

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

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.

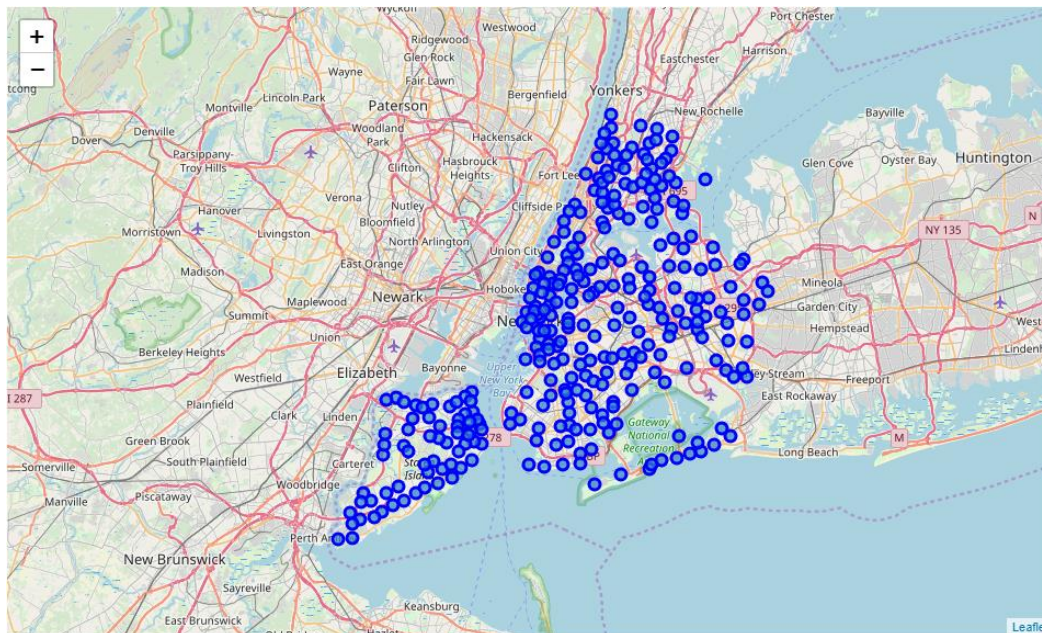
Data preparation

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.

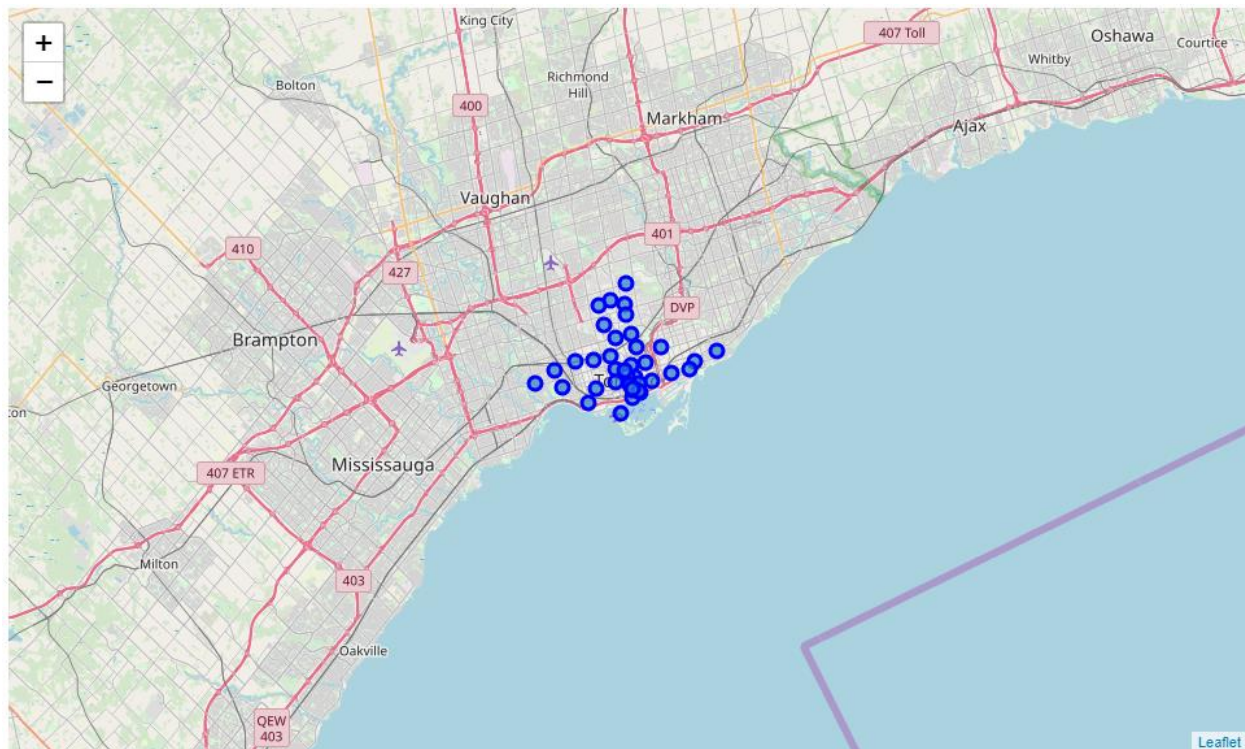
	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

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