The Battle of Neighborhoods

Introduction

It is interesting to compare the neighborhoods of the two cities and determine how similar or dissimilar they are. The result can provide useful information on those cities. For example, people can choose a neighborhood similar to the one they lived or like. People can also identify business opportunities as similar demand could exist in similar neighborhood.

In this project, I take New York and Toronto as an example. New York and Toronto are very diverse, multicultural and the financial capitals of their respective countries. By using Foursquare location data and clustering machine learning techniques, I segment the neighborhoods into different groups by their venues information. The goal is to find the most relevant neighborhoods between the two cities given the preference.

The clustering technique is an unsupervised algorithm and divides the data into non-overlapping subsets or clusters without any cluster internal structure or labels. Objects within a cluster are very similar and objects across different clusters are very different or dissimilar.

Data

Data sources

The Borough, Neighborhood, Latitude and Longitude information for New York and Toronto are required and obtained from the following data source.

New York: *https://cocl.us/new_york_dataset*

Toronto: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M (https://en.wikipedia.org/w/index.php?title=List_of_postal_codes_of_Canada:_M&oldid=945633050)

Foursquare is a technology company that built a massive dataset of location data. Currently its location data is the most comprehensive and accurate that it powers location data for many popular services like Apple Maps, Uber, Snapchat, Twitter and many others, and is currently being used by over 100,000 developers. In this project, I use Foursquare API to obtain venues information, such as Restaurant, Museum, Airport, Gallery, etc. This is done by constructing a URL to send a request to the API to search for venues by latitude and longitude.

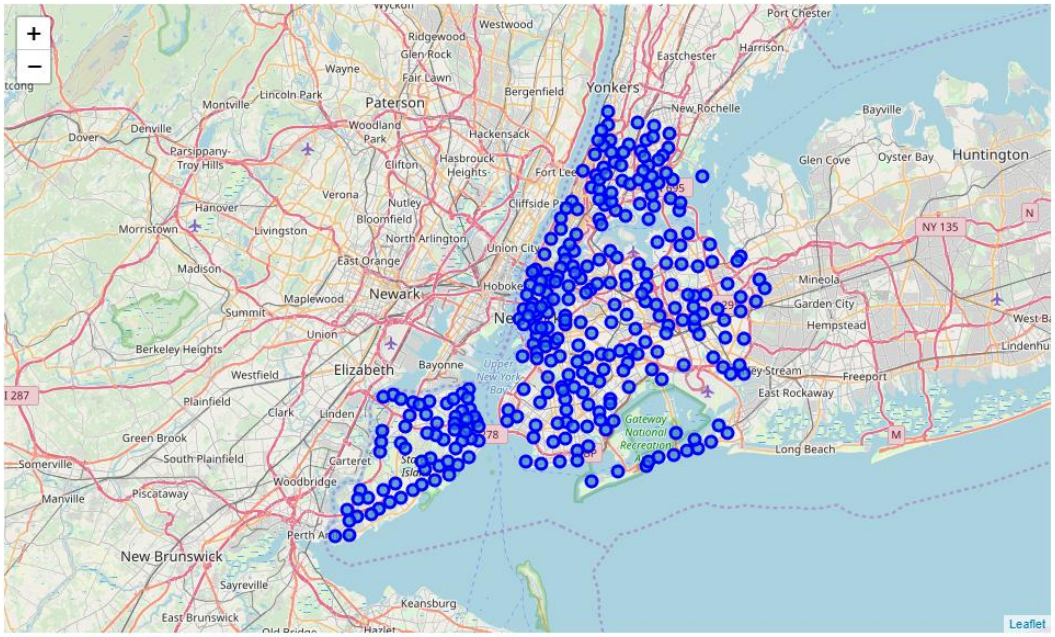
Data preparation

The dataset for New York is a JSON file which contains the 5 boroughs and the neighborhoods (306 in total) that exist in each borough as well as the latitude and longitude coordinates of each neighborhood. Notice how all the relevant data is in the *features* key, which is basically a list of the neighborhoods.

A snapshot of the dataset is shown below.

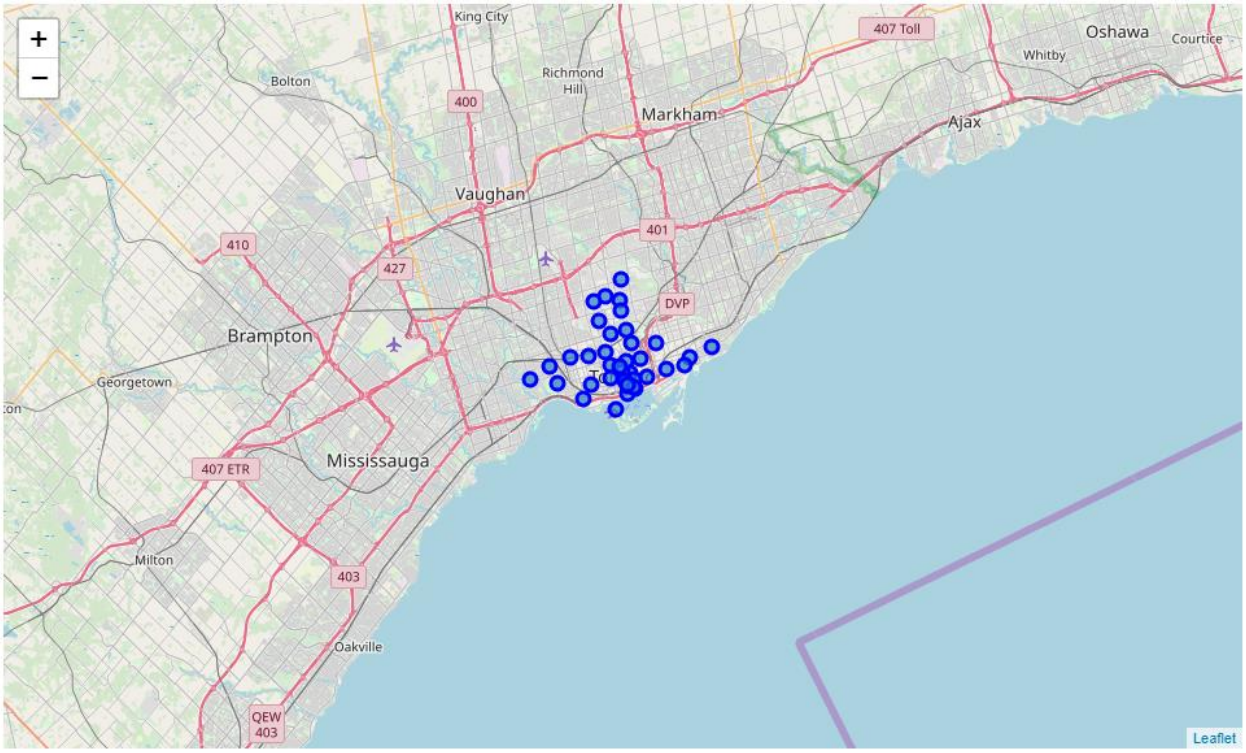
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Use Geopy library to get the latitude and longitude values of New York, 40.7127281 and -74.0060152. Use the Folium library to visualize the neighborhoods.



The Borough and Neighborhood information for Toronto is scraped from a webpage using Pandas library. Not-assigned borough is ignored. If more than one neighborhood can exist in one Borough, these are combined into one neighborhood separated with a comma. Not assigned neighborhood is assigned the same as the Borough. To obtain geographical coordinates, I use a csv file (Geospatial_Coordinates.csv) from the web. Toronto contains 4 boroughs and 39 neighborhoods.

	Postcode	Borough	Neighbourhood	Latitude	Longitude
37	M4E	East Toronto	The Beaches	43.676357	-79.293031
41	M4K	East Toronto	The Danforth West, Riverdale	43.679557	-79.352188
42	M4L	East Toronto	The Beaches West, India Bazaar	43.668999	-79.315572
43	M4M	East Toronto	Studio District	43.659526	-79.340923
44	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790



I use Foursquare API to obtain venues information for each neighborhood. A snapshot is shown below. The venue information (i.e. Venue Category) is then used for neighborhood clustering, how neighborhoods are similar or dissimilar.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop
4	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop

Methodology

To simplify the analysis (number of functional calls to Foursquare API), Manhattan from New York and Downtown Toronto from Toronto are selected. Note that the method can be generalized to any number of cities. The final dataset contains 59 neighborhoods.

Based on the venues information pulled from Foursquare API for each neighborhood, means of the frequency of occurrence of each venue category is calculated. K-means clustering is then applied to divide the neighborhoods into different clusters. K-means is one of the most popular clustering algorithms. Each cluster represents neighborhoods with similar features. Different K values are used to confirm the results.

Results

Five clusters are first segmented. The only cluster containing both Manhattan and Downtown Toronto neighborhoods is shown below. All other fours only contain neighborhoods from either Manhattan or Downtown Toronto. As you can see the following Downtown Toronto neighborhoods are similar to Manhattan:

Harbord, University of Toronto, China Town, Grange Park, Kensington Market

	Borough	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
2	Manhattan	Washington Heights	Café	Bakery	Grocery Store	Chinese Restaurant	Deli / Bodega	Mobile Phone Shop	Pizza Place	Donut Shop	Coffee Shop	Spanish Restaurant
3	Manhattan	Inwood	Mexican Restaurant	Lounge	Pizza Place	Restaurant	Café	Chinese Restaurant	Deli / Bodega	American Restaurant	Park	Frozen Yogurt Shop
4	Manhattan	Hamilton Heights	Pizza Place	Coffee Shop	Deli / Bodega	Café	Mexican Restaurant	Yoga Studio	Bakery	School	Sandwich Place	Caribbean Restaurant
7	Manhattan	East Harlem	Mexican Restaurant	Bakery	Thai Restaurant	Deli / Bodega	Latin American Restaurant	Pizza Place	Park	Spa	Convenience Store	Pharmacy
20	Manhattan	Lower East Side	Chinese Restaurant	Coffee Shop	Café	Art Gallery	Pizza Place	Sandwich Place	Japanese Restaurant	Bakery	Cocktail Bar	Ramen Restaurant
22	Manhattan	Little Italy	Café	Bakery	Bubble Tea Shop	Sandwich Place	Salon / Barbershop	Mediterranean Restaurant	Cocktail Bar	Italian Restaurant	Yoga Studio	Hotel
25	Manhattan	Manhattan Valley	Pizza Place	Bar	Coffee Shop	Indian Restaurant	Yoga Studio	Mexican Restaurant	Playground	Thai Restaurant	Korean Restaurant	Szechuan Restaurant
36	Manhattan	Tudor City	Café	Park	Mexican Restaurant	Coffee Shop	Pizza Place	Deli / Bodega	Diner	Thai Restaurant	Garden	Dog Run
12	Downtown Toronto	Harbord, University of Toronto	Café	Bakery	Japanese Restaurant	Italian Restaurant	Restaurant	Bookstore	Bar	Bank	Beer Store	Gym
13	Downtown Toronto	Chinatown, Grange Park, Kensington Market	Bar	Vietnamese Restaurant	Café	Coffee Shop	Vegetarian / Vegan Restaurant	Bakery	Mexican Restaurant	Dumpling Restaurant	Comfort Food Restaurant	Farmers Market

When ten clusters are segmented, additional mixed cluster is observed. You can see Queen's Park in Downtown Toronto is similar to Manhattan Ville in Manhattan.

	Borough	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
5	Manhattan	Manhattanville	Coffee Shop	Seafood Restaurant	Park	Mexican Restaurant	Italian Restaurant	Spanish Restaurant	Falafel Restaurant	Gastropub	Sushi Restaurant	Lounge
18	Downtown Toronto	Queen's Park	Coffee Shop	Park	Yoga Studio	Sandwich Place	Bar	Sushi Restaurant	Beer Bar	Fried Chicken Joint	Mexican Restaurant	Burger Joint

Note that only key result is highlighted in the report. Please refer to the Jupyter notebook for the entire analysis.

Discussion

It is well known that the desired number of clusters (K) plays a critical role in the K-mean clustering algorithm. To determine the optimal K, people can calculate total within-cluster squared distance by different K values. However, in this project, since the purpose is to find similar neighborhoods (and you may have a neighborhood to compare with), the optimal K value is not that critical, a large K value along with some trials (smaller K values) can give some good insight.

This work by comparing two major cities in USA and Canada is not a comprehensive study, but it can provide a methodology prototype. The approach can be easily extended to any number of cities.

Conclusion

The neighborhoods in Manhattan in New York and Downtown Toronto are compared for similarity. Harbord, University of Toronto, China Town, Grange Park, Kensington Market and Queen's Park are found similar to Manhattan. The work provides a prototype and can be easily extended any number of cities.