Battle of the Neighborhoods

by Mark

1.0 Introduction

Mr. Smith was originally from the United States and he is starting a new job in Toronto. He will be moving to Toronto together with his wife and their two children. He is now looking for a place to stay in Toronto but he does not know where to start. As there are so many neighborhoods within Toronto City, he wishes to narrow down the scope. This means coming out with a list of neighborhoods that fit his preferences best and he can do further filtering from there.

He prefers to choose to stay in a neighborhood that are close to the following venues:

1. Ease of access to a public transport
2. Close to schools for his children
3. Close to supermarkets for regular grocery shopping
4. A park or a garden

He does not prefer to the following close to his neighborhood:

1. A bar or a pub (He rarely drinks)
2. Restaurants or coffee shops
3. Hotel
4. Event halls

By leveraging location data from Foursquare, a K-means clustering approach will be used to determine the neighborhood that fits best to his preferences.

2.0 Data Requirements

The data analysis via K-means clustering algorithm will require the following:

1. The postal codes, boroughs and neighborhoods in Toronto, Canada
    - This data will be extracted from Wikipedia ('https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M)
    - Do note that all postal codes start with M are located within the city of Toronto.
    - For example:

Postal Code	Borough	Neighborhood
M3A	North York	Parkwoods

2. Geospatial data
    - Data extracted from Wikipedia contain postal codes, boroughs and neighborhoods
    - However, the results do not contain latitude and longitude which are important for Foursquare API to determine the venues around the neighborhood.
    - The data will be extracted from http://cocl.us/Geospatial_data
    - Postal codes found from Wikipedia will be matched with the corresponding latitude and longitude tabulated in the csv file.
    - For example:

Postal Code	Borough	Neighborhood	Latitude	Longitude
M3A	North York	Parkwoods	43.7532586	-79.3296565

3. Venues available within 1 km from the center of the neighborhood.
    - This data will be extracted from Foursquare API
    - Venue category is used to determine the type of venue located within 1 km radius of the neighborhood.
    - For example: A school, a supermarket and a school within the 1 km radius of M3A – North York.
    - The frequency of each venue category can be calculated.

3.0 Methodology

Data cleaning is first performed to remove postal codes which do not contain any boroughs or neighborhoods. Since there are no neighborhoods in these postal codes, there is no houses and Mr. Smith cannot choose to stay there in anyway.

The longitudes and latitudes of each postal code is determined and linked to the postal codes. These geospatial coordinates is inputted to Foursquare API to determine venues that are located within 1 km (1000 m) of the center of the coordinates. Since most of Mr. Smith's preferences are considered to be venue categories, this attribute will be used as the independent variable for the analysis. Venue categories for each venue located within 1000 m of the center of each neighborhood will be extracted from Foursquare API and used for the K-means clustering approach. Number of venues extracted for each neighborhood is limited to 100.

After the venue categories for each venue in the neighborhood have been extracted, one hot encoding is used to transform each venue category to categorical variables. Then, the mean frequency of each venue category in the neighborhood is calculated. We then extract the top ten (most common) and bottom ten (least common) venue category for each neighborhood. This reduces the effect of noise in the clustering algorithm, and thus, producing better results.

Iteration procedure is performed to determine the best value of K that can differentiate the neighborhood types based on the provided data. If the value of K is too large, it may result in too much noises and clusters with very small data sets, making the screening approach more difficult.

If the value of K is too small, it may result in clusters that are relatively too big, making the further screening approach much difficult as the data in each cluster is too many.

This approach is performed by manually increasing K and looking at the results. If the datasets show no significant difference between two clusters, the value of K is increased until some differences in the cluster characteristics can be observed. Based on the analysis performed, K=7 shows to have the best performance.

4.0 Results

103 neighborhoods in Toronto have been clustered into 7 respective clusters using K-means algorithm. The results of each of the cluster are shown in the table and figure below.

Cluster 1

	Borough	Cluster Labels	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
3	North York	0.0	Fast Food Restaurant	Coffee Shop	Restaurant	Accessories Store	Dessert Shop	Vietnamese Restaurant	Fried Chicken Joint	Sushi Restaurant
6	Scarborough	0.0	Fast Food Restaurant	Trail	Coffee Shop	Hobby Shop	Restaurant	Bus Station	Supermarket	Caribbean Restaurant
7	North York	0.0	Restaurant	Coffee Shop	Japanese Restaurant	Asian Restaurant	Burger Joint	Gym	Supermarket	Bank
10	North York	0.0	Grocery Store	Fast Food Restaurant	Gym	Pizza Place	Gas Station	Park	Coffee Shop	Bank
13	North York	0.0	Restaurant	Coffee Shop	Japanese Restaurant	Asian Restaurant	Burger Joint	Gym	Supermarket	Bank
16	York	0.0	Convenience Store	Pizza Place	Coffee Shop	Sushi Restaurant	Grocery Store	Sandwich Place	Bakery	Field
18	Scarborough	0.0	Pizza Place	Bank	Coffee Shop	Fast Food Restaurant	Grocery Store	Sports Bar	Food & Drink Shop	Electronics Store

Based on the Cluster 1 as figure above, it can be seen that this cluster contains neighborhood with lots of restaurants from various cuisines. There are total of 20 neighborhoods with these characteristics. Most common venues in these boroughs are restaurants, which are not what Mr. Smith wanted. Neighborhoods of this cluster should not be recommended to Mr. Smith.

Cluster 2

	Borough	Cluster Labels	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
2	Downtown Toronto	1.0	Coffee Shop	Theater	Diner	Café	Park	Pub	Breakfast Spot	Italian Restaurant
4	Downtown Toronto	1.0	Coffee Shop	Park	Sushi Restaurant	Italian Restaurant	Japanese Restaurant	Gastropub	Ramen Restaurant	Restaurant
9	Downtown Toronto	1.0	Coffee Shop	Gastropub	Japanese Restaurant	Café	Theater	Italian Restaurant	Restaurant	Plaza
15	Downtown Toronto	1.0	Coffee Shop	Café	Restaurant	Gastropub	Seafood Restaurant	Hotel	Italian Restaurant	Theater
19	East Toronto	1.0	Pub	Coffee Shop	Pizza Place	Breakfast Spot	Bakery	Beach	Japanese Restaurant	Burger Joint
20	Downtown Toronto	1.0	Coffee Shop	Café	Hotel	Japanese Restaurant	Restaurant	Seafood Restaurant	Park	Gastropub
24	Downtown Toronto	1.0	Coffee Shop	Ramen Restaurant	Diner	Café	Clothing Store	Sushi Restaurant	Japanese Restaurant	Gastropub
25	Downtown Toronto	1.0	Korean Restaurant	Café	Coffee Shop	Grocery Store	Ice Cream Shop	Mexican Restaurant	Cocktail Bar	Indian Restaurant
30	Downtown Toronto	1.0	Coffee Shop	Café	Theater	Hotel	Japanese Restaurant	American Restaurant	Restaurant	Concert Hall

Based on the Cluster 2 as figure above, it can be seen that this cluster contains neighborhood with lots of coffee shops and some hotels. There are total of 42 boroughswith these characteristics. This cluster seem like neighborhoods with lots of tourists attractions, definitely not something that Mr. Smith wanted. Therefore, neighborhoods in this cluster should not be recommended to Mr. Smith too.

Cluster 3

	Borough	Cluster Labels	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
12	Scarborough	2.0	Breakfast Spot	Park	Playground	Burger Joint	Italian Restaurant	Yoga Studio	Event Space	Dumpling Restaurant
101	Etobicoke	2.0	Park	Italian Restaurant	Eastern European Restaurant	Ice Cream Shop	Gym / Fitness Center	Ethiopian Restaurant	Donut Shop	Dry Cleaner

Based on the Cluster 3 as figure above, it can be seen that this cluster contains neighborhood with less restaurants compared to Cluster 1 and 2, which is good for Mr. Smith. However, apart from the park, there are not much other characteristics that fit into Mr. Smith's preferences. There are a total of 2 boroughs in this cluster. We'll keep this cluster just in case no other cluster has better characteristics than this one.

Cluster 4

	Borough	Cluster Labels	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
0	North York	3.0	Park	Convenience Store	Pharmacy	Shopping Mall	Bus Stop	Food & Drink Shop	Laundry Service	Caribbean Restaurant
1	North York	3.0	Coffee Shop	Boxing Gym	Hockey Arena	Sporting Goods Shop	Park	Portuguese Restaurant	Pizza Place	Golf Course
5	Etobicoke	3.0	Pharmacy	Grocery Store	Skating Rink	Bank	Convenience Store	Playground	Bakery	Shopping Mall
8	East York	3.0	Pizza Place	Brewery	Fast Food Restaurant	Athletics & Sports	Soccer Stadium	Bus Line	Breakfast Spot	Café
11	Etobicoke	3.0	Park	Hotel	Pizza Place	Theater	Restaurant	Bank	Fish & Chips Shop	Clothing Store

Based on the Cluster 4 as figure above, it can be seen that this cluster contains neighborhood do not have much restaurants and coffee shops. In fact, most of them have convenience stores, shopping malls, parks and bus stops/ lines. Neighborhoods in this cluster definitely fit best to Mr. Smith preferences. In fact, we can see that the first row (Index 0) meets the three out of four attributes that Mr. Smith wanted, a park, a shopping mall and a bus stop. That should be good for Mr. Smith! There are 35 neighborhoods in these cluster and Mr. Smith can do further screening on each neighborhood, focusing on this cluster.

Cluster 5

	Borough	Cluster Labels	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
45	North York	5.0	Park	Pool	Yoga Studio	Falafel Restaurant	Dry Cleaner	Dumpling Restaurant	Eastern European Restaurant	Electronics Store

Based on Cluster 5 as figure above, there is only one borough in this cluster. Similar to Cluster 3, although this cluster do not have much restaurants and coffee shops, there are not many attributes that are to the liking of Mr. Smith. There is only a park that Mr. Smith prefers. This cluster will be kept as a backup, just like Cluster 3.

Cluster 6

	Borough	Cluster Labels	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
94	Etobicoke	6.0	Hotel	Rental Car Location	Coffee Shop	Yoga Studio	Falafel Restaurant	Dry Cleaner	Dumpling Restaurant	Eastern European Restaurant

Based on Cluster 6 as figure above, this cluster is definitely not what Mr. Smith ones. The most common venue is the hotel, what Mr. Smith wants to stay out off. Therefore, this cluster should not be recommended to Mr. Smith.

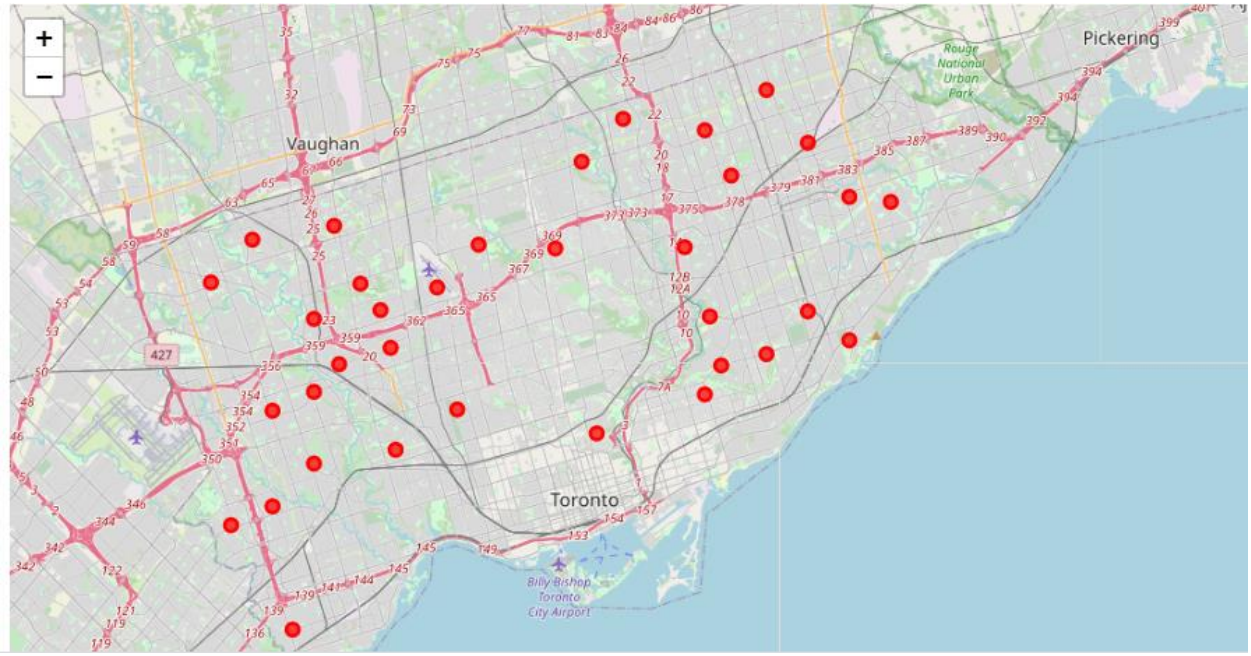
5.0 Discussion

Based on the K-means clustering performed for the 103 neighborhoods in Toronto, the results show that Cluster 1 and 2 consist of neighborhoods that has most restaurants and coffee shops. Although most of Cluster 2 neighborhoods have a park, it only fits into one preference of Mr. Smith. Therefore, Cluster 1 and 2 neighborhoods are definitely not where Mr. Smith wants to stay. There are a total of 20 and 42 boroughs in Cluster 1 and 2, respectively.

For Cluster 3 and 5, the number of coffee shops and restaurants are considerably less than Cluster 1 and 2. However, apart from the presence of the park as one of the most common venue categories, there are no other attributes that fit into Mr. Smith's preferences. There are only a total of 2 and 1 boroughs classified under Cluster 3 and 5, respectively. It is better to keep this as backup clusters if Mr. Smith couldn't find a better one.

Most common venue category in Cluster 6 is a hotel, exactly what Mr. Smith wants to stay out off. There are no other attributes in this category that fit to Mr. Smith preferences. So, this will not be recommended to Mr. Smith.

Based on the clustering algorithm performed, Cluster 4 seemed to be the most promising cluster, with neighborhoods that have less restaurants and do not have a hotel. There are 35 neighborhoods in this cluster and Mr. Smith should focus his further screening approach on this cluster. In fact, most of the neighborhoods in this cluster contain convenience store/shopping mall, parks and bus stops/lines. This is meeting 3 out of 4 of Mr. Smith's preferences. The figure below shows the location of each neighborhood under Cluster 4, which looks to be quite a distance away from Toronto City Centre. Mr. Smith will have to do further evaluations to pick the best one from here.



6.0 Conclusion

103 neighborhoods in Toronto have been clustered based on the most common and least common venue categories in each neighborhood. Results have shown that Cluster 4 contains neighborhoods that fit best to Mr. Smith's preferences. Although not all the neighborhoods are 100% fitting Mr. Smith's preferences, it has reduced the total number of 103 neighborhoods to 35 only. Mr. Smith can do further analysis, focusing only on these 35 neighborhoods, reducing his effort and time to look at the other neighborhoods that definitely do not fit into his preferences. Further analysis can still be performed with Cluster 4, using other attributes such as average house prices and crime index to determine the one that is most suited for Mr. Smith.