### The Battle of Neighborhoods

**Data Scicence Capstone Project Coursera**

**Content:**

1. Introduction

2. Big Picture of the problem

3. Getting Data

4. Exploring the Data

5. Data Preparation

6. Models Building

7. Result

8. Conclusion

**1. Introduction**

In this project I'm going to compare the neighborhoods of cities of Toronto and New York and determine how similar or dissimilar they are. Those cities are very diverse and are the financial capitals of their respective countries. Is New York City more like Toronto?

Let's imagine a scenario where a bank in New York wants to relocate in Toronto, in which neighborhood can we advice them to move on. How can we use data science to recommend them the right place to relocate?

My model will be generalized for any other venues in New York that wants to relocate in any other neighborhood In Toronto. Based on the current location of any venue in New York, the model will assign similar venue in Toronto and vice versa.

For this project, I will frame the problem as an offline unsupervised leaning since we I will not use labeled online data.

I will use K-mean algorithm to cluster the neighborhood based on similar venues.

According to my expertise having visited both cities, Fox Hills in New York is most likely to Parkwoods in Toronto. Let’s see if our model will confirm this assumption.

**2. Getting Data**

The first challenge is to list and get data we need for the project. Here is the list of data we will need.

- Boroughs , Postal code, longitude and latitude for each neighborhood in Toronto

- Boroughs , Postal code, longitude and latitude for each neighborhood in New York and

- An API to build location-aware experiences

New York data are available for free of use at location: [https://geo.nyu.edu/catalog/nyu\_2451\_3457](https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)

I wasn’t able to find similar location for Toronto Data. We will then scrap the web in order to get Toronto data.

For that purpose l used BeautifulSoup4 which is a very good library for web scrapping.

Using the following Wikipedia page as our data source we can extract Postal code, Boroughs and Neighborood of city of Toronto.

**3. Exploring the Data**

New York Data:

After downloading the data in a Pandas Data frame, here is the first five rows of New York data.

|  | 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 |

The shape of data set is the follow: (306, 4). 306 samples and 4 features: Borough, Neighborhood, Latitude, Longitude.

Borough, Neighborhood are text type. Latitude, Longitude are float number.

There are no missing values

Toronto Data:

Here is the first five rows of Toronto data frame after web scrapping.

|  | Postcode | Borough | Neighbourhood |
| --- | --- | --- | --- |
| 0 | M1A | Not assigned | Not assigned/n |
| 1 | M2A | Not assigned | Not assigned/n |
| 2 | M3A | North York | Parkwoods/n |
| 3 | M4A | North York | Victoria Village/n |
| 4 | M5A | Downtown Toronto | Harbourfront/n |

The shape of the data is (289, 3). 289 samples and 3 features

The features Postcode, Borough, Neighbourhood are all from Text Type.

We’re still missing geographical coordinates of each postal code. This can be found using Geocoder package. This will be done in the data preparation section.

We noticed that there are also missing values that we need to address.

**4. Data Preparation**

In this section we need to prepare our dataset.

In order to clean the datasets here are the list of action I did:

- Ignored cells with a borough that is Not assigned to delete missing values.

- Combined into one row the neighborhoods existing in one postal code. The goal here is to eliminate duplicate neighborhoods.

#### - Added latitude and the longitude coordinates of each neighborhood. We saw in previous section that those feature was missing for Toronto dataset.

- Feature selection: here I choose only the necessary columns. Post code from New York dataset.

- Concatenated dataset of Toronto and New York

After the data preparation, the final dataset consist of 4 features and 409 rows.

Here are five first rows.

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

**5. Exploring data set with Foursquare API**

In this section we are utilizing the Foursquare API to explore the neighborhoods and segment them.

For that purpose, we created using Foursquare API a function that return the top 10 venues for each neighborhood in a radius of 500 meters.

Here is how the dataframe returned looks like for “Parkwoods” for instance.

|  | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | Parkwoods | 43.753259 | -79.329656 | Brookbanks Park | 43.751976 | -79.332140 | Park |
| 1 | Parkwoods | 43.753259 | -79.329656 | KFC | 43.754387 | -79.333021 | Fast Food Restaurant |
| 2 | Parkwoods | 43.753259 | -79.329656 | TTC stop #8380 | 43.752672 | -79.326351 | Bus Stop |
| 3 | Parkwoods | 43.753259 | -79.329656 | Variety Store | 43.751974 | -79.333114 | Food & Drink Shop |
| 4 | Parkwoods | 43.753259 | -79.329656 | TTC stop - 44 Valley Woods | 43.755402 | -79.333741 | Bus Stop |

For each neighborhood we can check how many venues were returned

|  | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
| --- | --- | --- | --- | --- | --- | --- |
| Neighborhood |  |  |  |  |  |  |
| Adelaide , King , Richmond | 100 | 100 | 100 | 100 | 100 | 100 |
| Agincourt | 4 | 4 | 4 | 4 | 4 | 4 |
| Agincourt North , L'Amoreaux East , Milliken , Steeles East | 2 | 2 | 2 | 2 | 2 | 2 |
| Albion Gardens , Beaumond Heights , Humbergate , Jamestown , Mount Olive , Silverstone , South Steeles , Thistletown | 10 | 10 | 10 | 10 | 10 | 10 |
| Alderwood , Long Branch | 11 | 11 | 11 | 11 | 11 | 11 |

There are 456 uniques categories from all the returned venues

We then transformed each categorical data to numeric using one hot encoding.

Here is a subset of resulting data frame

| 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 | Coffee Shop | Café | American Restaurant | Steakhouse | Hotel | Gym | Breakfast Spot | Bar | Thai Restaurant | Restaurant |
| 1 | Agincourt | Skating Rink | Breakfast Spot | Clothing Store | Lounge | Women's Store | Falafel Restaurant | Electronics Store | Empanada Restaurant | English Restaurant | Ethiopian Restaurant |
| 2 | Agincourt North , L'Amoreaux East , Milliken ,... | Park | Playground | Women's Store | Falafel Restaurant | Egyptian Restaurant | Electronics Store | Empanada Restaurant | English Restaurant | Ethiopian Restaurant | Event Service |
| 3 | Albion Gardens , Beaumond Heights , Humbergate... | Grocery Store | Pizza Place | Fast Food Restaurant | Liquor Store | Fried Chicken Joint | Sandwich Place | Pharmacy | Coffee Shop | Beer Store | Eye Doctor |
| 4 | Alderwood , Long Branch | Pizza Place | Pool | Bank | Skating Rink | Pharmacy | Sandwich Place | Coffee Shop | Dance Studio | Gym | Pub |
| 5 | Allerton | Pizza Place | Chinese Restaurant | Supermarket | Deli / Bodega | Spa | Department Store | Donut Shop | Pharmacy | Fast Food Restaurant | Discount Store |

**6. Model Building**

Now that we have one dataset consisting of top 10 most common venue for each Neighborhood we can run k-means to cluster the neighborhood.

Let choose for simplicity K = 5. So our dataset will be segmented in 5 clusters.

We will use sci-kit learn package to build our model.

Her is the python code

*# import k-means from clustering stage*

*from sklearn.cluster import Kmeans*

*# set number of clusters*

*kclusters = 5*

*toronto\_grouped\_clustering = toronto\_grouped.drop('Neighborhood', 1)*

*# run k-means clustering*

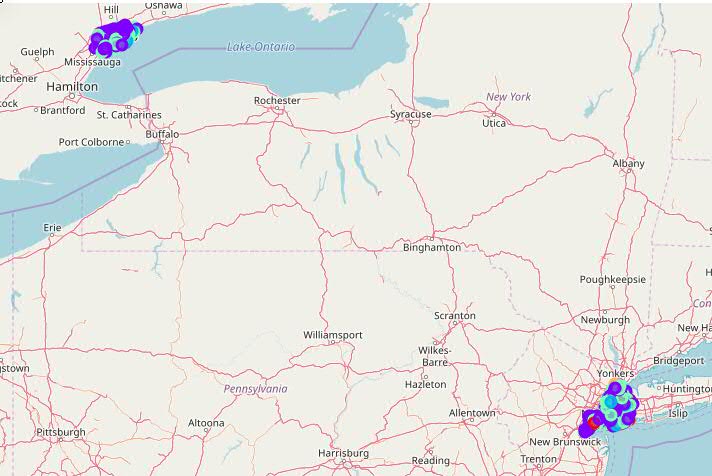
*kmeans = KMeans(n\_clusters=kclusters, random\_state=0).fit(toronto\_grouped\_clustering)*

*# check cluster labels generated for each row in the dataframe*

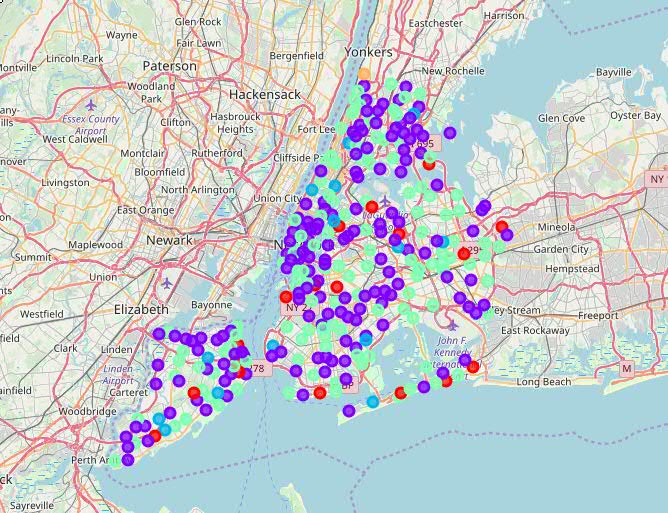
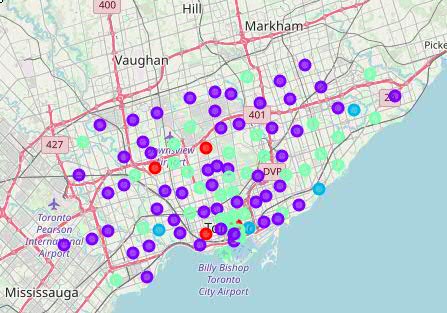
*kmeans.labels\_[0:10]*

**7. Result**

Let’s visualize the result.

****

**New York Toronto**

****

We can see that many Neighborhoods from Toronto and New York are similar.

Cluster 1 in red consist on 22 Neighborhoods. Here are some of them

| 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 | Parkwoods | Bus Stop | Park | Food & Drink Shop | Fast Food Restaurant | Women's Store | Farm | Empanada Restaurant | English Restaurant | Ethiopian Restaurant | Event Service |
| 1 | Victoria Village | Intersection | Portuguese Restaurant | Coffee Shop | Hockey Arena | Falafel Restaurant | Egyptian Restaurant | Electronics Store | Empanada Restaurant | English Restaurant | Ethiopian Restaurant |
| 10 | Cloverdale , Islington , Martin Grove , Prince... | Bank | Women's Store | Eastern European Restaurant | Electronics Store | Empanada Restaurant | English Restaurant | Ethiopian Restaurant | Event Service | Event Space | Exhibit |
| 11 | Highland Creek , Rouge Hill , Port Union | Moving Target | Bar | Women's Store | Fast Food Restaurant | Empanada Restaurant | English Restaurant | Ethiopian Restaurant | Event Service | Event Space | Exhibit |
| 13 | Woodbine Heights | Cosmetics Shop | Curling Ice | Skating Rink | Bus Stop | Video Store | Beer Store | Park | Flower Shop | Food Court | Empanada Restaurant |

Cluster 2 in purple color consist of 205 Neighborhoods is the most common venues in both New York and Toronto.

Here are some of cities in cluster 2

| 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 | Parkwoods | Bus Stop | Park | Food & Drink Shop | Fast Food Restaurant | Women's Store | Farm | Empanada Restaurant | English Restaurant | Ethiopian Restaurant | Event Service |
| 1 | Victoria Village | Intersection | Portuguese Restaurant | Coffee Shop | Hockey Arena | Falafel Restaurant | Egyptian Restaurant | Electronics Store | Empanada Restaurant | English Restaurant | Ethiopian Restaurant |
| 10 | Cloverdale , Islington , Martin Grove , Prince... | Bank | Women's Store | Eastern European Restaurant | Electronics Store | Empanada Restaurant | English Restaurant | Ethiopian Restaurant | Event Service | Event Space | Exhibit |
| 11 | Highland Creek , Rouge Hill , Port Union | Moving Target | Bar | Women's Store | Fast Food Restaurant | Empanada Restaurant | English Restaurant | Ethiopian Restaurant | Event Service | Event Space | Exhibit |

Cluster 3 in blue consist on 18 neighborhood in both Toronto and New York.

| 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 | Harbourfront , Regent Park | Coffee Shop | Pub | Bakery | Park | Café | Theater | Mexican Restaurant | Breakfast Spot | Farmers Market | Bank |
| 21 | Woburn | Coffee Shop | Korean Restaurant | Farmers Market | Electronics Store | Empanada Restaurant | English Restaurant | Ethiopian Restaurant | Event Service | Event Space | Exhibit |
| 57 | Birch Cliff , Cliffside West | General Entertainment | Skating Rink | Café | College Stadium | Farmers Market | Empanada Restaurant | English Restaurant | Ethiopian Restaurant | Event Service | Event Space |

Cluster 4 in green color consist of 157 neighborhood in both Toronto and New York.

| 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 |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 3 | Lawrence Heights , Lawrence Manor | Clothing Store | Women's Store | Furniture / Home Store | Arts & Crafts Store | Miscellaneous Shop | Vietnamese Restaurant | Event Space | Sporting Goods Shop | Boutique | Accessories Store |
| 4 | Queen's Park | Coffee Shop | Diner | Sushi Restaurant | Burger Joint | Gym | Japanese Restaurant | Yoga Studio | Nightclub | Bar | Creperie |
| 5 | Rouge , Malvern | Fast Food Restaurant | Women's Store | Farm | Egyptian Restaurant | Electronics Store | Empanada Restaurant | English Restaurant | Ethiopian Restaurant | Event Service | Event Space |

Cluster 5 consist of only one neighborhood “North Riverdale” located in Toronto

**8. Conclusion**

Cities of Toronto and New York are very similar in terms of neighborhoods. Both cities are financial capitals of their respective countries.

A bank from New York can easily move to Toronto in the Neighborhood which is in the same cluster. In the same way venue in Toronto can easily move to similar neighborhood in New York.

We can also confirm our initial assumption. Fox Hills and Parkwoods appeards in the same cluster.