# Content-Based Neighborhood Recommendation System

Diego Roncancio

29 April, 2021

## 1. Introduction

### 1.1. Background

Nowadays neighborhoods are an important factor that may raise or decrease the likelihood of a person/family to move to a specific place. Thus, a recommendation system that suggest neighborhoods based on their surrounding top venues can significantly reduce the scope of search for a new place to live and bring a satisfying new home. As satisfaction has been repeatedly linked to efficiency and prosperity, helping people to find suitable neighborhoods can help improve the city. Furthermore, overtime this may give valuable information of what venues drive customer satisfaction within a neighborhood.

### 1.2. Problem

To build this recommendation system, location data pertaining the neighborhoods of a city and their most common surrounding venues will be needed. This project aims to recommend a neighborhood to a customer based on their levels of satisfaction with their previous neighborhoods.

### 1.3. Interest

Real estate agents would be interested in the information this recommendation system will provide to narrow their search. Also, there is the possibility to develop an app with this type of system that could also be economically interesting.

## 2. Data acquisition and cleaning

### 2.1. Data Sources

Now, to simplify matters this recommendation system will be constructed for the city of Toronto. The information of the neighborhoods for this city was scrapped from the list of postal codes in Wikipedia (here) and the information pertaining to the surrounding venues will be obtained using the Foursquare API for each of the neighborhoods. As for the necessary coordinates for each neighborhood a .csv file was found to assign each one of them.

### 2.2. Data cleaning

The information scrapped from Wikipedia had to be parsed in order to be useful, this was achieved using the BeautifulSoup package for Python. Thus, the table in the webpage was extracted and information was recorded in a pandas dataframe with the following features: Postal Code, Borough, Neighborhood, Longitude and Latitude.

### 2.3. Feature selection

After data cleaning, the data frame had 103 neighborhoods with 5 features. Clearly for the recommendation system, we need the types of venues surrounding the neighborhoods and assign them to the dataframe to build the recommendation matrix. Thus, a function was defined to return the venues of a neighborhood in a 500m radius, this radius was defined to not obtain null cells in faraway neighborhoods. Each venue was hot coded and grouped by neighborhood. Since this is a content-based system, the user must have assigned a rating to their current (and previous if available) neighborhoods.

## 3. Exploratory Data Analysis

### 3.1. Exploring Toronto

The first step I wanted to take was to visualize all the neighborhoods in Toronto in a map. Using the folium library, I grouped the neighborhoods by borough and coded them in different colors as can be seen in Figure 1. This allows us to have an initial idea of how many neighborhoods per borough there are, how closely together each neighborhood is, and how are they distributed throughout the city.

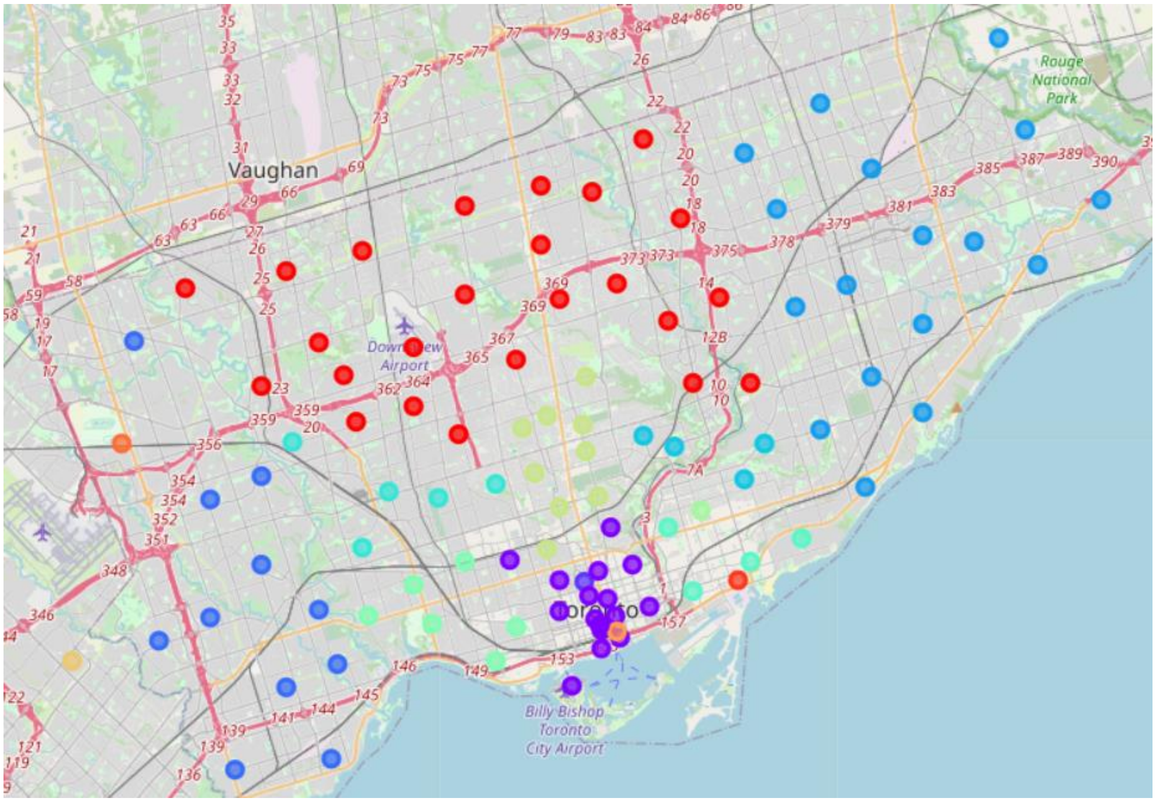


*Figure 1. Neighborhoods colored by Borough in Toronto.*

As we can see in the map, all the neighborhoods are fairly distributed across the city of Toronto, only a slight agglomeration can be perceived in the Downtown Toronto borough. Now to get a more exact idea of the neighborhood distribution per borough in Toronto a car chart was made (Figure 2).

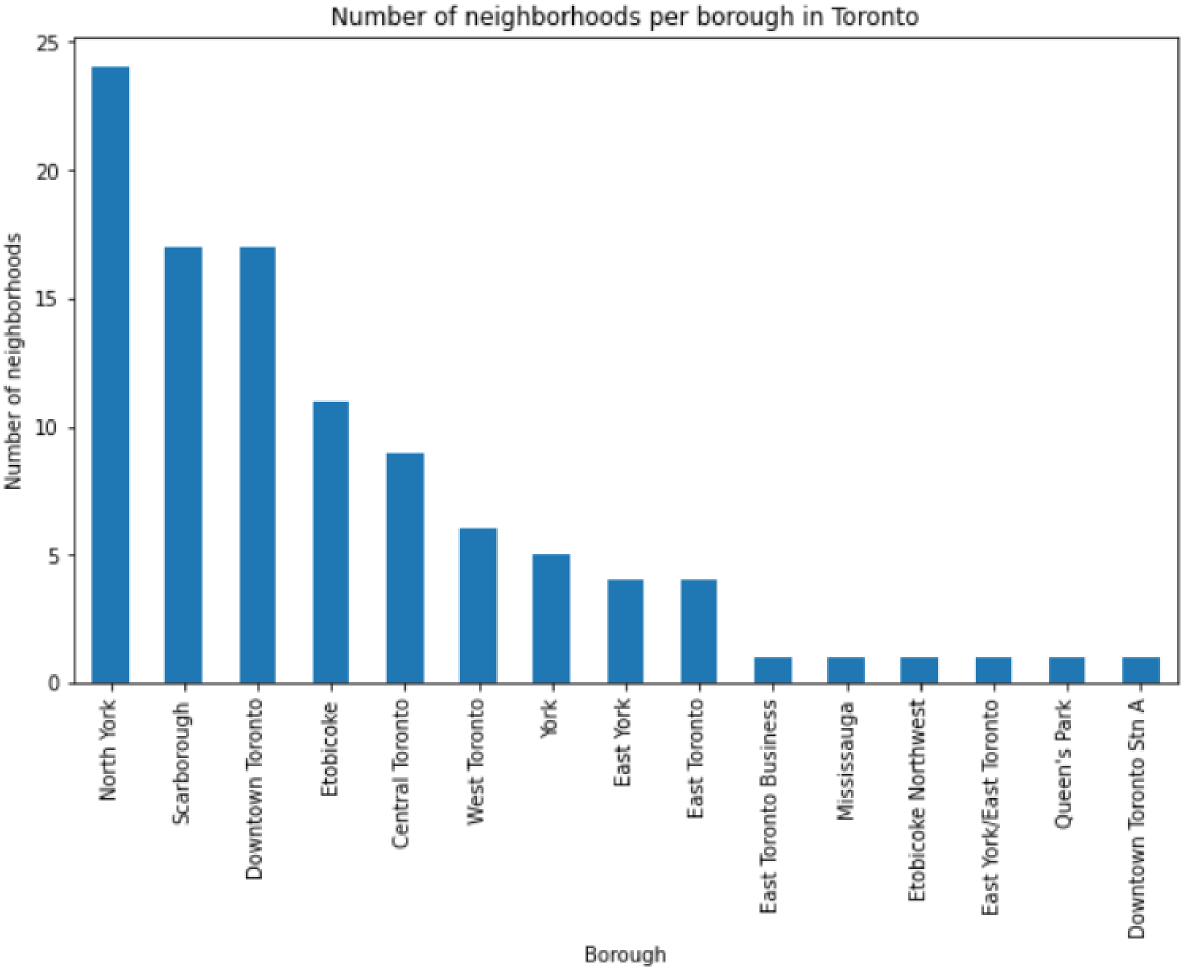


*Figure 2. Comparison between Boroughs regarding number of neighborhoods.*

As can be seen in this chart, three of the fifteen boroughs (North York, Scarborough, and Downtown Toronto) account for over half of the neighborhoods in Toronto. In the cases of Scarborough and North York, we could see in the map that they are extensive boroughs so having so many neighborhoods is not surprising. As for Downtown Toronto, this borough is the main central business district of Toronto which results in a coveted place to reside. Hence the large number of neighborhoods. Lastly, I wanted to look at the number of venues returned per borough, consider that the venues are returned in a 500m radius of each neighborhood, the distribution of the returned venues can be seen in Figure 3.

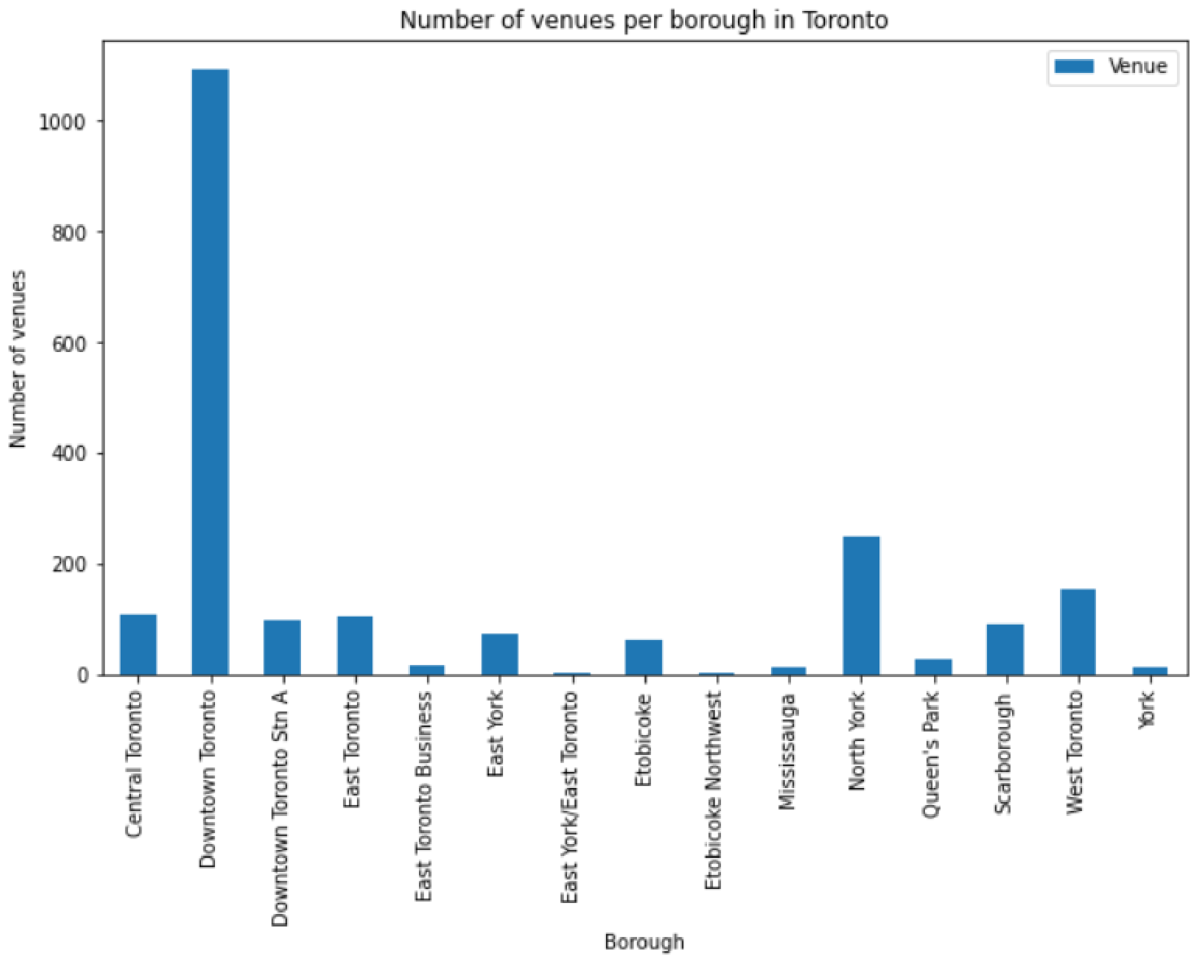


*Figure 3. Comparison between Boroughs regarding the number of venues.*

Looking into Figure 3, we can see a large difference in numbers between Downtown Toronto and the other venues. This can be mainly explained by 2 things: 1. Downtown Toronto being the main central business district of Toronto will result in a large concentration of venues in the area, and 2. The agglomeration that we can see in Figure 1 may result in venues repeating themselves per neighborhood, although I considered this redundancy necessary for the following steps of my recommendation system.

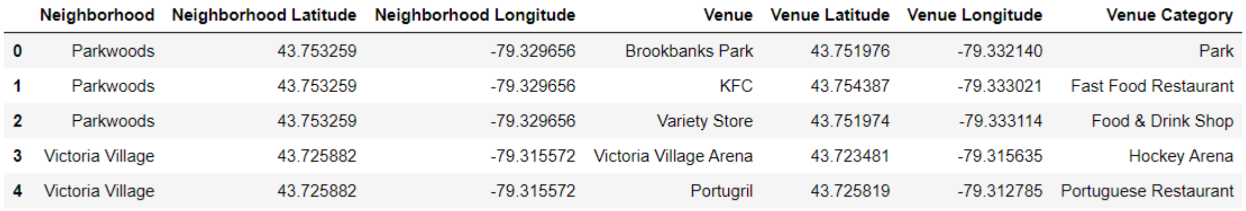
## 4. Content-Based Recommendation System

### 4.1. Preparing the data

*The first thing I needed to do was to collect the data of the venues found in a 500m radius in each neighborhood. Thus, resulting in the sniped version of the dataframe we can see in*

Table 1.

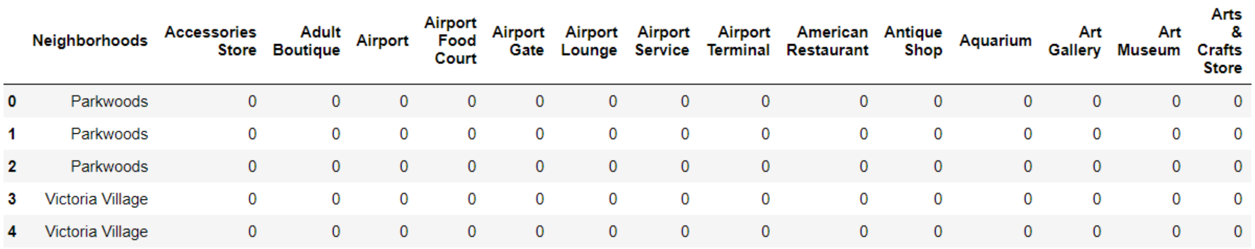
*Table 1. First 5 rows of the dataframe obtained with the venues per neighborhood.*



| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | Parkwoods | 43.753259 | -79.329656 | Brookbanks Park | 43.751976 | -79.332140 | Park |
| 1 | Parkwoods | 43.753259 | -79.329656 | KFC | 43.754387 | -79.333021 | Fast Food Restaurant |
| 2 | Parkwoods | 43.753259 | -79.329656 | Variety Store | 43.751974 | -79.333114 | Food & Drink Shop |
| 3 | Victoria Village | 43.725882 | -79.315572 | Victoria Village Arena | 43.723481 | -79.315635 | Hockey Arena |
| 4 | Victoria Village | 43.725882 | -79.315572 | Portugril | 43.725819 | -79.312785 | Portuguese Restaurant |

The most important feature in this case is the Venue Category since I will use this feature to construct my recommendation matrix. The first step was to drop all the other information since I do not really care about the exact coordinates and specific name of each venue. Then I converted the Venue Category into several columns with the unique categories present in my dataframe obtaining 269 features. These features have binary values that signify what type of venue it is. The result can be seen in Table 2.

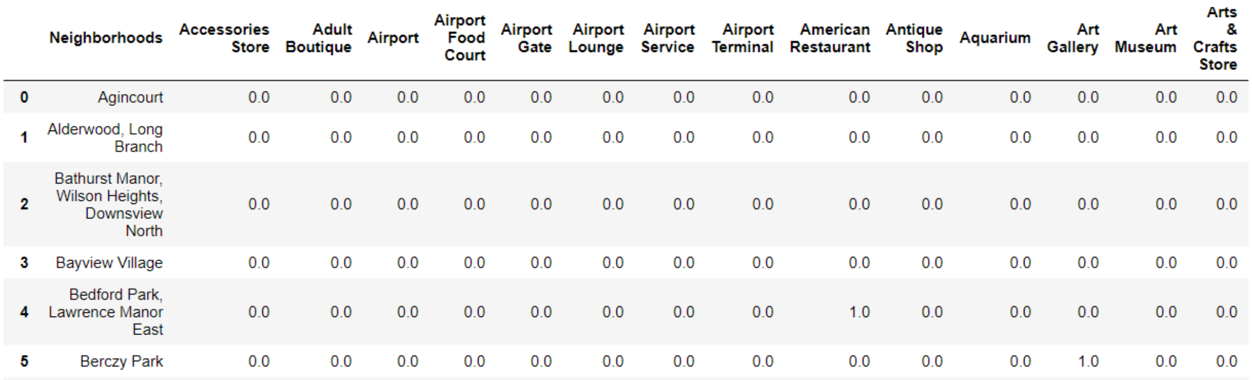
*Table 2. First 5 rows of the dataframe one-hot encoded by venue category.*



| | Neighborhoods | Accessories Store | Adult Boutique | Airport | Airport Food Court | Airport Gate | Airport Lounge | Airport Service | Airport Terminal | American Restaurant | Antique Shop | Aquarium | Art Gallery | Art Museum | Arts & Crafts Store |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Parkwoods | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | Parkwoods | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | Parkwoods | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | Victoria Village | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | Victoria Village | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Lastly, as we can see in Table 2 the dataframe is still not grouped by neighborhood. Thus, the next step is concise all the venue types present in each neighborhood. After doing this I finally obtained a dataframe that has information on which venue types each neighborhood has and can be seen in Table 3.

*Table 3. First 6 rows of the dataframe with venue categories per neighborhood.*

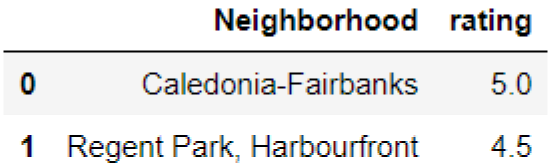


| | Neighborhoods | Accessories Store | Adult Boutique | Airport | Airport Food Court | Airport Gate | Airport Lounge | Airport Service | Airport Terminal | American Restaurant | Antique Shop | Aquarium | Art Gallery | Art Museum | Arts & Crafts Store |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Agincourt | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 1 | Alderwood, Long Branch | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 2 | Bathurst Manor, Wilson Heights, Downsview North | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 3 | Bayview Village | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 4 | Bedford Park, Lawrence Manor East | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 5 | Berczy Park | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 |

## 4.2. Defining the user profile

Finally, the last step to construct the recommendation system is to create the user profile. For the purpose of getting results I created a user that wishes to move to a particular borough (North York) to be closer to his job. This user inputted his two previous neighborhoods with their respective rating (Table 4).

*Table 4. User inputted neighborhoods.*

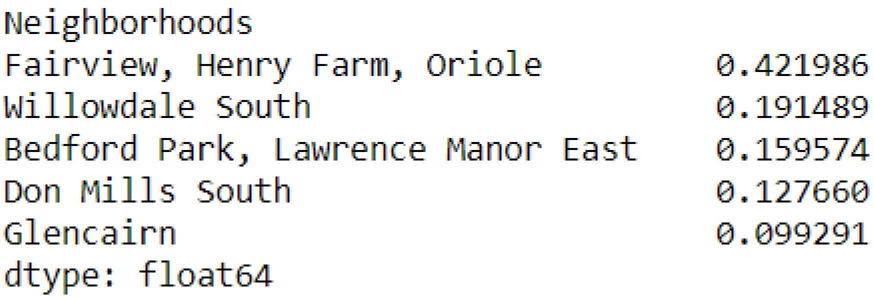


| | Neighborhood | rating |
|---|---|---|
| **0** | Caledonia-Fairbanks | 5.0 |
| **1** | Regent Park, Harbourfront | 4.5 |

Having these inputs by the user, I extracted the categories venues for his rated neighborhoods and appended the weighted results in a dataframe. With this dataframe I have the weighted preference in venue category of the user. The next step was to extract from the original dataframe in Table 3 the neighborhoods that belong to the North York borough. With these steps I have both the defined user profile and the search scope.

### 4.3. Resulting recommendation

Now passing the user profile and the dataframe with the borough of interest I obtained the new dataframe with the weighted preference of the user for each neighborhood in North York. After sorting it, I passed the top 5 recommended neighborhoods, and their weight as can be seen in Figure 4.



```
Neighborhoods
Fairview, Henry Farm, Oriole           0.421986
Willowdale South                       0.191489
Bedford Park, Lawrence Manor East      0.159574
Don Mills South                        0.127660
Glencairn                              0.099291
dtype: float64
```

*Figure 4. Top 5 neighborhood recommendations in North York.*

Finally, I wanted to display in the Toronto map the recommended neighborhoods. So I searched them in the dataframe that contained the coordinate information and created the following map (Figure 5).

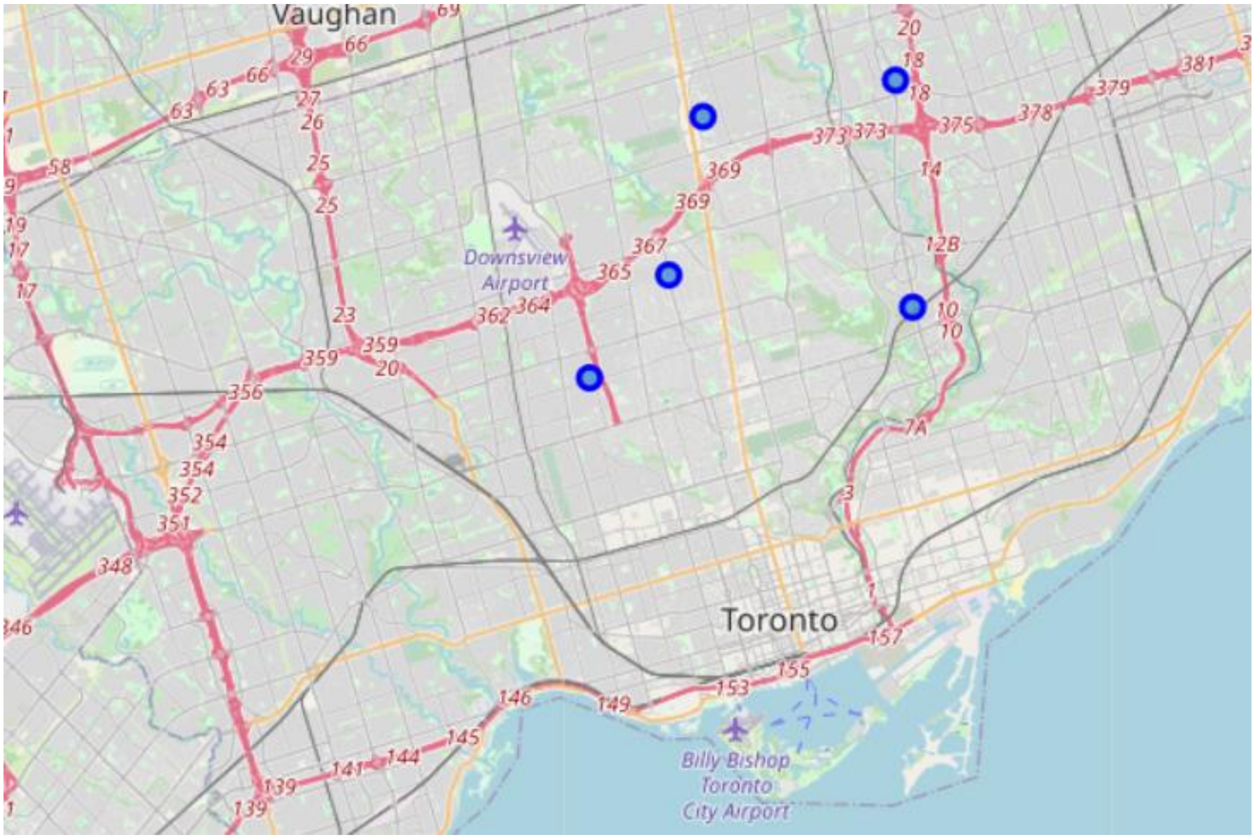


*Figure 5. Map of the recommended neighborhoods in North York.*

## 5. Discussion

Although the content-based system recommendation successfully sorted the neighborhoods in North York by the user preferences, if we observe the weights of the 5 recommendations, they are not very high. Only the Fairview neighborhood has an acceptable weight (0.42) for a recommendation. This can be explained by the truly vast number of unique categories present in the venues collected, each neighborhood having 269 features truly makes it difficult to find matches to the user preferences by rating their previous neighborhoods.

This problem may be solved by implementing a collaborative filtering recommendation system, but the problem that a huge amount of user data must be collected first. Another possible solution is to simplify things a bit and ask the user what specific venue categories are important to him and assign them ratings, instead of inputting neighborhoods. This would inconvenience the user a bit for inputting more information, but the recommendations could have a better tailored result.

## 6. Conclusions

- The content-based recommendation system successfully gave the recommended neighborhoods in the targeted borough, but it is limited in giving high weighted user preferences for a neighborhood due to the large number of features.

- Possible future work can be done to refine this recommendation system. The options proposed to better the service consist in either use another recommendation system (collaborative filtering or maybe a hybrid) or asking the user for their preferred venues instead of the rating for their past neighborhoods.