Toronto Business Opporunities

CHEN CHENG

IBM DATA SCIENCE PROFESSIONAL CERTIFICATE CAPSTONE PROJECT

JULY 27, 2019

Introduction

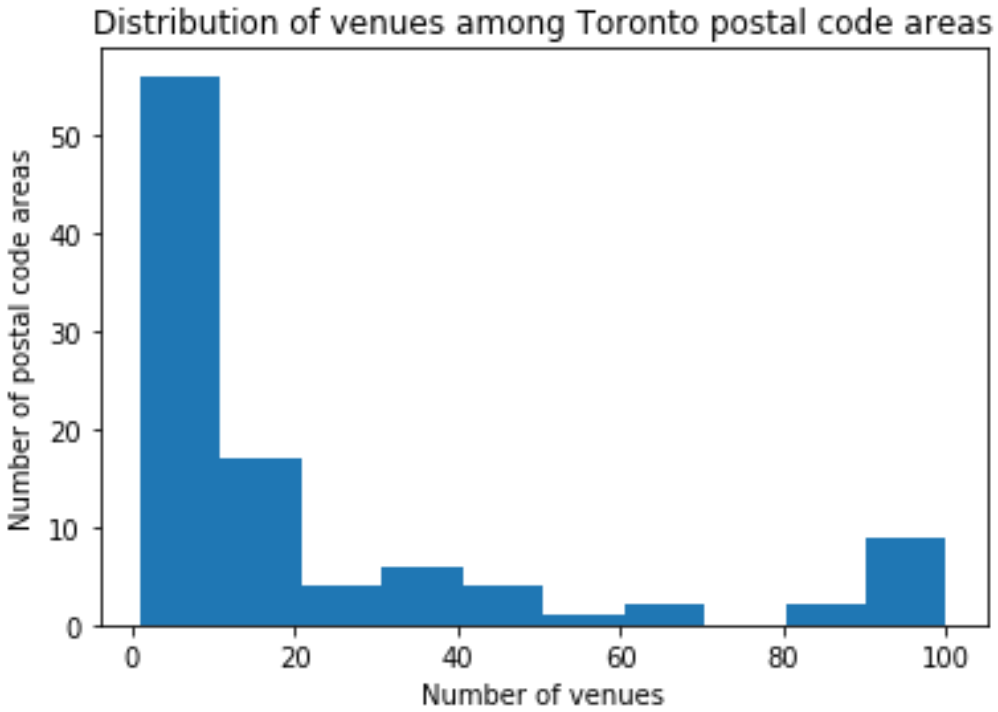
This study proposes to use a user-based collaborative filtering mechanism to recommend new business opportunities for Toronto neighborhoods that currently have few venues (<20), where the neighborhoods (defined by postal codes) are the "users", and the frequency of various kind of venues are the "scores" of the venues in each neighborhood.

This study hopes to uncover new business opportunities in areas where those businesses are currently absent or under-represented but should be in high demand based on other similar areas.

Data acquisition and formatting

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.

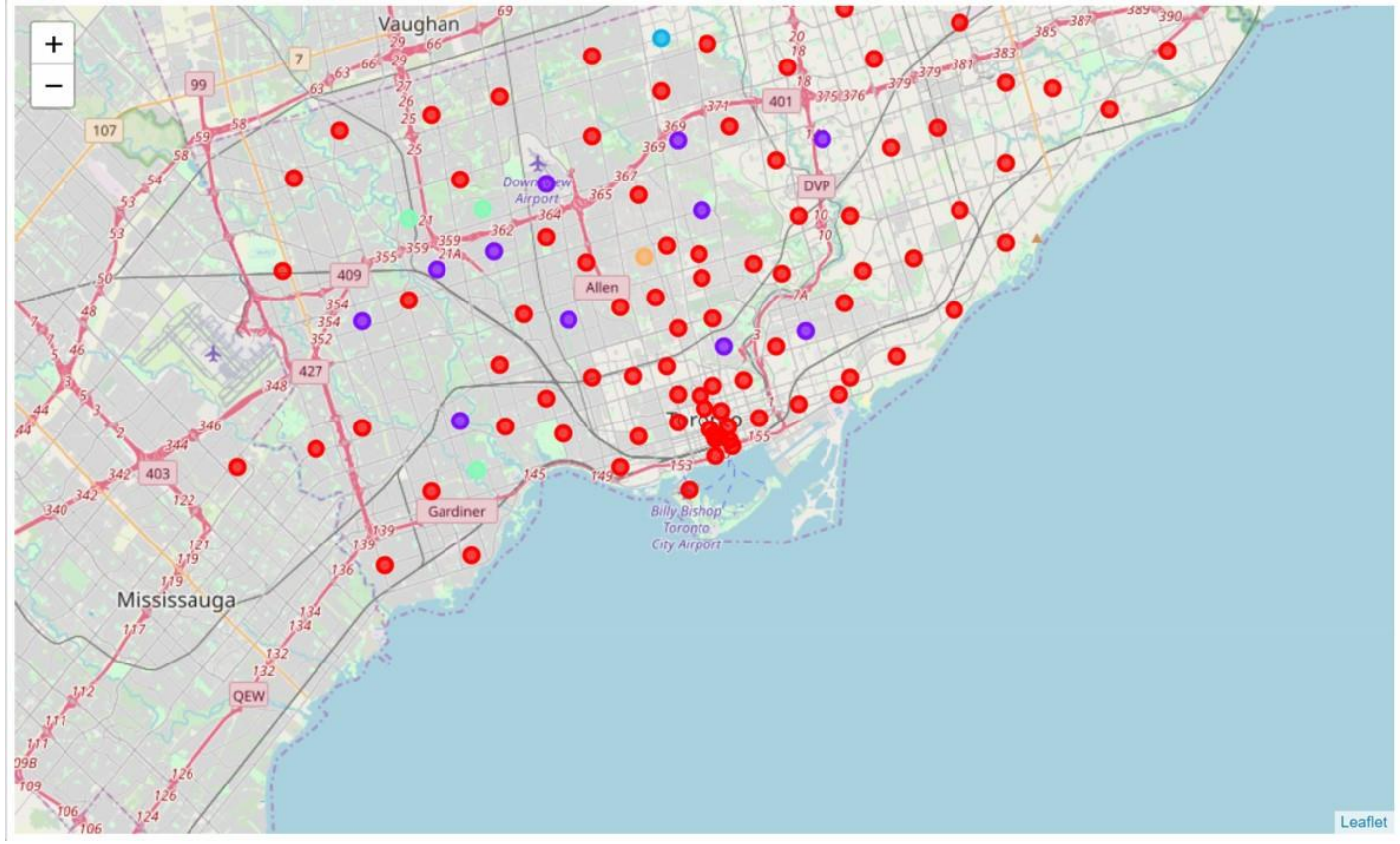
Venue density in various areas



Histogram of venue distribution shows that the majority of the areas have <20 venues, and less than 20 areas have >80 venues.

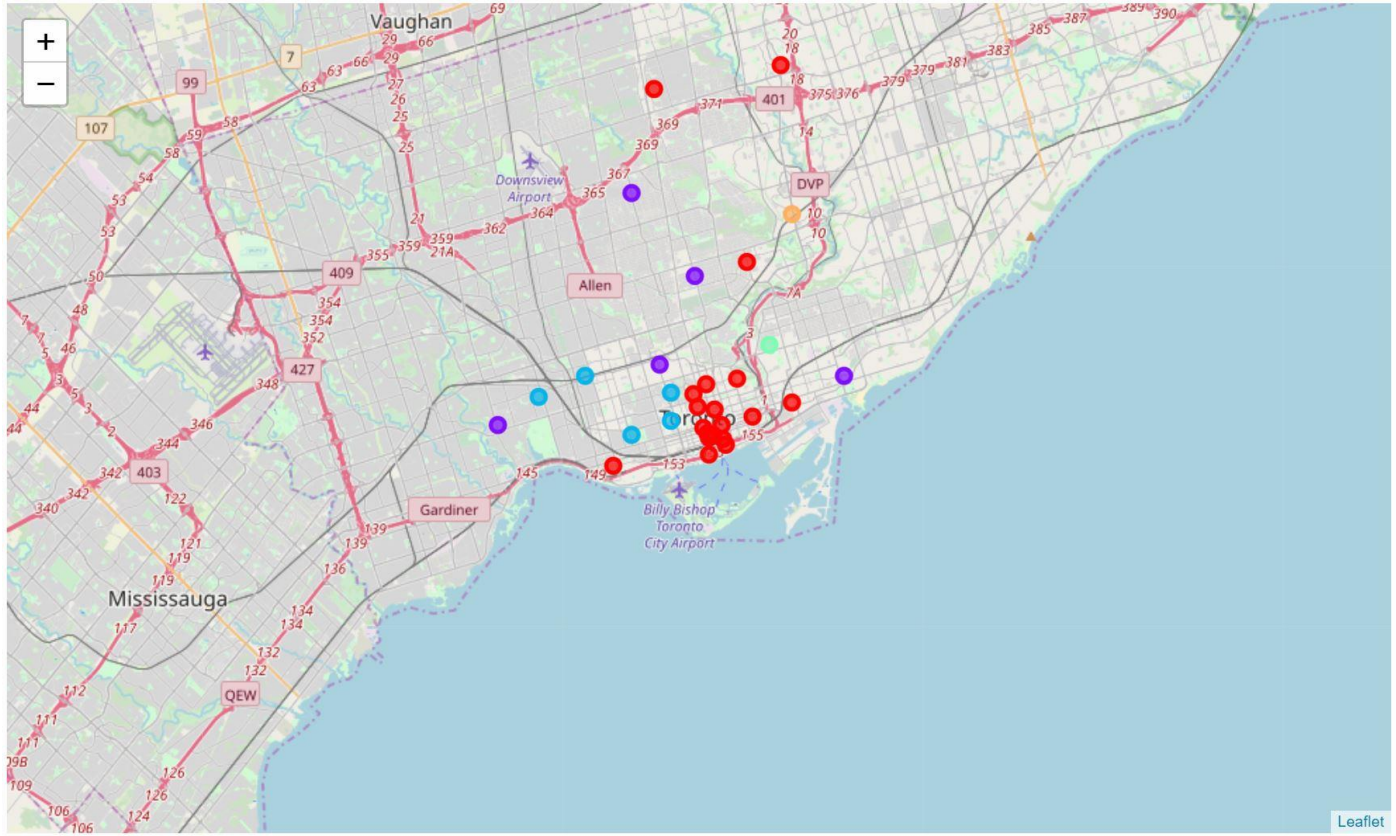
Clustering of the areas

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.



Clustering of the areas (cont.)

Because most peripheral postal code areas have few venues (<20), clustering was re-done with only postal code areas with >=20 venues. As shown here, most downtown areas still fall into one cluster (red), which suggests that the peripheral areas shown in previous figure (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 other similar postal code areas.



Clustering result discussion

Most neighborhoods fall into the same category as the downtown areas, contrary to the assumption 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 system

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 results for the “sparse areas”

Postal Code	1st, score	2nd, score	3rd, score	4th, score	5th, score
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)

1. Restaurants (Jewish, Portuguese, Mediterranean, Ramen), climbing gyms, and some stadiums and intersections dominating the table.

2. Some top recommendations are infrastructure projects (intersection, stadium) and may not be good small business opportunities

Conclusions and future directions

- Most Toronto neighborhoods fall into the same category as the downtown areas.
- The recommendations for sparse areas 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.

References

https://github.com/chencheng23/Toronto/blob/master/Toronto_postcode_clustering.ipynb

https://github.com/chencheng23/Toronto/blob/master/Toronto_neighborhood_new_biz_rec.ipynb