Clustering of Neighborhoods for Relocation

Albert Olszewski

June 19, 2019

Contents

1 Introduction

1.1 Background

The average person in the United States moves around 11 times in their lifetime [1]. A person may have to move for family or a job opportunity. Relocation is especially prominent for large tech and consulting firms that have locations in multiple large cities. Sometimes relocation has to be done without a prior visit or enough time to research neighborhoods and can result in someone moving to a neighborhood lacking wanted amenities. If a given person has found a neighborhood in their current city that has what they need to be comfortable, it would be nice to find a similar neighborhood in the city that they are moving to. This neighborhood would have similar amenities such as parks, schools, restaurants, and businesses.

1.2 Problem

In this project, we will be using machine learning to group like neighborhoods in Toronto, Chicago, and New York City in order that persons moving between them can find the most familiar and comfortable living situation possible without exhaustive research. The information yielded from this project can be used by persons who must relocate quickly, or broker firms who are looking to provide customers with the most ideal living situation.

2 Data

Neighborhood data used in this project will be obtained using the Foursquare Developers API. We will be collecting the top 10 neighborhood venues from each neighborhood in Toronto, Chicago, and New York City. Listed neighborhoods will be scraped off of tables from a variety of wikipedia pages. All of this data will be merged and collected into a single data frame using the pandas library in Python.

Chicago Neighborhoods: `https://en.wikipedia.org/wiki/List_of_neighborhoods_in_Chicago`
Toronto Neighborhoods: `https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M`
New York Neighborhoods: `https://geo.nyu.edu/catalog/nyu_2451_34572`

The latitude and longitude data for each neighborhood can be found using the geocoder package in python. This package pulls latitude and longitude coordinates from Google. Documentation for this package can be found at `https://geocoder.readthedocs.io` Because this package has been known to be inconsistent at being able to gather data, some cities may have to have corresponding latitude and longitude values entered via csv file. Toronto Nieghborhoods GeoLoc: $Geospatial_Coordinates.csv$

2.1 Toronto

The neighborhoods of Toronto are tabulated on wikipedia, so the names and boroughs will be scraped off of the link in the above section. There were some issues with geocoders for getting the latitude and longitude info, so we will be loading it in from an attached CSV document.

	Borough	Neighborhood	Latitude	Longitude
0	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	Scarborough	Woburn	43.770992	-79.216917
4	Scarborough	Cedarbrae	43.773136	-79.239476

Figure 1: The first five entries of Toronto neighborhood location data.

2.2 Chicago

Neighborhood names and community areas of Chicago are tabulated on a wikipedia page found in the above section. The "community areas" are technically not "boroughs" but will be labeled as such in order to maintain consistency of data tabulation from city to city. The latitude and longitude values for each neighborhood will be found using the geocoders package in Python. If a neighborhood does not have available latitude and longitude data, it will be assigned the latitude and longitude values of their corresponding community area centers.

	Borough	Neighborhood	Latitude	Longitude
0	Albany Park	Albany Park	41.971937	-87.716174
1	Riverdale	Altgeld Gardens	41.654864	-87.600446
2	Edgewater	Andersonville	41.977139	-87.669273
3	Archer Heights	Archer Heights	41.811422	-87.726165
4	Armour Square	Armour Square	41.840033	-87.633107

Figure 2: The first five entries of Chicago neighborhood location data.

2.3 New York City

There is a '.json' file available for all of the New York City Data. The borough, neighborhood, latitude, and longitude data will be parsed into a data frame in Python.

	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

Figure 3: The first five entries of New York City neighborhood location data.

2.4 Foursquare

Now that we have data frames of each city with borough, neighborhood, latitude, and longitude data we can start to request venue data from FourSquare. Foursquare has made developer tools that allow us to input a location and recieve rich information such as nearby restaurants, parks and recreation, schools, and businesses. You can obtain reviews, websites, and more of these places. For the purpose of this project, we will only be tracking the types of amenities and the frequency of them in neighborhoods for future clustering.

2.5 Processing

In order to perform clustering analysis on the data we have collected, we must prepare it using one hot encoding [2]. One hot encoding is a process of using dummy variables to assign categorical values a numerical value without introducing a hierarchy to the categories. It is very useful for the machine

learning techniques that we will be using on the data. In our project, we will be using one hot encoding on the neighborhood venues in order to examine the top 10 venue types for each neighborhood.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue
0	Adelaide,King,Richmond	Coffee Shop	Café	American Restaurant	Steakhouse	Bar
1	Agincourt	Breakfast Spot	Sandwich Place	Lounge	Chinese Restaurant	Food Court
2	Agincourt North,L'Amoreaux East,Milliken,Steel...	Park	Asian Restaurant	Playground	Yoga Studio	Falafel Restaurant
3	Albany Park	Sandwich Place	Cocktail Bar	Pizza Place	Chinese Restaurant	Hookah Bar
4	Albion Gardens,Beaumond Heights,Humbergate,Jam...	Grocery Store	Pizza Place	Coffee Shop	Sandwich Place	Fried Chicken Joint

Figure 4: A portion of the final data that will be used for clustering. Commonality of venues is decided by frequency.

3 Methods

3.1 K-Means Clustering

In this project, we want to cluster neighborhoods together based on like characteristics. We want to use a clustering machine learning technique because we don't know exactly what type of categories these neighborhoods are going to be. K-Means clustering algorithms will be used because it is a simple algorithm that performs very well. The data set we are analyzing is large enough that we need a clustering algorithm that is efficient, but it is not so large that we need to use DBSCAN.

An ideal value of K will be chosen using the elbow method. We will run K-means clustering on the data while looping through different values of K. The sum of squared distances of each point to its assigned cluster center will be used as an error descriptor.

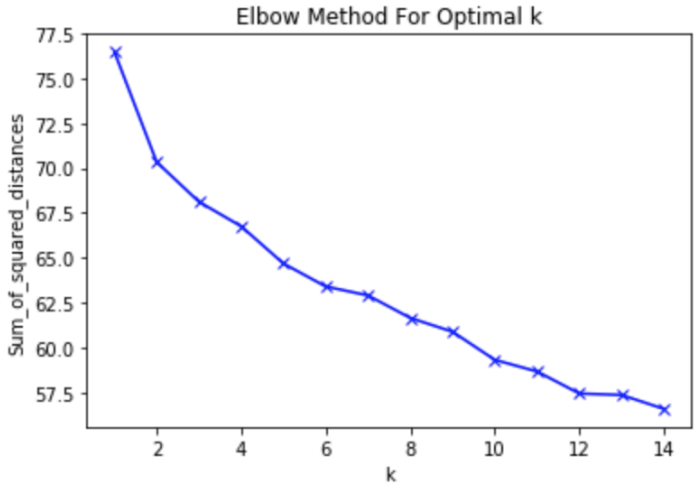

Figure 5: This is a plot of the sum squared distances of each data point to their assigned cluster center versus the value of K.

The "Sum of Squared Distances" will decrease as long as K increases. We will plot these values and choose a value of K where the rate at which error decreases seems to drop off (i.e. the elbow). We choose

K to be equal to 6.

3.2 Results

Now that we have ran the clustering algorithms on all of the neighborhoods together, we must now separate the cities again to plot and discuss them. For each data frame we want the city, borough, neighborhood, latitude, longitude, and cluster number.

After compiling the data into the data frames, we are able to plot the clustered neighborhoods on maps of the respective cities. The clusters are represented as follows:

- Cluster 0 : red
- Cluster 1 : purple
- Cluster 2 : blue
- Cluster 3 : cyan
- Cluster 4 : light green
- Cluster 5 : orange

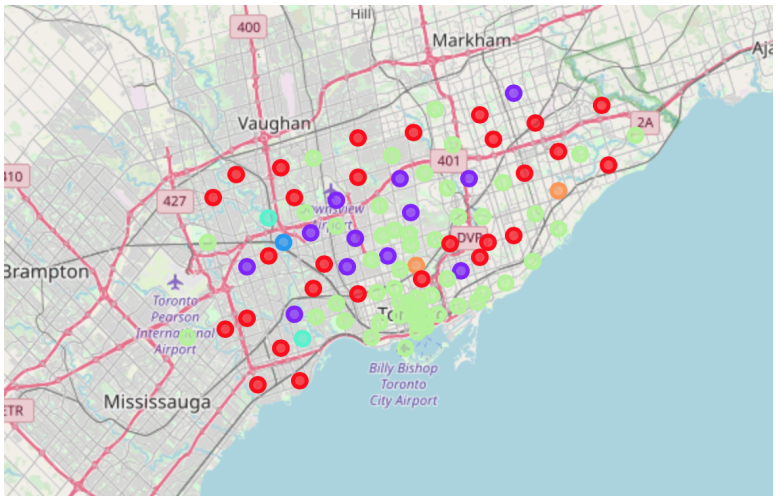

Figure 6: Map of Toronto with clustered neighborhoods plotted.

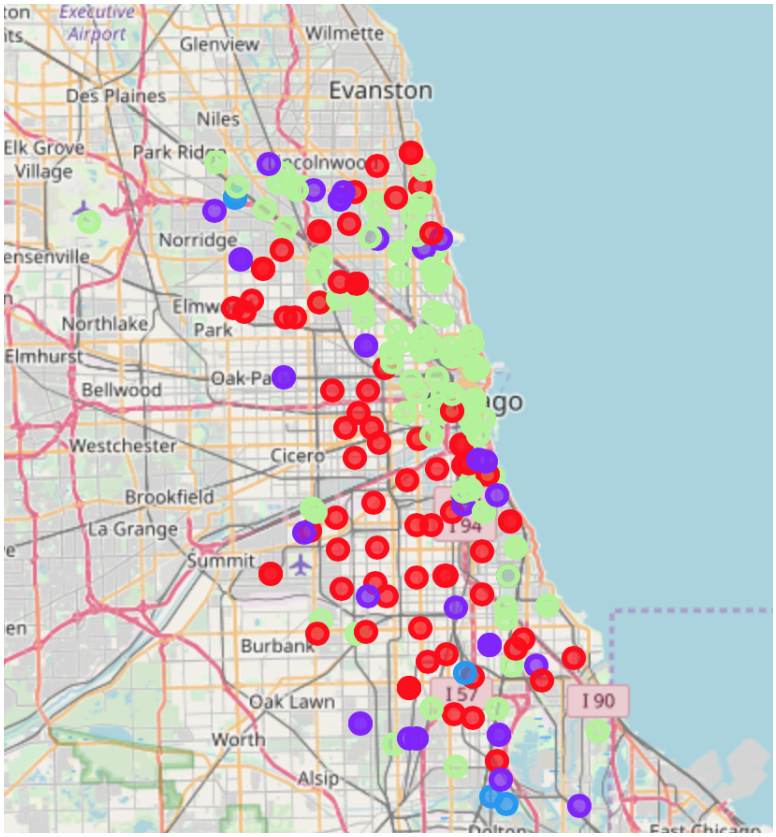

Figure 7: Map of Chicago with clustered neighborhoods plotted.

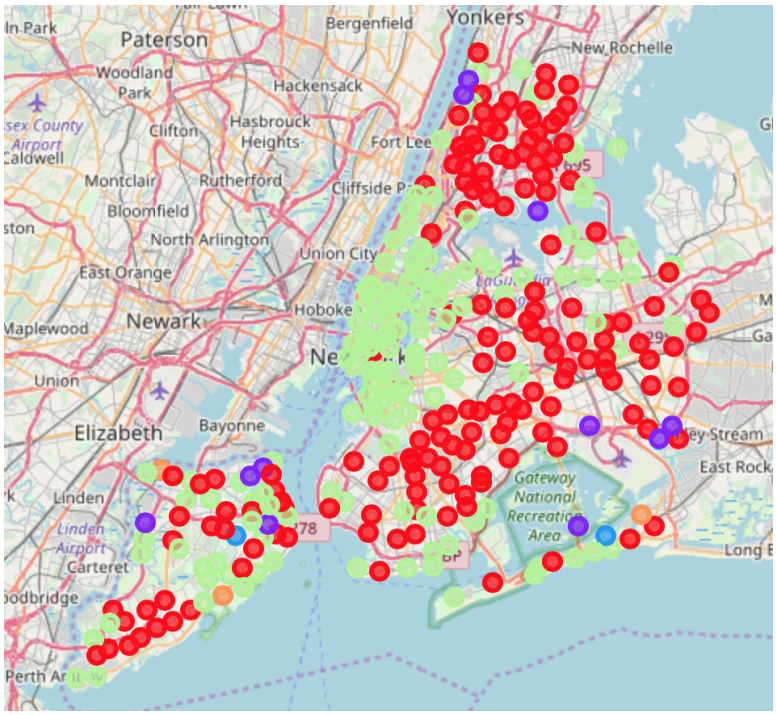

Figure 8: Map of New York City with clustered neighborhoods plotted.

4 Discussion

Now that we have clustered the data and mapped it, we should take a deeper look at what the clusters contain in order to describe it to the client.

Venue	Count
Pizza Place	80
Chinese Restaurant	48
Mexican Restaurant	34
Fast Food	32
Pharmacy	31

Table 1: A count of how many times a venue is the most common in a neighborhood for Cluster 0.

Neighborhoods that fall into Cluster 0 are heavily populated with pizza places, bodegas, chinese food, and pharmacies. The neighborhoods are located outside of the city centers or downtown areas. Due to the amenities available and the location of these neighborhoods, they are likely a little more affordable and residential than neighborhoods in Cluster 4. This cluster is the 2nd largest with 267 neighborhoods.

Venue	Count
Park	51
Fast Food	7
Bus Station	6
Deli / Bodega	6
Convenience Store	5

Table 2: A count of how many times a venue is the most common in a neighborhood for Cluster 1.

Cluster 1 is heavily populated with parks, bus stations, convenience stores, food trucks, and intersections. These neighborhoods offer more space, and would be very suitable for commuters because of the large availability of transportation options. Cluster 1 contains 63 neighborhoods from all of the cities involved in this study.

Venue	Count
Park	8
Yoga Studio	8
Fountain	4
Campground	2
Grocery Store	1

Table 3: A count of how many times a venue is the most common in a neighborhood for Cluster 2.

Cluster 2 is a smaller cluster with only 8 neighborhoods in it; however, it does have neighborhoods in all 3 cities. These neighborhoods are heavily populated with parks, yoga studios, campgrounds, and grocery stores. This cluster is very similar to Cluster 3. They both likely have a lot of schools and are good places to house families based off of their amenities.

Venue	Count
Electronics Store	2
Yoga Studio	2
Baseball Stadium	2

Table 4: A count of how many times a venue is the most common in a neighborhood for Cluster 3.

Cluster 3 is also a very small cluster with only 2 neighborhoods in it. It is a very specific type of neighborhood containing Egyptian restaurants, baseball fields, electronic stores, and yoga studios as the most common venues. The only neighborhoods in this cluster are in Toronto, so if someone is looking for a neighborhood similar to these two neighborhoods they would have to look at another cluster for Chicago or New York City.

Venue	Count
Coffee Shop	103
Bar	53
Italian Restaurant	41
c Sandwich Shop	c 40
c Cafe	c 34

Table 5: A count of how many times a venue is the most common in a neighborhood for Cluster 4.

Neighborhoods that fall into Cluster 4 are heavily populated with coffee shops, bars, and Italian restaurants. If we look at the maps with the clusters plotted on them, it looks like these locations are located closer to the downtown areas of each city. This cluster is likely higher end living and located close to businesses. Someone who live in a Cluster 4 of one neighborhood is likely to live in a busy part of the city that is heavily populated. Cluster 0 is also the largest cluster with 306 neighborhoods in it.

Venue	Count
Yoga Studio	3
Playground	3
Bar	2
Tennis Court	2
Park	1

Table 6: A count of how many times a venue is the most common in a neighborhood for Cluster 5.

Cluster 5 only contains neighborhoods in Toronto and New York City. There is only 5 total neighborhoods in this cluster. The most popular venues in this neighborhood are tennis courts, yoga studios, playgrounds, and parks. This neighborhood is also similar to Clusters 2 and 3 but might be more suitable for a young family (given that there are plenty of playgrounds.

5 Conclusion

In this project, we clustered neighborhoods together from Toronto, Chicago, and New York City based on the amenities they offered. Real time data was scraped off of the internet or gathered using the Foursquared Developer API. A K-Means clustering algorithm was used to explore the collected data. This model will always be current when the code is run. This model can be useful for those who are relocating between these cities on short notice, but want to keep a similar neighborhood that they are comfortable living in. Brokers who work in real estate can use this data to provide better services to their customers. There are neighborhoods in every city that offer similar living situations for families, commuters, physically active workers, and deep city lovers. The data presented in this project gives rich information to those looking to move to either Toronto, Chicago, or New York City making it easier for them to make a comfortable decision on what neighborhood to live in.

5.1 Future Works

The information gathered in this project is very useful and insightful. However, there is possibility for further development to make the material present more robust. There are more major cities in North America that can be included in this study such as San Francisco, Seattle, or Houston. By including these cities, we can expand our reach of audience. It is completely possible that neighborhoods have different amenities than the ones listed on Foursquare. Data can be gathered from more sources to include housing price, square footage, surrounding cultures, religious organizations, and other aspects considered during moving. As developers it is important for us to make this information more usable for the consumer, so a user friendly tool needs to be developed for cross-referencing neighborhoods. Overall, there are many ways to expand on this exciting insight and make it more useful to consumers.

6 References

References

[1] Chandler, Adam. "Why Do Americans Move So Much More Than Europeans?" October 2016. Web.

[2] Hacker Noon. "What Is One Hot Encoding? Why And When Do Your Have To Use It?" August 2017. Web.