IBM Data Science Professional Certification

The Battle of Neighborhoods

Capstone Project

CONTENT

1
INTRODUCTION

1.1 Backround

In the current situation, many people are losing their jobs because some companies are going bankrupt due to the corona situation. These people then have to find a new job as quickly as possible. Especially when they have a family to support. With the new job, it can quickly happen that you have to move to a new country. Moving to a new city is not easy because you want a similar or better infrastructure and quality of life. In addition, the new place of residence should not be too far away from the new job. Researching where you want to move so that these criteria are met is often tedious and time- consuming. A comparison between the districts of the new area and the hometown would be very helpful to find out which are similar and can therefore be prioritized.

1.2 Data Description

To be able to plan the move with all the factors, there is a lot of information that must be taken into account. It would be helpful to compare different cities to get an overview of their advantages and disadvantages.

Specifically, the problem is as follows: Mr. Smith is a real estate agent and specializes in supporting clients who want to leave their hometown and move to another unfamiliar area. His clients usually do not want to forego anything of their old environment when moving and therefore commission Mr. Smith to find a similar environment with all the previous advantages. Mr. Smith has expertise in machine learning algorithms and clustering. His new client is Mr. Miller. Mr. Miller is 45 years old, is married and has two children. He is a sole earner and lost his job in New York because his company went bankrupt. Now he has found a new job in Toronto and would like to move there with his family. The similarity of the cities is very

important to Mr. Miller and his family. To determine similarity, each neighborhood is described as a numeric vector so clustering is applied, such as DBSCAN and KMeans to group them into different groups.

1.3 Interest

The problem of finding a similar neighborhood is not limited only to the problem described above. This can also be applied to a wide range of applications and takes place in a wide variety of real situations. When a company, such as a restaurant, a cinema, or a supermarket, decides to expand somewhere in a new city or country or to open a new branch. It would be advantageous for the company to find a similar neighborhood because there are often interactive effects between the shops. For example, for a cinema it is beneficial to find a neighborhood where people can easily get to the train station, or a large car park can be built, or in the middle of the city centre within walking distance.

2
DATA DESCRIPTION

A total of three different data sources are required, data from venues, from New York's and Toronto's neighborhoods.

2.1 Venues of FourSquare

The first data source is FourSquare, this is also the main data source. The FourSquare database is used to obtain the advantages of a city. FourSquare is a location-based recommendation service in the form of application software for event locations such as restaurants. It also offers a free API for developers to access its databases. It contains global location data for 190 countries and territories. The venues are categorized, and the geographic coordinates can be accessed.

Above all, categories of event locations are required for the analysis. FourSquare has a total of ten categories, which are divided into a total of 470 sub-categories. The breakdown of the subcategories is as follows:

	Categorie	Amount of Subcategories
0	Arts & Entertainment	38
1	College & University	23
2	Event	12
3	Food	92
4	Nightlife Spot	7
5	Outdoors & Recreation	66
6	Professional & Other Places	44
7	Residence	5
8	Shop & Service	147
9	Travel & Transport	36

It can be seen in the graphic that the main categories are as follows:

'Arts & Entertainment', 'College & University', 'Event', 'Food', 'Nightlife Spot', 'Outdoors & Recreation', 'Professional & Other Places', 'Residence', 'Shop & Service', 'Travel & Transport'

2.2 Neighborhoods and Coordinates of New York

The second data source is a JSON file that contains the names and coordinates of New York's neighborhoods. The names of geographical coordinates can be extracted and saved in a data frame. This is to provide an overview of Mr. Miller's current environment. There is a total of five boroughs with a total of 306 neighborhoods. The boroughs are the following:

'Bronx', 'Manhattan', 'Brooklyn', 'Queens', 'Staten Island'

A small overview of this looks like this:

	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

In addition, the geographical coordinates of the centre of New York are required, which reflect the starting point and the current place of residence of Mr. Miller. These coordinates are obtained using the Python geopy library, which is a Python client for several popular geocoding web services.

Mr. Miller's coordinates are as follows: latitude = 40.7127281, longitude = -74.0060152

2.3 Neighborhoods and Coordinates of Toronto

The third and final data source is Wikipedia. The encyclopedia provides an overview of the Toronto neighborhoods. This is to be the new home of Mr. Miller and his family. The URL is the following:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The table looks like this:

	Postal Code	Borough	Neighbourhood
0	M1A	Not assigned	Not assigned
1	M2A	Not assigned	Not assigned
2	M3A	North York	Parkwoods
3	M4A	North York	Victoria Village
4	M5A	Downtown Toronto	Regent Park, Harbourfront

After cleaning up the data, the geographical coordinates of the neighborhoods can be added again. These can be determined and added using the zip code and the Python library "geocoder". Accordingly, the overview looks like this:

	PostalCode	Borough	Neighborhood	Latitude	Longitude
0	M1B	Scarborough	Malvern, Rouge	43.8114	-79.1966
1	M1C	Scarborough	Rouge Hill, Port Union, Highland Creek	43.7857	-79.1587
2	M1E	Scarborough	Guildwood, Morningside, West Hill	43.7658	-79.1747
3	M1G	Scarborough	Woburn	43.7681	-79.2176
4	M1H	Scarborough	Cedarbrae	43.7694	-79.2389

Toronto has a total of ten boroughs, with a total of 103 neighborhoods. The ten boroughs are as follows:

'Scarborough', 'North York', 'East York', 'East Toronto', 'Central Toronto', 'Downtown Toronto', 'York' 'West Toronto', 'Mississauga', 'Etobicoke'

Finally, we need the geographic coordinates of the centre of Toronto, which is Mr. Miller's new place of residence, in which his new job is. These coordinates are obtained again with the Python library "geopy".

The coordinates of his new job look like this: latitude = 43.6534817, longitude = -79.3839347

3
METHODOLOGY

3.1 Exploratory Data Analysis

For the exploratory data analysis, the individual geographic coordinates of the neighborhoods, which were previously determined, are very important. This can be used to determine the venues of the neighborhoods that are within a radius of 1km from them. Thus, the number of individual venues per neighborhood is different. There is a total of 333 unique sub-categories in Toronto. An overview of the number of venues per neighborhood was created. The following is an excerpt from it:

Neighborhood	Venue	Venue Category
Agincourt	43	43
Alderwood, Long Branch	27	27
Bathurst Manor, Wilson Heights, Downsview North	29	29
Bayview Village	7	7
Bedford Park, Lawrence Manor East	38	38
Berczy Park	100	100
Birch Cliff, Cliffside West	16	16
Brockton, Parkdale Village, Exhibition Place	100	100

For example, the “Malvern, Rouge” neighborhood has the following four unique subcategories. 'Zoo Exhibit', 'Fast Food Restaurant', 'Trail', 'Hobby Shop'

Whereby there are a total of 5 entries of venues. An overview of the venues looks like this:

	Neighborhood	Venue	Venue Category
0	Malvern, Rouge	Canadiana exhibit	Zoo Exhibit
1	Malvern, Rouge	Wendy's	Fast Food Restaurant
2	Malvern, Rouge	Grizzly Bear Exhibit	Zoo Exhibit
3	Malvern, Rouge	Upper Rouge Trail	Trail
4	Malvern, Rouge	Lee Valley	Hobby Shop

3.2 Machine Learning

To find similar neighborhoods, Toronto neighborhoods need to be grouped by venue. This is necessary so that the new hometown Toronto has similar advantages as the current hometown New York. In order to divide the neighborhoods of Toronto into groups, clusters must be formed. A cluster is a group of data points or objects in a data set that are similar to other objects in the group and are not similar to data points in other clusters. The machine learning algorithm clustering is used to form these clusters. This is an unsupervised learning algorithm that are trained on the data set so that conclusions can be drawn about unlabelled data. Little or no information about the data or the expected results is available. For clustering, two different techniques are used to determine the feasibility of this problem. These are KMeans and DBSCAN. DBSCAN stands for Density-Based Spatial Clustering. These clustering methods are two of the most popular unattended algorithms that can be used to solve the current problem. KMeans belongs to the partitioned-based clustering and DBSCAN belongs to the density- based clustering.

3.2.1 DBSCAN

DBSCAN is useful for studying spatial data. It locates high density regions and separates outliers. The density describes the number of points within a specified radius. The algorithm creates clusters of arbitrary shape without being influenced by noise.

The advantages here are that the algorithm is relatively efficient for medium-sized and large data sets and forms clusters of any shape. It can even find clusters that are

completely surrounded by another cluster. In addition, it has an idea of noise and is robust to outliers. Finally, it does not need any information about the number of clusters.

However, this algorithm also has disadvantages. It's a little slower than KMeans in terms of time and complexity. It doesn't work well with clusters of different densities. In this project it turned out that it is not a good cluster algorithm because it could not group the districts of Toronto. This will be because the procedure will return half of the neighborhood as an outlier and the other half in a single cluster.

3.2.2 K-Means

KMeans is mainly used for segmenting customers. It only groups unsupervised data based on how similar it is to each other. In addition, it belongs to the partitioning clustering, which groups the data in K non-overlapping subsets or clusters without a cluster-internal structure or labels. Accordingly, the intra-cluster distances are minimized, and the inter-cluster distances are maximized. Overall, a local optimum is found here.

The advantages of the algorithm are that it is relatively efficient for medium and large data sets. It also creates sphere-like clusters as the clusters are shaped around the centroids.

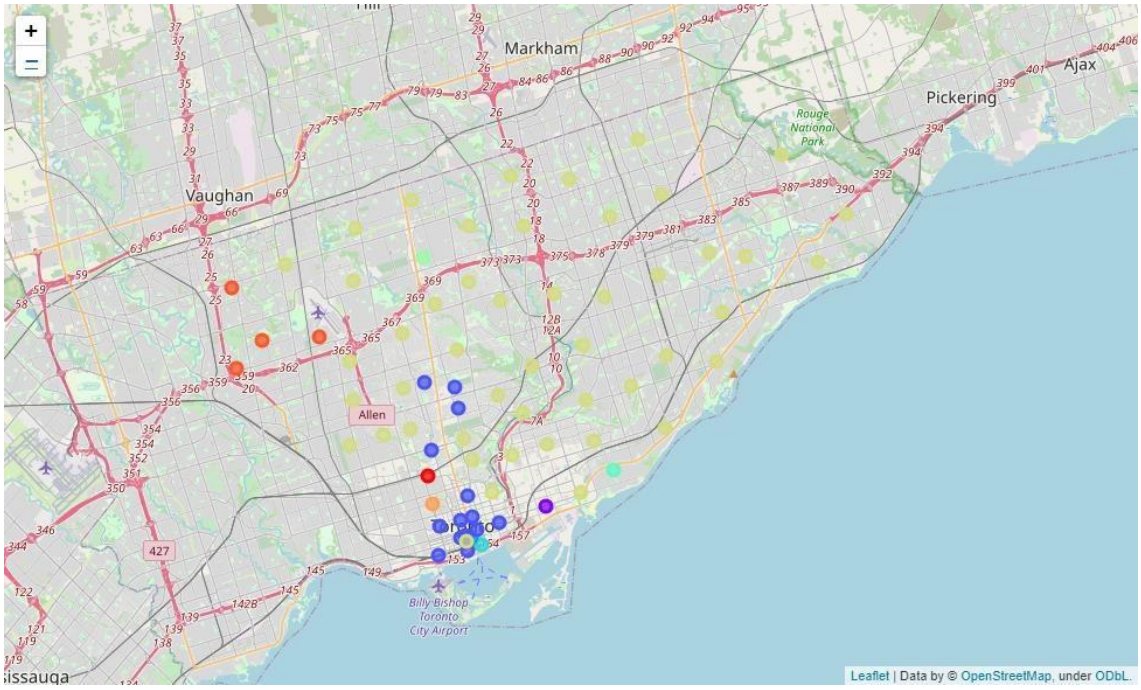
A disadvantage here is that the number of clusters must be specified beforehand.

4
RESULT

The clustering DBSCAN could not group the data set because it only specified one cluster. Accordingly, only the result of the KMeans clustering result is considered in the following. KMeans has formed a total of 10 clusters, with the distribution of how many neighborhoods were assigned in which clusters, as follows:

Cluster	Amount of neighborhood's
0	1
1	1
2	15
3	1
4	1
5	1
6	1
7	47
8	1
9	1

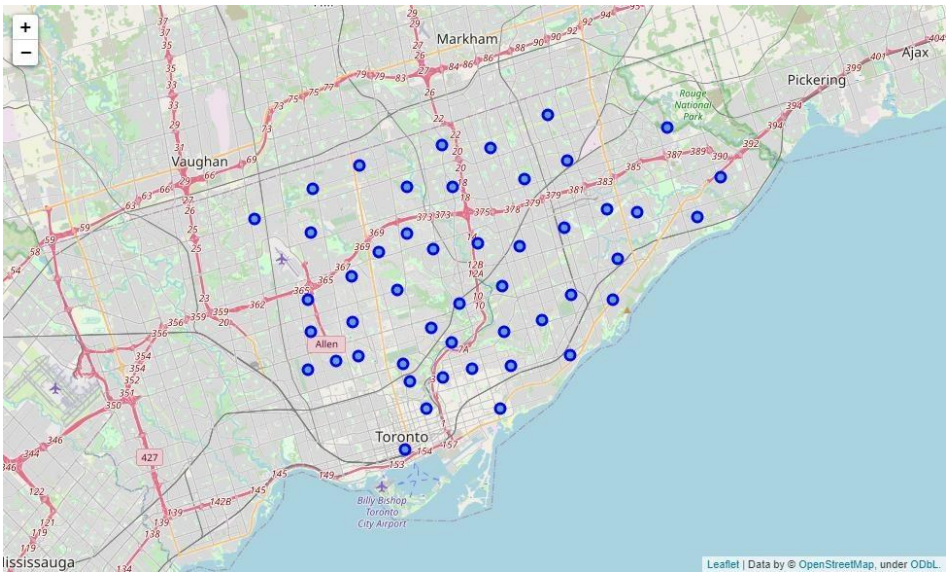
The following map shows exactly where the neighborhoods of the individual clusters are. Each color represents one cluster.



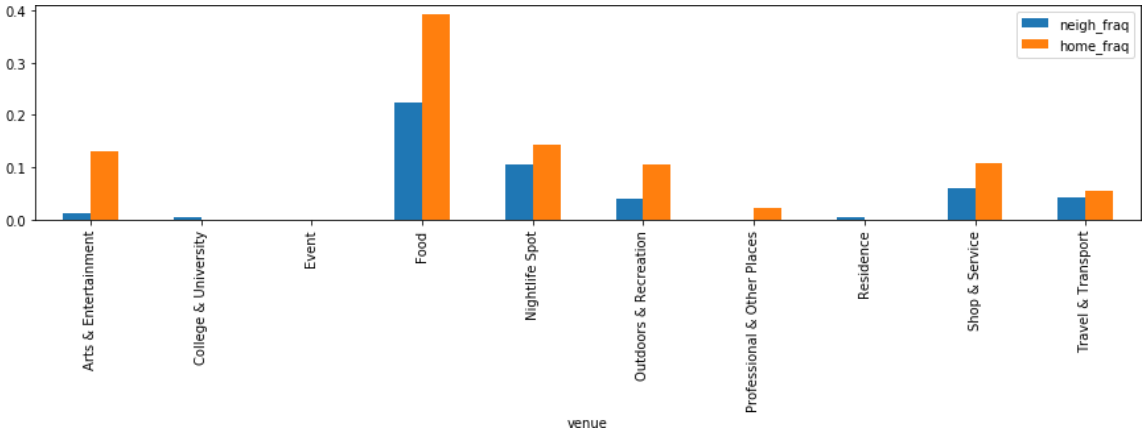
As a result, cluster seven is the most similar. Thus, Mr. Miller's current neighborhoods of New York is most similar to the neighborhoods in cluster No. 7 in Toronto. 47 neighborhoods belong to this cluster. These are the following:

'Agincourt', 'Bathurst Manor, Wilson Heights, Downsview North', 'Bayview Village', 'Bedford Park, Lawrence Manor East', 'Birch Cliff, Cliffside West', 'Caledonia-Fairbanks', 'Cedarbrae', 'Clarks Corners, Tam O'Shanter, Sullivan', 'Cliffside, Cliffcrest, Scarborough Village West', 'Don Mills', 'Dorset Park, Wexford Heights, Scarborough Town Centre', 'East Toronto, Broadview North (Old East York)', 'Fairview, Henry Farm, Oriole', 'Forest Hill North & West, Forest Hill Road Park', 'Glencairn', 'Golden Mile, Clairlea, Oakridge', 'Guildwood, Morningside, West Hill', 'Hillcrest Village', 'Humewood-Cedarvale', 'India Bazaar, The Beaches West', 'Kennedy Park, Ionview, East Birchmount Park', 'Lawrence Manor, Lawrence Heights', 'Lawrence Park', 'Leaside', 'Malvern, Rouge', 'Milliken, Agincourt North, Steeles East, L'Amoreaux East', 'Moore Park, Summerhill East', 'Northwood Park, York University', 'Parkview Hill, Woodbine Gardens', 'Parkwoods', 'Rosedale' 'Roselawn', 'Rouge Hill, Port Union, Highland Creek', 'Scarborough Village', 'St. James Town, Cabbagetown', 'Steeles West, L'Amoreaux West', 'The Danforth West, Riverdale', 'Thorncliffe Park', 'Toronto Dominion Centre, Design Exchange', 'Victoria Village', 'Wexford, Maryvale', 'Willowdale, Newtonbrook', 'Willowdale, Willowdale West', 'Woburn', 'Woodbine Heights', 'York Mills West', 'York Mills, Silver Hills'

Illustrated graphically on the map, the most similar neighborhoods look like this:



Interestingly, these neighborhoods in Toronto are located outside of the downtown area. To validate the result, the frequency of the ten main categories is illustrated. These are compared with Mr. Miller's home based on the mean values of the frequency of the neighborhoods. Illustrated graphically, the comparison of the frequency of the neighborhoods looks like this:



What is particularly noticeable in the graphic is that some categories, such as “Arts & Entertainment” or “Food”, appear much more often in Mr. Miller's current home, so in New York, than in the new area around Toronto. Otherwise, all other categories are relatively close to each other. This result is quite useful and can be used as a reference.

However, in order to make Mr. Smith's work easier, in that he only has to look for a new apartment for Mr. Miller in a certain neighborhood, the closest neighborhood of cluster No. 7 is searched for. To do this, the frequency of the five most common venues is explored by Mr. Miller's current environment. So far, Mr. Miller and his family have placed the most emphasis on the following event categories:

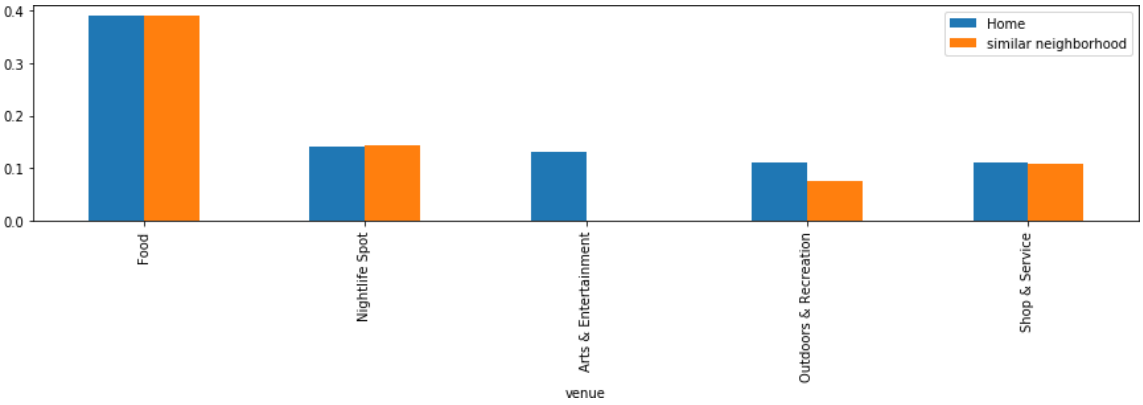
----Home----

	venue	freq
0	Food	0.39
1	Nightlife Spot	0.14
2	Arts & Entertainment	0.13
3	Outdoors & Recreation	0.11
4	Shop & Service	0.11

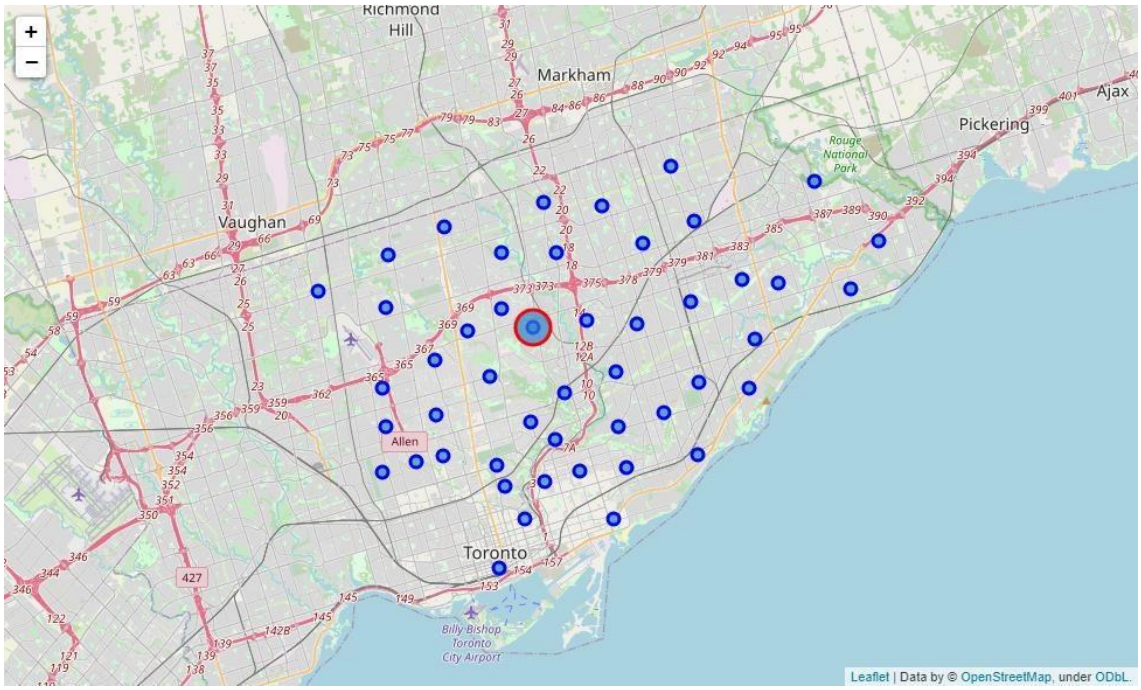
In order to find out the closest neighborhood, the next step is to determine the frequency of the five most common venues. The deviation can then be determined using the mean absolute error. The neighborhood with the smallest deviation can then be selected as the new home for Mr. Miller. An overview of the neighborhoods with the five smallest deviations looks like this:

	Neighborhood	Error
0	Don Mills	0.033912
1	Moore Park, Summerhill East	0.047920
2	Thorncliffe Park	0.068719
3	East Toronto, Broadview North (Old East York)	0.070652
4	Woodbine Heights	0.075870

Based on this, the most similar neighborhood can now be selected. This is "Don Mills". A graphical comparison of the frequency of the five most common venues shows the similarity of the old and new environments.



It is noticeable in the graphic that the individual categories are very similar, with exception of the “Arts & Entertainment” category. If Mr. Miller doesn't mind, Mr. Smith can find a new place to live in this neighborhood. The neighborhood "Don Mills" is in the red marker.


Richmond Hill
Markham
Ajax
Pickering
Rouge National Park
Vaughan
Allen
Toronto
427
Billy Bishop Toronto City Airport
Leaflet | Data by © OpenStreetMap, under ODbL.

5
DISCUSSION

The results of an unsupervised learning model are sometimes difficult to assess. However, these can be reflected on with current knowledge. The result can also be validated with various methods, for example with validation in the sample or with descriptive statistics.

DBSCAN seemed to have a bigger advantage in the beginning. However, it is not robust enough for clusters with different densities over a high-dimensional data space.

For this problem it is also advantageous to decide on a neighborhood since this way the preferred wishes of the client can be better taken into account.

6
CONCLUSION

The problem of finding a similar neighborhood is an important problem as it cannot only be applied to people who want to move. Instead, it can also be used for companies that want to open a new branch and are very satisfied in their current neighborhood so far.

The FourSquare APIs made it easy to get the information they wanted. The individual venues are even subdivided into categories and sub-categories, which facilitated clarity and grouping. In this project, Toronto venues could be broken down into 10 main categories and 333 sub-categories.

The local data from New York and Toronto could be used to verify their venues. With this data, two different cluster algorithms could be used to form clusters. These two cluster algorithms have been DBSCAN and KMeans. It turned out that KMeans clustering has established itself for this problem and has delivered better results. Ten different clusters were formed, so that cluster seven was most similar to the hometown. This cluster includes 47 neighborhoods.

After the cluster with its neighborhoods was illustrated, the most important event categories were verified and compared with those of the hometown. A neighborhood was then selected that was most similar to the hometown. Thus, Mr. Smith has a point of reference for Mr. Miller to hold onto when looking for an apartment. The new neighborhood for Mr. Miller will be "Don Mills". So that he has guaranteed all of his advantages from his current home (New York) in his new home (Toronto).