Your new neighbourhood - recommender system (The Battle of Neighbourhoods)

Alexandr Ignatchenko

February 9, 2021

1. Introduction

1.1 Background

When people relocate to unfamiliar city/town it takes significant amount of time and effort to find the right place to live and choose the right neighbourhood to be happy in. In some cases we can get guidance from people we might know, or the employer, if this is work related relocation. Unfortunately, for many people this help is not available and they have to rely on their own research of the new location.

1.2 Problem

Many real estate web sites, apart form property selection and prices, often provide an additional information such as demographics, schools and transportation availability in the area. However, many important aspects and characteristics of the desired neighbourhood are missing. For example, availability of sport and outdoor activities that can be enjoyed on the daily bases, cafes, restaurants, community centres, banks, gas stations, supermarkets and so on. The list can be pretty long and in order to make a right choice we have to spend hour and hour of researching the area just to realise it is not the right one.

1.3 Interest

The goal of this study is to create a system based on the Foursquare data which will use user preferences to recommend the top neighbourhoods which closely match with majority of desired point of interest of the user. The system can complement real estate search engines or can be used as stand alone tool for the first discovery step in the users journey to find the desired neighbourhood.

2. Data

2.1 Data sources

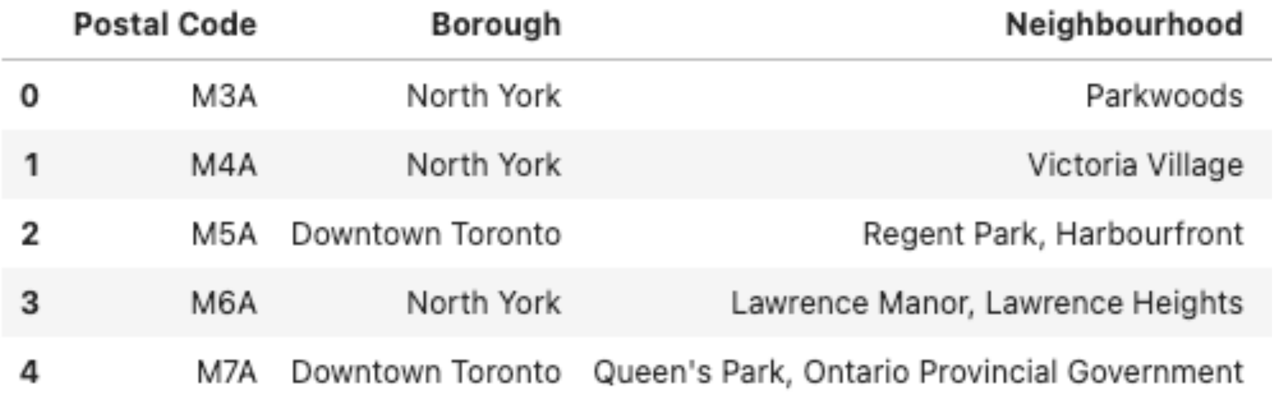
The system described in the introduction section will use following data sources.

- **Neighbourhood information**

In this pilot study we'll be using neighbourhood data for Toronto, Canada which postal codes starts with letter M and can be obtained here:

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

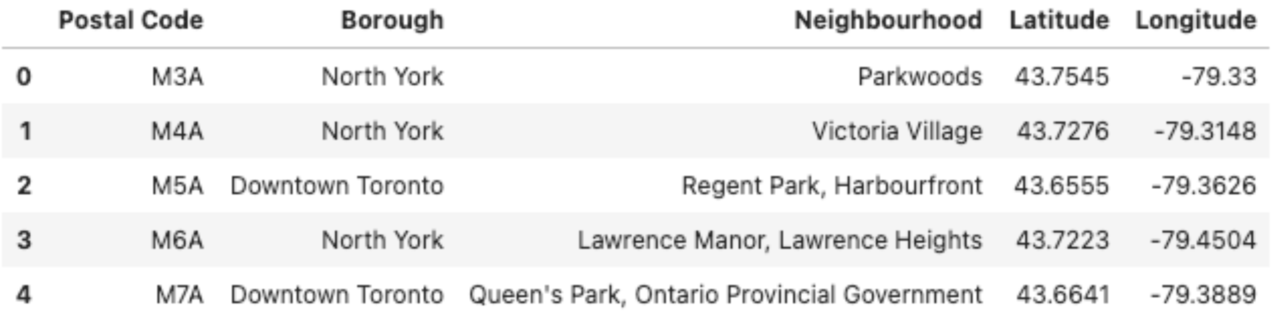
Which can be converted to following data frame



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

- **Geolocation information**

The latitude and longitude data for each postal code extracted by "pgeocode" python package which resulted in data frame bellow



	Postal Code	Borough	Neighbourhood	Latitude	Longitude
0	M3A	North York	Parkwoods	43.7545	-79.33
1	M4A	North York	Victoria Village	43.7276	-79.3148
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.6555	-79.3626
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.7223	-79.4504
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.6641	-79.3889

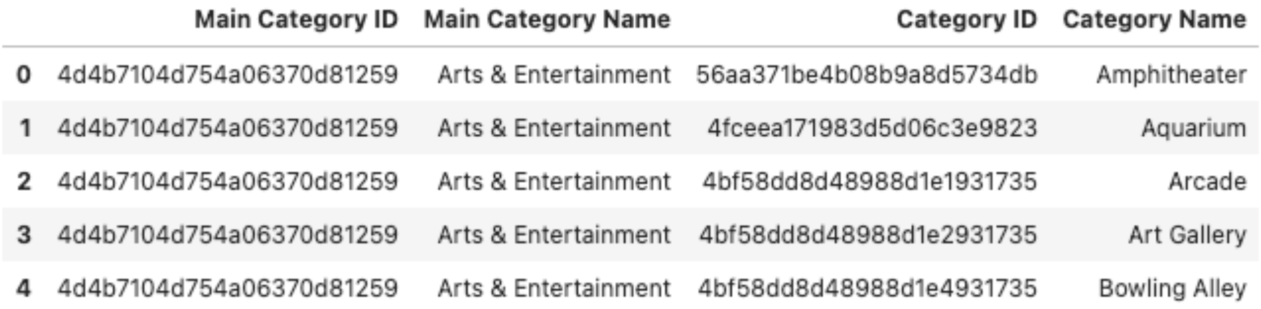
Read more about "pgeocode" here https://pypi.org/project/pgeocode/

- **Points of interests**

The study uses two endpoints of the Foursquare API.

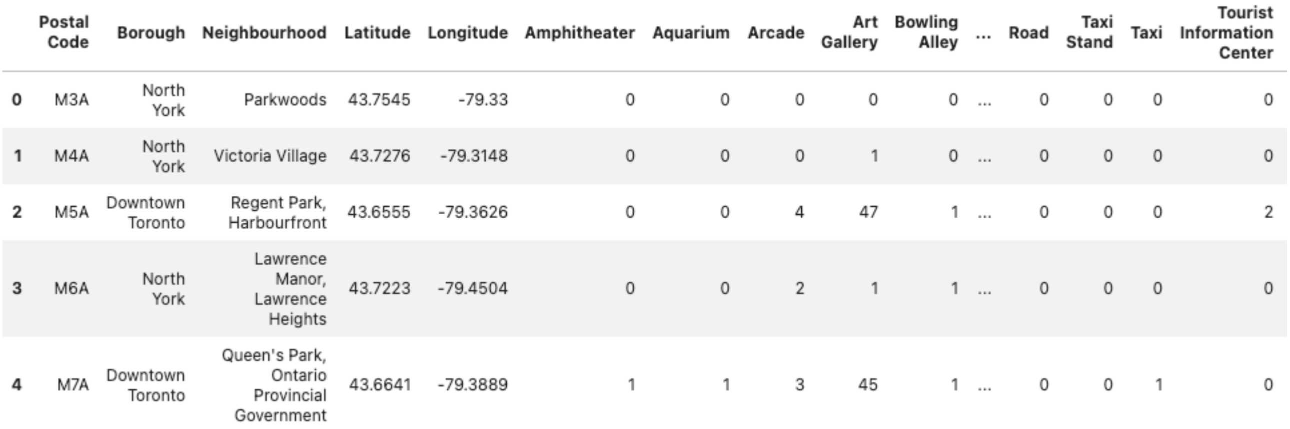
https://developer.foursquare.com/docs/places-api/endpoints/

The the full list of points of interest which are available from Foursquare extracted from “venue categories” and resulted in following data frame



	Main Category ID	Main Category Name	Category ID	Category Name
0	4d4b7104d754a06370d81259	Arts & Entertainment	56aa371be4b08b9a8d5734db	Amphitheater
1	4d4b7104d754a06370d81259	Arts & Entertainment	4fceea171983d5d06c3e9823	Aquarium
2	4d4b7104d754a06370d81259	Arts & Entertainment	4bf58dd8d48988d1e1931735	Arcade
3	4d4b7104d754a06370d81259	Arts & Entertainment	4bf58dd8d48988d1e2931735	Art Gallery
4	4d4b7104d754a06370d81259	Arts & Entertainment	4bf58dd8d48988d1e4931735	Bowling Alley

The Foursquare API “search” endpoint was used to obtain extensive points of interest data for each neighbourhood resulted in summarised data frame bellow



	Postal Code	Borough	Neighbourhood	Latitude	Longitude	Amphitheater	Aquarium	Arcade	Art Gallery	Bowling Alley	...	Road	Taxi Stand	Taxi	Tourist Information Center
0	M3A	North York	Parkwoods	43.7545	-79.33	0	0	0	0	0	...	0	0	0	0
1	M4A	North York	Victoria Village	43.7276	-79.3148	0	0	0	1	0	...	0	0	0	0
2	M5A	Downtown Toronto	Regent Park, Harbourfront	43.6555	-79.3626	0	0	4	47	1	...	0	0	0	2
3	M6A	North York	Lawrence Manor, Lawrence Heights	43.7223	-79.4504	0	0	2	1	1	...	0	0	0	0
4	M7A	Downtown Toronto	Queen's Park, Ontario Provincial Government	43.6641	-79.3889	1	1	3	45	1	...	0	0	1	0

2.2 Data cleaning and processing

- Neighbourhood data and coordinates

Neighbourhood data for Canadian postal codes which start with M was downloaded and converted not pandas data frame using native read_html method. Unknown values and outliers for ‘Borough’ column such as ‘Not assigned’ or ‘Mississauga’ had been removed. This resulted to clean data frame of 3 columns and 102 rows. The dataset get added columns for latitude and longitude data from ‘pgeocode’ package.

- Foursquare API data

The Foursquare API was used for two sources of data. The points of interests obtained from venue “categories” end point as json object and converted to the pandas data frame object. This data frame resulted in 4 columns and 470 rows of the venue categories.
The ‘venue’ data for each of 470 venue category was extracted as json object using the Foursquare API ‘search’ end point. Due to massive amount of request send to Foursquare API the resulting data frame was also saved into the csv file for later use. The details how data was processed will be explained later in the following section.

3. Analysis methodology

The purposes of this study is to evaluate **Content Based Filtering** system as primary algorithm in the neighbourhood recommender.

3.1 Data normalisation

The data frame which was acquired and pre-processed on the previous converted to in a the way so Postal Code become its index and all extra information columns apart from Foursquare API data were dropped.

The data then was normalise data using min-max method which is one of the most common.

$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Main advantage of min-max normalisation method that it guarantees all features will have the exact same scale. However one of the drawbacks is that it does not handle outliers very well.

More information on different normalisation methods can be found here:

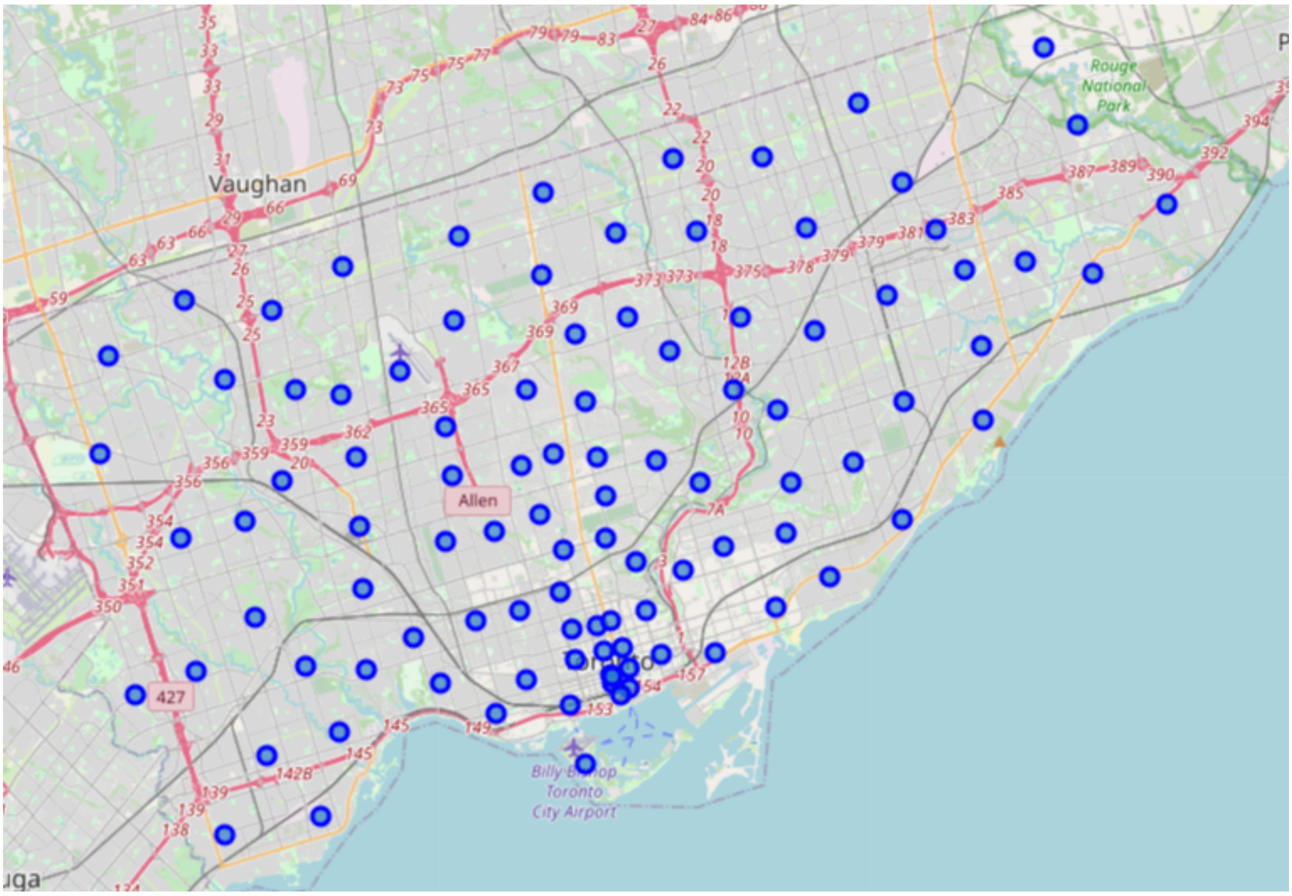
https://en.wikipedia.org/wiki/Feature_scaling

3.2 Data visualisation

Maps and geolocation positions data was visualised using Folium library. The library is well known for it’s scalability and interactivity.

More on Folium library here (https://python-visualization.github.io/folium/)

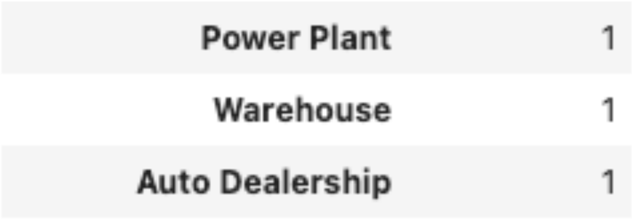
Bellow you can see map of Toronto with plotted blue circles for each neighbourhood. The map is centred on mean latitude and longitude values acquired from ”pgeocode” package.



3.3 Test set selection

- Random selection

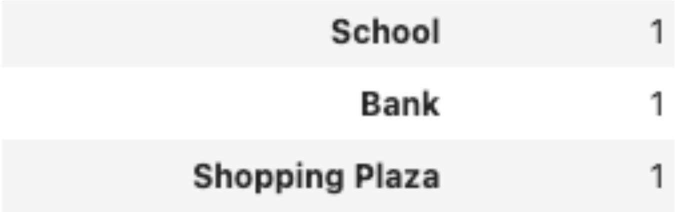
The sample set of 20 POIs will be generated randomly from full set of POIs, which can include very unlikely choses for the users, see bellow.



Power Plant	1
Warehouse	1
Auto Dealership	1

- Custom choice

This set has hand picked POIs from the full list and it looks more realistic as the selection for the Neighbourhood recommender.



School	1
Bank	1
Shopping Plaza	1

- Random neighbourhood top 20 POI

In random neighbourhood test we selected M9C postal code. The top 20 POIs from it were chosen after descendant sort by normalised venue count values. The neighbourhood on the far west side of the city of Toronto has following POIs at the bottom part of its top 20.

- Car Wash
- Pet Store
- Spiritual Center
- Pool
- Salon / Barbershop

3.4 Content based filtering

The neighbourhood recommender system, such we building here, can be based on a variety of different algorithms and machine learning methods. We chose to evaluate content based filtering approach because is simplicity, high level of user personalisation, and ability to learn from user preferences. Also simple, the content based filtering can utilise wide range of statistical methods and technics. For more information on content based filtering please see:

https://en.wikipedia.org/wiki/Recommender_system#Content-based_filtering

and

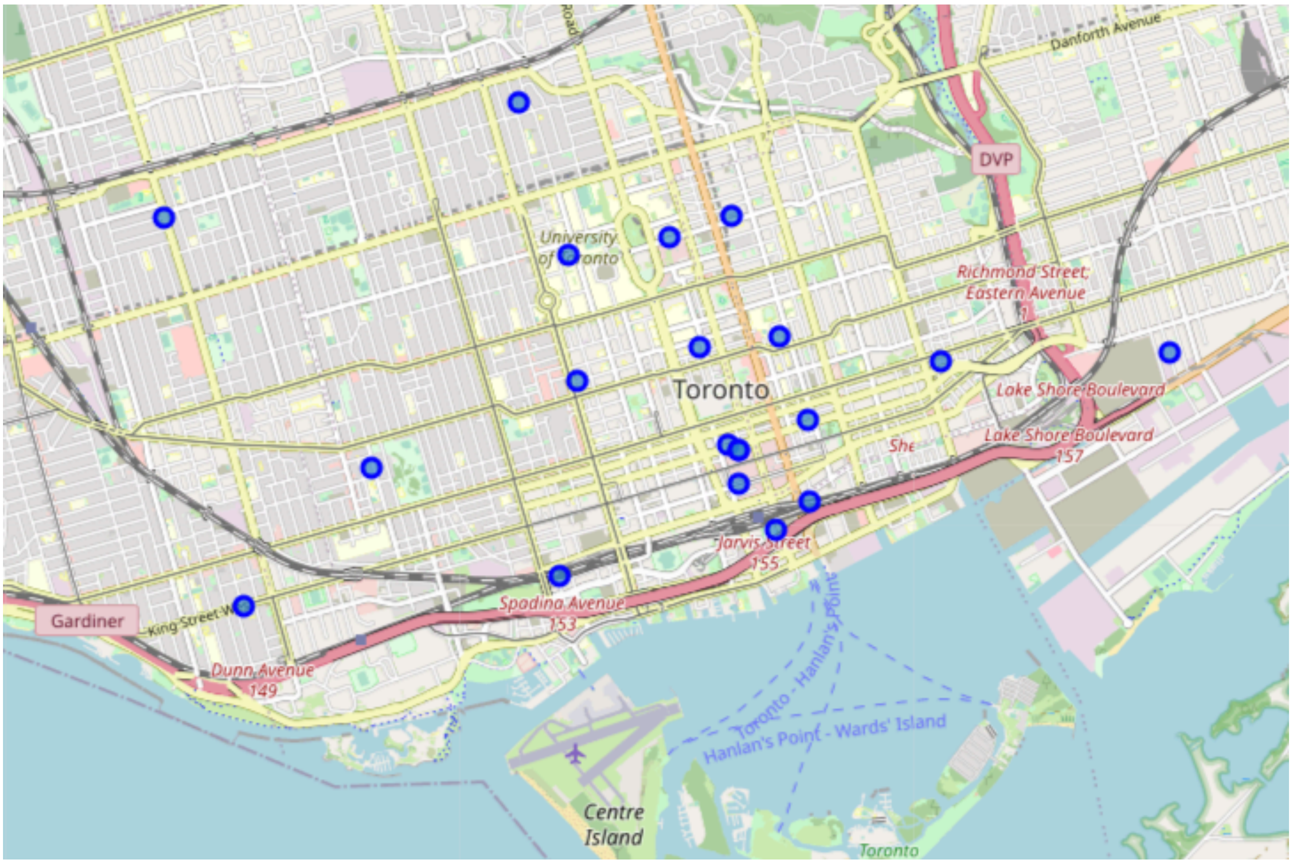
https://towardsdatascience.com/introduction-to-recommender-systems-1-971bd274f421

4. Results

- Random selection

The first test set which was used to test the neighbourhood recommender system was generated randomly from full set of POIs. The set includes very unlikely choses which were shown in the previous section. Even though, the set high not represent a user selection, it is valuable for testing and evaluation of the content based filtering methodology.

Here is the map of top 20 recommended neighbourhoods.

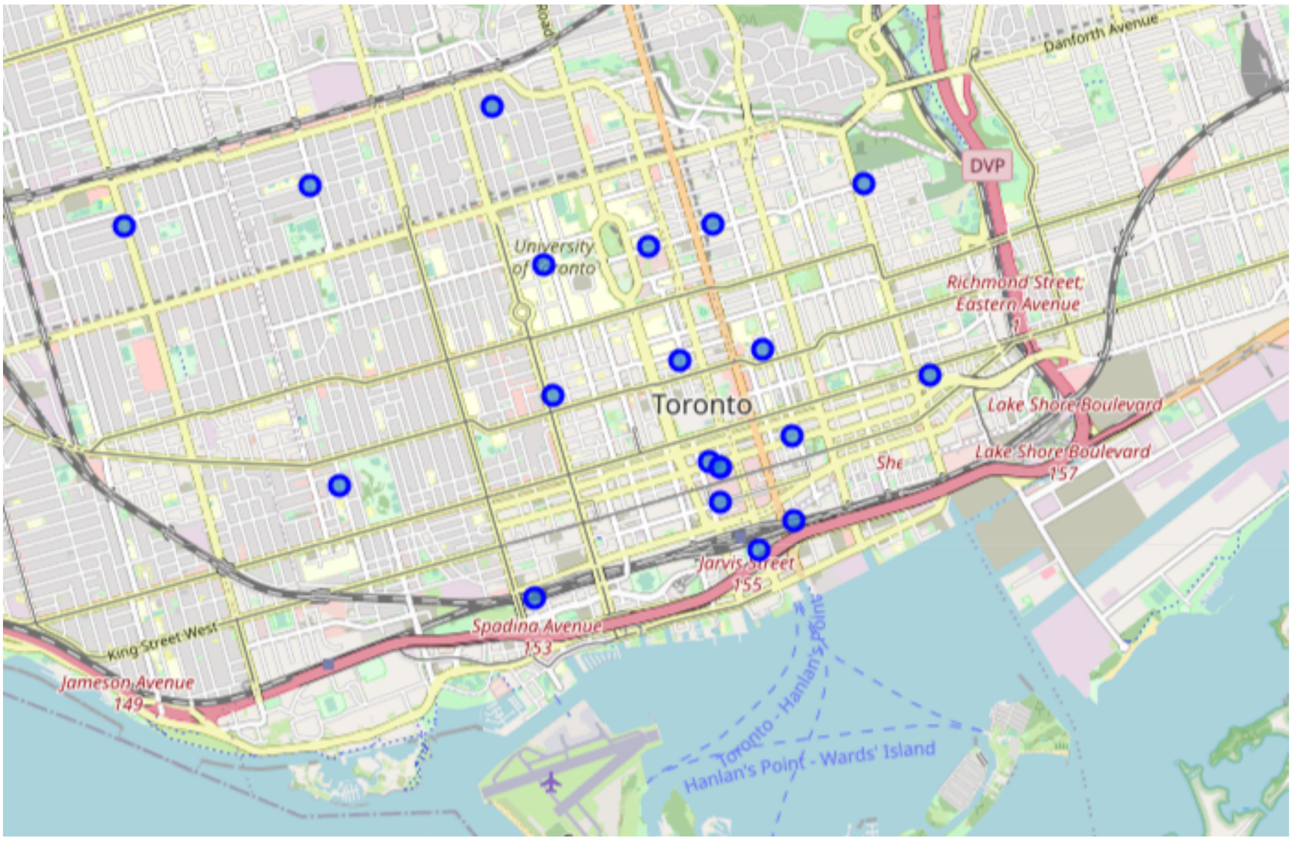


From resulted map we see that the majority of the recommended neighbourhoods located in the city centre or in a close proximity. This is might not be surprising because many amenities indeed can be found in the city centres. However, as it will be shown later in discussion section, the selection of the neighbourhoods can also have a bios in the type and frequency of data collected by Foursquare, which used in this study.

In the random selection test set the top 10 recommended neighbourhoods has predicted values ranges from 0.60 and 0.44.

- Custom choice

The map of top 20 recommended neighbourhoods for hand picked 20 POIs

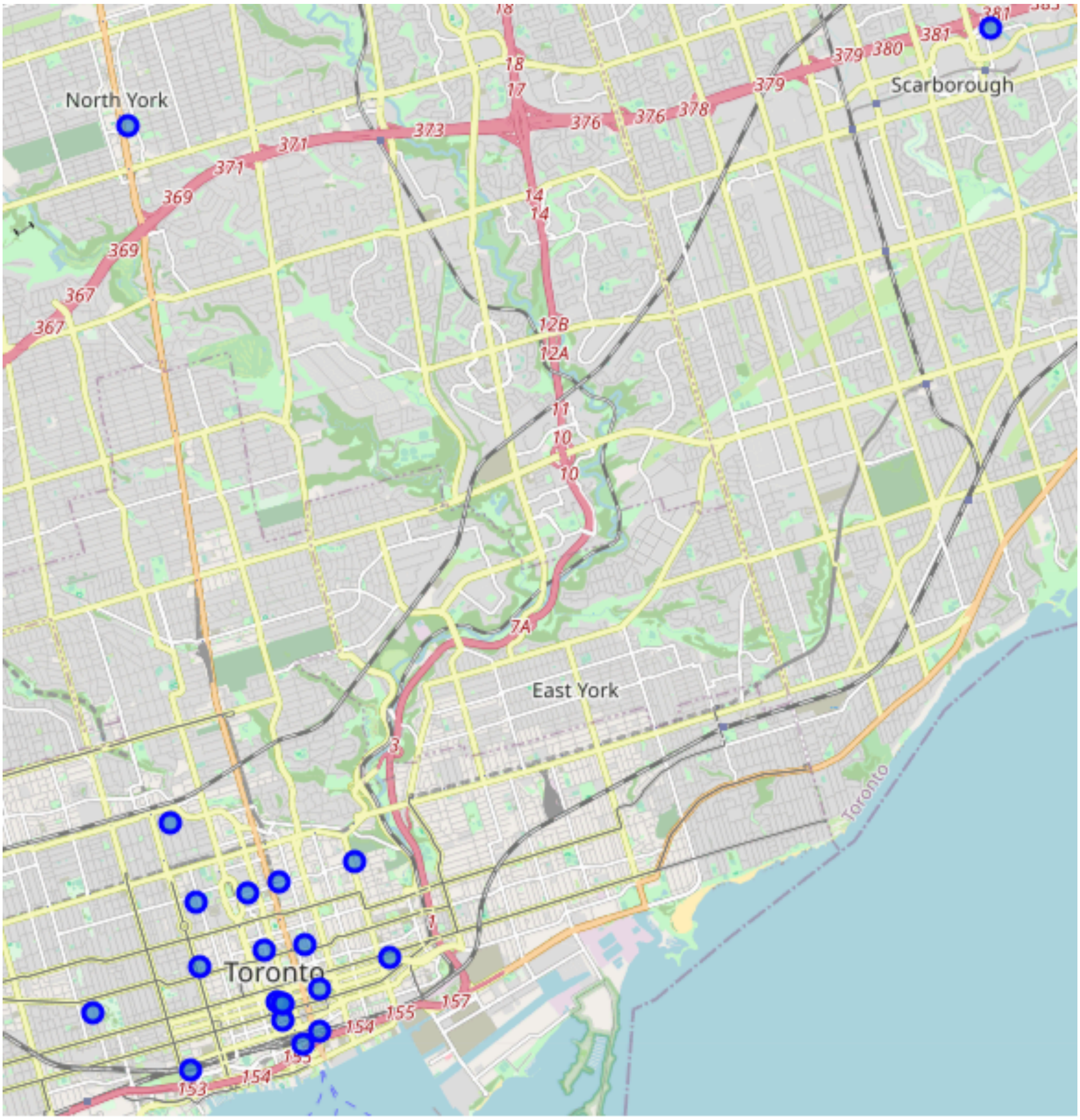


The recommended neighbourhoods from custom selection of 20 POIs are different, but also centred around the city centre.

In contrast to the random test set the predicted values from content based recommendation system are much higher, and ranges between 0.80 and 0.61

- **Random neighbourhood top 20 POI**

The map of top 20 recommendation based on postal code M9C



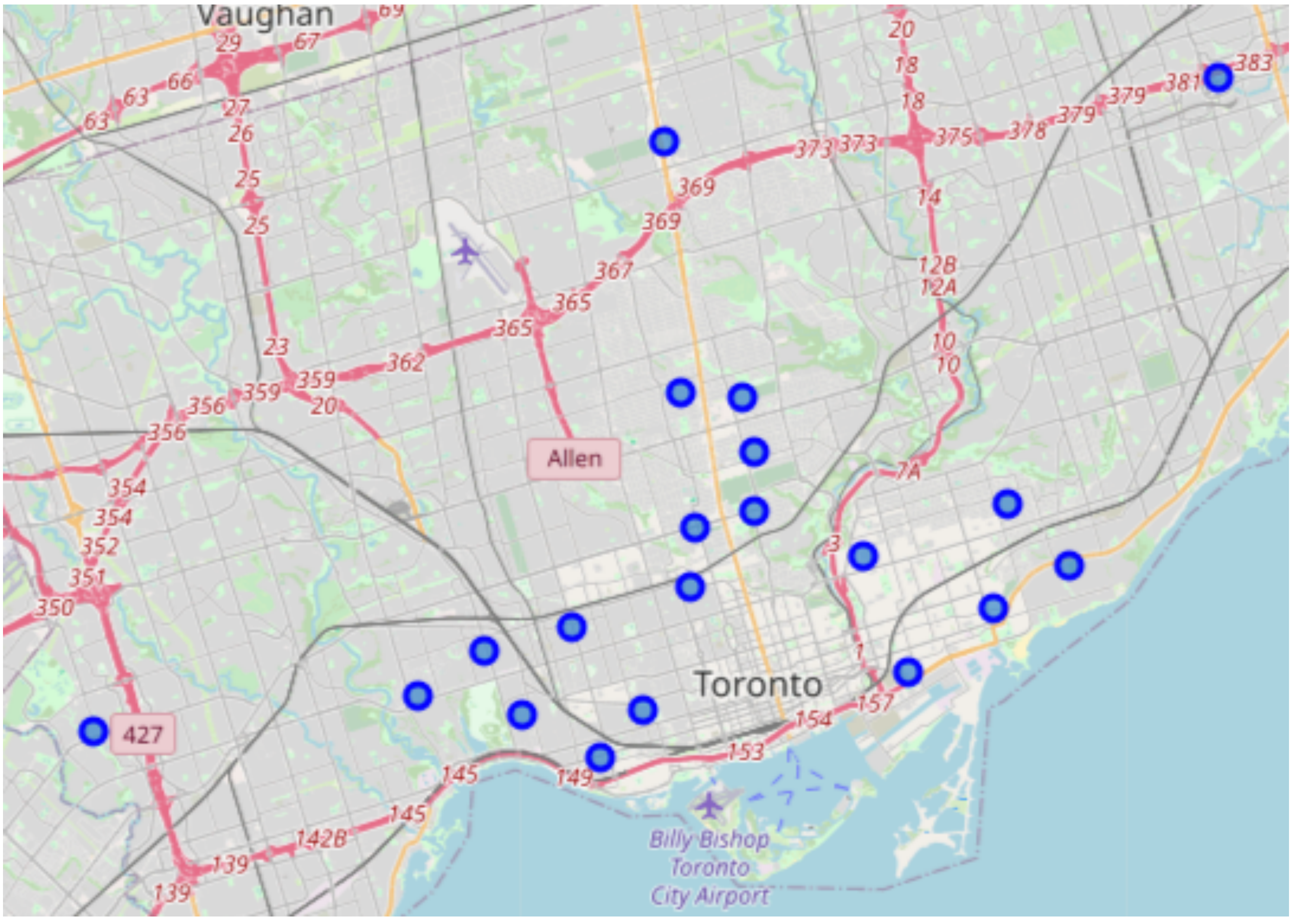
The resulting map, apart from city core, is very different than other recommendations. One thing to point is that M9C Neighbourhood is actually missing from top 20 recommendations plotted on the map. It could simple be due to the other neighbourhoods having even higher normalised values for the POIs selected.

For the M9C postal code test set, the top 10 recommended neighbourhoods has predicted values ranges from 0.73 and 0.57.

- M9C top 20 POI without Downtown

Here, the results from another test on the set selected for M9C postal code. In this test time the neighbourhoods data for the ‘Borough’ equal 'Downtown Toronto’ was removed.

The map of top 20 recommendation for M9C without downtown



In this test set the top 10 recommended neighbourhoods has predicted values ranges from 0.50 and 0.33.

5. Discussion

This study evaluated the content based filtering as the appropriate methodology for neighbourhood recommender system. The results are based on evaluation of four different test sets. The data from top 10 neighbourhood scores of the each test set shows that system performs the best for user selected POI set. It’s neighbourhood scores significantly higher in comparison to other random sample sets.

Interestingly, for any given test set the city centre neighbourhoods seems to score the best. As it was stated in result section, many desired amenities indeed can be found in the city centre. However, close evaluation of the dataset acquired from Foursquare API shows that the city centre area also better annotated, and certain POIs is better annotated in general. For example, if we search for Bus Stop we find that according to Foursquare API data 14 neighbourhoods in Toronto doesn’t have any bus stops in 1000m radius. An alternative search in Google maps shows multiple bus stops in the area. The Foursquare API is a good source of POI information an alternative data source should also be evaluated.

With all of the discussed drawbacks the content based filtering method seems to work in the neighbourhood recommender well. The model can be refined and improved with use of alternative normalisation technic or an additional prediction alygotighms. There are plenty of alternative unsupervised approaches which can be used as an alternative to content based filtering. Unsupervised machine learning method such as K-means clustering can also be evaluated, but its not a part of this study.

6. Conclusion

This study evaluated the content based filtering as an appropriate methodology for neighbourhood recommender system. Based on four different test results we can see that system performs well for the user selected set of POIs in comparison to the other random sample selections. Even though, the content based filtering method seems to suit the neighbourhood recommender well, the alternative approaches, such as K-means clustering or other unsupervised methods should also be evaluated.