Recommending business opportunities in Toronto neighborhoods

Chen Cheng

IBM Data Science Professional Certificate Capstone project

July 27, 2019

Introduction

The purpose of this study is to use a recommender system to suggest new business opportunities for Toronto neighborhoods that currently have few venues (<20). The study proposes to use a user-based collaborative filtering mechanism where the neighborhoods (defined by postal codes) are the "users", and the frequency of various kind of venues are the "scores" of the venues in each neighborhood. Based on similarity among neighborhoods, this study seeks to recommend venues that are characteristic of but currently absent from certain neighborhoods, thus potentially good business opportunities. The underlying assumption is that neighborhoods that are similar to each other should have similar venues.

Some caveats of this study include:

- Some of the venues in from the dataset are public infrastructure, such as bus stops or airport, and should not be considered business opportunities. The study will include those venues when determining the similarity among neighborhoods, because it was thought that public infrastructure is also a part of the characteristics of a neighborhood and thus should be included)
- Some venues may be mutually exclusive or competitive with each other, such as different kinds of restaurants. This should be factored in when analyzing the recommendation results.
- Obviously, other factors and analyses, such as population, population density, type of the neighborhood (urban vs suburban), socioeconomic factors, projected costs and revenues, etc. need to be factored in when making business decisions. This study merely assumes that neighborhoods that similar to each other should have similar kinds of venues.

Methodology

Data source and cleaning up

1. Toronto neighborhoods data (name, postal code, borough) were scraped from Wikipedia (https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M) using pandas. Then, postal codes with “not assigned” neighborhoods were deleted, and neighborhoods with the same postal codes were grouped by the postal codes. There were 103 unique postal codes, and those postal codes were used to define an area/neighborhood.
2. Geographic coordinates of each postal code were obtained using the Python geocoder library, and added to the data table.

3. Up to 100 venues (venue name, category, geographic coordinates) within each postal code area were obtained using Foursquare API.

Data exploration

The venue counts in each postal code area were calculated (Fig 1), and while there are quite a few areas with >80 venues (presumably the busy downtown areas), most postal code areas have <20 venues, which could correspond to suburbs or industrial areas.

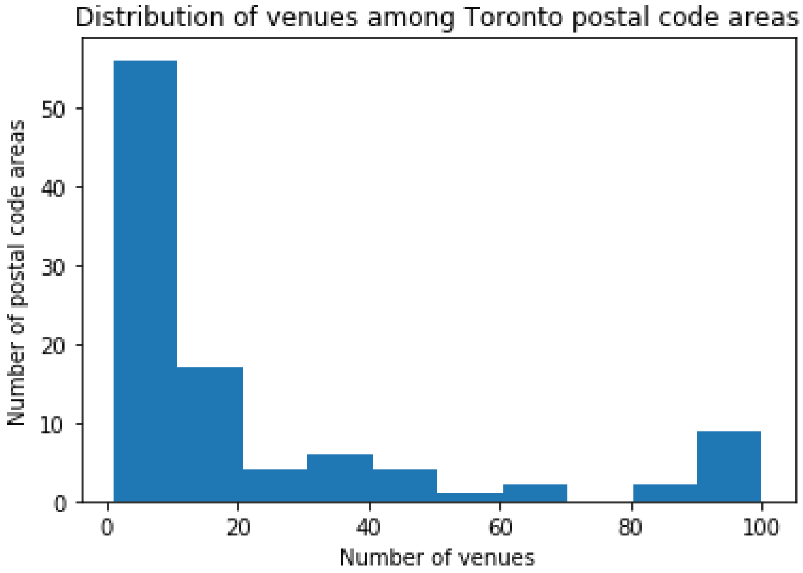


Fig 1. Histogram of the number of venues in Toronto postal code areas.

Clustering

The postal code areas were clustered based on the presence or absence of shared venue types. The venue category column was one-hot encoded and fed into the K-means clustering algorithm. Different numbers of cluster were explored. For all postal areas, 6 clusters were used. The results were plotted using folium. As shown in Fig 2, most areas fall into one cluster (red), which also happen to include the downtown Toronto area.

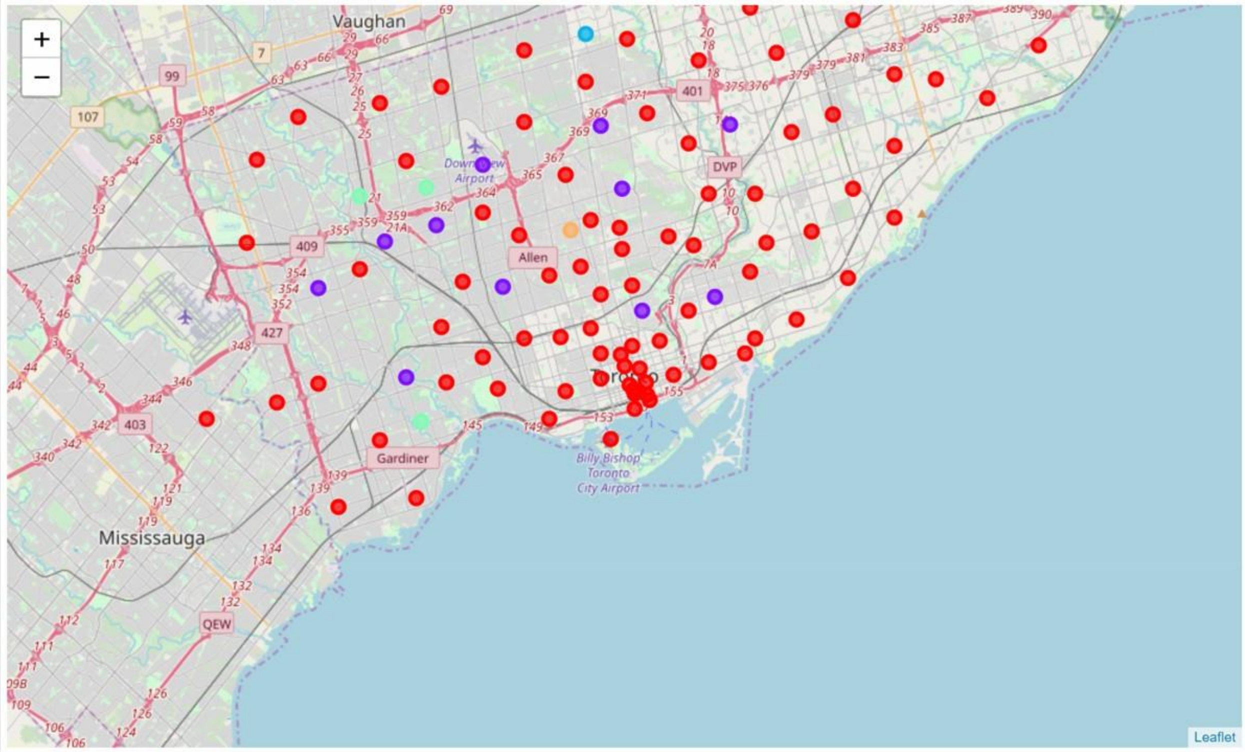


Fig 2. Toronto postal code areas clustered using K-means clustering based on shared venue categories.

Because most peripheral postal code areas have few venues (<20), clustering was re-done with only postal code areas with >=20 venues. As shown in Fig 3, most downtown areas still fall into one cluster (red), which suggests that the peripheral areas shown in Fig 2 (which are in the same cluster as downtown areas) could be similar to the downtown areas in terms of the venue types and could benefit from a recommendation system that recommend new venues based on similar postal code areas.

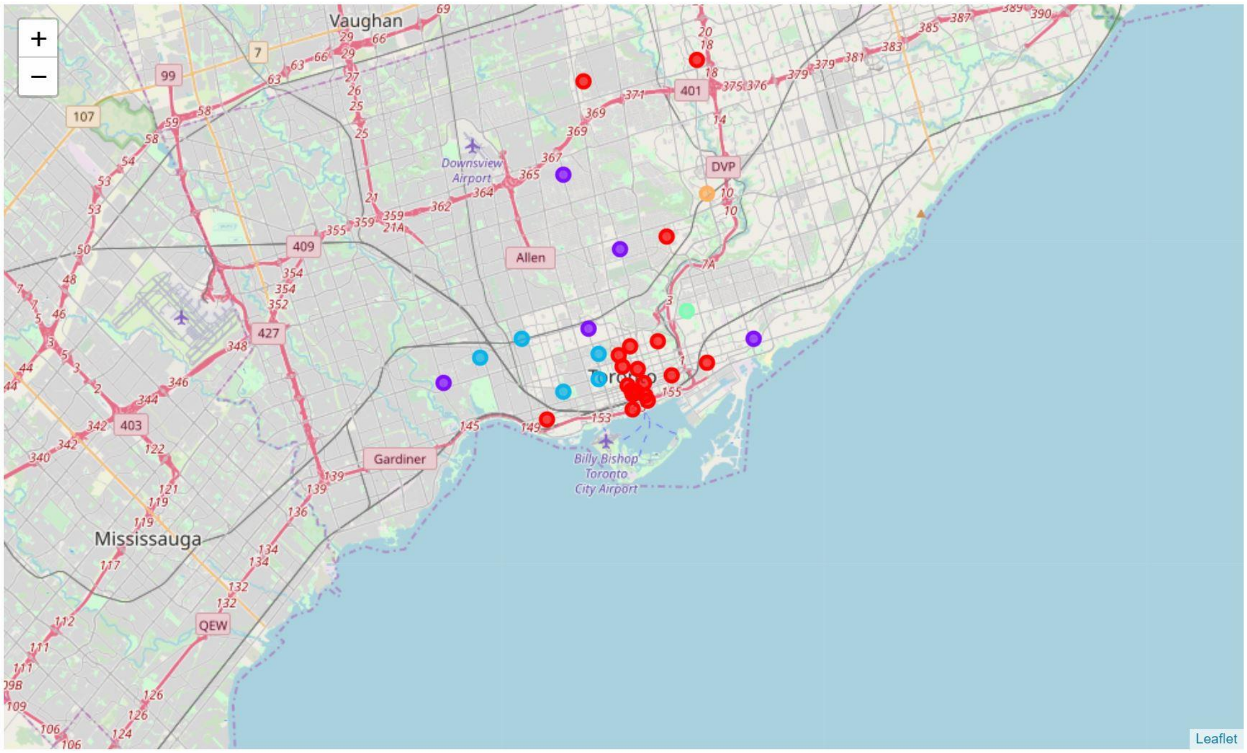


Fig 3. Toronto postal code areas with >=20 venues clustered using K-means based on shared venue categories.

Date source and cleaning up, data exploration, and clustering were done in the following notebook: https://github.com/chencheng23/Toronto/blob/master/Toronto_postcode_clustering.ipynb

Recommendation system

Because few additional data for the venue categories are available, and because the target areas (postal code areas with 10-19 venues) have relatively few venues, a user-based collaborative filtering system was chosen. The specific steps are:

1. Postal code areas with >=20 venues were selected as the dense areas, and postal code areas with 10-19 venues as neighborhoods this study aims to recommend venues to (sparse areas).
2. From the one-hot venue category table (used above for clustering), the average frequency (score) of each venue category in each postal code areas were calculated.
3. For each of sparse areas, a similarity coefficient with all the dense areas were calculated.
4. A weighted score was calculated for each venue category was calculated for each of the sparse areas, and top 5 scores were selected. The results were stored in a data table.

Recommendation system was done in the following notebook: https://github.com/chencheng23/Toronto/blob/master/Toronto_neighborhood_new_biz_rec.ipynb

Results and Discussion

Clustering

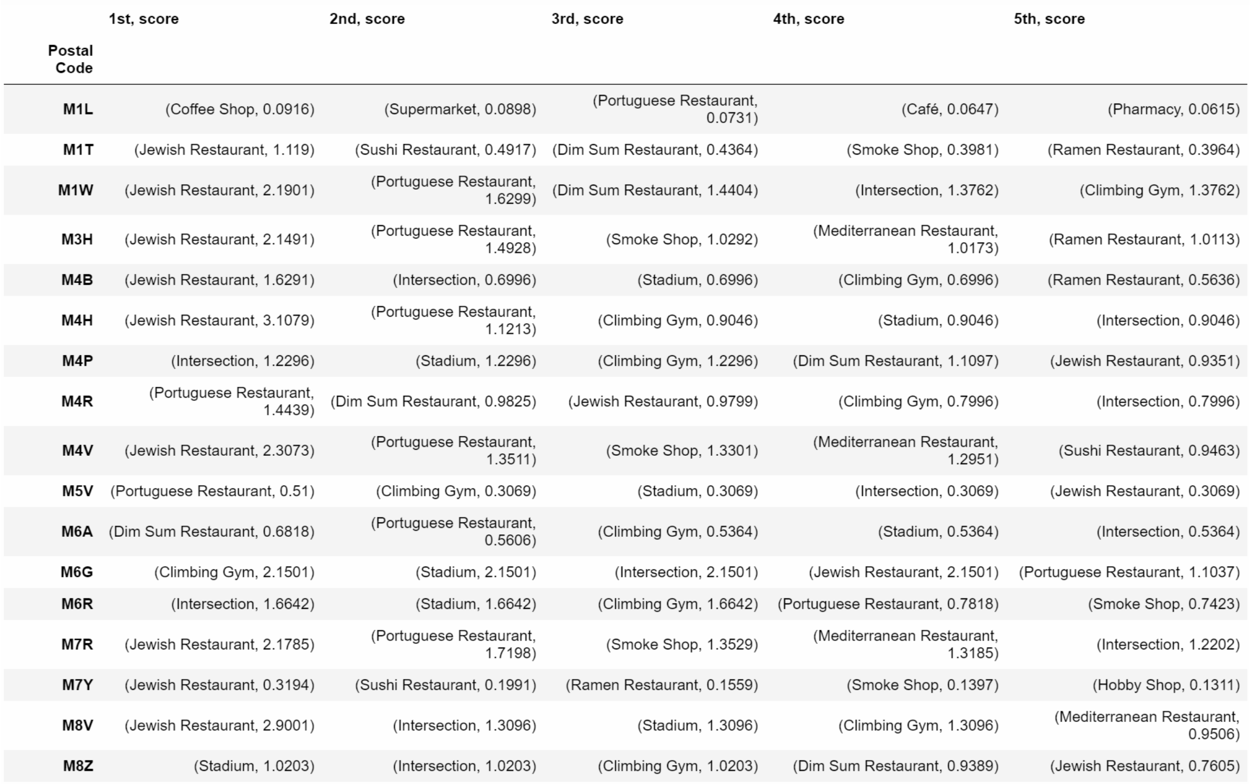
As shown in Figs 2 and 3, most neighborhoods fall into the same category as the downtown areas, which was a little surprising, given that it was assumed that there should be a different set of venues in downtown areas than in the suburbs. One reason for this could be that with a venue category of 279 and a limit of 100 venues per area, there are more zeros in the one-hot encoded venue category table than none-zero values, and two areas might be similar to each other based on the many zero values rather than the none-zero values (i.e. if two areas both lack many types of venues, they can have a very high similarity coefficient, but venues present in one of those areas are not necessarily characteristic of the other area).

Recommendation

The top five recommended venues for each of the sparse areas are shown in Fig 4. Some areas, like M5V, have overall low scores, and upon closer inspection, M5V is likely an airport area, so it's not too surprising that there aren't many businesses around. Jewish restaurants appear to be good opportunities in many areas, some with high scores, such as M4H or M8V, and it may be worthwhile to follow up on those areas and figure out if there's market demand for additional restaurants in those areas, and in particular Jewish restaurants.

Overall though, the top 5 recommendation results are quite homogeneous, with Restaurants (Jewish, Portuguese, Mediterranean, Ramen), climbing gyms, and some stadiums and intersections dominating the table. This is likely a result from the fact that most of the sparse areas highly resemble the dense downtown areas (they fall into the same cluster, see Figs 2 and 3), and whatever is popular in the dense areas are recommended to the sparse areas.

Finally, some top recommendations are infrastructure projects and may not be small business opportunities, but the city planning might benefit from the recommendations when, for example, considering installing additional intersections in certain neighborhoods.



	1st, score	2nd, score	3rd, score	4th, score	5th, score
Postal Code					
M1L	(Coffee Shop, 0.0916)	(Supermarket, 0.0898)	(Portuguese Restaurant, 0.0731)	(Café, 0.0647)	(Pharmacy, 0.0615)
M1T	(Jewish Restaurant, 1.119)	(Sushi Restaurant, 0.4917)	(Dim Sum Restaurant, 0.4364)	(Smoke Shop, 0.3981)	(Ramen Restaurant, 0.3964)
M1W	(Jewish Restaurant, 2.1901)	(Portuguese Restaurant, 1.6299)	(Dim Sum Restaurant, 1.4404)	(Intersection, 1.3762)	(Climbing Gym, 1.3762)
M3H	(Jewish Restaurant, 2.1491)	(Portuguese Restaurant, 1.4928)	(Smoke Shop, 1.0292)	(Mediterranean Restaurant, 1.0173)	(Ramen Restaurant, 1.0113)
M4B	(Jewish Restaurant, 1.6291)	(Intersection, 0.6996)	(Stadium, 0.6996)	(Climbing Gym, 0.6996)	(Ramen Restaurant, 0.5636)
M4H	(Jewish Restaurant, 3.1079)	(Portuguese Restaurant, 1.1213)	(Climbing Gym, 0.9046)	(Stadium, 0.9046)	(Intersection, 0.9046)
M4P	(Intersection, 1.2296)	(Stadium, 1.2296)	(Climbing Gym, 1.2296)	(Dim Sum Restaurant, 1.1097)	(Jewish Restaurant, 0.9351)
M4R	(Portuguese Restaurant, 1.4439)	(Dim Sum Restaurant, 0.9825)	(Jewish Restaurant, 0.9799)	(Climbing Gym, 0.7996)	(Intersection, 0.7996)
M4V	(Jewish Restaurant, 2.3073)	(Portuguese Restaurant, 1.3511)	(Smoke Shop, 1.3301)	(Mediterranean Restaurant, 1.2951)	(Sushi Restaurant, 0.9463)
M5V	(Portuguese Restaurant, 0.51)	(Climbing Gym, 0.3069)	(Stadium, 0.3069)	(Intersection, 0.3069)	(Jewish Restaurant, 0.3069)
M6A	(Dim Sum Restaurant, 0.6818)	(Portuguese Restaurant, 0.5606)	(Climbing Gym, 0.5364)	(Stadium, 0.5364)	(Intersection, 0.5364)
M6G	(Climbing Gym, 2.1501)	(Stadium, 2.1501)	(Intersection, 2.1501)	(Jewish Restaurant, 2.1501)	(Portuguese Restaurant, 1.1037)
M6R	(Intersection, 1.6642)	(Stadium, 1.6642)	(Climbing Gym, 1.6642)	(Portuguese Restaurant, 0.7818)	(Smoke Shop, 0.7423)
M7R	(Jewish Restaurant, 2.1785)	(Portuguese Restaurant, 1.7198)	(Smoke Shop, 1.3529)	(Mediterranean Restaurant, 1.3185)	(Intersection, 1.2202)
M7Y	(Jewish Restaurant, 0.3194)	(Sushi Restaurant, 0.1991)	(Ramen Restaurant, 0.1559)	(Smoke Shop, 0.1397)	(Hobby Shop, 0.1311)
M8V	(Jewish Restaurant, 2.9001)	(Intersection, 1.3096)	(Stadium, 1.3096)	(Climbing Gym, 1.3096)	(Mediterranean Restaurant, 0.9506)
M8Z	(Stadium, 1.0203)	(Intersection, 1.0203)	(Climbing Gym, 1.0203)	(Dim Sum Restaurant, 0.9389)	(Jewish Restaurant, 0.7605)

Fig 4. Top five recommended venues for postal areas with 10-19 venues (sparse areas).

Conclusions and future directions

Toronto neighborhoods were clustered based on the frequency of various venue types, and most neighborhoods fall into the same category as the downtown areas. Based on this, the venue-sparse peripheral areas were given recommendations of new venues based on their similarity to the venue-dense areas, and the recommendations comprise mostly of restaurants, gyms, and infrastructure. To further refine the recommendation results, one can:

1. Increase the limit of venues (currently set at 100) returned by Foursquare so that the one-hot venue category table is less sparse, and the similarity coefficients calculated would make more sense (see Results and discussion section).
2. Calculate similarity coefficients based only on the overlapping none-zero venue frequencies to mitigate the problem that currently many neighborhoods are considered similar not because they have common venues, but because they all lack many common venues.
3. Exclude infrastructure type venues in the recommendation output to make better recommendation of small business opportunities.