A Recommender System for Groceries Contractor

Abhishek Kumar

1. Introduction

Problem Description:
There is a groceries contractor in one of the boroughs of Toronto (Scarborough). 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.

For example, if the warehouse is close to those old and famous restaurants, then the vegetables and other groceries would be delivered to the restaurant in the right time and there would be no delay so the restaurant cooks can start their job from the morning and the Quality of Service will be high and this contractor will gain more reputation and income.

The contractor should build this warehouse where it is closest to its customers in order to minimize the cost of transportation in addition to the example above. which neighbourhood (in that borough) would be a better choice for the contractor to build the warehouse in that neighbourhood. Finding the right neighbourhood is our mission and our recommender system will provide this contractor with a sorted list of neighbourhoods in which the first element of the list will be the best suggested neighbourhood.

2. Data acquisition

We will need geo-locational information about that specific borough and the neighbourhoods 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 for us. I fact we will first find neighbourhoods inside Scarborough by their corresponding Postal Codes.

We will need data about different venues in different neighbourhoods 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 neighbourhoods. As basic information, we can obtain its precise latitude and longitude and also its distance from the center of the neighbourhood. 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:

```
scarborough_venues.head()
```

	Postal Code	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Summary	Venue Category	Distance
0	M1W	Steeles West	43.799525	-79.318389	Mr Congee Chinese Cuisine 龍粥記	This spot is popular	Chinese Restaurant	72
1	M1W	Steeles West	43.799525	-79.318389	Agincourt Bakery	This spot is popular	Bakery	759
2	M1W	Steeles West	43.799525	-79.318389	Little Sheep Mongolian Hot Pot 小肥羊	This spot is popular	Hotpot Restaurant	972
3	M1W	Steeles West	43.799525	-79.318389	Phoenix Restaurant 金鳳餐廳	This spot is popular	Chinese Restaurant	147
4	M1W	Steeles West	43.799525	-79.318389	Price Chopper	This spot is popular	Grocery Store	16

[Postal Code]	**[Neighbourhood(s)]**	**Latitude**	**Longitude**	**Venue**	**Venue Summary**	**Venue Category**	**Distance (meter)**
M1L	Clairlea, Golden Mile, Oakridge	43.711112	79.284577	Tim Hortons	This spot is popular	Coffee Shop	592

Some Notes about "Foursquare": https://foursquare.com

Foursquare is a local search-and-discovery service mobile app which provides search results for its users (Wikipedia).
Founded: New York City, New York, U.S
Users: 60 million
Date launched: March 11, 2009
Employees: Over 200
Founders: Dennis Crowley, Naveen Selvadurai
Owner: Foursquare Labs, Inc.

3. Exploratory Data Analysis

3.1 Identifying Neighbourhoods inside "Scarborough"

We will use Postal Codes of different regions inside Scarborough to find the list of neighbourhoods. 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.

```
scarborough_data
```

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1W	Scarborough	Steeles West	43.799525	-79.318389
1	M1J	Scarborough	Scarborough Village	43.744734	-79.239476
2	M1G	Scarborough	Woburn	43.770992	-79.216917
3	M1C	Scarborough	Highland Creek, Rouge Hill, Port Union	43.784535	-79.160497
4	M1N	Scarborough	Birch Cliff	43.692657	-79.264848
5	M1R	Scarborough	Maryvale, Wexford	43.750072	-79.295849
6	M1V	Scarborough	Agincourt North, Milliken	43.815252	-79.284577
7	M1H	Scarborough	Cedarbrae	43.773136	-79.239476
8	M1T	Scarborough	Tam O'Shanter	43.781638	-79.304302
9	M1M	Scarborough	Cliffcrest, Cliffside	43.716316	-79.239476
10	M1E	Scarborough	Morningside, West Hill	43.763573	-79.188711
11	M1B	Scarborough	Rouge, Malvern	43.806686	-79.194353
12	M1S	Scarborough	Agincourt	43.794200	-79.262029
13	M1K	Scarborough	Ionview, Kennedy Park	43.727929	-79.262029
14	M1P	Scarborough	Dorset Park, Scarborough Town Centre, Wexford ...	43.757410	-79.273304
15	M1X	Scarborough	Upper Rouge	43.836125	-79.205636
16	M1L	Scarborough	Clairlea, Golden Mile, Oakridge	43.711112	-79.284577

Identification of postal codes

```
The geograpical coordinate of "Scarborough" are: 43.773077, -79.257774.
```

3.2 Connecting to Foursquare and Retrieving Locational Data for Each Venue in Every Neighbourhood

After finding the list of neighbourhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighbourhood. For each neighbourhood, 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 neighbourhood. (I think distance is measured by latitude and longitude of venues and neighbourhoods, and it is not the walking distance for venues).

3.3 Processing the Retrieved Data and Creating a DataFrome for All the Venues inside the Scarborough

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.

3.4 Applying one of Machine Learning Techniques (K-Means Clustering)

Here we cluster neighbourhoods 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 neighbourhood.

3.5 Final Metadata:

Out[42]:

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
Agincourt	2	1	0	0	0	0	0	1	1	2	20	0
Agincourt North, L'Amoreaux East, Milliken, Steeles East	1	0	0	0	0	0	0	1	2	0	10	1
Birch Cliff, Cliffside West	0	0	1	0	0	0	0	0	0	0	3	0
Cedarbrae	4	0	0	0	0	0	1	0	1	0	7	3
Clairlea, Golden Mile, Oakridge	2	0	1	0	0	0	1	0	1	1	3	0
Clarks Corners, Sullivan, Tam O'Shanter	0	0	0	0	0	0	1	1	1	2	13	1
Cliffcrest, Cliffside, Scarborough Village West	0	0	0	0	0	0	0	0	3	0	3	2
Dorset Park, Scarborough Town												

3.6 Clustering:

- K-Means Clustering used
- Number of clusters considered- 5

Showing Centers of Each Cluster

In [44]:

```
means_df = pd.DataFrame(kmeans.cluster_centers_)
means_df.columns = scarborough_onehot.columns
means_df.index = ['G1','G2','G3','G4','G5']
means_df['Total Sum'] = means_df.sum(axis = 1)
means_df.sort_values(axis = 0, by = ['Total Sum'], ascending=False)
```

Out[44]:

	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
G4	2.000000	1.000000	0.0	0.00	0.00	0.000000	0.000000	1.0	1.000000	2.000000	20.0	0.000000	27.000000
G1	0.500000	0.250000	0.0	0.25	0.00	0.000000	1.250000	0.5	1.500000	0.750000	11.0	1.750000	17.750000
G2	2.333333	0.333333	0.0	0.00	0.00	0.333333	0.666667	0.0	0.666667	0.666667	7.0	1.333333	13.333333
G5	0.000000	0.250000	0.0	0.00	0.25	0.000000	0.500000	0.0	2.000000	0.750000	4.5	1.250000	9.500000
G3	0.500000	0.250000	0.5	0.00	0.00	0.000000	0.250000	0.0	0.250000	0.250000	2.5	0.250000	4.750000

Clustering

4. 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.

Results:

Based on this analysis, the best recommended neighbourhood will be:
{'Neighbourhood': 'Agincourt',
'Postal Code': 'M1S',
'Neighbourhood Latitude': 43.7942003,
'Neighbourhood Longitude': -79.26202940000002}