Recommendations for Business Opportunities in Toronto

F.Topsever

20.03.2021

1. Introduction

1.1. Background

Toronto is one of the most multicultural cities in the world. Its population consists of many visible minorities. It is easy to recognize the trend, that Toronto is an attractive city for many foreigners around the world, not only for touristic purpose, but also for living, working or education. In 2016, 51,5% of the residents of the city belonged to a visible minority group, with a total population of 2,7 million residents. The population grew by 4,3% from 2011 to 2016 (Source: https://en.wikipedia.org/wiki/Demographics_of_Toronto)

In this capstone project, it will be therefore assumed that a new investment in Toronto is an attractive opportunity due to growing population. Since we are working with Foursquare data, I will discuss the investments in terms of new venues in the city.

Due to new development in the city, I want to especially focus on the areas with a high increase of population in the last years. As an example, I will focus on new investment opportunities in the neighborhood "Willowdale East" in the scope of this project, because this neighborhood had one of highest population increase in the last years, due to the demographic statistics in Wikipedia (62,3% change in population since 2001, https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods).

Our approach in this project can be also easily applied to other neighborhoods. But I will restrict our recommendation only to "Willowdale East" for simplicity.

1.2. Problem

As the neighborhood "Willowdale East" is identified as an area which may provide opportunities for the investors or entrepreneurs. The next question is, which venue category will promise the highest potential for them. Therefore, this project aims to provide recommendations in terms of venue categories. Of course, these recommendations must be evaluated by further criteria (e.g. average income of the area, return on investment, etc.), but I aim to provide recommendations as a starting point for further discussions.

1.3. Approach

To answer the question described in the last section, I will use collaborative filtering approach, introduced in the week 5 of the machine learning module in the IBM Data Science Course.

This means, I will search for similar neighborhoods based on the existing venue categories and their frequency (number of venues in each category and neighborhood). Then, I will determine the venue categories which are not existing in Willowdale East yet, but in the most similar neighborhoods. At the last step, I will also exclude the categories, which already exist in the closest neighborhoods to Willowdale East.

As result, a list of venue categories will be provided as recommendation, which does not exist in Willowdale East, and also does not exist in the closest neighborhoods of Willowdale East.

2. Data Acquisition and Preprocessing

In the week 3 of the capstone project, we already used Foursquare data, also the neighborhood data and the geospatial data, to explore and cluster the neighborhoods in Toronto, based on the different venue categories.

In the scope of this project I will reuse the same data, but I will make minor changes in preprocessing and apply a different approach after preprocessing.

2.1. Loading borough information and geospatial data for Toronto

List of postal codes of Canada:

The postal codes, boroughs and related neighborhoods of Toronto will be loaded from the following Wikipedia link:

https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

This data will be transferred into the pandas data frame "toronto_data", which will have the three columns "Postal Code", "Borough" and "Neighborhood". An extract from the data frame can be found below:

Postal Code	Borough	Neighborhood
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park, Harbourfront
M6A	North York	Lawrence Manor, Lawrence Heights
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government

Figure 1: List of postal codes

Geospatial data:

The geospatial data for the postal codes of Toronto will be loaded from the following link, as also requested in the week 3 of the capstone project:

http://cocl.us/Geospatial_data

This data will be transferred to the pandas data frame "df_geo", which will have the three columns "Postal Code", "Latitude" and "Longitude" for all the postal codes of Toronto.

Postal Code	Latitude	Longitude
M1B	43.806686	-79.194353
M1C	43.784535	-79.160497
M1E	43.763573	-79.188711
M1G	43.770992	-79.216917
M1H	43.773136	-79.239476

Figure 2: Geospatial data for the postal codes

Merging data frames:

After this step, the geospatial data will be merged with "toronto_data" imported in the first step, using Postal Code as a common attribute. The merged data frame will contain latitude and longitude data for all the neighborhoods of Toronto.

Postal Code	Borough	Neighborhood	Latitude	Longitude
M3A	North York	Parkwoods	43.753259	-79.329656
M4A	North York	Victoria Village	43.725882	-79.315572
M5A	Downtown Toronto	Regent Park, Harbourfront	43.654260	-79.360636
M6A	North York	Lawrence Manor, Lawrence Heights	43.718518	-79.464763
M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.662301	-79.389494

Figure 3: Merged data frame

2.2. Loading Foursquare data for venues

The "explore" endpoint of Foursquare will be used, to capture all the venues including id, name, location, category id and category information for every neighborhood (analogue to the week 3 assignment in capstone project). This data will be saved in the pandas data frame "toronto_venues".

Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue ID	Venue	Venue Latitude	Venue Longitude	Category ID	Venue Category
Parkwoods	43.753259	-79.329656	4e8d9dcdd5fbbbb6b3003c7b	Brookbanks Park	43.751976	-79.332140	4bf58dd8d48988d163941735	Park
Parkwoods	43.753259	-79.329656	4cb11e2075ebb60cd1c4caad	Variety Store	43.751974	-79.333114	4bf58dd8d48988d1f9941735	Food & Drink Shop
Victoria Village	43.725882	-79.315572	4c633acb86b6be9a61268e34	Victoria Village Arena	43.723481	-79.315635	4bf58dd8d48988d185941735	Hockey Arena
Victoria Village	43.725882	-79.315572	4f3ecce6e4b0587016b6f30d	Portugril	43.725819	-79.312785	4def73e84765ae376e57713a	Portuguese Restaurant
Victoria Village	43.725882	-79.315572	4bbe904a85fbb713420d7167	Tim Hortons	43.725517	-79.313103	4bf58dd8d48988d1e0931735	Coffee Shop

Figure 4: Toronto venues, source: Foursquare

In the first step of preprocessing, this data frame will be grouped by the three attributes "Neighborhood", "Category ID" and "Venue Category". The rest of the attributes will be counted. The attribute "Venue ID" will be renamed as "Frequency", which will represent the number of the venues in every category and every neighborhood. This grouped data frame will be named as "venues_by_category".

2.3. Preparation of input venues

For the selected neighborhood "Willowdale East", an input data frame will be created. For this purpose, the data records from the data frame "venues_by_category" will be taken over to a new data frame "inputNeigh", where the neighborhood is filtered due to "Willowdale East" in the data records.

In the second step, the columns „Neighborhood", "Neighborhood Latitude", "Neighborhood Longitude", "Venue", "Venue Latitude", "Venue Longitude" will be dropped from inputNeigh. At the end, only the three columns "Category ID", "Venue Category" and "Frequency" will remain in the data frame.

This input data frame represents all the venue categories and their frequencies (number venues in this category) for the selected neighborhood "Willowdale East".

First five lines of inputNeigh can be find below:

Category ID	Venue Category	Frequency
52e81612bcbc57f1066b7a0c	Bubble Tea Shop	1
52dea92d3cf9994f4e043dbb	Discount Store	1
4bf58dd8d48988d1fd941735	Shopping Mall	1
4bf58dd8d48988d1fa931735	Hotel	1
4bf58dd8d48988d1e0931735	Coffee Shop	2

Figure 5: Input data frame - venue categories in Willowdale East and their frequency

2.4. Preparation of venue subset

A venue subset will be created by taking over all records from the data frame "venues_by_category" to a new data frame, considering the condition that the categories

are included also in inputNeigh. This new data frame will be named as "venueSubset" and will include only the categories involved in inputNeigh.

In the second step, the columns "Neighborhood Latitude", "Neighborhood Longitude", "Venue", "Venue Latitude", "Venue Longitude" will be dropped from venueSubset. As result, only the four columns "Neighborhood", "Category ID", "Venue Category" and "Frequency" will remain in the data frame.

In the third step, the data frame "venueSubset" will be grouped by "Nieghborhood", so that it can be used easily in the for-loop in the next step, to calculate the Pearson coefficient for the neighborhoods. This grouped data frame will be named as "venueSubsetGroup". Some examples of this data frame can be seen below:

Neighborhood	Category ID	Venue Category	Frequency
Agincourt	4bf58dd8d48988d121941735	Lounge	1
Alderwood, Long Branch	4bf58dd8d48988d1ca941735	Pizza Place	2
Alderwood, Long Branch	4bf58dd8d48988d1c5941735	Sandwich Place	1
Alderwood, Long Branch	4bf58dd8d48988d1e0931735	Coffee Shop	1
Bathurst Manor, Wilson Heights, Downsview North	4bf58dd8d48988d1ca941735	Pizza Place	1

Figure 6: Venue categories in the neighborhoods with their frequencies

3. Methodology

Collaborative filtering approach will be used in this project, in order to find similar neighborhoods. Based on the similarity of the neighborhoods, most relevant venue categories will be identified, as recommendation.

In the subsections below, Pearson correlation coefficient will be calculated for all the neighborhoods, which will be used as similarity index. Based on this similarity index, recommendations will be prepared.

3.1. Calculation of the Pearson Correlation Coefficient

As an indicator for the similarity between Willowdale East and other neighborhoods, Pearson correlation coefficient will be calculated.

The calculated values vary between -1 and 1, where "1" indicates a strong correlation between the neighborhoods.

For the calculation, the neighborhoods will be picked up from the data frame "venueSubsetGroup" by using a for-loop. The groups will represent the neighborhoods, since the data frame is grouped by neighborhood in the previous step, to create the data frame "venueSubsetGroup". The following steps will be executed for the calculation of every neighborhood's coefficient:

- The length the group will be calculated and saved as "nRatings"

- The categories in every neighborhood will be found first, which are also included in Willowdale East, and saved in the separate data frame "temp_df".
- The frequency of every category in the neighborhood will be saved in "tempRatingList".
- Then the frequencies of all categories in the neighborhood will be saved in "tempGroupList".
- Pearson correlation coefficient will be calculated by using tempRatingList, tempGroupList and nRatings

3.2. Identification of the most relevant venue categories

The pearson correlation coefficients of all neighborhoods will be sorted in descending order. Then the first 20 neighborhoods will be taken over to a separate data frame "topNeigh".

Then, the data frame topNeigh will be merged with venues_by_category, by using Neighborhood as common attribute. This merged data frame will be saved as topNeighFreq.

In the last step, the topNeighFreq will be grouped by categories and the similarity index will be summed up and saved as tempTopNeighFreq. This data frame will be sorted in descending order.

As result, tempTopNeighFreq contains 183 unique venue categories.

In the next step, the venue categories will be dropped from tempTopNeighFreq, which already exists in Willowdale East. After this step, 157 unique venue categories remain in the data frame tempTopNeighFreq.

Then, also the venue categories which already exist in the closest neighborhoods "Willowdale West", "Willowdale Newtonbrook" and "Bayview Village" will be dropped. As result, 153 unique venue categories will remain in tempTopNeighFreq, as a list of recommendations.

4. Results

The first 20 results from tempTopNeighFreq are shown in the table below.

Venue Category	sum_similarityIndex
Bakery	8.670215
Italian Restaurant	7.965109
Thai Restaurant	6.574352
Gastropub	5.990489
Breakfast Spot	5.556232
Diner	5.500840
Art Gallery	4.920902
Seafood Restaurant	4.804869
Bookstore	4.712067
Bar	4.657645
Gym	4.471747
Vegetarian / Vegan Restaurant	4.361945
Pub	4.318102
Clothing Store	4.311917
Asian Restaurant	4.032791
French Restaurant	3.800281
Gym / Fitness Center	3.557120
Yoga Studio	3.505968
Farmers Market	3.434536
Jazz Club	3.386072

Figure 7: Top 20 recommendations for venue categories

5. Discussion

These recommendations represent a starting point for further evaluations. We can conclude from the list for example, bakeries exist in many similar neighborhoods, although it doesn't exist in Willowdale East and also not in the closest neighborhoods. Therefore, it can be a good recommendation, to open a new bakery in Willowdale East.

The second recommendation "Italian restaurant" can be also taken into consideration, whereas we can recognize from the input data frame that there are already two venues in the category "Pizza Place". If we assume that the category Italian restaurant provides a higher quality service comparing to "Pizza Place", then it may be a good idea to analyze the average income of the neighborhood.

The third recommendation "Thai restaurant" is also an interesting option, since we can already recognize the interest for Asian cuisine, because there are already Japanese and Vietnamese restaurants in Willowdale East.

6. Conclusion

This analysis can be extended by further steps, to achieve a more detailed analysis. These steps are not considered here for the sake of simplicity and also due to the restrictions in the premium calls of Foursquare and availability of the data.

- Number of total check-ins or likes can be used as rating of the venue categories, so that a weighted similarity index can be calculated for the neighborhoods.
- Users can be grouped according to age and gender. Then the venue categories can be evaluated due to these groups based on their total number of check-ins, to find the best fit for venue category and focus group.
- Based on the number of the total check-ins over the last years, a regression analysis can be conducted to find the trending venues in the city and also in different neighborhoods.