Finding Neighborhoods in Toronto that are Most Similar to a Given Location

Navami Shenoy
July 4, 2019

1. Introduction

1.1 The Problem

This project is aimed at helping stakeholders who wish to move to Toronto in a neighborhood most similar to their current locality. As an instance, we are going to pick a suitable neighborhood for a client currently residing in Hamilton Heights, New York.

Using data science we will generate a cluster of neighborhoods grouped by their similarity, so that the stakeholders can choose the best possible neighborhood based on their preferences.

1.2 Background

Every year, thousands of people move to Toronto for a variety of reasons related to work, education, recreation etc. If they knew which places in Toronto are the most similar to their current place of residence, such that when they move in they will be able to find the same type of venues (bakeries, gyms, etc) around them, they can make a move with convenience and still feel at home. Besides people looking to move, this project will benefit anyone who wishes to go to Toronto for a temporary stay, or city enthusiasts who just wish to feed their curiosity.

2. Data

2.1 Data Description and Acquisition

Data required for this project should be able to provide accurate description of neighborhoods in order to facilitate comparison. For this, we would need to know what kinds of places exist in a neighborhood; then we can say a neighborhood is analogous to another based on which venues are most common among them. Therefore, the data should contain geospatial and venue information, including, but not limited to: names of neighborhoods and their geographical coordinates (latitudes and longitudes), types of venues found in those neighborhoods and their names, coordinates, frequency of occurrence etc.

Data on Toronto's neighborhoods was scraped from Wikipedia. Foursquare API was used to look up venues in both Toronto and Hamilton Heights. Toronto geospatial data was obtained through Cognitive Class. Here is a detailed description of each data source:

- Wikipedia - contains a list of postal codes, boroughs and neighborhoods in Canada
- Foursquare API - has an exhaustive record of the name and type of venues and their locations in a neighborhood
- Cognitive Class - contains a link to the geographical coordinates of each Toronto neighborhood
- geopy - Python package that will be used to obtain coordinates of an individual location

2.2 Data Cleaning

The Toronto post code data contained many boroughs and neighborhoods that had not been assigned a name. These were treated as missing data or null values and were eliminated from the data as they do not provide useful information. However, boroughs that do have a name assigned but not a neighborhood (for example, Queen's Park) could not be dropped; in such cases the neighborhood was assigned the same name as its borough.

Moreover, there were several neighborhoods listed on multiple rows that have the same postal code, i.e., many postal codes have multiple neighborhoods. Therefore, those neighborhoods were all grouped under one postcode. Not only this, the geospatial data for Toronto had to be acquired and added to our dataframe. Since the geocoder package at the time of this project was unreliable, the link from Cognitive Class mentioned in the previous section was used to load the coordinates. Lastly, data scraped or downloaded from multiple sources were combined into one table.

3. Methodology

3.1 Exploratory Data Analysis

After acquiring and cleaning the data, some initial analysis was performed to explore and get better acquainted with the data. It was found that, postcode-vise, there were about 103 neighborhoods in Toronto, which were visualized on a map to observe how they were distributed geographically (Figure 1). It was noted that the neighborhoods appear clumped at the center (where downtown Toronto is) and radially spread out as one moves outwards. However, it was decided that the sample size of 103 neighborhoods would prove to be cumbersome for this project, primarily because Foursquare has a limit over the number of API calls that can be made; it may go well over the limit while retrieving venues for *each* of the 103 neighborhoods.

Figure 1. Map showing all neighborhoods in Toronto (in blue)

Therefore, we needed to reduce the sample space. Moreover, since the client's location was in Manhattan which has the most hubbub in NYC, neighborhoods in Toronto closer to the city center were a better match for our client's tastes.

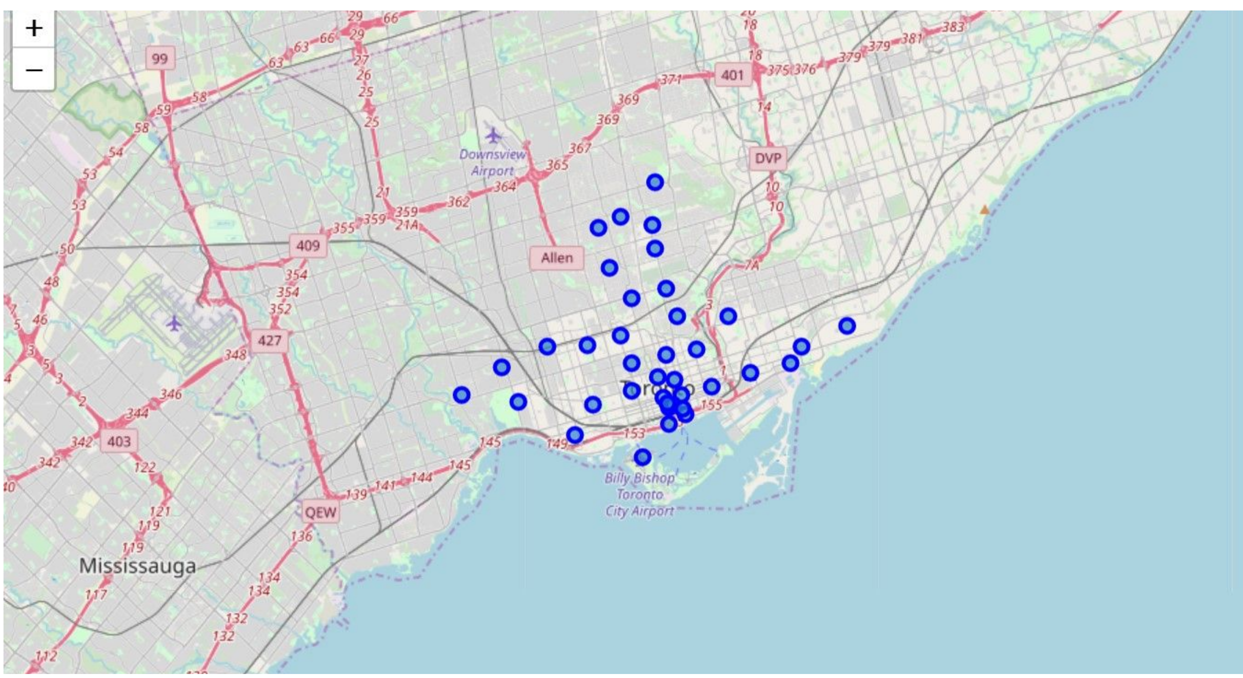

Figure 2. Map showing neighborhoods around downtown Toronto

Keeping these points in mind, the data was filtered to contain only boroughs with 'Toronto' in their name, for example, Central Toronto, West Toronto, etc. which are all close to Toronto's downtown (Figure 2). Now this left us with 38 neighborhoods to work with. After this, in order to facilitate the machine learning algorithm we would use later, our client's address was appended to the Toronto data, so our dataframe looked like this (Figure 3):

	Borough	Postcode	Neighbourhood	Latitude	Longitude
34	West Toronto	M6K	Brockton, Exhibition Place, Parkdale Village	43.636847	-79.428191
35	West Toronto	M6P	High Park, The Junction South	43.661608	-79.464763
36	West Toronto	M6R	Parkdale, Roncesvalles	43.648960	-79.456325
37	West Toronto	M6S	Runnymede, Swansea	43.651571	-79.484450
38	Manhattan	10032	Hamilton Heights	40.824145	-73.950062

Figure 3. Dataframe showing the last five rows (includes Hamilton Heights, Manhattan)

3.2 Foursquare

With the Foursquare API we can find what kind of venues exist in each neighborhood, and how many of them a neighborhood has. This allows us to compare two neighborhoods— if one contains more gyms and yoga clubs than the other, than a fitness enthusiast is going to prefer the former to the latter.

Using Foursquare we obtained the venues found in each neighborhood, setting a limit of 100 to the number of venues retrieved and a radius of 500 meters per neighborhood. Then we calculated the frequency of occurrence for each type of venue, using which we were able to find the ten most common venues for each neighborhood (Figure 4).

	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 Most Common Venue
0	Adelaide, King, Richmond	Steakhouse	Café	Pizza Place	Hotel	Asian Restaurant	Gastropub	Monument / Landmark	Noodle House	Opera House	Concert Hall
1	Berczy Park	Coffee Shop	Cocktail Bar	Farmers Market	Café	Beer Bar	Seafood Restaurant	Park	Bistro	Bakery	Basketball Stadium
2	Brockton, Exhibition Place, Parkdale Village	Coffee Shop	Breakfast Spot	Café	Convenience Store	Pet Store	Performing Arts Venue	Italian Restaurant	Intersection	Gym	Furniture / Home Store
3	Business Reply Mail Processing Centre 969 Eastern	Yoga Studio	Pizza Place	Restaurant	Skate Park	Recording Studio	Smoke Shop	Brewery	Spa	Farmers Market	Fast Food Restaurant
4	CN Tower, Bathurst Quay, Island airport, Harbo...	Airport Lounge	Airport Service	Airport Terminal	Boat or Ferry	Coffee Shop	Sculpture Garden	Boutique	Bar	Harbor / Marina	Plane

3.3 Clustering

To discover neighborhoods that are the most alike, a type of machine learning algorithm called clustering was used. Specifically, we used **K-Means Clustering**, which divides the data into 'k' number of non-overlapping clusters.

Why K-Means?

- because it is an unsupervised algorithm
- because objects within a cluster a very similar, and objects across different clusters are very different

Figure 4. Sample showing top ten venues for each neighborhood

- selects initial cluster centers for k-means clustering in a smart way when its init parameter is set to 'k-means++'

This means that the neighborhoods K-Means puts in a cluster are the most similar to each other, while the neighborhoods in different clusters are very different from each other. This allows us to discover which neighborhoods are most similar to Hamilton Heights. Since K-Means group similar neighborhoods into clusters we can easily them visualize on a map, which will make it easier to spot which cluster our client's neighborhood falls in. Therefore, this enables us to analyze other members of the clusters as potential candidates for the solution.

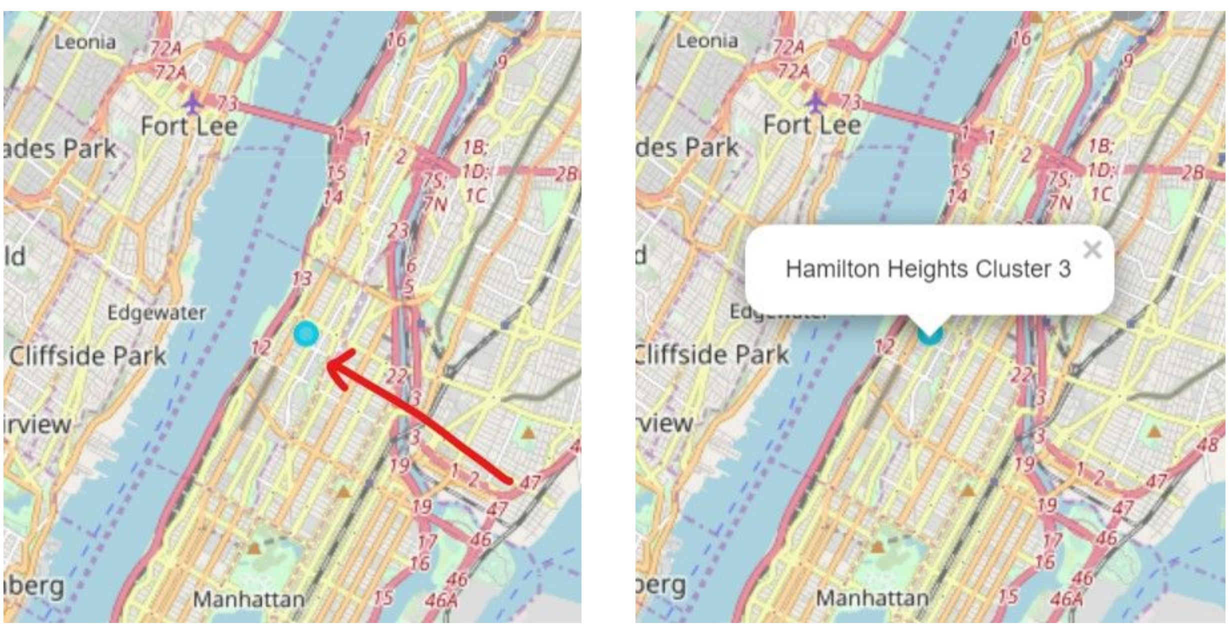

Figure 5 & 6. Hamilton Heights, left (in blue),
Cluster Label, right

After performing K-Means and visualizing the clusters, Hamilton Heights in New York appeared as a light blue dot on the map (Figure 5). Moving the cursor over the dot revealed that Hamilton Heights was assigned into **Cluster 3** by the K-Means algorithm (Figure 6). The other member neighborhoods in Cluster 3 were located in the map of Toronto, all represented by the same color, light blue (Figure 7).

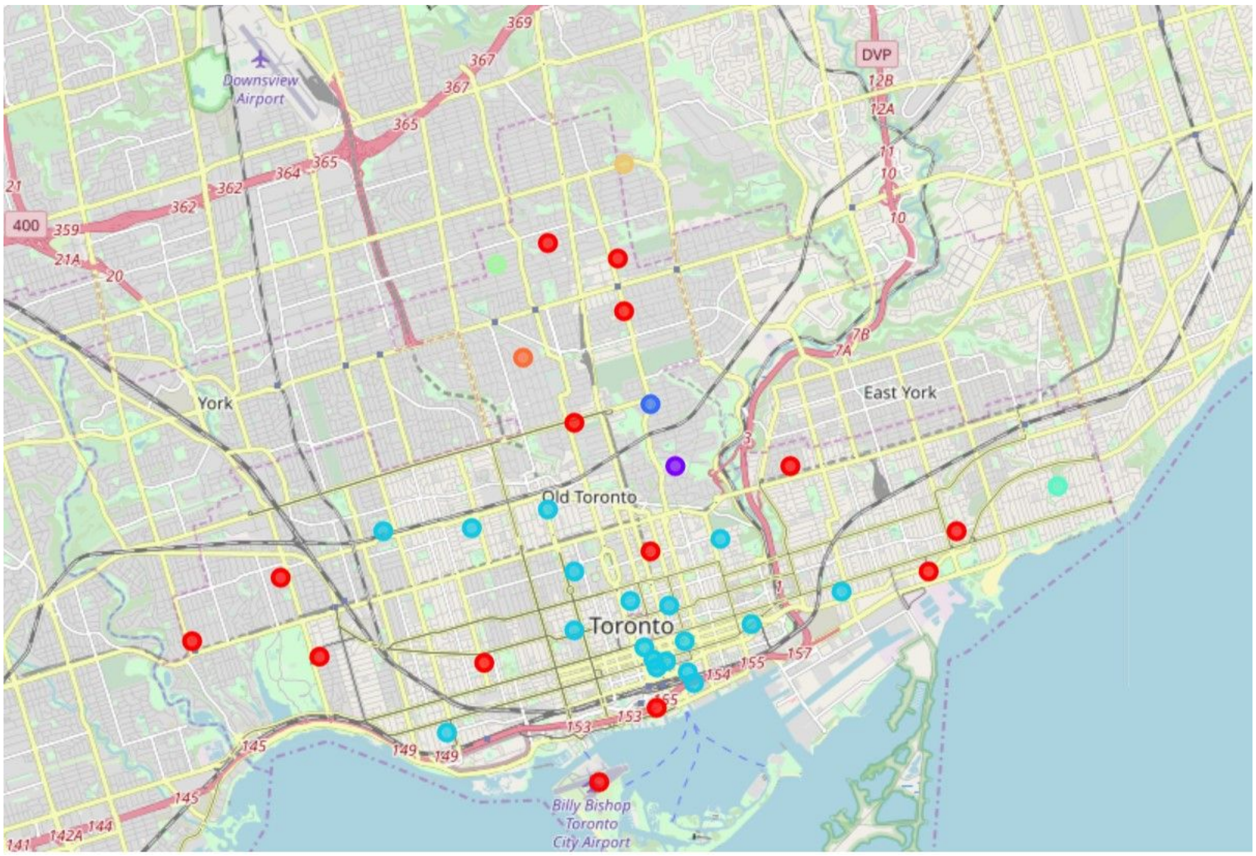

Figure 7. Map showing eight clusters of neighborhoods;
Cluster 3 is represented in light blue

There were 19 neighborhoods in Cluster 3, including Hamilton Heights. This means that there are **18 neighborhoods** that are very similar to Hamilton Heights, New York. The neighborhoods in rest of the clusters (colored in red, purple, orange, etc) are very dissimilar from the ones in Cluster 3, and hence can be disregarded. Therefore, the client has 18 options to choose from. However, these neighborhoods can be further filtered by determining what other factors influence the stakeholder's choice, which, of course, can only be specified by the stakeholders themselves.

The following dataframe shows a list of all 18 neighborhoods and their most common venues, and how they compare with Hamilton Heights (Figure 8):

Neighbourhood	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	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
The Annex, North Midtown, Yorkville	3	Café	Coffee Shop	Sandwich Place	Pizza Place	History Museum	Liquor Store	Jewish Restaurant	Burger Joint	Pub	Indian Restaurant
Cabbagetown, St. James Town	3	Restaurant	Coffee Shop	Italian Restaurant	Bakery	Café	Playground	Diner	Liquor Store	Jewelry Store	Pub
Harbourfront, Regent Park	3	Coffee Shop	Bakery	Park	Mexican Restaurant	Café	Breakfast Spot	Yoga Studio	Gym / Fitness Center	Historic Site	French Restaurant
Ryerson, Garden District	3	Café	Clothing Store	Diner	Fast Food Restaurant	Spa	Movie Theater	Sandwich Place	Beer Bar	Japanese Restaurant	Ramen Restaurant
St. James Town	3	Gastropub	Coffee Shop	Hotel	Japanese Restaurant	Italian Restaurant	Restaurant	Hostel	Food Truck	Latin American Restaurant	Speakeasy
Berczy Park	3	Coffee Shop	Cocktail Bar	Farmers Market	Café	Beer Bar	Seafood Restaurant	Park	Bistro	Bakery	Basketball Stadium
Central Bay Street	3	Coffee Shop	Spa	Bubble Tea Shop	Italian Restaurant	Seafood Restaurant	Café	Poke Place	Japanese Restaurant	Sushi Restaurant	Pizza Place
Adelaide, King, Richmond	3	Steakhouse	Café	Pizza Place	Hotel	Asian Restaurant	Gastropub	Monument / Landmark	Noodle House	Opera House	Concert Hall
Design Exchange, Toronto Dominion Centre	3	Coffee Shop	Restaurant	Deli / Bodega	Café	Gastropub	Hotel	Pub	Sandwich Place	Bookstore	Japanese Restaurant
Commerce Court, Victoria Hotel	3	Café	Coffee Shop	Restaurant	Hotel	Deli / Bodega	Gastropub	Gym	Ice Cream Shop	Pub	Sandwich Place
Harbord, University of Toronto	3	Café	Italian Restaurant	Bookstore	Bar	Japanese Restaurant	Bakery	Restaurant	Gym	Theater	Video Game Store
Chinatown, Grange Park, Kensington Market	3	Café	Mexican Restaurant	Vietnamese Restaurant	Vegetarian / Vegan Restaurant	Caribbean Restaurant	Bakery	Comfort Food Restaurant	Belgian Restaurant	Cheese Shop	Coffee Shop
Stn A PO Boxes 25 The Esplanade	3	Café	Seafood Restaurant	Beer Bar	Cocktail Bar	Farmers Market	Park	Art Gallery	Coffee Shop	Museum	Concert Hall
First Canadian Place, Underground city	3	Café	Coffee Shop	Restaurant	Steakhouse	Food Court	Gastropub	Pizza Place	Deli / Bodega	Pub	Salad Place
Christie	3	Grocery Store	Café	Park	Italian Restaurant	Nightclub	Convenience Store	Restaurant	Diner	Baby Store	Coffee Shop
Studio District	3	Café	Coffee Shop	American Restaurant	Italian Restaurant	Bakery	Yoga Studio	Ice Cream Shop	Sandwich Place	Bookstore	Seafood Restaurant
Dovercourt Village, Dufferin	3	Bakery	Supermarket	Pharmacy	Pool	Middle Eastern Restaurant	Brewery	Liquor Store	Fast Food Restaurant	Bar	Bank
Brockton, Exhibition Place, Parkdale Village	3	Coffee Shop	Breakfast Spot	Café	Convenience Store	Pet Store	Performing Arts Venue	Italian Restaurant	Intersection	Gym	Furniture / Home Store
Hamilton Heights	3	Yoga Studio	Bakery	Mexican Restaurant	Café	Caribbean Restaurant	Cocktail Bar	Coffee Shop	Indian Restaurant	Park	Mediterranean Restaurant

Figure [illegible]: Results

4. Results and Discussion

Through our analysis we have discovered that although Toronto and New York are in two different countries, there is a promising number of neighborhoods comparable to Hamilton Heights. Since the client wishes to move to a place with a similar pace as Manhattan, we focused our attention to neighborhoods in and around Downtown Toronto. While other boroughs could have been great potential candidates (such as North York), but due to the limitations on Foursquare's API calls as well as an intent to reduce our sample space our attention was directed mainly to neighborhoods that have 'Toronto' in their names.

After narrowing our area of interest we gathered all venues within a radius of 500m of all neighborhoods, including Hamilton Heights. We filtered this dataset further by finding ten most common venues in each neighborhood. These venues were then clustered at their respective locations in their corresponding neighborhoods on the map to identify the cluster containing our client's neighborhood. Neighborhoods most similar to Hamilton Heights were put in the same cluster; we were able to point out **18** such **neighborhoods** for our client.

This, of course, does not imply that these neighborhoods are actually suitable for our client to move into. Purpose of this analysis was to only provide information on neighborhoods that are analogous to the client's current place of residence in terms of the most commonly found venues there. It is possible that the reason the client prefers Hamilton Heights is not because of the yoga studios commonly found there. Also, the preference of one neighborhood over another also depends on other factors such as lower crime rates, ease of commute to work, etc. Therefore, the recommended neighborhoods should only be considered as a starting point for a more detailed analysis which could eventually lead to finding the most ideal and convenient location for our stakeholders.

5. Conclusion

The purpose of this data science project was to identify neighborhoods in Toronto for stakeholders who wish to move to the Canadian city. Specifically, we examined downtown Toronto's areas that have similar venues in comparison to Hamilton Heights in Manhattan, New York, the place where our client currently resides. Those locations were then clustered in order to create a visual for our client to use as a starting point in further selection of the most optimal neighborhood for them.

The ultimate decision will of course be made by the stakeholders based on the characteristics of each recommended neighborhood, taking into account factors such as personal venue preferences, affordability, house prices, safety, access to major roads, proximity to work or school etc.