My Final Capstone FourSquare Project

Description of the problem and a discussion of the background:
Problem Statement: Finding a home suitable to personal tastes in relative to price expectation is getting harder and harder every other day. This research is done for real estate agencies to find houses for the people whom are eager to migrate or move into Toronto, CA from New York that are living in similar places.
Background: Toronto is a city in Canada growing fast and is provincial capital of Ontario and the most populous city in Canada. A lot of people has the curiosity to discover what Toronto holds for them. Healtcare and wages are also well above expectations where for the most people in US which makes this city a target. Since immigration to Canada is more welcoming than most countries according to here; Toronto has the most immigration ratio compared to other cities in Canada. Toronto is a culturally rich and diverse city as a result. Low crime rates and great cusine also a deciding factor for many and New Yorkers too. This paper is helping agencies to boost on these kind of factors, diversities, likes when there is a potention customer. My solution is creating an algorithm to find the best possible house for the customer according to where he lives and like in his/her homeplace (New York in our case) and makes a suggestion using machine learning and geo locator services.

A description of the data and how it will be used to solve the problem.
Toronto and New York location data will be used to compare the neighborhoods, and then we rank the neighborhoods from the selected boroughs for the cities.
I used the previous assignments to retrieve the neighborhood and geo co-ordinates for New York and Toronto. Unsupervised learning is used to cluster the neighborhoods of the two cities. Data sources are Toronto Wikipedia WebPage and New York. Geo locations for Toronto will use following source file.
Implementation logic is as follows:

- Create a dataset that holds the Geo-cordinates for Toronto's neighborhood
- Use Foursquare APIs to get venues for each of the Toronto's neighborhood
- Sort through the data to identify top 10 common venue categories for each of the Toronto's neighborhood
- Perform the above for New York's neighborhoods
- This step wil be the input. Select a neighborhood in NY for which we are looking for similar places in Toronto
- Use K-means on dataset that has all Toronto's neighborhoods plus this neighhborhood
- Then find the cluster which has the NY neighborhood in it and list all Toronto neighborhoods there

Methodology

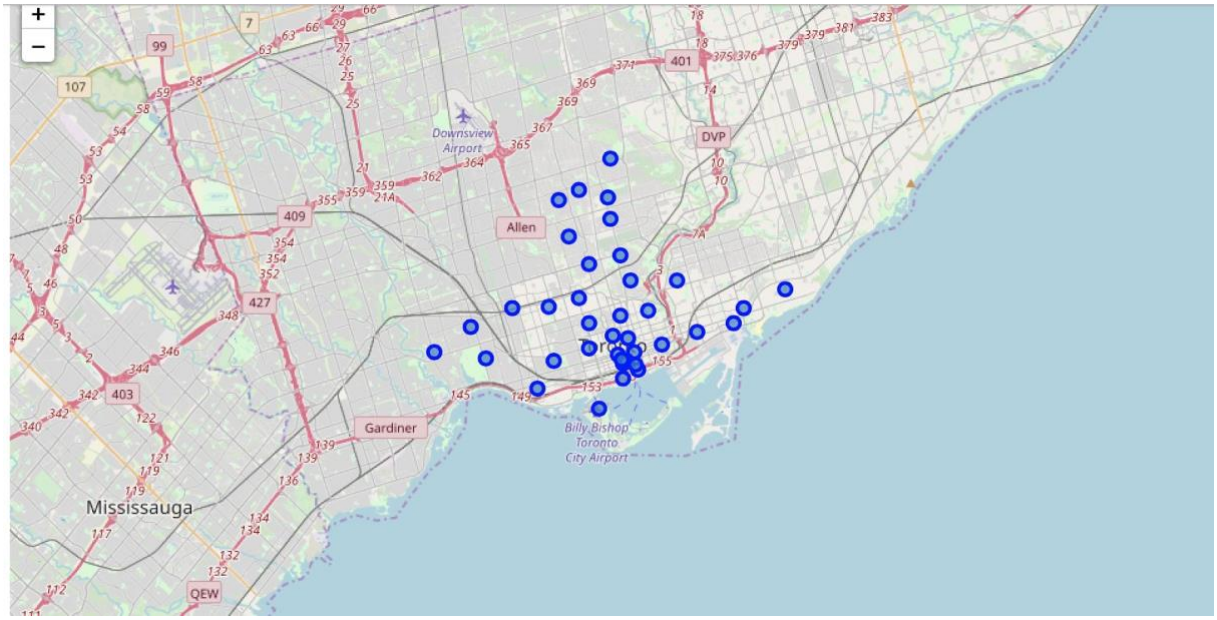
I followed the following steps:

1. Create a dataset that holds the Geo-coordinates for Toronto's neighborhood

Use BeautifulSoup package to perform web scraping on Toronto Wiki page to build the following dataset.

	Borough	Neighbourhood	Latitude	Longitude
37	East Toronto	The Beaches	43.676357	-79.293031
41	East Toronto	The Danforth West,Riverdale	43.679557	-79.352188
42	East Toronto	The Beaches West,India Bazaar	43.668999	-79.315572
43	East Toronto	Studio District	43.659526	-79.340923
44	Central Toronto	Lawrence Park	43.728020	-79.388790

Then, I further explored the city of Toronto. The following unclustered map view is a representation of it.



2. Use Foursquare APIs to get venues for each of the Toronto's neighborhoods

Next, we use the Foursquare's venues API to get nearby venues. We will capture the following data points: 'Neighbourhood', 'Neighbourhood Latitude', 'Neighbourhood Longitude', 'Venue', 'Venue Latitude', 'Venue Longitude', and 'Venue Category'. A snapshot of the dataset looks like the following. This dataset will form the basis for k-means clustering which I will illustrate in following sections.

	Neighbourhood	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	...	Theme Restaurant	Thrift / Vintage Store	Toy / Game Store	Trail	Tra Stati
0	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	1	
1	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
2	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
3	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	0	
4	The Beaches	0	0	0	0	0	0	0	0	0	...	0	0	0	0	

5 rows × 194 columns

3. Sort through the data to identify top 10 common venue categories for each of the Toronto's neighborhoods

Dataset snapshot is given below for your reference.

	Neighbourhood	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 Most Common Venue
0	Adelaide,King,Richmond	Steakhouse	Asian Restaurant	Café	Pizza Place	Hotel	Neighborhood	Lounge	Burger Joint	Seafood Restaurant	Smoke Shop
1	Berczy Park	Seafood Restaurant	Coffee Shop	Cocktail Bar	Beer Bar	Café	Farmers Market	Greek Restaurant	Jazz Club	Basketball Stadium	Fish Market
2	Brockton,Exhibition Place,Parkdale Village	Coffee Shop	Breakfast Spot	Café	Climbing Gym	Stadium	Burrito Place	Restaurant	Caribbean Restaurant	Pet Store	Bakery
3	Business Reply Mail Processing Centre 969 Eastern	Yoga Studio	Fast Food Restaurant	Park	Comic Shop	Pizza Place	Butcher	Burrito Place	Recording Studio	Restaurant	Brewery
4	CN Tower,Bathurst Quay,Island airport,Harbourf...	Airport Lounge	Airport Service	Airport Terminal	Harbor / Marina	Sculpture Garden	Airport Food Court	Airport Gate	Bar	Boat or Ferry	Boutique

4. Perform the above for New York's neighborhoods

Same steps are performed for the city of New York. For conciseness, the final datasets will look like the the following

	Neighbourhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Animal Shelter	Antique Shop	Arcade	Arepa Restaurant	...	Warehouse Store	Waste Facility	Wate
0	Wakefield	0	0	0	0	0	0	0	0	0	...	0	0	
1	Wakefield	0	0	0	0	0	0	0	0	0	...	0	0	
2	Wakefield	0	0	0	0	0	0	0	0	0	...	0	0	
3	Wakefield	0	0	0	0	0	0	0	0	0	...	0	0	
4	Wakefield	0	0	0	0	0	0	0	0	0	...	0	0	

5 rows × 380 columns

	Neighbourhood	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 Most Common Venue
0	Allerton	Pizza Place	Pharmacy	Spa	Deli / Bodega	Supermarket	Department Store	Fried Chicken Joint	Breakfast Spot	Bus Station	Gas Station
1	Annadale	Pizza Place	Park	Sports Bar	Restaurant	Food	Diner	Train Station	Pharmacy	Field	Event Space
2	Arden Heights	Pharmacy	Coffee Shop	Pizza Place	Bus Stop	Yoga Studio	Financial or Legal Service	Factory	Falafel Restaurant	Farm	Farmers Market
3	Arlington	Bus Stop	Deli / Bodega	American Restaurant	Boat or Ferry	Food	Grocery Store	Fish Market	Farm	Farmers Market	Fast Food Restaurant
4	Arrochar	Deli / Bodega	Pizza Place	Italian Restaurant	Bus Stop	Athletics & Sports	Middle Eastern Restaurant	Bagel Shop	Liquor Store	Supermarket	Hotel

5. This step will is the input. Select a neighborhood in NY for which we are looking for similarities in Toronto. Here, we chose Chelsea, NY.
6. Use K-means on dataset that has all Toronto's neighborhoods plus this neighborhood. Then find the cluster which has the NY neighborhood in it and list all Toronto neighborhoods there. Following dataset shows each neighborhood in Toronto with its assigned cluster label.

	Neighbourhood	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	Antique Shop	Aquarium	...	Veterinarian	Video Store	Warehouse Store	Waste Facility
38	Chelsea	0.0	0.0	0.0	0.0	0.0	0.0	0.029412	0.0	0.0	...	0.0	0.0	0.0	0.0

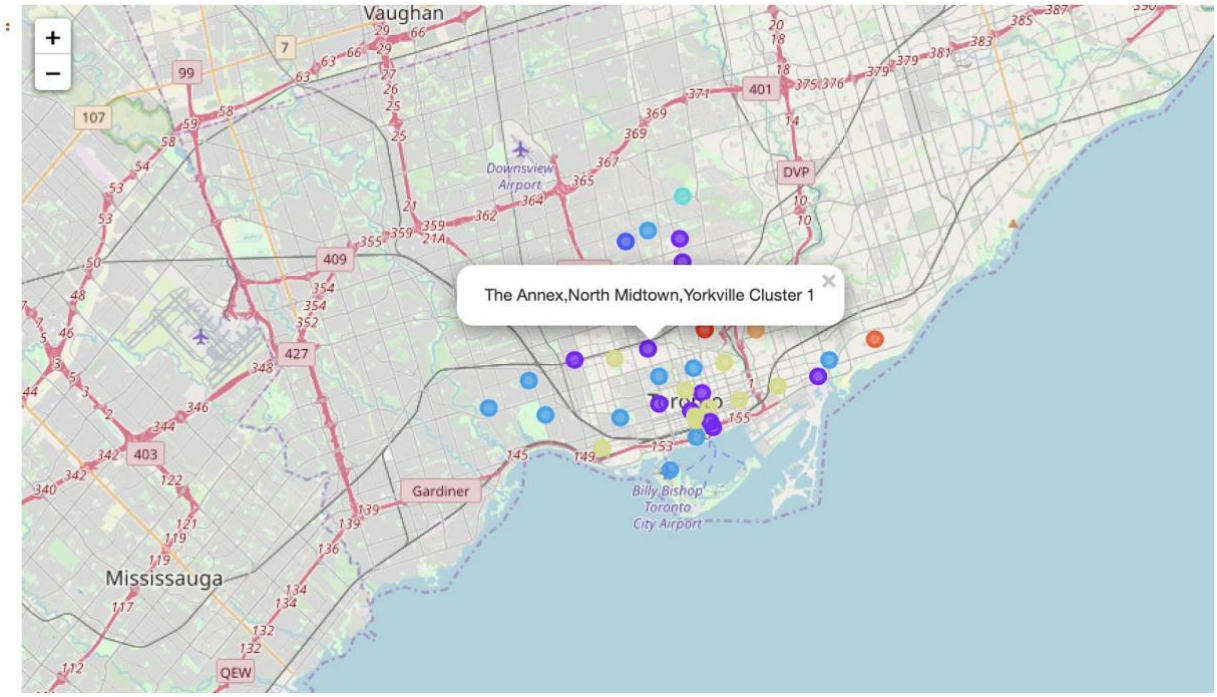
	Borough	Neighbourhood	Latitude	Longitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8 C
0	East Toronto	The Beaches	43.676357	-79.293031	9	Neighborhood	Other Great Outdoors	Health Food Store	Trail	Pub	Cuban Restaurant	Ethiopian Restaurant	E Re
1	East Toronto	The Danforth West,Riverdale	43.679557	-79.352188	8	Greek Restaurant	Ice Cream Shop	Italian Restaurant	Yoga Studio	Bookstore	Restaurant	Spa	
2	East Toronto	The Beaches West,India Bazaar	43.668999	-79.315572	3	Park	Pet Store	Ice Cream Shop	Liquor Store	Sandwich Place	Burger Joint	Fast Food Restaurant	
3	East Toronto	Studio District	43.659526	-79.340923	7	Café	Coffee Shop	Bakery	Italian Restaurant	American Restaurant	Middle Eastern Restaurant	Stationery Store	
4	Central Toronto	Lawrence Park	43.728020	-79.388790	4	Bus Line	Park	Swim School	Dance Studio	Falafel Restaurant	Ethiopian Restaurant	Eastern European Restaurant	D Re

Results

For Chelsea, NY, I found 10 similar neighborhoods in Toronto. They are listed below:

data_index	
0	Adelaide,King,Richmond
1	Berczy Park
3	Business Reply Mail Processing Centre 969 Eastern
7	Chinatown,Grange Park,Kensington Market
11	Davisville
12	Davisville North
15	Dovercourt Village,Dufferin
30	Ryerson,Garden District
32	Stn A PO Boxes 25 The Esplanade
34	The Annex,North Midtown,Yorkville

Also, if we want to visualize them in maps, we can see them as below (marked in purple)



Discussion

K-means is a very powerful yet simple unsupervised learning technique to compare neighborhoods. For this algorithm to be successful, we need as much data as possible available for both the cities that speaks about the lifestyle of those areas. Then, we can run ML unsupervised K-means algorithm to find similars.

This algorithm can be generalized to compare any two cities, given the data is available as mentioned above and will be used for Real estate agencies for all over the World.

Conclusion

For Chelsea,NY we were able to find similar neighborhoods in Toronto, ON. We based it off several lifestyle parameters for the neighborhoods and leveraged Foursquare APIs for the same. This is a powerful algorithm that will help relocation agencies zero in on localities that will fit the bill for their clients. If correctly incorporated in their client onboarding process, this algorithm can have potential positive impacts on the bottom line.