Groceries Warehouse Recommender System

Nader Asadi

1 Introduction

This project is about a groceries contractor in Toronto city. This contractor provides places such as: Different types of Restaurants, Bakery, Breakfast Spot, Brewery and Café with fresh and high-quality groceries. The contractor wants to build a warehouse for the groceries it buys from villagers and farmers inside the borough, so that they will support more customers and also bring better "Quality of Service" to the old customers. It is important where to select as the location of warehouse. For example, if the warehouse location is selected near a famous restaurant, not only the cost of transportation diminishes but also the quality of service increases. Which neighborhood (in that borough) would be a better choice for the contractor to build the warehouse in that neighborhood. Finding the right neighborhood is our mission and our recommender system will provide this contractor with a sorted list of neighborhoods in which the first elemnt of the list will be the best suggested neighborhood.

2 Data

We will need geo-locational information about that specific borough and the neighborhoods in that borough. We specifically and technically mean the latitude and longitude numbers of that borough. We assume that it is "Scarborough" in Toronto. This is easily provided for us by the contractor, because the contractor has already made up his mind about the borough. The Postal Codes that fall into that borough (Scarborough) would also be sufficient fo us. I fact we will first find neighborhoods inside Scarborough by their corresponding Postal Codes.
We will need data about different venues in different neighborhoods of that specific borough. In order to gain that information we will use "Foursquare" locational information. By locational information for each venue we mean basic and advanced information about that venue. For example there is a venue in one of the neighborhoods. As basic information, we can obtain its precise latitude and longitude and also its distance from the center of the neighborhood. But we are looking for advanced information such as the category of that venue and whether this venue is a popular one in its category or maybe the average price of the services of this venue. A typical request from Foursquare will provide us with the following information:

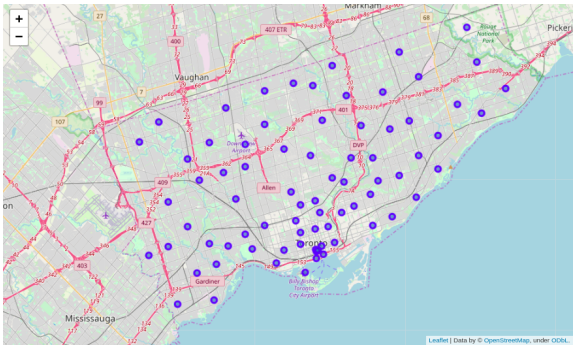
[Postal Code] [Neighborhood(s)] [Neighborhood Latitude] [Neighborhood Longitude] [Venue] [Venue Summary] [Venue Category] [Distance (meter)]

3 Methodology

3.1 Identifying Neighborhoods inside "Scarborough"

We will use Postal Codes of different regions inside Scarborough to find the list of neighborhoods. We will essentially obtain our information from $https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada :_M$ and then process the table inside this site. Images from dataframes and also from maps will be provided in the presentation. Here we only present our strategy and how we got the mission accomplished.

	Unnamed: 0	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	0	M8W	Etobicoke	Alderwood, Long Branch	43.602414	-79.543484
1	1	M4N	Central Toronto	Lawrence Park	43.728020	-79.388790
2	2	M6H	West Toronto	Dovercourt Village	43.669005	-79.442259
3	3	M2L	North York	York Mills	43.757490	-79.374714
4	4	M2H	North York	Hillcrest Village	43.803762	-79.363452



3.2 Connecting to Foursquare and Retrieving Location Data

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 1000 meter. It means that we have asked Foursquare to find venues that are at most 1000 meter far from the center of the neighborhood. (I think distance is measured by latitude and longitude of venues and neighborhoods, and it is not the walking distance for venues.)

3.3 Processing the Retrieved Data

When the data is completely gathered, we will perform processing on that raw data to find our desirable features for each venue. Our main feature is the category of that venue. After this stage, the column "Venue's Category" will be One-hot encoded and different venues will have different feature-columns. After On-hot encoding we will integrate all restaurant columns to one column "Total Restaurants" and all food joint columns to "Total Joints" column. We assumed that different restaurants use the Same raw groceries. This assumption is made for simplicity and due to not having a very detailed dataset about different venues. Now, the dataset is fully ready to be used for machine learning (and statistical analysis) purposes.

Neighborhood	Postal Code	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Summary	Venue Category	Distance
Agincourt	50	50	50	50	50	50	50
Agincourt North, Milliken	25	25	25	25	25	25	25
Birch Cliff	15	15	15	15	15	15	15
Clairlea, Golden Mile, Oakridge	29	29	29	29	29	29	29
Cliffcrest, Cliffside	13	13	13	13	13	13	13

	Unnamed: 0	Postal Code	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Summary	Distance	American Restaurant	Asian Restaurant	Automotive Shop	BBQ Joint	Badminton Court
0	0	M1V	Agincourt North, Milliken	43.815252	-79.284577	Fahmee Bakery & Jamaican Foods	This spot is popular	669	0	0	0	0	0
1	1	M1V	Agincourt North, Milliken	43.815252	-79.284577	Jim Chai Kee Wonton Noodle 沾仔記	This spot is popular	689	0	0	0	0	0
2	2	M1V	Agincourt North, Milliken	43.815252	-79.284577	Lotus Pond Vegetarian Restaurant 蓮花素食	This spot is popular	934	0	0	0	0	0
3	3	M1V	Agincourt North, Milliken	43.815252	-79.284577	DaanGo Cake Lab	This spot is popular	809	0	0	0	0	0
4	4	M1V	Agincourt North, Milliken	43.815252	-79.284577	The Brighton Convention & Event Centre	This spot is popular	890	0	0	0	0	0

3.4 Applying K-Means Clustering

Here we cluster neighborhoods via K-means clustering method. We think that 5 clusters is enough and can cover the complexity of our problem. After clustering we will update our dataset and create a column representing the group for each neighborhood.

	Bakery	Breakfast Spot	Diner	Fish Market	Food & Drink Shop	Fruit & Vegetable Store	Grocery Store	Noodle House	Pizza Place	Sandwich Place	Total Restaurants	Total Joints	Total Sum
G3	2.000000	1.000000	0.000000	0.000000e+00	0.000000e+00	0.000000	1.000000	1.0	2.000000	2.000000	21.000000	1.000000	31.000000
G4	1.000000	0.000000	0.000000	0.000000e+00	0.000000e+00	0.000000	1.000000	0.5	1.500000	1.500000	12.500000	2.000000	20.000000
G1	1.500000	0.500000	0.000000	5.000000e-01	0.000000e+00	0.000000	1.500000	0.5	2.500000	0.000000	8.500000	1.000000	16.500000
G5	0.000000	0.000000	0.000000	0.000000e+00	3.333333e-01	0.000000	1.000000	0.0	3.000000	0.666667	3.666667	1.666667	10.333333
G2	0.333333	0.166667	0.333333	1.387779e-17	1.387779e-17	0.166667	0.166667	0.0	0.333333	0.500000	3.833333	0.166667	6.000000

3.5 Decision Making and Reporting Results

Now, we focus on the centers of clusters and compare them for their "Total Restaurants" and their "Total Joints". The group which its center has the highest "Total Sum" will be our best recommendation to the contractor. Note: Total Sum = Total Restaurants + Total Joints + Other Venues. This algorithm although is pretty straightforward yet is strongly powerful.