Identifying similar neighborhoods in different cities.

Aibek Chokotaev March 26, 2019

1. Introduction.

1.1. Problem.

When moving from one city to another the transition and acclimatization might be quite a challenge for a person. Living in one location for several years one can get used to the convenience of the are in terms of presence of gyms, attractions, restaurants, parks, shops, etc. Thus, when looking for a new place to live in a new city one will try to find a place like his previous one. Therefore, it would be convenient to have a handy table which would compare neighborhoods in two cities and show neighborhood pairs which are alike in these two cities. In my analysis, I decided to compare New York and Toronto.

1.2. Interest

I am sure that people who are planning to move from New York to Toronto or vise-versa would be interested in this type of information as it would allow them to avoid the anxiety of researching the neighborhoods and worrying about whether it will be similar to the one that a person has already got used to.

2. Data description.

2.1. Data sources.

For my analysis, I have used multiple sources to obtain the data. Particularly for New York City, I used data from here (New York), and for Toronto, I have used census csv dataset that I have found at Toronto. For venues around the neighborhoods, I scraped foursquare.com and combined it to two data frames for both cities.

2.2. Data cleaning.

To obtain the dataset which will be suitable for my analysis I cleaned and transformed certain columns. First of all, not all the columns that were supposed to provide numeric data were in fact in numeric format, so I had to remove some characters such as “,” and “$” to obtain numeric values. Further, since New York data was in different units compared to Toronto, I had to convert square miles columns into square kilometers. After getting all the values in the numeric and proper form, I normalized the data since the absolute values might distort our analysis due to the differences in the size of the two cities.

Once I have cleaned the data about neighborhoods I have obtained GPS coordinates of neighborhoods and append them to the data frame with all the available information. There were some neighborhoods which I was not able to obtain coordinates for. After checking with a map of the cities on google maps and confirming that those neighborhoods were already captured by bigger other neighborhoods already existing in the dataset, I decided to drop them.

Further, I extracted the data regarding venues around the given neighborhood and populated new pandas data frame with the required information that I obtained. As I could not use text data for my analysis, I created dummy variables for each of the venue categories I observed in foursquare. After creating a dummy variable for each of the venue categories, I grouped it by neighborhoods and took the sum of each type of venues and normalized it to make it comparable between the two cities.

I have identified many types of venues in these two cities, but since I needed to compare two cities, I decided to use an only similar type of venues and drop those columns which were not common for both cities. After completing this part, I obtained two tables with the same set of columns (around 200) and appended them together.

3. Data description.

3.1. Exploratory data analysis.

There different types of venues in both cities but to cluster neighborhoods based on the venues, there should be some similarity between them. Therefore, I decided to get the top 10 most common venues in two cities and see whether there is a similarity. After I grouped the results by type of venues, I obtained the following tables:

Toronto

	Venues	Number of venues
0	Coffee Shop	518
1	Café	269
2	Park	227
3	Pizza Place	227
4	Bakery	185
5	Italian Restaurant	184
6	Sandwich Place	167
7	Grocery Store	149
8	Fast Food Restaurant	136
9	Restaurant	135

New York

	Venues	Number of venues
0	Coffee Shop	169
1	Park	163
2	Italian Restaurant	158
3	American Restaurant	128
4	Pizza Place	128
5	Theater	127
6	Bakery	112
7	Gym	108
8	Hotel	96
9	Café	90

As you can see from the table above, the top 10 types of venues are generally the same although not identical. But it gives us some certainty that clustering based on these parameters might be reasonable.

There are of course other factors that usually influence people's decision when choosing where to live. I have added such information as area, population and income level of the neighborhoods. Some other information as real estate prices, the rating of schools in the neighborhood, rent rates could have improved the quality of clustering, but unfortunately, I was not able to obtain such information for both cities. I have compared the ones that I was able to find, for details, please refer to the table below:

Toronto	New York
Median normalized population	
0.6% of the total population per neighborhood	0.7% of the total population per neighborhood
Median normalized household income in a neighborhood	
18.7% of the most wealthy neighborhood	52.5% of the most wealthy neighborhood
Median normalized area of a neighborhood	
8% of the largest neighborhood	3.2% of the largest neighborhood

Based on these observations I can conclude that the population distribution in these two cities is relatively similar, while based on income Manhattan has either more wealthy neighborhoods compared to Toronto or the gap between richest and other neighborhoods in Toronto is bigger compared to Manhattan.

Although some of the parameters that I have chosen are not similar, I still think that typical person would consider the household income of the neighborhood when choosing where to move in, as the income level of the neighborhood would affect the type of venues and overall quality of the neighborhood.

4. Modeling.

4.1. Model selection.

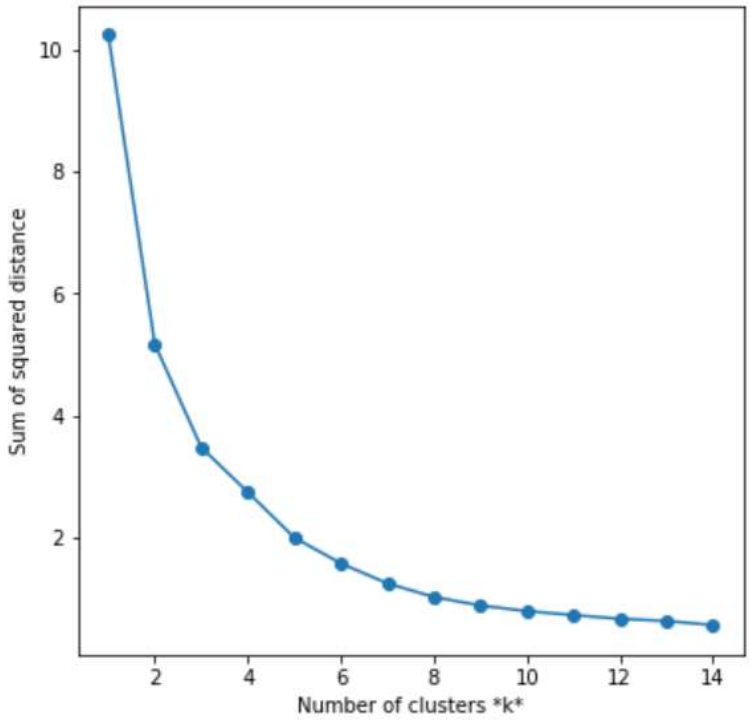
Considering that the dataset does not include and since I want to identify similar groups of neighborhoods, the obvious choice was unsupervised machine learning model, and particularly I decided to use k-means clustering model, which considers all the parameters that I will feed in and assign a cluster to each neighborhood.

In terms of the parameters selection, I decided to use normalized area, population, and income levels and after making sure the types of venues are closely similar in New York and Toronto, I decided to use bigger categories as assigned by Foursquare, where all the venues fall into one of the eight main categories.

To evaluate the accuracy of the model and decide how many clusters to select I used the sum of the squared distance to the cluster centroids. Additionally, as naturally the sum of the squared distance to cluster centroids become smaller with the increasing number of clusters I used elbow method to pick the optimal number of clusters.

4.2. Evaluation and choosing the number of clusters.

As with any unsupervised model, it is hard to intuitively evaluate the performance of the model. In my case, I used the iteration with a different number of clusters and looked for elbow where the slope of change in the sum of the squared distance between points in the cluster and the centroid of a cluster. I plot the results to decide which number of clusters would be the optimal option for my classification problem.



So based on the graph above I decided to use 5 clusters to cluster neighborhoods in New York and Toronto.

4.3. Model results.

After running the model, I observed the following results:

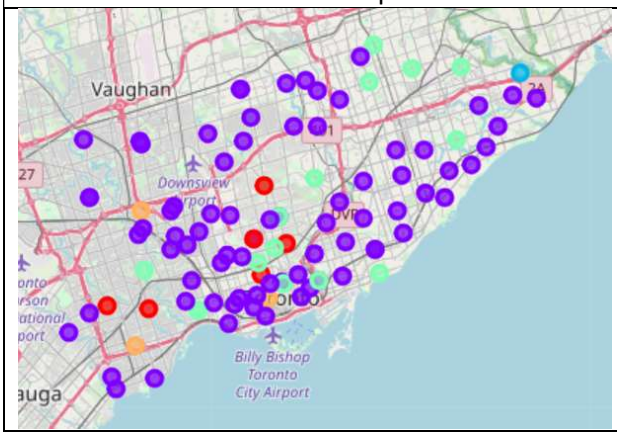
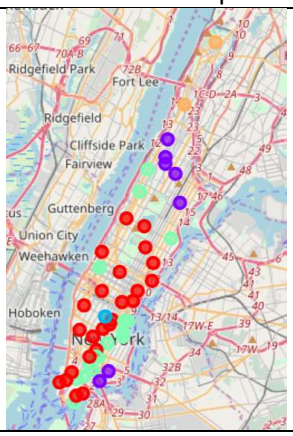
Toronto	New York
Cluster 0	
Annex Forest Hill South Islington-City Centre West Kingsway South Lawrence Park South Rosedale-Moore Park	Battery Park City Central Park Chelsea Garment District Gramercy Park Gramercy-Flatiron Greenwich Village Lenox Hill Midtown Murray Hill NoHo Seaport Soho Sutton Place Tribeca Tudor City

	Turtle Bay Union Square Upper East Side Upper West Side Wall Street West Side West Village World Trade Center
Cluster 1	
Alderwood Bathurst Manor Bay Street Corridor Bayview Village Bayview Woods-Steeles Bendale Black Creek Broadview North Caledonia-Fairbank Centennial Scarborough Clanton Park Cliffcrest Corso Italia-Davenport Danforth Danforth East York Don Valley Village Dorset Park Dufferin Grove East End-Danforth Eglinton East Elms-Old Rexdale Englemount-Lawrence Etobicoke West Mall Flemingdon Park Forest Hill North Glenfield-Jane Heights Guildwood Henry Farm Highland Creek Hillcrest Village Humber Heights-Westmount Humber Summit Ionview Junction Area	East Harlem Hamilton Heights Harlem Lower East Side Manhattanville St.Nicholas Terrace Two Bridges

Keelesdale-Eglinton West Kennedy Park Lansing-Westgate Little Portugal Long Branch Maple Leaf Markland Wood Morningside Moss Park Mount Dennis Mount Pleasant East New Toronto Newtonbrook East Newtonbrook West Niagara North Riverdale North St. James Town O'Connor-Parkview Oakridge Oakwood Village Old East York Palmerston-Little Italy Pleasant View Regent Park Rexdale-Kipling Roncesvalles Runnymede-Bloor West Village Rustic Scarborough Village South Parkdale Steeles Taylor-Massey Thorncliffe Park Trinity-Bellwoods University Victoria Village West Hill Weston Willowdale West Wychwood Yonge-Eglinton York University Heights Yorkdale-Glen Park	

Cluster 2	
Rouge	Uptown
Cluster 3	
Agincourt North Banbury-Don Mills Casa Loma Church-Yonge Corridor Edenbridge-Humber Valley High Park North High Park-Swansea L'Amoreaux Lawrence Park North Malvern Milliken Mount Pleasant West South Riverdale The Beaches Willowdale East Woburn Yonge-St.Clair	Bellevue Bowery Chinatown City Hall Civic Center Downtown East Village Hell's Kitchen Kips Bay Little Italy Manhattan Valley Midtown West Morningside Heights Nolita Peter Cooper Village Stuyvesant Town Yorkville
Cluster 4	
Kensington-Chinatown Pelmo Park-Humberlea Stonegate-Queensway	Inwood Washington Heights

Further, I have plotted the clustered neighborhoods to see if the location in terms of downtown seems to be similar:

Toronto map	New York map
	

It is hard to precisely evaluate whether the results of clustering based on the parameters that I have fed into the model is correct or not as there are no true labels. But as I have tried different parameters, I observed that the model is very sensitive to each parameter.

5. Conclusion.

In this study, I analyzed the neighborhoods in the cities of Toronto and New York with an intent to identify similarities among them and generate recommendations for people moving between these two cities and create a list of similar neighborhoods to make it easier for them to decide where to move. I have observed a high sensitivity to the parameters that are fed into the model. Thus, it is very important to obtain a similar set of data for both cities and decide which parameters are important in choosing the neighborhood to move in. The accuracy of the model would be much better if I would be able to obtain such crucial information for cities such as crime rates, school availability and the average score for them in the neighborhood, housing prices by neighborhoods, rent rates by neighborhood, etc.

Based on the parameters that I have used I have generated a list broken down by clusters and cities with similarities on the parameters fed to the model.