

Capstone Project - The battle of Neighborhood

Introduction/Business Problem

The aim of this report is to study and analyze the neighborhoods of Toronto city and group them into similar clusters and, to analyze those clusters to gather meaningful information. That information can be used to find out neighborhoods that are same as your current neighborhood or at least similar. This information provided by this report would be valuable for people who are interested in relocating to different part of the city and are looking to find new neighborhoods that are very similar to their existing neighborhood.

2. Problem Description:

Now let me explain the context of this Capstone project through a scenario. Say you live on the west side of the city of Toronto in Canada. You love your neighborhood, mainly because of all the great amenities and other types of venues that exist in the neighborhood, such as gourmet fast food joints, pharmacies, parks, graduate schools and so on. Now say you receive a job offer from a great company on the other side of the city with great career prospects. However, given the far distance from your current place you unfortunately must move if you decide to accept the offer.

Wouldn't it be great if you are able to determine neighborhoods on the other side of the city that are the same as your current neighborhood, and if not perhaps similar neighborhoods that are at least closer to your new job?

2.1 Data Description:

To consider the objective stated above, we can list the below data sources used for the analysis.

a) Toronto Neighborhood Data:

The following Wikipedia page was scraped to pull out the necessary information: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The information obtained i.e. the table of postal codes, borough and neighborhood was transformed into a pandas data frame for further analysis.

b) Coordinate data for each Neighborhood in Toronto:

The following csv file was used to get the latitude and longitude for the Neighborhood http://cocl.us/Geospatial_data

3.0 Methodology

a. Scrape the Wikipedia page to get data and convert into Pandas dataframe

To start with our analysis, we used the **Beautiful Soup** package to transform the Wikipedia data into a pandas dataframe.

Kindly see figure below;

| | PostalCode | Borough | Neighbourhood |
|---|------------|------------------|------------------|
| 0 | M3A | North York | Parkwoods |
| 1 | M4A | North York | Victoria Village |
| 2 | M5A | Downtown Toronto | Harbourfront |
| 3 | M6A | North York | Lawrence Heights |
| 4 | M6A | North York | Lawrence Manor |

We also got the coordinate data for all neighborhoods in Toronto using the csv file and converted into a pandas dataframe.

| | Postal Code | Latitude | Longitude |
|---|-------------|-----------|------------|
| 0 | M1B | 43.806686 | -79.194353 |
| 1 | M1C | 43.784535 | -79.160497 |
| 2 | M1E | 43.763573 | -79.188711 |
| 3 | M1G | 43.770992 | -79.216917 |
| 4 | M1H | 43.773136 | -79.239476 |

The dataframe's were merged i.e. adding the coordinate dataframe to the original dataframe.

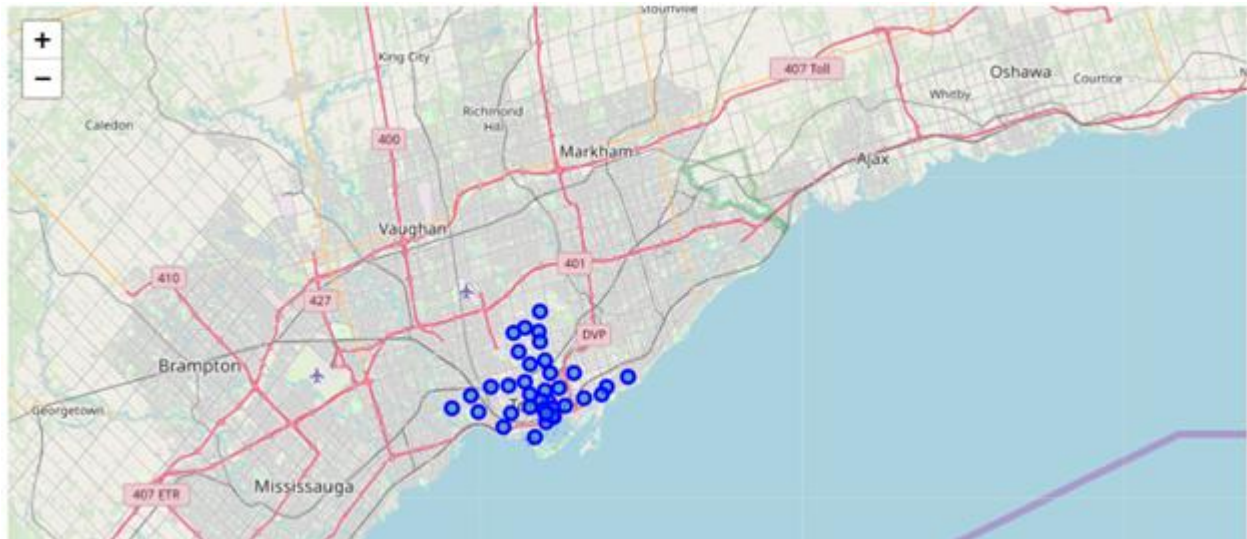
| | PostalCode | Borough | Neighborhood | Latitude | Longitude |
|---|------------|-------------|--|-----------|------------|
| 0 | M1B | Scarborough | Rouge, Malvern | 43.806686 | -79.194353 |
| 1 | M1C | Scarborough | Highland Creek, Rouge Hill, Port Union | 43.784535 | -79.160497 |
| 2 | M1E | Scarborough | Guildwood, Morningside, West Hill | 43.763573 | -79.188711 |
| 3 | M1G | Scarborough | Woburn | 43.770992 | -79.216917 |
| 4 | M1H | Scarborough | Cedarbrae | 43.773136 | -79.239476 |

b. Generating a map of Toronto and plotting the Neighborhood data on it

We first filter the data to find boroughs containing the word "Toronto",

| | PostalCode | Borough | Neighborhood | 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 |

We then use the python **folium** library to visualize geographic details of Toronto and its boroughs. I created a map of Toronto with boroughs superimposed on top using the latitude and longitude values to get the visual as below:



C. Utilizing Foursquare API to explore the neighborhoods

Next, start utilizing the Foursquare API to explore the neighborhoods and segment them. We set the LIMIT parameter to **100**, which would limit the number of venues returned by the Foursquare API and the radius of 50 meter. Below is the first five (5) list of Nearby Venues for the various Towns (Borough) i.e. McCowan Park, Price Chopper

| | Borough | Borough Latitude | Borough Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|-----------------|------------------|-------------------|-------------------------------------|----------------|-----------------|----------------------|
| 0 | Scarborough | 43.744734 | -79.239476 | McCowan Park | 43.745089 | -79.239336 | Playground |
| 1 | Scarborough | 43.799525 | -79.318389 | Price Chopper | 43.799445 | -79.318563 | Grocery Store |
| 2 | North York | 43.803762 | -79.363452 | New York Fries | 43.803664 | -79.363905 | Fast Food Restaurant |
| 3 | Central Toronto | 43.704324 | -79.388790 | Jules Cafe Patisserie | 43.704138 | -79.388413 | Dessert Shop |
| 4 | Central Toronto | 43.704324 | -79.388790 | Thobors Boulangerie Patisserie Café | 43.704514 | -79.388616 | Café |

d. Analyze each neighborhood

Using One Hot Encoding, the data frame was standardized and the data was grouped by neighborhoods, we created a new data frame consisting of the top 10 venues in each neighborhood.

| | 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 |
|---|------------------|-----------------------|-----------------------|------------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|-----------------------|
| 0 | Central Toronto | Supermarket | Liquor Store | Furniture / Home Store | Dessert Shop | Park | Italian Restaurant | Coffee Shop | Seafood Restaurant | Bowling Alley |
| 1 | Downtown Toronto | Coffee Shop | Café | Performing Arts Venue | Gym | Deli / Bodega | Sushi Restaurant | Art Gallery | Concert Hall | Bakery |
| 2 | East Toronto | Trail | Yoga Studio | Diner | Coffee Shop | College Gym | Concert Hall | Cosmetics Shop | Dance Studio | Deli / Bodega |
| 3 | East York | Coffee Shop | Sporting Goods Shop | Sandwich Place | Housing Development | Indian Restaurant | College Gym | Concert Hall | Cosmetics Shop | Dance Studio |
| 4 | Etobicoke | Dance Studio | Coffee Shop | Pizza Place | Diner | College Gym | Concert Hall | Cosmetics Shop | Deli / Bodega | Dessert Shop |

From the results there were some common venue categories in the neighborhoods. So there is a need to segment these similarities for easy identification.

e. K-means Algorithm

K-means is vastly used for clustering and useful when you need to discover insights of unlabeled data.

K-means uses Euclidean distance hence doesn't work well with categorical variables The unsupervised K-means algorithm was used to cluster the neighborhoods. K-Means algorithm is one of the most common method for clustering and discovering insights.

We use a k-cluster value of 5 to split the neighborhoods into 5 different clusters based on the similarity of the venues.

4.0 Results:

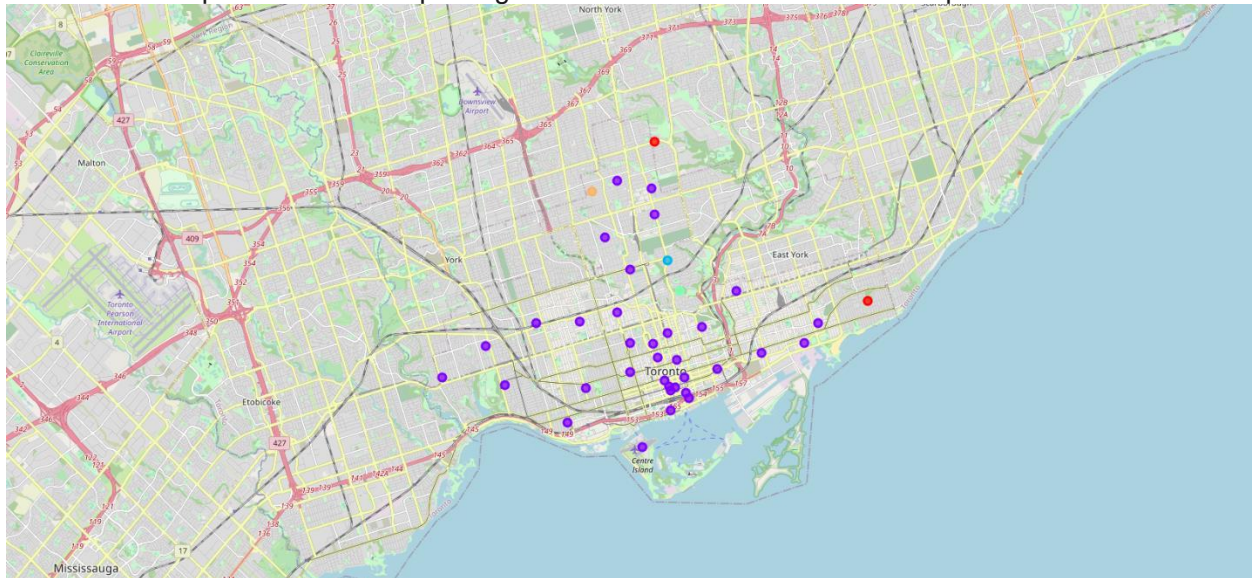
a. Adding the Cluster Labels to the Venue Data

After using K-means algorithm for clustering of the data, the cluster labels were then added to easily identify when neighborhood belong to the different clusters. The below table depicts the clustered data along with the top 10 most common venues in that cluster.

| | Cluster Labels | 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 | 1 | Berczy Park | Coffee Shop | Cocktail Bar | Bakery | Cheese Shop | Cafe | Restaurant | Beer Bar | Seafood Restaurant | Greek Restaurant | Gourmet Shop |
| 1 | 1 | Brockton, Parkdale Village, Exhibition Place | Cafe | Coffee Shop | Breakfast Spot | Stadium | Bar | Italian Restaurant | Bakery | Restaurant | Climbing Gym | Furniture / Home Store |
| 2 | 1 | Business reply mail Processing Centre, South C. | Yoga Studio | Auto Workshop | Skate Park | Light Rail Station | Smoke Shop | Spa | Farmers Market | Fast Food Restaurant | Burrito Place | Restaurant |
| 3 | 1 | CN Tower, King and Spadina, Railway Lands, Har... | Airport Service | Airport Lounge | Airport Terminal | Harbor / Marina | Bar | Plane | Coffee Shop | Rental Car Location | Sculpture Garden | Boat or Ferry |
| 4 | 1 | Central Bay Street | Coffee Shop | Italian Restaurant | Japanese Restaurant | Sandwich Place | Cafe | Department Store | Salad Place | Bubble Tea Shop | Burger Joint | Korean Restaurant |

b. Visualizing the resulting Clusters

We use the matplotlib and folium packages to visualize the clusters on a map of Toronto.



5.0 Discussion:

The goal of this analysis was carried out to find out similar neighborhoods for a person relocating within the city of Toronto.

As we analyze the results section, we can analyze the clusters and see similar neighborhoods in different parts of the city. For example, if we compare the different neighborhoods clustered in cluster 2.

As seen in the table above, if a person wished to move from a suburb region in Downtown Toronto to Central Toronto. If a person's current location were in the Neighborhood of Studio District in Downtown Toronto, which has venues like cafes, Gym, Art Gallery and Sushi restaurants nearby, the person, would like to relocate to a neighborhood like North York which also has venues like Coffee Shops and Restaurants. This is just one example of how our data analysis can help people relocate from one part of the city to another.

6.0 Conclusion

In a world like ours driven by data, there are many real-life problems or scenarios where data can be used to find solutions to those problems. As seen in the example above, data was used to cluster neighborhoods in Toronto based on the most common venues in those neighborhoods hence someone that is searching for a neighborhood where there are shops, restaurants and gym places can use this data as a guide to relocate.

References:

- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
- CSV for Coordinate data: http://cocl.us/Geospatial_data
- Foursquare API