Are neighbourhoods of New york and Toronto similar?

Table of Contents

1. Introduction/Business problem

We know that the cities New York and Toronto are very diverse and are the financial capitals of their respective countries. One interesting idea would be to compare the neighborhoods of the two cities and determine how similar or dissimilar they are. Is New York City more like Toronto? Can we identify how similar are the neighbourhoods of New York to the neighbourhoods of toronto? We are going to compare the similarity of neighbourhoods in the two cities based on the venues present in each neighbourhood.

1.1 Background and Importance

We are going to explore the neighbourhoods of the two cities using *Foursquare API* and apply the *k-means clustering* algorithm for grouping the neighbourhoods of the two cities. By forming the clusters, we can understand how many neighbourhoods of New York are similar to Toronto.

This problem would be helpful to the people who are having trouble in taking a decision of choosing a neighbourhood of the city to take residence or choosing a neighbourhood of the city to have breakfast,lunch or dinner or choosing a neighbourhood of the city to travel as some people choose to travel to cities which are different from the cities they have already visited while some people choose to travel to cities which are similar to the cities they have already visited. In our case, we are going to explore the neighbourhoods of Toronto which are similar to New York.

2. Data Understanding

As the neighbourhoods in New York are very large in number, I have chosen only the neighbourhoods in the borough **Manhattan** for representing New York. We can collect the location

data of the neighbourhoods of New York and Toronto and explore the top 100 venues around each neighbourhood using the Foursquare API. Moreover, we are exploring the top 100 venues around each neighbourhood for both cities because we need some means of comparison between the neighbourhoods for grouping. By using these venues for the neighbourhoods we can apply one hot encoding and build a metric of comparison based on the category types of each venue.

2.1 Data Processing

The packages required for exploring the neighbourhoods are pandas, numpy, matplotlib, folium, sklearn.

Step 1

Create a dataframe which shows the location coordinates of the neighbourhoods.It looks as shown below.

	Borough	Neighborhood	Latitude	Longitude
0	Manhattan	Marble Hill	40.876551	-73.910660
1	Manhattan	Chinatown	40.715618	-73.994279
2	Manhattan	Washington Heights	40.851903	-73.936900
3	Manhattan	Inwood	40.867684	-73.921210
4	Manhattan	Hamilton Heights	40.823604	-73.949688

Step 2

Get the top 100 venues around each neighbourhood location into a new dataframe for both cities.It looks as shown below.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Marble Hill	40.876551	-73.91066	Arturo's	40.874412	-73.910271	Pizza Place
1	Marble Hill	40.876551	-73.91066	Bikram Yoga	40.876844	-73.906204	Yoga Studio
2	Marble Hill	40.876551	-73.91066	Tibbett Diner	40.880404	-73.908937	Diner
3	Marble Hill	40.876551	-73.91066	Starbucks	40.877531	-73.905582	Coffee Shop
4	Marble Hill	40.876551	-73.91066	Dunkin'	40.877136	-73.906666	Donut Shop

Step 3

Perform one hot encoding on the dataframe obtained from step 2 for each neighbourhood based on the category type for both cities into a new dataframe. It consists of all category types as the columns.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	...	Vietnamese Restaurant	Volleyball Court
0	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0
1	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0
2	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0
3	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0
4	Marble Hill	0	0	0	0	0	0	0	0	0	...	0	0

Step 4

Perform the group by neighbourhood operation on the dataframe in Step 3 and perform the mean operation on each group which will give the mean occurrence/frequency of each category in the respective neighbourhood.Perform it for both cities.

	Neighborhood	Accessories Store	Adult Boutique	Afghan Restaurant	African Restaurant	American Restaurant	Antique Shop	Arcade	Arepa Restaurant	Argentinian Restaurant	...	Vietnamese Restaurant	Volleyball Court	Waterfr
0	Battery Park City	0.000000	0.00	0.00	0.000000	0.010526	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
1	Carnegie Hill	0.000000	0.00	0.00	0.000000	0.010101	0.00	0.00	0.000000	0.010101	...	0.020202	0.00	0.000
2	Central Harlem	0.000000	0.00	0.00	0.068182	0.045455	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
3	Chelsea	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
4	Chinatown	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	...	0.030000	0.00	0.000
5	Civic Center	0.000000	0.00	0.00	0.000000	0.030000	0.01	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
6	Clinton	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
7	East Harlem	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
8	East Village	0.000000	0.00	0.00	0.000000	0.020000	0.01	0.00	0.020000	0.010000	...	0.020000	0.00	0.000
9	Financial District	0.010000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
10	Flatiron	0.000000	0.00	0.00	0.000000	0.040000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
11	Gramercy	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.01	0.000000	0.000000	...	0.020000	0.00	0.000
12	Greenwich Village	0.000000	0.00	0.00	0.000000	0.020000	0.00	0.00	0.000000	0.000000	...	0.020000	0.00	0.000
13	Hamilton Heights	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
14	Hudson Yards	0.000000	0.00	0.00	0.000000	0.060241	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
15	Inwood	0.000000	0.00	0.00	0.000000	0.037037	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
16	Lenox Hill	0.000000	0.00	0.01	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
17	Lincoln Square	0.000000	0.00	0.00	0.000000	0.030000	0.00	0.00	0.000000	0.000000	...	0.000000	0.00	0.000
18	Little Italy	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000	...	0.010000	0.00	0.000
19	Lower East Side	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.019608	...	0.019608	0.00	0.000
20	Manhattan	0.000000	0.00	0.00	0.000000	0.000000	0.00	0.00	0.000000	0.000000		0.020833	0.00	0.000

Step 5

Create a new dataframe which shows neighbourhood along with top 10 venues in the respective neighbourhood. Perform it for both cities.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Battery Park City	Park	Coffee Shop	Hotel	Wine Shop	Women's Store	Clothing Store	Gym	Memorial Site	Pizza Place	Grocery Store
1	Carnegie Hill	Coffee Shop	Pizza Place	Cosmetics Shop	Yoga Studio	Bakery	Gym	Bookstore	Café	Japanese Restaurant	Wine Shop
2	Central Harlem	African Restaurant	Chinese Restaurant	American Restaurant	Bar	Cosmetics Shop	Seafood Restaurant	French Restaurant	Fried Chicken Joint	Bookstore	Caribbean Restaurant
3	Chelsea	Coffee Shop	Bakery	Italian Restaurant	Ice Cream Shop	Nightclub	Theater	American Restaurant	Hotel	Tapas Restaurant	Cocktail Bar
4	Chinatown	Chinese Restaurant	Cocktail Bar	American Restaurant	Salon / Barbershop	Bakery	Spa	Optical Shop	Vietnamese Restaurant	Hotpot Restaurant	Sandwich Place

Step 6

Concatenate the two dataframes of two cities obtained from step 4. It simply combines the neighbourhoods of New York and Toronto. Each row shows a category metric of each neighbourhood based on which we perform k-means clustering.

	Neighborhood	Yoga Studio	Afghan Restaurant	Airport	Airport Food Court	Airport Gate	Airport Lounge	Airport Service	Airport Terminal	American Restaurant	...	Turkish Restaurant	Udon Restaurant	U Booksl
0	Adelaide,King,Richmond	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.02	...	0.0	0.0	
1	Berczy Park	0.000000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	...	0.0	0.0	
2	Brockton,Exhibition Place,Parkdale Village	0.083333	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	...	0.0	0.0	
3	Business Reply Mail Processing Centre 969 Eastern	0.055556	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00	...	0.0	0.0	
4	CN Tower,Bathurst Quay,Island airport,Harbourf...	0.000000	0.0	0.058824	0.058824	0.058824	0.117647	0.176471	0.117647	0.00	...	0.0	0.0	

Step 7

Add a City column to the two dataframes obtained from Step 5. Concatenate the two dataframes of two cities obtained from step 5. It simply combines the neighbourhoods of new york and toronto. Each row shows the top 10 venues of the respective neighbourhood.

	Neighborhood	City	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Adelaide,King,Richmond	Toronto	Coffee Shop	Café	Bar	Hotel	Restaurant	Steakhouse	Burger Joint	Cosmetics Shop	Asian Restaurant	Thai Restaurant
1	Berczy Park	Toronto	Coffee Shop	Cocktail Bar	Cheese Shop	Seafood Restaurant	Farmers Market	Beer Bar	Bakery	Steakhouse	Café	Gourmet Shop
2	Brockton,Exhibition Place,Parkdale Village	Toronto	Café	Yoga Studio	Breakfast Spot	Coffee Shop	Gym	Pet Store	Performing Arts Venue	Italian Restaurant	Intersection	Gym / Fitness Center
3	Business Reply Mail Processing Centre 969 Eastern	Toronto	Light Rail Station	Yoga Studio	Garden	Butcher	Fast Food Restaurant	Auto Workshop	Farmers Market	Burrito Place	Spa	Pizza Place
4	CN Tower,Bathurst Quay,Island airport,Harbourf...	Toronto	Airport Service	Airport Terminal	Airport Lounge	Sculpture Garden	Plane	Coffee Shop	Boat or Ferry	Boutique	Harbor / Marina	Airport Gate

Now, we can use the dataframe obtained from step 6 for k-means clustering which has been done in the methodology section.

3. Methodology

We have collected the location data of the neighbourhoods of New York and Toronto and explored the top 100 venues around each neighbourhood using the **Foursquare API**. Moreover, we are exploring the top 100 venues around each neighbourhood for both cities because we need some means of comparison between the neighbourhoods for grouping. By using these venues for the neighbourhoods we can apply one hot encoding and build a metric of comparison based on the category types of each venue.

After completing one-hot encoding based on venue category we have grouped the dataframe with neighbourhood and performed the mean operation which gave the mean occurrence of the category in the respective neighbourhood. It has been done for both the cities. Later we have merged the dataframes of the two cities.

In brief, the information related to the final dataframe and the methodology we are going to follow is-

1. The dataframe has 79 rows and 380 columns. Rows represent neighbourhood and columns represent different venue categories.

2. Out of the 79 neighbourhoods, 39 of them correspond to Toronto and the rest of them correspond to Manhattan(New York).

3. Now we have merged both cities neighbourhoods.

4. The metric which we are going to use for comparing different neighbourhoods is the venue category metric along the neighbourhood's row.

5. We can find out what are the neighbourhoods in New York(Manhattan) which are similar to the neighbourhoods in Toronto and vice versa.

6. The similarity can be calculated using the category metric of each neighbourhood.

7. The machine learning algorithm which is best suitable for our situation is k-means clustering.

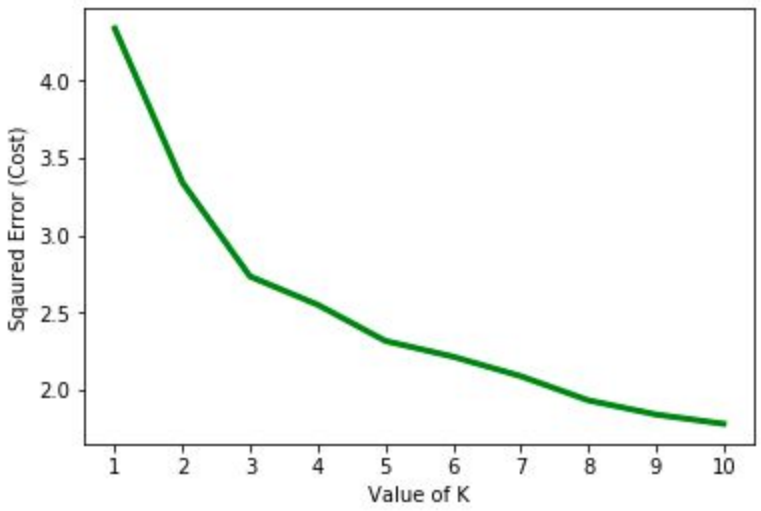
8. We are going to apply the k-means clustering using our final dataframe as the input data and group all the neighbourhoods.

9. What is the best value of k we need to use for the k-means clustering? We are going to decide it using elbow method.

10. By choosing the optimal value of k, we train the model and visualize all the neighbourhoods grouped based on cluster labels on the folium Map object.

11. We can list out all the neighbourhoods in each cluster for a clear picture(Results section).

Get the optimal value of k using the elbow method



The optimal value of k is 5

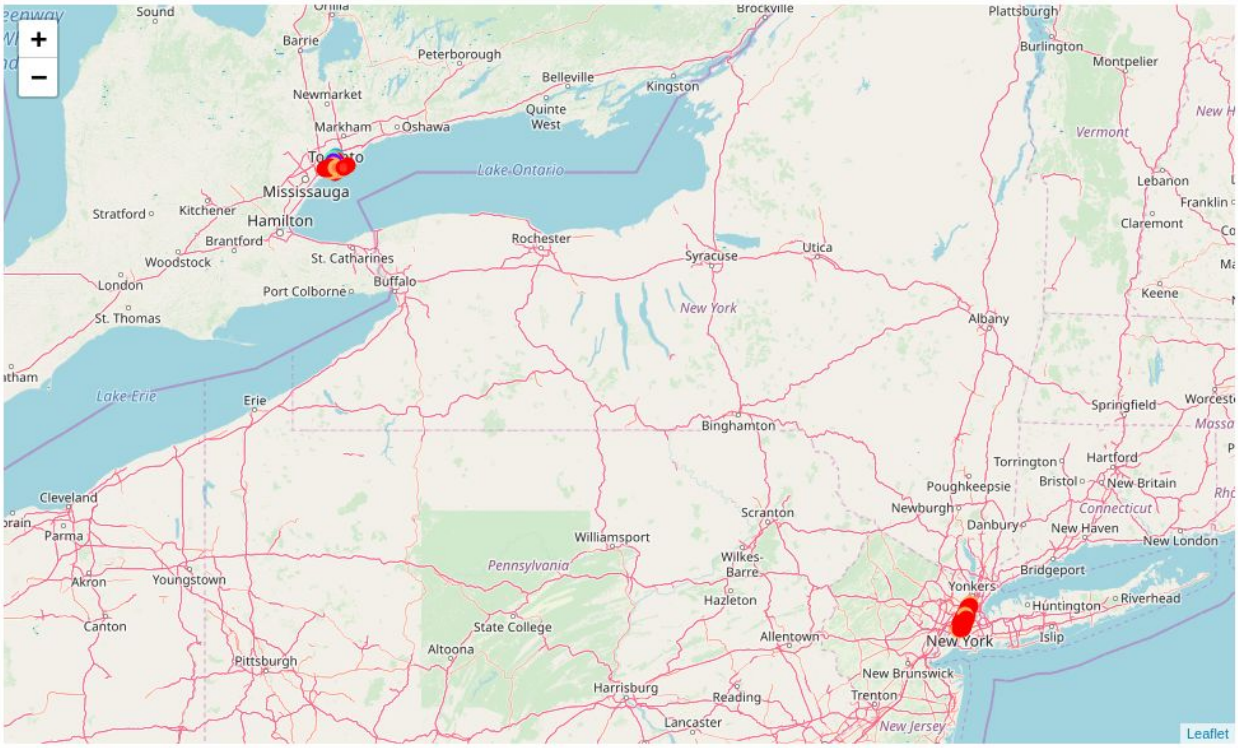
We can apply the k-means clustering model by using sklearn package. After grouping them into clusters we can visualise them on a folium map using the neighbourhoods coordinates.

The counts of neighbourhoods of each city distributed along each cluster have been shown below. As we can we see, cluster 1 and cluster 5 contain neighbourhoods from both the cities whereas cluster 2,3 and 4 contain neighbourhoods only from Toronto.

Cluster analysis

	New York	Toronto	Total neighbourhoods
Cluster 1	35	14	49
Cluster 2	0	3	3
Cluster 3	0	1	1
Cluster 4	0	1	1
Cluster 5	5	20	25
Aggregate	40	39	79

Visualize the cluster neighbourhoods on a folium map



4. Results

We are going to list out the neighbourhoods in each cluster.

Cluster 1

We can observe that there are neighbourhoods corresponding to Toronto and New York. There are 49 neighbourhoods in cluster 1, 14 of them correspond to Toronto and the rest to New York.

Toronto:

1.High Park,The Junction South
2.Davisville
3.Runnymede,Swansea
4.Chinatown,Grange Park,Kensington Market
5.Little Portugal,Trinity
6.Studio District
7.Harbord,University of Toronto
8.The Beaches West,India Bazaar
9.Dovercourt Village,Dufferin
10.CN Tower,Bathurst Quay,Island airport,Harbourfront West,King and Spadina,Railway Lands,South Niagara
11.Parkdale,Roncesvalles
12.Business Reply Mail Processing Centre 969 Eastern
13.The Danforth West,Riverdale
14.The Beaches

New York:

1.Turtle Bay
2.Hudson Yards
3.Upper East Side
4.Flatiron
5.West Village
6.Manhattanville
7.Little Italy
8.East Harlem
9.Inwood
10.Greenwich Village
11.Sutton Place
12.Gramercy
13.Hamilton Heights

14.Central Harlem
15.Tudor City
16.Noho
17.Stuyvesant Town
18.Chelsea
19.Tribeca
20.Chinatown
21.Lower East Side
22.Midtown
23.Clinton
24.Upper West Side
25.Washington Heights
26.Carnegie Hill
27.Civic Center
28.Murray Hill
29.Lenox Hill
30.Lincoln Square
31.Yorkville
32.Midtown South
33.East Village
34.Manhattan Valley
35.Soho

Cluster 2

We can observe that there are neighbourhoods corresponding to Toronto but not New York. There are 3 neighbourhoods of Toronto which have been listed out.

1.Rosedale
2.Moore Park,Summerhill East
3.Forest Hill North,Forest Hill West

Cluster 3

We can observe that there are only neighbourhoods corresponding to Toronto but not New York. There is only one neighbourhood.

1.Lawrence Park

Cluster 4

We can observe that there are only neighbourhoods corresponding to Toronto but not New York. There is only one neighbourhood.

1.Roselawn

Cluster 5

We can observe that there are neighbourhoods corresponding to Toronto and New York. There are 25 neighbourhoods in cluster 5, 20 of them correspond to Toronto and the rest to New York.

Toronto:

1.Christie
2.Berczy Park
3.Queen's Park
4.Harbourfront East,Toronto Islands,Union Station
5.Deer Park,Forest Hill SE,Rathnelly,South Hill,Summerhill West
6.Adelaide,King,Richmond
7.Harbourfront
8.Commerce Court,Victoria Hotel
9.Church and Wellesley
10.Brockton,Exhibition Place,Parkdale Village
11.St. James Town
12.North Toronto West
13.Central Bay Street
14.The Annex,North Midtown,Yorkville
15.Davisville North
16.Cabbagetown,St. James Town
17.Design Exchange,Toronto Dominion Centre
18.First Canadian Place,Underground city
19.Ryerson,Garden District
20.Stn A PO Boxes 25 The Esplanade

New York:

1.Financial District
2.Morningside Heights
3.Marble Hill
4.Roosevelt Island
5.Battery Park City

5. Discussion

We can observe that cluster 2, cluster 3 and cluster 4 consists of neighbourhoods corresponding to Toronto.There are no neighbourhoods corresponding to New York. The only clusters which consists of neighbourhoods of both the cities are cluster 1 and cluster 5.Hence, the observations/suggestions which we can make are-

1. There are 14 neighbourhoods in Toronto which are similar to 35 neighbourhoods in Manhattan(New York) in cluster 1.
2. There are 20 neighbourhoods in Toronto which are similar to 5 neighbourhoods in Manhattan(New York) in cluster 5.
3. There is no similarity between the other 5 neighbourhoods of Toronto present in cluster 2,cluster 3,cluster 4 and the neighbourhoods of Manhattan(New York).
4. Hence, we can recommend the above observations we have made to the people who are looking for neighbourhoods of Toronto which are similar to New York or vice versa.

6. Conclusion

We can make the below conclusions from the clustering analysis-

1. There are 14 neighbourhoods in Toronto which are similar to 35 neighbourhoods in New York.
2. There are 20 neighbourhoods in Toronto which are similar to 5 neighbourhoods in New York.

Hence, the below 14 neighbourhoods of Toronto are similar to the below mentioned 35 neighbourhoods of New York.

Toronto:

1. High Park,The Junction South
2. Davisville
3. Runnymede,Swansea
4. Chinatown,Grange Park,Kensington Market
5. Little Portugal,Trinity
6. Studio District
7. Harbord,University of Toronto
8. The Beaches West,India Bazaar
9. Dovercourt Village,Dufferin
10. CN Tower,Bathurst Quay,Island airport,Harbourfront West,King and Spadina,Railway Lands,South Niagara
11. Parkdale,Roncesvalles
12. Business Reply Mail Processing Centre 969 Eastern

13. The Danforth West,Riverdale
14. The Beaches

New York:

1. Turtle Bay
2. Hudson Yards
3. Upper East Side
4. Flatiron
5. West Village
6. Manhattanville
7. Little Italy
8. East Harlem
9. Inwood
10. Greenwich Village
11. Sutton Place
12. Gramercy
13. Hamilton Heights
14. Central Harlem
15. Tudor City
16. Noho
17. Stuyvesant Town
18. Chelsea
19. Tribeca
20. Chinatown
21. Lower East Side
22. Midtown
23. Clinton
24. Upper West Side
25. Washington Heights
26. Carnegie Hill
27. Civic Center
28. Murray Hill
29. Lenox Hill
30. Lincoln Square
31. Yorkville
32. Midtown South
33. East Village
34. Manhattan Valley
35. Soho

Hence, the below 20 neighbourhoods of Toronto are similar to the below mentioned 5 neighbourhoods of New York.

Toronto:

1. Christie
2. Berczy Park
3. Queen's Park
4. Harbourfront East,Toronto Islands,Union Station
5. Deer Park,Forest Hill SE,Rathnelly,South Hill,Summerhill West
6. Adelaide,King,Richmond
7. Harbourfront
8. Commerce Court,Victoria Hotel
9. Church and Wellesley
10. Brockton,Exhibition Place,Parkdale Village
11. St. James Town
12. North Toronto West
13. Central Bay Street
14. The Annex,North Midtown,Yorkville
15. Davisville North
16. Cabbagetown,St. James Town
17. Design Exchange,Toronto Dominion Centre
18. First Canadian Place,Underground city
19. Ryerson,Garden District
20. Stn A PO Boxes 25 The Esplanade

New York:

1. Financial District
2. Morningside Heights
3. Marble Hill
4. Roosevelt Island
5. Battery Park City