# Recommendations for Business Opportunities in Toronto

F. TOPSEVER

20.03.2021

# Introduction

- Toronto is a multicultural city with growing population (4,3% increase between 2011-2016). It is an attractive city for new investments.
- Fast growing areas in the city may have high potential for investors and entrepreneurs.
- Willowsdale East recognized as a fast-growing area with 62,3% change in population between 2001 and 2016. Business opportunities in this neighborhood are analyzed in this project.
- Investment for a new venue in Willowdale East will be evaluated, since we will use Foursquare data for this project.
- **Problem:** Which venue categories promise the best business opportunities?
- **Approach:** Collaborative filtering approach is applied in this project, in order to find the most similar neighborhoods and identify opportunities by comparing Willowdale East with most similar neighborhoods.

# Data Acquisition

- Following data sources will be used :
  - List of postal codes in Canada (incl. boroughs and related neighborhoods)
  - Geospatial data for the postal codes in Canada
  - „Explore“ endpoint of Foursquare data

# Data Preprocessing (1)

1. The list of the postal codes is merged with the geospatial data. The merged data represents the coordinates of every neighborhoods in 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 |

# Data Preprocessing (2)

2. After loading the Foursquare data by using the explore endpoint:
   a. All the venues in every neighborhood are saved in a data frame. This data frame will be grouped by “Neighborhood”, “Category ID” and “Venue Category”.
   b. The other attributes are counted for grouping.
   c. The column “Venue ID” is renamed as “Frequency”. This column represents the number of venues in each category and neighborhood.
   d. An input data frame is created by filtering the grouped data frame for the neighborhood “Willowdale East”. All the columns except “Category ID”, “Category Name” and “Frequency” are dropped.

| 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 |

# Data Preprocessing (3)

3. A venue subset is prepared based on the grouped data frame in the last step:
   a. All the columns except “Neighborhood”, “Category ID”, “Category Name” and “Frequency” are dropped.
   b. The subset is grouped by neighborhood:

| 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 |

# Methodology (1)

- As mentioned in the introduction, collabrative filtering approach is used in this project, to generate recommendations for the business.
- In the first step, Pearson correlation coefficient is calculated for all the neighborhoods in the venue subset, to find the similarity index, which represents the similarity between the neighborhood and Willowsdale East.
- Only the venue categories are considered for calculation, which are common in Willowsdale East and in the neighborhood from the subset.
- After calculating the coefficients, most similar 20 neighborhoods are filtered from results.
- The data frame of most similar 20 neighborhoods is merged with the data frame, which is created in the step 2c of preprocessing, to find the venue categories and their frequencies for the most similar 20 neighborhoods.
- In the last step, the data frame is grouped by categories and the similarity index is summed up.

# Results (1)

- The result is sorted in descending order.
- The top 10 results are visible in the table.
- For example, „Cafe“ is the most common venue category in the most similar neighborhoods.
- However, there are already two cafes in Willowdale East due to the input data frame.

| Venue Category | |
|---|---|
| Café | 10.866809 |
| Coffee Shop | 9.366809 |
| Bakery | 8.670215 |
| Italian Restaurant | 7.965109 |
| Park | 7.715135 |
| Restaurant | 7.261499 |
| Japanese Restaurant | 6.647363 |
| Thai Restaurant | 6.574352 |
| Gastropub | 5.990489 |

# Results (2)

- It will be assumed that the venue categories which does not exist in Willowdale East and in the closest neighborhoods can represent the best opportunuties.
- Therefore, the venue categories in Willowdale East and in the closest neighborhoods are excluded from the data frame.
- The top 10 results can be seen in the table.

| 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 |

# 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.

# Conclusion

The results represent recommendations based on collaborative filter approach. 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.