Welcome to Toronto!

A Neighborhood Recommender System

Shan Ali

April 29, 2020

1. Introduction

A Note from the New Arrivals Department of the City of Toronto Government:

Dear Newest Resident,

Welcome to the beautiful City of Toronto!

We know moving can be challenging and stressful. To help your transition to our great city, we provide a free service to help connect you with the right neighborhood for you. The service utilizes your distance from work, similarity to previous neighborhoods, and preferred lifestyle amenities nearby. Just input the required information and we will take care of the rest. Thank you and welcome home!

1.1. Background & Problem

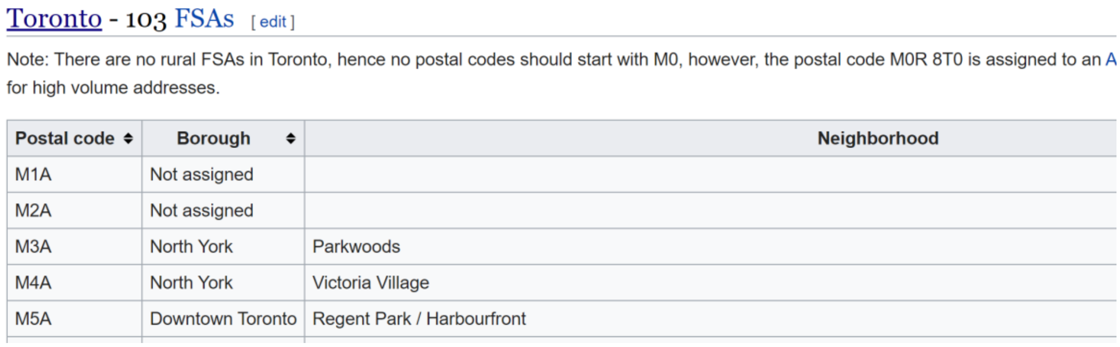
Moving to a new city can be stressful. New roads, new neighborhoods, and new employment are all challenges when attempting to move. One of the most challenging tasks, however, is choosing your next home, which can be a time consuming and costly endeavor. In an effort to ease this burden, I have developed a classic recommender system that utilizes location data to recommend you the best neighborhoods in Toronto, Canada. This simple and easy to use system builds your recommendations based on the similarity to your previous neighborhood and distance of your commute. That’s all it takes! This system will save new citizens precious time, money, and resources, all while enabling the most priceless thing of all: peace of mind.

2. Data Acquisition and Cleaning

2.1. Data Sources

The data comes in two forms: 1) system dataset on neighborhoods in Toronto and 2) input dataset on the user's neighborhood data.

The system dataset contains the postal code, borough, name, longitude, latitude, and venue data about each neighborhood in Toronto. The dataset pulls its data from three sources. Postal code, borough and name data for each neighborhood was scraped off this Wikipedia entry on postal codes in Toronto. The longitude and latitude data are called from the Geopy location dataset using the neighborhood name. The venue data was called using the Foursquare API.



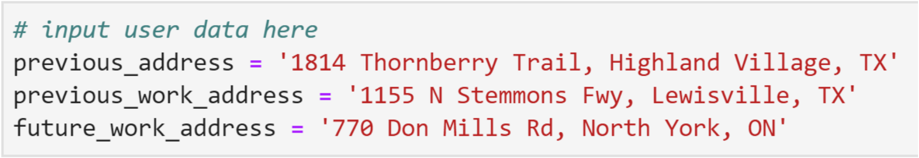
Toronto - 103 FSAs [edit]

Note: There are no rural FSAs in Toronto, hence no postal codes should start with M0, however, the postal code M0R 8T0 is assigned to an A for high volume addresses.

Postal code ♦	Borough ♦	Neighborhood
M1A	Not assigned	
M2A	Not assigned	
M3A	North York	Parkwoods
M4A	North York	Victoria Village
M5A	Downtown Toronto	Regent Park / Harbourfront

Fig 1. Screen shot sample of the Toronto postal code Wikipedia webpage

The input dataset contains the address, neighborhood, longitude, and latitude of the user's former address, former work address, and future work address. The address data is inputted by the user. The longitude and latitude data are call from the Goopy location dataset using the neighborhood name.

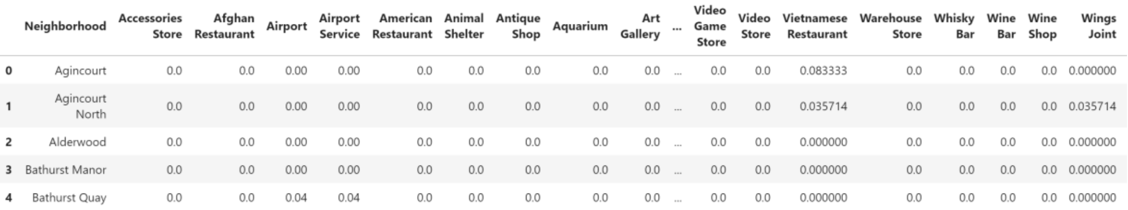


```
# input user data here
previous_address = '1814 Thornberry Trail, Highland Village, TX'
previous_work_address = '1155 N Stemmons Fwy, Lewisville, TX'
future_work_address = '770 Don Mills Rd, North York, ON'
```

Fig 2. Example of input data

2.2. Data Cleaning

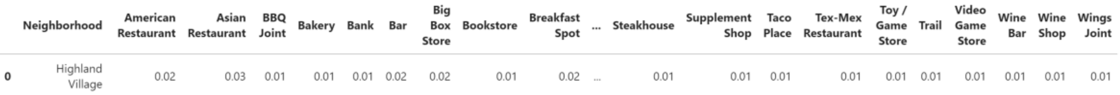
The system dataset was cleaned considerably before compilation. The borough, neighborhood, and postal code data of the system dataset was scraped off Wikipedia using the BeautifulSoup python library. The function returns the html of the page so additional processing was required. The <tr> features of the page were isolated and then organized to generate the beginning of the system dataset. The neighborhoods come grouped by borough and require splitting and isolation to generate unique rows for each entry. The dataset was then run through the Geopy Nominatim function to pull the latitude and longitude of each neighborhood based on their address, for example “Downtown Toronto, ON”. Then using the latitude and longitude, neighborhoods are explored to discover the nearby venue information using the foursquare API. This data pulls the 100 most popular venues within 500 m of a location. The venues were then isolated and had their venue category one-hotted, summed, and normalized. This gives the venue frequency of each neighborhood and will be used as the comparator for the recommender system.



	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Service	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Art Gallery	...	Video Game Store	Video Store	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wine Shop	Wings Joint
0	Agincourt	0.0	0.0	0.00	0.00	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.083333	0.0	0.0	0.0	0.0	0.000000
1	Agincourt North	0.0	0.0	0.00	0.00	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.035714	0.0	0.0	0.0	0.0	0.035714
2	Alderwood	0.0	0.0	0.00	0.00	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000
3	Bathurst Manor	0.0	0.0	0.00	0.00	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000
4	Bathurst Quay	0.0	0.0	0.04	0.04	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.000000

Fig 3. First 5 entries in the system dataset after being loaded with venue data.

The input dataset required less processing. The address was split to isolate the neighborhood name. The address was then run through the Geopy Nominatim function to pull longitude and latitude data for each of the three inputted addresses.



	Neighborhood	American Restaurant	Asian Restaurant	BBQ Joint	Bakery	Bank	Bar	Big Box Store	Bookstore	Breakfast Spot	...	Steakhouse	Supplement Shop	Taco Place	Tex-Mex Restaurant	Toy / Game Store	Trail	Video Game Store	Wine Bar	Wine Shop	Wings Joint
0	Highland Village	0.02	0.03	0.01	0.01	0.01	0.02	0.02	0.01	0.02	...	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Fig 4. Input dataset loaded with venue data for previous neighborhood

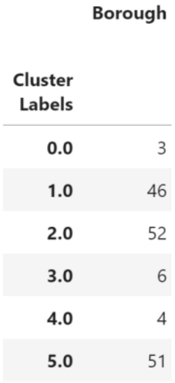
2.3. Feature Selection

Generating the system dataset build a detailed trove of information about each neighborhood. However, not all is useful in generating the recommendations. Along with the nearby venue data, distance to work was also needed to filter neighborhood options. Using the inputted former home and work address’s longitude and latitude, the Euclidean distance is found from home to work, i.e. the commute distance. This was then added to the input dataset. The work to home distance was then calculated between the Toronto neighborhoods and future work’s longitude and latitude. This data was then normalized and added to a clone of the system dataset. The distance in the input dataset was normalized based on the farthest distance of the system dataset. To save on memory, the postal code, neighborhoods name, borough name, longitude, and latitude data was dropped from the clone. That left the venue and commute data.

The venue data and commute data were chosen as the content features for each neighborhood. This is because these data give unique and comparable information about each neighborhood and can be easily compared quantitatively to generate a similarity/dissimilarity score. This score will directly compare the former neighborhood to potential Toronto neighborhoods. These features will be used as the feed for a content-based recommendation model.

3. Exploratory Data Analysis

Before implementing the solution, the system dataset was explored to determine the variety of neighborhoods in the City of Toronto. Using k-means clustering, the dataset was divided into six similar clusters (k = 6). The neighborhoods split into two groups of three clusters, with each cluster having a similar number of neighborhoods as the other two in its group. This demonstrates the diversity in neighborhoods across Toronto and ensures the solution will filter out at least 2/3 of the dissimilar suggestions.



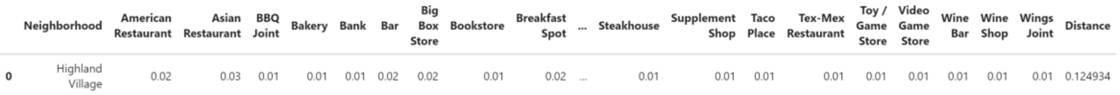
Cluster Labels	Borough
0.0	3
1.0	46
2.0	52
3.0	6
4.0	4
5.0	51

Fig 5. Number of neighborhoods in each cluster

4. Recommendation Model

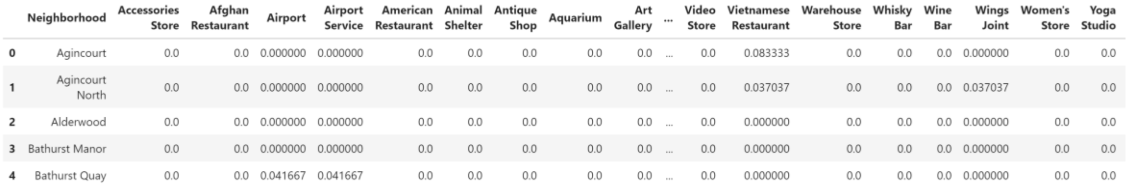
4.1. Solution Implementation

To implement the solution, the system had to provide a recommendation for similar neighborhoods in Toronto to the user’s previous neighborhood. A content-based recommendation system was built as the solution. The cloned dataset was used for the recommender. To ensure calculations were efficient, the venue categories not present in the previous neighborhood were dropped. The system calculated the multi-variable Euclidean distance between the venue and commute distance data of each neighborhood and the former neighborhood. This data was then summed and was representative of the dissimilarity between two neighborhoods. The neighborhood name and dissimilarity score were then compiled into a results dataset and ordered in ascending order to represent the neighborhoods most similar to the input neighborhood. The borough, longitude, and latitude data were merged into the results dataset.



	Neighborhood	American Restaurant	Asian Restaurant	BBQ Joint	Bakery	Bank	Bar	Big Box Store	Bookstore	Breakfast Spot	...	Steakhouse	Supplement Shop	Taco Place	Tex-Mex Restaurant	Toy / Game Store	Video Game Store	Wine Bar	Wine Shop	Wings Joint	Distance
0	Highland Village	0.02	0.03	0.01	0.01	0.01	0.02	0.02	0.01	0.02	...	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.124934

Fig 6. Input neighborhood content dataset. Includes venue and distance data.



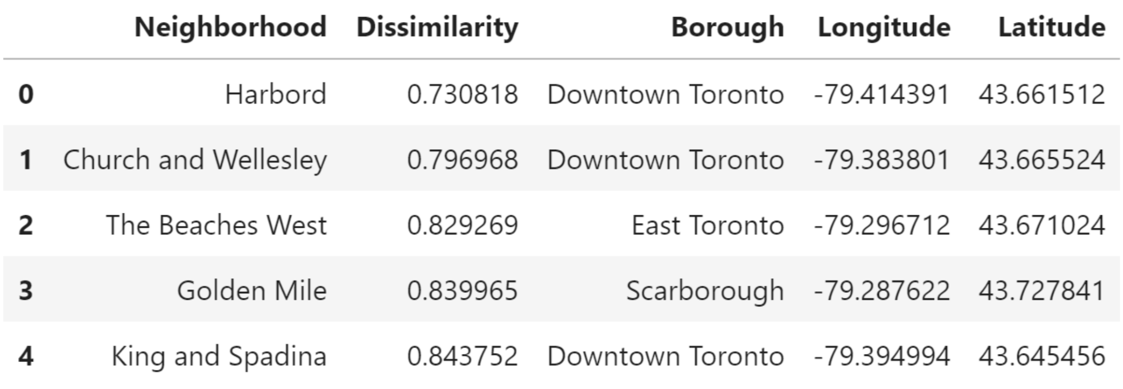
	Neighborhood	Accessories Store	Afghan Restaurant	Airport	Airport Service	American Restaurant	Animal Shelter	Antique Shop	Aquarium	Art Gallery	...	Video Store	Vietnamese Restaurant	Warehouse Store	Whisky Bar	Wine Bar	Wings Joint	Women's Store	Yoga Studio
0	Agincourt	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.083333	0.0	0.0	0.0	0.000000	0.0	0.0
1	Agincourt North	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.037037	0.0	0.0	0.0	0.037037	0.0	0.0
2	Alderwood	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0
3	Bathurst Manor	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0
4	Bathurst Quay	0.0	0.0	0.041667	0.041667	0.0	0.0	0.0	0.0	0.0	...	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0

Fig 7. System neighborhood content dataset. Includes venue and distance data.

To visualize the data, a map of Toronto was generated using the Folium python library. The map was populated with all the neighborhood and borough data of Toronto. It was overlaid with the top 5 most similar neighborhoods, the future work location, and a ring showing the distance of the user's previous commute. This map and the results dataset were displayed as the final output of the system.

4.2. Solution Results & Discussion

Due to the elegance of the recommender, the system operates quickly to calculate results. In the test example, the top five most similar neighborhoods are: Harbord, Church and Wellesley, the Beaches West, Golden Mile, and King and Spadina. These locations feature the most similar venues and are similarly distanced from the user's future work.



	Neighborhood	Dissimilarity	Borough	Longitude	Latitude
0	Harbord	0.730818	Downtown Toronto	-79.414391	43.661512
1	Church and Wellesley	0.796968	Downtown Toronto	-79.383801	43.665524
2	The Beaches West	0.829269	East Toronto	-79.296712	43.671024
3	Golden Mile	0.839965	Scarborough	-79.287622	43.727841
4	King and Spadina	0.843752	Downtown Toronto	-79.394994	43.645456

Fig 8. Results table with top 5 most similar neighborhoods.

The map helps to visualize the best neighborhoods to where to live based on similarity to previous neighborhood and distance to work. The map populates the neighborhoods of Toronto (Black), the top 5 most similar neighborhoods (Red), the user’s future work location (Blue), and the radius of the user’s previous commute.

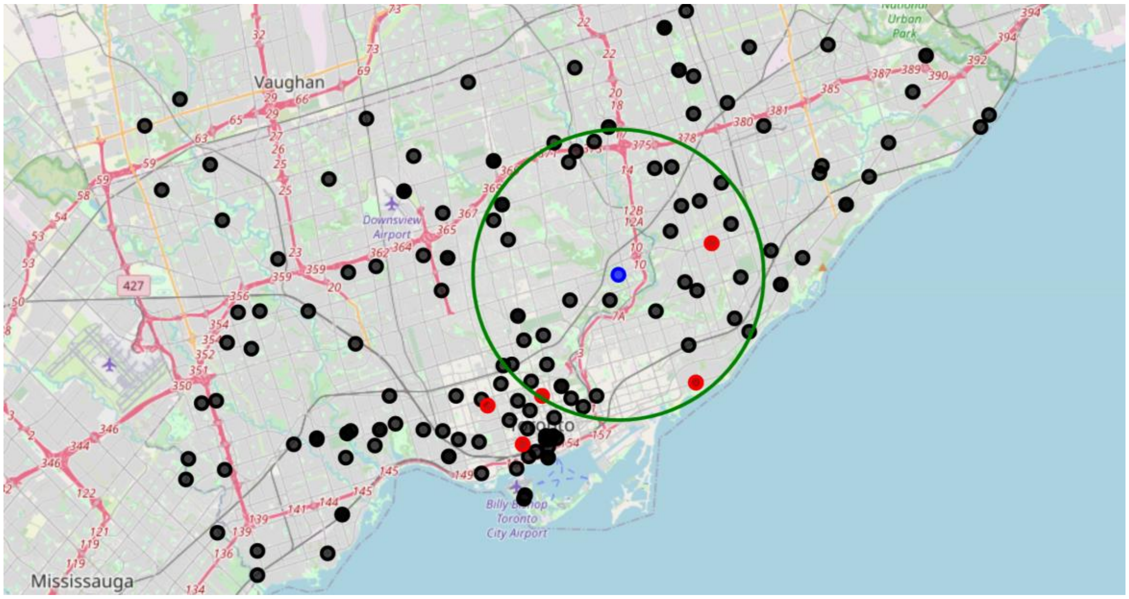


Fig 9. Results map featuring neighborhoods of Toronto, top 5 neighborhoods, future work location, and previous commute distance

4.3. Solution Limitations

The system had some limitations. To save on memory and user input requirements, the system only considers historical venue and commute data. The system also assumes the user likes their former neighborhood and that nearby venues are predictive of living standards. To make the system more tailored to the user, a more detailed questionnaire would be useful. Questions like, what are your top 10 venue categories (i.e. gym, Asian restaurant, trials, etc.)? What is your ideal commute to work? What price range for housing are you looking at? These questions would give more detail into the user’s preference; however, the simplicity of the current format does allow the system to be quick and easy to use while still providing quality results and maximizing privacy.

5. Conclusion

In conclusion, the system provides a robust and simple solution to recommending neighborhoods in Toronto for a new resident. It considers the venues of the previous neighborhood and commute distance to provide the most simple and complete solution. This system could be helpful to and easily implemented by the City of Toronto as a tool for its new or moving residents.