.847201
1	New York	Bronx	Co-op City	40.874294	-73.829939
2	New York	Bronx	Eastchester	40.887556	-73.827806
3	New York	Bronx	Fieldston	40.895437	-73.905643
4	New York	Bronx	Riverdale	40.890834	-73.912585



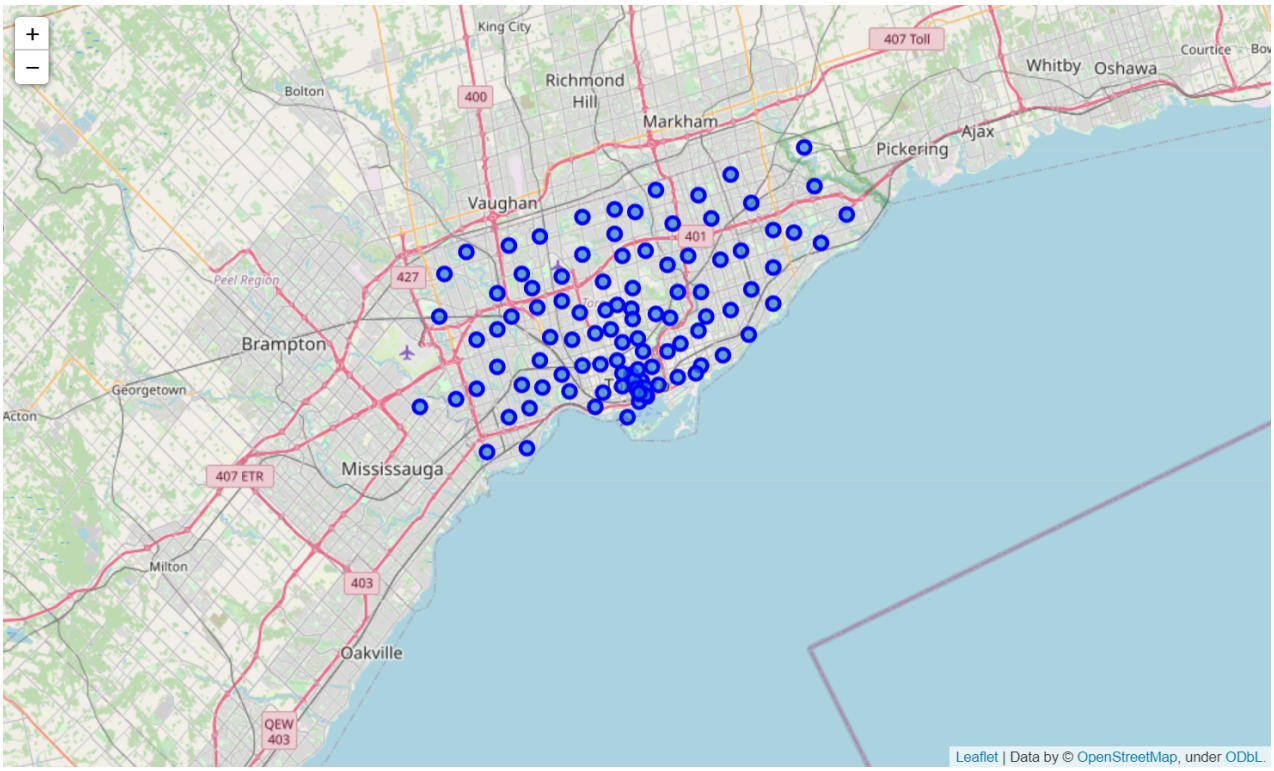
Above is a Folium map centered at New York city, with blue dots denoting all the neighborhoods in New York city.

2.1.3 Toronto

We have already created a location data frame in our exercise project of segmenting and clustering neighborhoods in Toronto. There, we obtain the data of boroughs and neigh-

borhoods by scraping the Wikipedia page `https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M` and we obtain the latitude and longitude data at `https://cocl.us/Geospatial_data`. I repeat the procedures in that exercise project. The first five rows in the location data frame for Toronto, `Toronto_Data`, is shown below.

	City	Borough	Neighborhood	Latitude	Longitude
0	Toronto	North York	Parkwoods	43.753259	-79.329656
1	Toronto	North York	Victoria Village	43.725882	-79.315572
2	Toronto	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
3	Toronto	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
4	Toronto	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494



Above is a Folium map centered at Toronto city, with blue dots denoting all the neighborhoods in Toronto city.

Here, I would like to point out that we have obtained the data frame `Toronto_Data` by scraping the Wikipedia page of a list of postal codes in Canada where the first letter is M. There are cases where several different neighborhoods share the same postal code. In each of these cases, the several different neighborhoods have been treated as one single big neighborhood in our data frame `Toronto_Data`. For example, Regent Park and Harbourfront are two different neighborhoods sharing the same postal code, thus being treated as a single big neighborhood in our data frame `Toronto_Data`. This is not a concern for me because I do not need very fine results. Our data frame `Toronto_Data` has ten boroughs and 103 neighborhoods.

2.1.4 Summary

Finally, we create the data frame `TwoCities` by merging the data of New York and Toronto together. The data frame `TwoCities` has 306+103=409 rows. Further, we create the data frame `ThreeCities` by merging `TwoCities` and the data of Sha Tin, Hong Kong. The data frame `ThreeCities` has 409+6=415 rows. Creating these two master data frames will greatly facilitates later data processing.

New York has 306 neighborhoods and Toronto has 103 neighborhoods. The number of neighborhoods in New York is roughly three times the number of neighborhoods in Toronto. Thus, if the 20 recommended neighborhoods are randomly distributed in these 409 neighborhoods, the number of recommended neighborhoods in New York should be about three times the number of recommended neighborhoods in Toronto. This should be keep in mind for later discussion.

2.2 Venue Information Data

We use Foursquare API to fetch the venue information data, same as what we did in a previous exercise project. In order to get the same number of venue categories for each neighborhood, we need to fetch the venue data for all three cities at the same time. Thus, we use our grand master `ThreeCities` data frame here. We store the venue information data which we obtain from Foursquare in a data frame named `ThreeCities_venues`. The first five rows of this data frame is shown below. It has a total of 12001 rows, which is the total number of returned venues in all three cities. There're 461 unique venue categories.

	City	Borough	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	New York	Bronx	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	New York	Bronx	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
2	New York	Bronx	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
3	New York	Bronx	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
4	New York	Bronx	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

We then group rows by neighborhood. After that, we take the mean of the frequency of occurrence of each category to obtain the data frame `venues_grouped`. First five rows and first 14 columns are shown below. The data frame `venues_grouped` has 406 rows, which means that 409-406=3 neighborhoods are lost. This means that Foursquare did not return any data of these three neighborhoods. We just delete these three neighborhoods from our analysis.

Now with the data frame `venues_grouped` at hand, we have finished data collection and data preprocessing. Now we are ready to proceed to perform our data analysis.

	City	Borough	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant
0	Hong Kong	Sha Tin	Fo Tan	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Hong Kong	Sha Tin	Ma Liu Shui	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Hong Kong	Sha Tin	Ma On Shan	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Hong Kong	Sha Tin	Sha Tin	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Hong Kong	Sha Tin	Tai Wai	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

3 Methodology

In this section, we carry out the main analysis of this project. The procedure can be divided into three steps. First, we build the data frame of user features, `UserFeatures` data frame. Second, we build the data frame of features of neighborhoods in New York and Toronto, `NewNeighborFeatures` dataframe. Third, we build the data frame providing recommendation, `RecommendedNeighborhoods` data frame. Readers may refer to my Jupiter Notebook for more implementation details.

3.1 Building the Data Frame of User Features: `UserFeatures`

We split the data frame `venues_grouped` into two daughter data frames. The first one is for neighborhoods in Sha Tin, Hong Kong, named `UserVenues`. Below is the first 14 columns of the data frame `UserVenues`.

	City	Borough	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant
0	Hong Kong	Sha Tin	Fo Tan	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	Hong Kong	Sha Tin	Ma Liu Shui	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	Hong Kong	Sha Tin	Ma On Shan	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	Hong Kong	Sha Tin	Sha Tin	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	Hong Kong	Sha Tin	Tai Wai	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	Hong Kong	Sha Tin	Wu Kai Sha	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

We then average over the rows to get a `UserFeatures` data frame, which defines the characteristics of my current working and living place. It is used essentially as a user profile. Below is the first 20 rows of the list `UserFeatures`.

```
Accessories Store                 0.000000
Adult Boutique                    0.000000
Afghan Restaurant                 0.000000
African Restaurant                0.000000
Airport                           0.000000
Airport Food Court                0.000000
Airport Gate                      0.000000
Airport Lounge                    0.000000
Airport Service                   0.000000
Airport Terminal                  0.000000
American Restaurant               0.000000
Animal Shelter                    0.000000
Antique Shop                      0.000000
Aquarium                          0.000000
Arcade                            0.000000
Arepa Restaurant                  0.000000
Argentinian Restaurant            0.000000
Art Gallery                       0.000000
Art Museum                        0.033333
Arts & Crafts Store               0.000000
```

3.2 Building the Data Frame of Features of Neighborhoods in New York and Toronto: `NewNeighborVenues`

The second daughter data frame of `venues_grouped` is for neighborhoods in New York and Toronto, named `NewNeighborVenues`. Below is the first five rows and first 14 columns of the data frame `NewNeighborVenues`.

	City	Borough	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	A S
0	New York	Bronx	Allerton	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.033333	0
1	New York	Bronx	Baychester	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0
2	New York	Bronx	Bedford Park	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0
3	New York	Bronx	Belmont	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.010309	0
4	New York	Bronx	Bronxdale	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0

From it, we drop the columns of city, borough, and neighborhood to obtain the data frame `NewNeighborFeatures`. The first five rows and first 14 columns of the data frame `NewNeighborFeatures` are as follows.

	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Animal Shelter	Antique Shop	Aquarium	A
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.033333	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.010309	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0

3.3 Providing Recommendation: `RecommendedNeighborhoods`

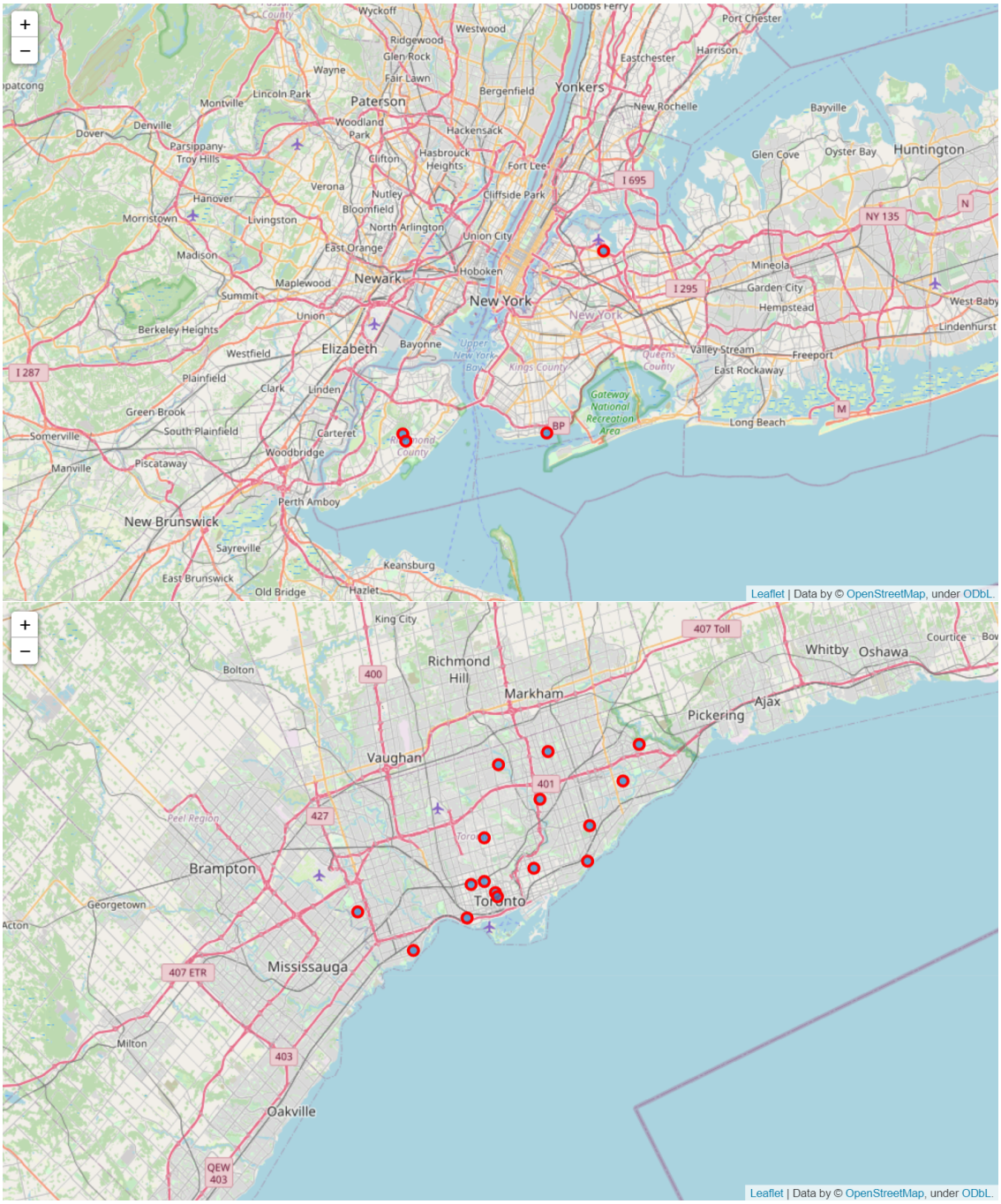
With data frames `UserFeatures` and `NewNeighborFeatures` at hand, we are ready to calculate the overall similarity score of each neighborhood in New York and Toronto. We define the similarity score of a neighborhood as the average of features of that neighborhood weighted by the user features. The implementation of this calculation is straight forward. After implementing the calculation and sorting the results, we obtain the resulting data frame `NewNeighborRating`. The first 20 rows of `NewNeighborRating` are the recommended neighborhoods that are most similar to my current working and living place.

We have these rows of neighborhoods joined with their latitude data and longitude data, forming the data frame of our final results, `RecommendedNeighborhoods`. This final results data frame is shown below.

	City	Borough	Neighborhood	Similarity Score	Latitude	Longitude
0	Toronto	Scarborough	Malvern, Rouge	0.082222	43.806686	-79.194353
1	Toronto	North York	Bayview Village	0.055000	43.786947	-79.385975
2	Toronto	Scarborough	Woburn	0.037500	43.770992	-79.216917
3	Toronto	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	0.035171	43.605647	-79.501321
4	Toronto	North York	Parkwoods	0.032037	43.753259	-79.329656
5	Toronto	Scarborough	Birch Cliff, Cliffside West	0.031111	43.692657	-79.264848
6	Toronto	Downtown Toronto	Christie	0.030131	43.669542	-79.422564
7	Toronto	Scarborough	Steeles West, L'Amoreaux West	0.030040	43.799525	-79.318389
8	Toronto	East York	East Toronto	0.029630	43.685347	-79.338106
9	Toronto	Etobicoke	Eringate, Bloordale Gardens, Old Burnhamthorpe...	0.028889	43.643515	-79.577201
10	New York	Staten Island	Lighthouse Hill	0.028519	40.576506	-74.137927
11	New York	Staten Island	Richmond Town	0.026222	40.569606	-74.134057
12	Toronto	Downtown Toronto	Queen's Park, Ontario Provincial Government	0.025359	43.662301	-79.389494
13	Toronto	Scarborough	Kennedy Park, Ionview, East Birchmount Park	0.025278	43.727929	-79.262029
14	Toronto	Central Toronto	North Toronto West	0.025028	43.715383	-79.405678
15	New York	Queens	East Elmhurst	0.024444	40.764073	-73.867041
16	Toronto	Downtown Toronto	Central Bay Street	0.024435	43.657952	-79.387383
17	New York	Brooklyn	Manhattan Beach	0.024343	40.577914	-73.943537
18	Toronto	Central Toronto	The Annex, North Midtown, Yorkville	0.024120	43.672710	-79.405678
19	Toronto	West Toronto	Brockton, Parkdale Village, Exhibition Place	0.023816	43.636847	-79.428191

4 Results

Our final results, the 20 recommended neighborhoods are shown in the above data frame. Here we visualize these neighborhoods using the Folium map. The recommended neighborhoods in New York and Toronto are shown in red dots below.



5 Discussion

We count the number of recommended neighborhoods in New York and Toronto respectively. It turns out that there are 4 recommended neighborhoods in New York and 16 in Toronto. That is a ratio of 1:4. We recall that the total number of neighborhoods in New York is roughly three times the number of neighborhoods in Toronto. We should get a ratio of about 3:1 if the recommended neighborhoods were randomly chosen.

Furthermore, we notice that all of the first ten recommended neighborhoods, with top ten highest similarity scores, are neighborhoods in Toronto. **Thus, we conclude that Toronto is much more similar to Sha Tin, Hong Kong than New York.**

The neighborhood Malvern-Rouge in Scarborough has a similarity score significantly higher than any of the rest of neighborhoods. **Thus, we recommend Malvern-Rouge, Scarborough, Toronto as the most familiar neighborhood.**

Next, we want to gain some intuition about the recommended neighborhoods. To do so, we find out the top ten most common venue categories in each recommended neighborhood and use the data to form the data frame `Discussion`. The first ten rows and first five most common venue categories are shown below.

	City	Borough	Neighborhood	Similarity Score	Latitude	Longitude	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Toronto	Scarborough	Malvern, Rouge	0.082222	43.806686	-79.194353	Fast Food Restaurant	Yoga Studio	Exhibit	Dumpling Restaurant	Duty-free Shop
1	Toronto	North York	Bayview Village	0.055000	43.786947	-79.385975	Chinese Restaurant	Bank	Japanese Restaurant	Café	Yoga Studio
2	Toronto	Scarborough	Woburn	0.037500	43.770992	-79.216917	Coffee Shop	Korean Restaurant	Soccer Field	Yoga Studio	Dosa Place
3	Toronto	Etobicoke	New Toronto, Mimico South, Humber Bay Shores	0.035171	43.605647	-79.501321	Café	Gym	Fried Chicken Joint	Liquor Store	Seafood Restaurant
4	Toronto	North York	Parkwoods	0.032037	43.753259	-79.329656	Food & Drink Shop	Park	Fast Food Restaurant	Event Space	Dumpling Restaurant
5	Toronto	Scarborough	Birch Cliff, Cliffside West	0.031111	43.692657	-79.264848	Skating Rink	College Stadium	Café	General Entertainment	Yoga Studio
6	Toronto	Downtown Toronto	Christie	0.030131	43.669542	-79.422564	Grocery Store	Café	Park	Diner	Coffee Shop
7	Toronto	Scarborough	Steeles West, L'Amoreaux West	0.030040	43.799525	-79.318389	Chinese Restaurant	Fast Food Restaurant	Coffee Shop	Grocery Store	Breakfast Spot
8	Toronto	East York	East Toronto	0.029630	43.685347	-79.338106	Convenience Store	Park	Coffee Shop	Event Space	Dumpling Restaurant
9	Toronto	Etobicoke	Eringate, Bloordale Gardens, Old Burnhamthorpe...	0.028889	43.643515	-79.577201	Beer Store	Pizza Place	Shopping Plaza	Café	Coffee Shop

The information of the recommended neighborhoods revealed by the above data frame is consistent with our intuition. As we can see, in most of the top ten most recommended neighborhoods, there are many Asian-style restaurants, including dumpling restaurants, Chinese restaurants, Japanese restaurants, and Korean restaurants. This is reasonable and in line with expectations because I used data of Sha Tin, Hong Kong, which is a borough in East Asia, to create the user profile. There are indeed many Asian-style restaurants in Sha Tin, Hong Kong.

In addition, we may recall that Sha Tin is not close to the city center of Hong Kong. We can observe that among the top ten most recommended neighborhoods in Toronto, nine of these neighborhoods are not close to the city center of Toronto. The exception is the neighborhood of Christie located at Downtown Toronto. Among the four recommended neighborhoods in New York, none is close to the city center of New York. All of these observations are consistent with intuition, thus suggesting that the algorithm we have built works well.

6 Conclusion

In this project, I have built a recommendation system that recommends neighborhoods in given new cities which are most similar to the benchmark location. I have chosen the benchmark location to be Sha Tin district of Hong Kong and have chosen two new cities which are New York and Toronto. The recommendation system returns 20 neighborhoods in these two cities based on their similarity scores with respect to the benchmark location.

Among these recommended 20 neighborhoods, 16 are in Toronto and four are in New York. If we restrict our attention to the top ten most recommended neighborhoods, we find that they are all in Toronto. Since the number of neighborhoods in New York is roughly three times the number of neighborhoods in Toronto, we conclude that when venue categories are concerned, Toronto is much more similar to Sha Tin, Hong Kong than New York.

The top one most recommended neighborhood is Malvern-Rouge, Scarborough, Toronto. The five most common venue categories in this neighborhood are: fast food restaurant, yoga studio, exhibit, dumpling restaurant, and duty-free shop.

In the future, the analysis done here can be expanded to include more cities. One general commercial application might be a more comprehensive recommendation system which includes all major cities around the world. In order to accomplish this goal, we will need to maintain our own database which stores all the location data for neighborhoods in different cities. We will also need a commercial account on Foursquare so that we can fetch unlimited amount of venue data.

While doing this project, I have noticed that Foursquare does not always return the same venue information for the same input. Thus, for each time we run the same code, the result may be different slightly. To handle this issue, it would be good that we use our own database to store all the venue data returned by Foursquare every time, thus forming a complete data set of information of venues.

This concludes my project report.