Course: Coursera Data Science Professional by IBM
Title: Data Science Capstone Project
Author: Lee Aik Tuan

Introduction

In this project, we are working for a relocating firm that helps people relocate from Toronto, Ontario, Canada to Manhattan, New York, USA. As part of the effort to help customers ease into their new environment, we are starting a pilot project to help them find neighborhoods that match well with their existing neighborhood in Toronto. For example, people who stay near parks now would be interested in finding places with parks in Manhattan, or those who like sushi may prefer to have a few Japanese restaurants near their homes. In this case, the business question is "Which neighborhoods in Manhattan most closely matches with a given neighborhood in Toronto, in terms of venues and facilities?"

Methodology

To answer this question, we would need to develop a kind of recommendation system that recommends the neighborhoods of Manhattan based on the company's clients' existing neighborhoods in Toronto. We would need to neighborhood information about both Toronto and New York. Information about the New York neighborhoods is available on the internet (https://geo.nyu.edu/catalog/nyu_2451_34572) but for the project, a cached copy at the IBM server (https://cocl.us/new_york_dataset) was used. Information about Toronto neighborhoods can be scraped off following the Wikipedia page, https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M and the coordinates can be obtained either from Google Maps Geocoding API or the Geocoder Python package. In this project, the latitudes and longitudes are sourced from a cached copy on the IBM server (https://cocl.us/Geospatial_data).

We would also need information about the venues or places of interest, their categories and the number of such venues in these two cities. For such information, we would make use of the Four Square APIs.

Methodology[1]

Exploratory Data Analysis—Toronto Neighborhoods

From the source, there are a total of 39 Toronto neighborhoods with a diverse array of venues, making a total of 233 different unique venue categories. Figure 1 shows statistics about different venue categories in the neighborhoods.

[1] Data is based on results from the script run on 30 June 2020 10:22 AM (UTC+8).

	Venue Category
count	39.000000
mean	41.589744
std	33.520393
min	2.000000
25%	16.000000
50%	35.000000
75%	62.000000
max	100.000000

From the above, it can be seen that the categories in each neighborhood vary greatly, from 2 to 100 (or more). Most places would have about 41 venue categories. For this pilot project, we shall focus on the top 5 categories for each neighborhood in Toronto[2].

Figure 2 and Figure 3 show sample listings of Toronto neighborhoods and the most common venue in each.

Figure 1: Venue Category Statistic

	Toronto Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	1st Venue Weight	2nd Venue Weight	3rd Venue Weight	4th Venue Weight	5th Venue Weight	Total Wt
29	St. James Town	Café	Coffee Shop	Cocktail Bar	Restaurant	American Restaurant	0.263158	0.263158	0.157895	0.157895	0.157895	0.237500
0	Berczy Park	Coffee Shop	Cocktail Bar	Seafood Restaurant	Bakery	Restaurant	0.357143	0.214286	0.142857	0.142857	0.142857	0.241379
31	Stn A PO Boxes	Coffee Shop	Café	Seafood Restaurant	Cocktail Bar	Restaurant	0.416667	0.166667	0.166667	0.125	0.125	0.247423
13	Garden District, Ryerson	Clothing Store	Coffee Shop	Bubble Tea Shop	Middle Eastern Restaurant	Café	0.36	0.28	0.12	0.12	0.12	0.250000
30	St. James Town, Cabbagetown	Coffee Shop	Pizza Place	Pub	Italian Restaurant	Bakery	0.25	0.25	0.166667	0.166667	0.166667	0.272727
6	Church and Wellesley	Coffee Shop	Sushi Restaurant	Japanese Restaurant	Restaurant	Gay Bar	0.285714	0.238095	0.190476	0.142857	0.142857	0.272727
25	Richmond, Adelaide, King	Coffee Shop	Restaurant	Café	Gym	Hotel	0.37037	0.185185	0.185185	0.148148	0.111111	0.287234
17	Kensington Market, Chinatown, Grange Park	Café	Bakery	Mexican Restaurant	Vietnamese Restaurant	Coffee Shop	0.294118	0.176471	0.176471	0.176471	0.176471	0.288136
14	Harbourfront East, Union Station, Toronto Islands	Coffee Shop	Aquarium	Hotel	Café	Scenic Lookout	0.448276	0.172414	0.137931	0.137931	0.103448	0.290000
2	Business reply mail Processing Centre, South C...	Yoga Studio	Skate Park	Auto Workshop	Brewery	Burrito Place	0.2	0.2	0.2	0.2	0.2	0.294118

Figure 2: Top 5 common venue categories of diverse Toronto Neighborhoods

	Toronto Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	1st Venue Weight	2nd Venue Weight	3rd Venue Weight	4th Venue Weight	5th Venue Weight	Total Wt
10	Dufferin, Dovercourt Village	Bakery	Pharmacy	Bank	Supermarket	Bar	0.285714	0.285714	0.142857	0.142857	0.142857	0.500000
3	CN Tower, King and Spadina, Railway Lands, Har...	Airport Service	Airport Lounge	Airport Terminal	Boutique	Coffee Shop	0.333333	0.222222	0.222222	0.111111	0.111111	0.529412
9	Davisville North	Gym	Food & Drink Shop	Sandwich Place	Hotel	Dog Run	0.2	0.2	0.2	0.2	0.2	0.625000
5	Christie	Grocery Store	Café	Park	Baby Store	Nightclub	0.363636	0.272727	0.181818	0.0909091	0.0909091	0.687500
26	Rosedale	Park	Playground	Trail	None	None	0.5	0.25	0.25	0	0	1.000000
27	Roselawn	Music Venue	Garden	None	None	None	0.5	0.5	0	0	0	1.000000
12	Forest Hill North & West, Forest Hill Road Park	Park	Trail	Jewelry Store	Sushi Restaurant	None	0.25	0.25	0.25	0.25	0	1.000000
35	The Beaches	Trail	Neighborhood	Health Food Store	Pub	None	0.25	0.25	0.25	0.25	0	1.000000
20	Moore Park, Summerhill East	Gym	Trail	None	None	None	0.5	0.5	0	0	0	1.000000
18	Lawrence Park	Park	Swim School	Bus Line	None	None	0.333333	0.333333	0.333333	0	0	1.000000

Figure 3: Top 5 common venue categories of least diverse Toronto Neighborhoods

The initial weights were assigned based on the mean of the category under which the venues fall. For example, if there are 100 venues listed in the neighborhood, and 30 of them are bakeries and 10 of them are coffee shops, the weight will be 0.3 and 0.1 for Bakery and Coffee Shops respectively. The sum of the 5 weights (for the top 5 categories) are listed in the 'Total Wt' column. If the 'Total Wt' is low, it means that the neighborhood contains a wide variety of venues, and the top 5 only accounts for a small percentage of the venues. These are apparent in places like Berczy Park, St James Town and Garden District where the Total Wt values are between 0.24 and 0.25, which means that the top 5 categories account

[2] The top category is the category in which most venues in the neighborhood would fall under.

for a quarter of the venues in those neighborhoods. Other similar neighborhoods are shown in Figure 2.

Conversely, if the 'Total Wt' score is high, it means that most (or all) of the venues fall under the Top 5 categories. The neighborhoods that show such characteristics are illustrated in Figure 3. For example, Moore Park, Summerhill East only has two venue categories, gym and trail, while Lawrence Park only has park, swim school and bus line.

The columns '1st Venue Weight' to '5th Venue Weight' are further normalized so they add up to 1. This normalization is necessary because we will later use these to score the Manhattan neighborhoods on how well they match. We would not want a neighborhood to have poor matching scores due to the diversity of venues.

A neighborhood with a lower 'Total Wt' score typically has a lot of other venues in its vicinity. Intuitively these would be nearer city centres where density is higher. We can determine if this is is the case with a map. In Figure 4, a Toronto map is overlaid with neighborhood information, where a bigger circle indicates a neighborhood that has fewer venue categories and is likely to be sparser and a smaller circle indicates a neighboorhood that has more diverse venue categories and is likely to be in a dense area.

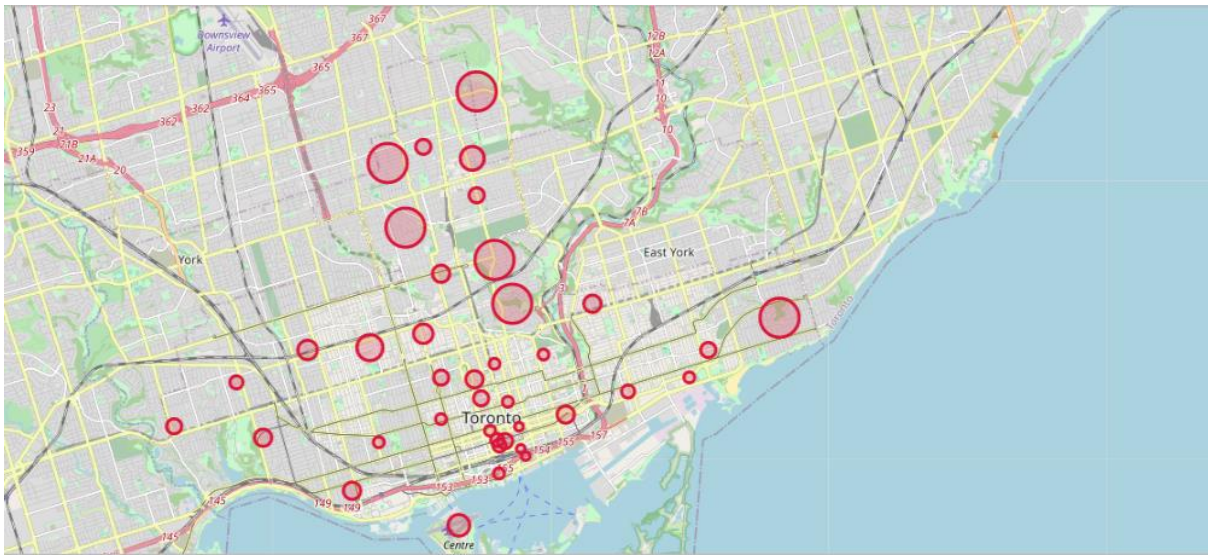

Figure 4: Bubble map of Toronto neighborhoods. Bigger bubbles mean fewer venue categories.

Exploratory Data Analysis—Manhattan Neighborhoods

New York comprises 5 boroughs and 306 neighborhoods, of which 40 fall into the Manhattan borough. The 40 Manhattan neighborhoods will be the focus of this project.

For the Manhattan neighborhoods, the Four Square search instead of explore API is used, as we are interested in the category that correspond to those in Toronto. To reduce the number of API calls, we first identify the unique categories that are present in the top 5 list of all the Toronto neighborhood.

There are a total of 71 unique categories—this means a total of 2840 API calls instead of 7800 API calls if we were to process Toronto neighborhood by neighborhood.

	Manhattan Neighborhood	Café	Coffee Shop	Clothing Store	Yoga Studio	Bar	Thai Restaurant	Sandwich Place	Park	Greek Restaurant	...	Baby Store	Gay Bar	Scenic Lookout	Burrito Place	Men's Store	Salad Place	Pet Store	Dog Run	Indian Restaurant	Nightclub
0	Marble Hill	3	6	16	1	3	1	7	3	0	...	3	0	2	0	0	0	3	3	1	0
1	Chinatown	49	49	48	4	50	16	18	18	5	...	1	1	7	2	21	0	15	3	6	28
2	Washington Heights	9	11	37	1	13	1	5	5	0	...	2	0	5	0	6	2	8	1	2	2
3	Inwood	8	7	17	2	10	2	4	7	0	...	0	0	3	0	3	2	5	1	0	6
4	Hamilton Heights	9	14	14	5	12	4	10	8	1	...	0	2	2	0	1	0	2	1	2	1
5	Manhattanville	8	8	7	3	7	2	4	7	0	...	0	0	2	0	0	1	0	1	2	4
6	Central Harlem	4	6	9	0	13	0	5	2	0	...	0	1	0	0	2	0	0	3	1	0
7	East Harlem	10	8	19	3	8	5	10	5	1	...	0	1	6	0	2	0	6	3	3	0
8	Upper East Side	27	34	48	11	17	6	6	9	0	...	3	0	10	3	5	5	6	4	5	3
9	Yorkville	15	21	15	5	37	11	11	8	0	...	1	1	12	0	0	0	17	3	7	7

Figure 5: Number of venues in the 71 categories in Manhattan neighborhoods

Figure 5 shows a sample of the results obtained from the search. As an example, within Marble Hill there are 6 coffee shops, 3 cafes, and 3 bars, whereas in Chinatown there are about 50 of each of those. So, if a client likes to get a cup of coffee in the morning and end the day with a stronger drink, he will most likely find something he likes in Chinatown rather than Marble Hill.

	Café	Coffee Shop	Clothing Store	Yoga Studio	Bar	Thai Restaurant	Sandwich Place	Park	Greek Restaurant	Pub	...	Baby Store	Gay Bar	Scenic Lookout	Burrito Place	Men's Store	Salad Place	Pet Store	Dog Run	Indian Restaurant	Nightclub
count	40.000000	40.000000	40.000000	40.00000	40.00000	40.000000	40.000000	40.000000	40.000000	40.000000	...	40.000000	40.000000	40.000000	40.00000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000
mean	28.925000	31.750000	35.725000	9.77500	30.12500	7.300000	17.725000	15.750000	2.900000	6.525000	...	1.125000	3.425000	11.425000	2.10000	12.975000	5.475000	7.600000	4.025000	8.600000	15.100000
std	17.375768	16.760378	15.980738	10.77149	16.99802	6.801961	13.739308	8.799038	2.667948	6.694765	...	1.435583	4.706174	8.214707	2.19323	15.967094	8.161786	5.067746	2.506274	8.613645	15.475042
min	0.000000	3.000000	0.000000	0.00000	0.00000	0.000000	2.000000	2.000000	0.000000	0.000000	...	0.000000	0.000000	0.000000	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	13.000000	14.750000	20.500000	2.75000	13.00000	1.750000	7.000000	8.000000	0.000000	1.750000	...	0.000000	0.000000	5.750000	0.00000	1.000000	0.750000	3.000000	2.000000	2.000000	3.750000
50%	29.000000	33.500000	45.000000	5.50000	32.00000	6.000000	15.000000	15.000000	2.000000	4.000000	...	0.500000	1.500000	10.000000	2.00000	6.000000	3.000000	7.000000	4.000000	6.000000	9.000000
75%	47.250000	48.000000	47.250000	11.25000	50.00000	10.000000	23.250000	21.250000	5.000000	9.250000	...	2.000000	5.000000	16.250000	3.00000	21.750000	6.250000	12.000000	5.250000	10.000000	28.500000
max	50.000000	50.000000	50.000000	48.00000	50.00000	27.000000	50.000000	35.000000	9.000000	26.000000	...	5.000000	19.000000	32.000000	9.00000	50.000000	38.000000	17.000000	10.000000	34.000000	47.000000

Figure 6: Statistic Overview of the Venue Categories in Manhattan Neighborhoods

Data Model

Figure 6 provides a statistic overview of the results. As can be seen from the figure, the maximum number of venues in the neighborhood can vary greatly across differently categories. For example, the maximum number of venues for Swim School and Dessert Shop are 2 and 50 respectively. We need to normalize these numbers so that no one category would overwhelm the score. A positive normalized number is preferred so that the eventual score would be positive, thus min-max scaling is used instead of a standard scaler.

The normalized score is given by

$$\widetilde{N} = \frac{N - N_{min}}{N_{max} - N_{min}}$$

where N is the number of venues in the category in the neighborhood, N_{min} and N_{max} are the minimum and maximum number of venues in the category respectively.

The score thus ranges from 0 to 1 linearly. With customer experience and feedback, this scaling can be further optimized. For example, while the difference of having 1 coffee shop and 10 coffee shops may make a big difference, having 50 compared to 20 may not. Figure 7 shows the normalized score.

	Manhattan Neighborhood	Café	Coffee Shop	Clothing Store	Yoga Studio	Bar	Thai Restaurant	Sandwich Place	Park	Greek Restaurant	...	Baby Store	Gay Bar	Scenic Lookout	Burrito Place	Men's Store	Salad Place	Pet Store	Dog Run	Indian Restaurant	Nightclub
0	Marble Hill	0.06	0.063830	0.32	0.020833	0.06	0.037037	0.104167	0.030303	0.000000	...	0.6	0.000000	0.06250	0.000000	0.00	0.000000	0.176471	0.3	0.029412	0.000000
1	Chinatown	0.98	0.978723	0.96	0.083333	1.00	0.592593	0.333333	0.484848	0.555556	...	0.2	0.052632	0.21875	0.222222	0.42	0.000000	0.882353	0.3	0.176471	0.595745
2	Washington Heights	0.18	0.170213	0.74	0.020833	0.26	0.037037	0.062500	0.090909	0.000000	...	0.4	0.000000	0.15625	0.000000	0.12	0.052632	0.470588	0.1	0.058824	0.042553
3	Inwood	0.16	0.085106	0.34	0.041667	0.20	0.074074	0.041667	0.151515	0.000000	...	0.0	0.000000	0.09375	0.000000	0.06	0.052632	0.294118	0.1	0.000000	0.127660
4	Hamilton Heights	0.18	0.234043	0.28	0.104167	0.24	0.148148	0.166667	0.181818	0.111111	...	0.0	0.105263	0.06250	0.000000	0.02	0.000000	0.117647	0.1	0.058824	0.021277

Figure 7: Normalized score for different venue categories in various Manhattan neighborhoods

Recommendation System

The recommendation score between the neighborhoods across the 2 cities can be derived. How well a Manhattan neighborhood m matches a Toronto neighborhood t is given by

$$S_{m,t} = \sum_{i=1}^{5} \tilde{N}_{m,i} \times W_{t,i}$$

where $\tilde{N}m,i$ is the normalized score of venue category i in neighborhood m, Wt,i is the weight of category i in Toronto neighborhood t.

Results

Figure 8 shows the statistics about the score of Manhattan neighborhoods for each Toronto neighborhood.

	St. James Town	Berczy Park	Stn A PO Boxes	Garden District, Ryerson	St. James Town, Cabbagetown	Church and Wellesley	Richmond, Adelaide, King	Kensington Market, Chinatown, Grange Park	Harbourfront East, Union Station, Toronto Islands	Business reply mail Processing Centre, South Central Letter Processing Plant Toronto	Little Portugal, Trinity	First Canadian Place, Underground city	Toronto Dominion Centre, Design Exchange
count	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000
mean	0.516720	0.450157	0.489284	0.544259	0.441317	0.425036	0.536833	0.441330	0.453077	0.228229	0.445943	0.535344	0.492545
std	0.322717	0.287701	0.295978	0.277428	0.268894	0.277667	0.318203	0.270373	0.267347	0.140817	0.269655	0.319190	0.306030
min	0.022654	0.006211	0.015435	0.016849	0.000000	0.010021	0.041473	0.017647	0.023491	0.000000	0.000000	0.038117	0.028560
25%	0.205385	0.169978	0.201258	0.276036	0.208079	0.164605	0.250582	0.217140	0.205808	0.112500	0.188064	0.256897	0.232923
50%	0.482120	0.376558	0.441772	0.612425	0.421005	0.402246	0.538225	0.406612	0.434832	0.234028	0.446883	0.537877	0.463388
75%	0.825258	0.688776	0.762926	0.786125	0.679695	0.659434	0.858806	0.699255	0.710632	0.307986	0.658032	0.857430	0.786003
max	0.986819	0.912799	0.905833	0.902647	0.910474	0.890927	0.987770	0.884447	0.799440	0.562500	0.956044	0.986556	0.989565

	Studio District	High Park, The Junction South	Commerce Court, Victoria Hotel	North Toronto West, Lawrence Park	University of Toronto, Harbord	Davisville	India Bazaar, The Beaches West	Runnymede, Swansea	Central Bay Street	Brockton, Parkdale Village, Exhibition Place	The Danforth West, Riverdale	Regent Park, Harbourfront	Queen's Park, Ontario Provincial Government
count	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000
mean	0.506852	0.370540	0.523986	0.468072	0.510296	0.427822	0.388579	0.457734	0.511906	0.490582	0.378781	0.463235	0.522445
std	0.307890	0.230542	0.320641	0.241961	0.312537	0.280418	0.230838	0.286375	0.317677	0.293219	0.264773	0.275617	0.297054
min	0.018462	0.015000	0.026787	0.002915	0.032843	0.019640	0.045960	0.012857	0.009600	0.018000	0.020300	0.028708	0.012987
25%	0.206766	0.170970	0.236958	0.262249	0.196219	0.212546	0.230068	0.184490	0.214676	0.237055	0.107570	0.206889	0.238894
50%	0.507829	0.326776	0.498355	0.510122	0.517017	0.398464	0.326364	0.480114	0.490922	0.474239	0.367360	0.450293	0.554550
75%	0.801582	0.579918	0.829525	0.653412	0.825763	0.692597	0.536661	0.686215	0.814979	0.735627	0.643565	0.696833	0.802195
max	0.987567	0.912619	0.990510	0.875009	0.940500	0.922153	0.898160	0.903878	0.955029	0.926779	0.826190	0.917755	0.926424

	Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park	Parkdale, Roncesvalles	The Annex, North Midtown, Yorkville	Dufferin, Dovercourt Village	CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	Davisville North	Christie	Rosedale	Roselawn	Forest Hill North & West, Forest Hill Road Park	The Beaches	Moore Park, Summerhill East	Lawrence Park
count	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000	40.000000
mean	0.408720	0.291955	0.478565	0.414356	0.175300	0.487487	0.370576	0.383958	0.262192	0.348865	0.235865	0.413138	0.326566
std	0.202194	0.220131	0.290713	0.223605	0.152649	0.253583	0.211775	0.201305	0.203451	0.224835	0.188289	0.240927	0.170433
min	0.020408	0.028571	0.016364	0.032070	0.000000	0.052611	0.056843	0.045455	0.010000	0.005319	0.009615	0.000000	0.060606
25%	0.237587	0.105464	0.189576	0.243290	0.069409	0.272843	0.215649	0.266477	0.127267	0.148826	0.092788	0.231915	0.178272
50%	0.421326	0.217945	0.464587	0.400791	0.159362	0.503543	0.326868	0.356818	0.182326	0.298862	0.190385	0.404255	0.317645
75%	0.537141	0.478853	0.705476	0.524054	0.218227	0.678416	0.525056	0.460606	0.327965	0.515749	0.351442	0.589362	0.435203
max	0.836735	0.796160	0.981283	0.929543	0.762222	0.896401	0.849663	0.847348	0.906977	0.818529	0.721154	0.968085	0.766495

Figure 8: Distribution of matching scores for different Toronto neighborhoods.

Toronto Neighborhood	1st Matching Neighborhood	1st Matching Score	2nd Matching Neighborhood	2nd Matching Score	3rd Matching Neighborhood	3rd Matching Score	4th Matching Neighborhood	4th Matching Score
Commerce Court, Victoria Hotel	Midtown South	0.99051	Midtown	0.988213	Financial District	0.937204	Murray Hill	0.929568
Toronto Dominion Centre, Design Exchange	Midtown	0.989565	Midtown South	0.96102	Murray Hill	0.917313	Financial District	0.890235
Richmond, Adelaide, King	Midtown South	0.98777	Midtown	0.98027	Financial District	0.939819	Murray Hill	0.937939
Studio District	Midtown South	0.987567	Soho	0.95375	Greenwich Village	0.946224	Midtown	0.916797
St. James Town	Midtown	0.986819	Midtown South	0.982041	Noho	0.945272	Flatiron	0.941845
First Canadian Place, Underground city	Midtown South	0.986556	Midtown	0.982103	Soho	0.940827	Financial District	0.939222
The Annex, North Midtown, Yorkville	Midtown	0.981283	Midtown South	0.951889	Financial District	0.914043	Murray Hill	0.890407
Moore Park, Summerhill East	Financial District	0.968085	Greenwich Village	0.878723	Flatiron	0.689362	Civic Center	0.678723
Little Portugal, Trinity	Little Italy	0.956044	Noho	0.859302	Greenwich Village	0.853575	Chinatown	0.84031
Central Bay Street	Financial District	0.955029	Midtown	0.934248	Noho	0.910914	Murray Hill	0.892248
University of Toronto, Harbord	Noho	0.9405	Midtown	0.928549	Financial District	0.915515	Greenwich Village	0.908763
Dufferin, Dovercourt Village	Chinatown	0.929543	Little Italy	0.924016	Soho	0.880425	Midtown South	0.740343
Brockton, Parkdale Village, Exhibition Place	Soho	0.926779	Flatiron	0.926382	Midtown South	0.913448	Little Italy	0.882159
Queen's Park, Ontario Provincial Government	Murray Hill	0.926424	Midtown South	0.918561	Midtown	0.87269	East Village	0.843046
Davisville	Midtown	0.922153	Financial District	0.891599	Midtown South	0.845475	Noho	0.832997
Regent Park, Harbourfront	Midtown	0.917755	Midtown South	0.915438	Greenwich Village	0.825821	Financial District	0.8157
Berczy Park	Midtown South	0.912799	Midtown	0.901464	Soho	0.886941	Little Italy	0.880731
High Park, The Junction South	Noho	0.912619	Flatiron	0.7303	Midtown	0.706916	East Village	0.69966
St. James Town, Cabbagetown	Midtown	0.910474	Noho	0.889761	Midtown South	0.846599	Financial District	0.82633
Roselawn	Noho	0.906977	East Village	0.7	Greenwich Village	0.634419	Soho	0.579535
Stn A PO Boxes	Midtown South	0.905833	Little Italy	0.900563	Midtown	0.898656	Soho	0.885562
Runnymede, Swansea	Midtown South	0.903878	Midtown	0.887886	Noho	0.874736	Financial District	0.84046
Garden District, Ryerson	Little Italy	0.902647	Soho	0.890885	Midtown	0.867867	Chinatown	0.863558
India Bazaar, The Beaches West	Financial District	0.89816	Midtown	0.883838	Midtown South	0.842797	Civic Center	0.723314
Davisville North	Financial District	0.896401	Midtown South	0.876656	Clinton	0.874356	Murray Hill	0.8505
Church and Wellesley	Midtown South	0.890927	Noho	0.847466	Midtown	0.812563	Murray Hill	0.784512
Kensington Market, Chinatown, Grange Park	Little Italy	0.884447	Chinatown	0.859122	Soho	0.840324	Noho	0.835598
North Toronto West, Lawrence Park	Soho	0.875009	Little Italy	0.843295	Flatiron	0.819634	Midtown South	0.795576
Christie	Soho	0.849663	Little Italy	0.825332	Chinatown	0.791404	Civic Center	0.641884
Rosedale	Financial District	0.847348	Greenwich Village	0.831439	West Village	0.809848	Flatiron	0.749242
Summerhill West, Rathnelly, South Hill, Forest Hill SE, Deer Park	Midtown South	0.836735	Noho	0.756476	Financial District	0.748332	Murray Hill	0.72763
The Danforth West, Riverdale	Midtown South	0.82619	Flatiron	0.782951	Midtown	0.713159	Greenwich Village	0.701335
Forest Hill North & West, Forest Hill Road Park	Financial District	0.818529	West Village	0.810059	Greenwich Village	0.752739	Midtown	0.68055
Harbourfront East, Union Station, Toronto Islands	Midtown South	0.79944	Soho	0.794083	Noho	0.790302	Murray Hill	0.778707
Parkdale, Roncesvalles	Noho	0.79616	Midtown South	0.741691	Midtown	0.685956	Murray Hill	0.612486
Lawrence Park	Financial District	0.766495	Civic Center	0.624113	Manhattan Valley	0.605416	Flatiron	0.601977
CN Tower, King and Spadina, Railway Lands, Harbourfront West, Bathurst Quay, South Niagara, Island airport	Midtown South	0.762222	Soho	0.44208	Midtown	0.428889	Murray Hill	0.406667
The Beaches	West Village	0.721154	Greenwich Village	0.688462	Midtown South	0.55	Gramercy	0.546154
Business reply mail Processing Centre, South Central Letter Processing Plant Toronto	Midtown	0.5625	Midtown South	0.493056	Flatiron	0.477778	Noho	0.4

Figure 9: Top 4 matching Manhattan neighborhoods for each Toronto neighborhood

The top 4 matching Manhattan neighborhoods along with the score are listed in Figure 9.

To further analyse a few samples of the neighborhood, consider the following neighborhoods: “The Beaches”, “Business reply … Letter Processing Plant Toronto”, “Christe” and “Commerce Court, Victoria Hotel”. Figure ? shows the weightage of the venue categories and the number of venues in the top 4 matching Manhattan neighborhoods.

The Beaches in Toronto has equal parts of trails, neighborhoods, health food stores and pubs. Its matching neighborhoods in Manhattan include West Village and Greenwich Village which are by the Hudson River and have trails and a good number of pubs too. Northeast Gramercy, with Stuy Town, Union Square and Gramercy Park nearby, also offers plenty of trails.

Commerce Court is near coffee shops, restaurants, cafes, hotel, and American restaurants. In Manhattan, areas in Murray Hill, Financial District, Midtown and Midtown South all have plenty of those in the neighborhoods.

	Nhood	Trail	Neighborhood	Health Food Store	Pub
0	Weightage	0.25	0.25	0.25	0.25
1	West Village	4.00	2.00	7.00	10.00
2	Greenwich Village	4.00	1.00	8.00	17.00
3	Midtown South	1.00	0.00	10.00	26.00
4	Gramercy	5.00	1.00	3.00	10.00

The Beaches

	Nhood	Coffee Shop	Restaurant	Café	Hotel	American Restaurant
0	Weightage	0.333333	0.194444	0.194444	0.166667	0.111111
1	Midtown South	50.000000	48.000000	49.000000	49.000000	49.000000
2	Midtown	50.000000	46.000000	50.000000	49.000000	50.000000
3	Financial District	50.000000	40.000000	48.000000	45.000000	48.000000
4	Murray Hill	50.000000	37.000000	47.000000	47.000000	49.000000

Commerce Court, Victoria Hotel

	Nhood	Yoga Studio	Skate Park	Auto Workshop	Brewery	Burrito Place
0	Weightage	0.2	0.2	0.2	0.2	0.2
1	Midtown	15.0	0.0	2.0	3.0	9.0
2	Midtown South	37.0	1.0	0.0	6.0	4.0
3	Flatiron	48.0	0.0	0.0	3.0	8.0
4	Noho	24.0	0.0	1.0	4.0	3.0

Business reply mail Processing Centre, South Central Letter Processing Plant Toronto

	Nhood	Grocery Store	Café	Park	Baby Store	Nightclub
0	Weightage	0.363636	0.272727	0.181818	0.090909	0.090909
1	Soho	41.000000	50.000000	22.000000	5.000000	37.000000
2	Little Italy	48.000000	50.000000	13.000000	3.000000	42.000000
3	Chinatown	49.000000	49.000000	18.000000	1.000000	28.000000
4	Civic Center	14.000000	47.000000	35.000000	4.000000	14.000000

Christie

Figure 10: Detailed results of selected neighborhoods

If any client stays in "Business reply … Letter Processing Plant Toronto", he or she would be near Yoga Studios, Skate Park, Auto Workshop and Brewery as well as Burrito Place. Over in Manhattan, there aren't too many of those—one skate park in Midtown South, one Auto workshop in Noho and some breweries and burrito places scattered around. With the matching score, it is easier to narrow the search.

Those who stay in Christie would be near grocery store, cafés, parks, baby stores and night clubs. They would find the best match in Soho which has all these in good numbers, followed by Little Italy, Chinatown and Civic Center.

Possible Future Enhancements

As mentioned earlier in the report, the scaling of the categories in the Manhattan neighborhood can potentially be further optimized with user input. The diversity of the venues could also be used as an additional criterion for improved matching in future. As indicated by the "Total Wt" column of the neighborhood score, a low score indicates a more diverse neighborhood, and the target neighborhood can be scored by using the number of unique categories found within the neighborhood. Finally this could form part of a bigger system where other factors such as pricing, housing types could be taken into consideration.

Conclusion

A recommendation system was developed for clients who are relocating from Toronto to Manhattan, to help them find neighborhoods that match their current (Toronto) neighborhood in terms of venues in the neighborhoods. The system was developed based on public domain neighborhood information and Four Square APIs for the places to go.

Initial results are promising and further possible enhancements are also discussed.