Comparison between neighborhoods of Toronto and New York

Tarini Ghosh

12-May-2019

INTRODUCTION: PROBLEM OVERVIEW

The assignment problem can be broadly described in the following points below.

1. Different cities have specific neighborhoods, which have specific characteristics/facilities/infrastructure, in terms of restaurants, hospitals, parks, etc. So, the question that remains to be answered is, given two cities and their neighborhood profiles, can we identify neighborhoods across the two cities that are similar in terms of the above characteristics? Identifying this will be important for individuals who wish to move from one city to the other, who in turn, would be looking for similar neighborhoods to settle in. For people visiting one city from the one from the other, who are specifically looking for neighborhoods similar to the ones that they like in their home town, such an analysis would provide the clues. For this we would need to identify the neighborhoods across the two cities that are similar.
2. The second question is can we identify the exact differences between the neighborhoods, in terms of which characteristics are specifically high or lower in numbers/frequencies in one group of neighborhoods as compared to the others? This will be important as it will help individuals to look for neighborhood based on their choices in terms of hospitals, restaurants, location, etc. This will also help people who are willing to start a business that targets specific neighborhoods or neighborhoods that lack specific infra in which the individual is interested to start a business.

DATA AND METHODS OVERVIEW

1. I used the neighborhood data from three different cities, namely New York and Toronto for which I already had the neighborhood data along with postal codes. Once, I had the neighborhood information, I could get the latitude and longitude data.

2. I then used the FourSquare API to get the neighborhood details/venues by categories. For each city, the venue-neighborhood dataframes that are thus obtained was then merged with an additional column containing city names.

3. Multivariate analysis, namely Principal Component analysis, was then performed to identify how similar are the three cities in terms of their neighborhood and if I could get specific inter-city clusters of neighborhoods.

4. Thus, using this, I could group neighborhoods based on the venue profiles.

5. For this I used k-means. I could identify three clusters, where in each cluster, I could see representations from both cities, but with distinct representation patterns.

RESULTS OVERVIEW AND DISCUSSION

1. Obtaining the venue types in each neighborhood from the two cities

Neighborhoods can be compared with each other based on the kinds of venues that they contain. Using Four-Square API, I downloaded the venuetypes for each neighborhood belonging to the two cities. The types of venues were then categorized as 'found in both cities', 'found only in Toronto' and 'Found only in New York'.

I first focused on those 'found in both cities'. Even within them, there were differences across cities (Table 1).

Table 1: The top 20 venues that are highly prevalent in (a) New York and (b) Toronto

(A) New York

	Toronto	New York	LogFoldChange
['Bus Station']	0.001947	0.027586	-2.650811
['Mobile Phone Shop']	0.000974	0.010345	-2.363129
['Martial Arts Dojo']	0.000974	0.006897	-1.957663
['Shoe Store']	0.002921	0.013793	-1.552198
['Supermarket']	0.005842	0.025862	-1.487660
['Deli / Bodega']	0.009737	0.036207	-1.313306
['Supplement Shop']	0.001947	0.006897	-1.264516
['Baseball Field']	0.001947	0.006897	-1.264516
['Moving Target']	0.000974	0.003448	-1.264516
['Metro Station']	0.003895	0.012069	-1.130985
['Playground']	0.002921	0.008621	-1.082195
['Market']	0.001947	0.005172	-0.976834
['Latin American Restaurant']	0.005842	0.015517	-0.976834
['Discount Store']	0.008763	0.022414	-0.939094
['Bank']	0.014606	0.032759	-0.807758
['Chinese Restaurant']	0.014606	0.032759	-0.807758
['Food Truck']	0.003895	0.008621	-0.794513
['Mexican Restaurant']	0.009737	0.018966	-0.666679
['Pharmacy']	0.018500	0.034483	-0.622662

(B) Toronto

	Toronto	New York	LogFoldChange
['Cheese Shop']	0.004869	0.001724	1.038069
['Art Gallery']	0.009737	0.003448	1.038069
['Café']	0.029211	0.010345	1.038069
['Indian Restaurant']	0.010711	0.003448	1.133379
['Bagel Shop']	0.005842	0.001724	1.220390
['Breakfast Spot']	0.017527	0.005172	1.220390
['Clothing Store']	0.012658	0.003448	1.300433
['Thai Restaurant']	0.012658	0.003448	1.300433
['Farmers Market']	0.006816	0.001724	1.374541
['Sporting Goods Shop']	0.007790	0.001724	1.508072
['Yoga Studio']	0.008763	0.001724	1.625855
['Greek Restaurant']	0.008763	0.001724	1.625855
['Sushi Restaurant']	0.017527	0.003448	1.625855
['Lounge']	0.008763	0.001724	1.625855
['French Restaurant']	0.009737	0.001724	1.731216
['Beer Bar']	0.009737	0.001724	1.731216
['Steakhouse']	0.011685	0.001724	1.913538
['Dessert Shop']	0.011685	0.001724	1.913538
['Bookstore']	0.014606	0.001724	2.136681
['Japanese Restaurant']	0.021422	0.001724	2.519673

Overall, it revealed certain interesting patterns specifically with respect to the types of restaurants. Deli, Latin American, Chinese and Mexican Restaurants were higher in New York. In contrast, Asian (Thai, Japanese, Indian, Sushi) and European (Greek, French, Bagel shop) were higher in Toronto. While, Baseball field, Food trucks, Supermarkets, Discount store and Banks were higher in New York, Farmers' markets, Art gallery, Yoga studio, Book store were higher in Toronto. Identification of such differences could help individuals interested for setting up businesses in specific cities.

2. Probing which neighborhoods are similar to each other across both cities

Using Multivariate Principal Component Analysis, we can identify similarities among neighborhoods. The objective was to first check if we could group neighborhoods across both cities into clusters. This will be important for someone who wants to move from one city to another and wants to explore/settle in similar neighborhoods. The PCA plot in Figure 1 shows these trends.

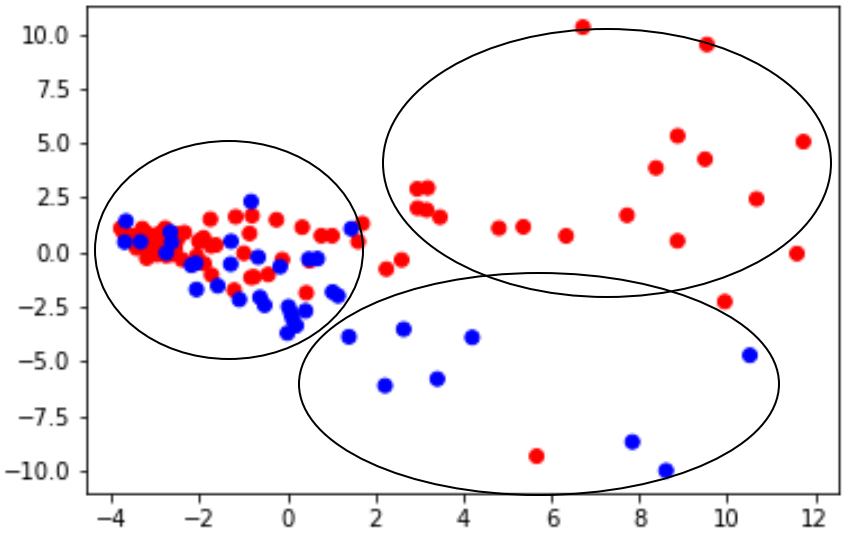

Figure 1: PCA plot of neighborhoods from Toronto (Red) and New York (Blue), based on the prevalence of different types of venues. Three different clusters of neighborhoods could be observed, similar in both cities, specific to Toronto and specific to New York.

Figure 1 shows the overall differences between the neighborhoods of Toronto (Red) and New York (Blue). There are three distinct groups, one covering neighborhoods from both cities, the other with a larger spread and containing only neighborhoods from Toronto, the third outlier group dominated by those from New York. This was also validated using k-means clustering (Figure 2).

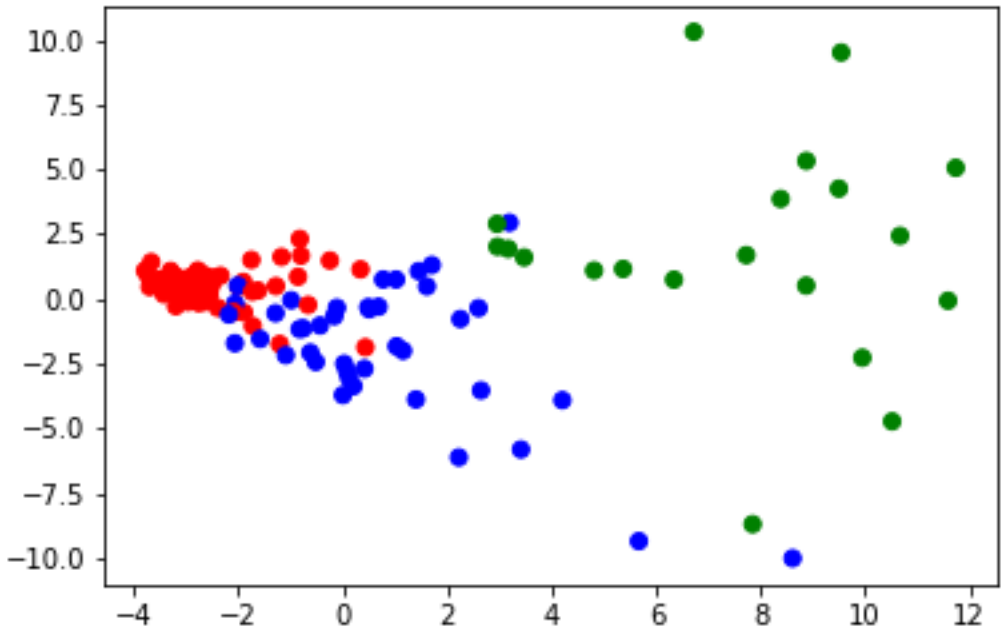

Figure 2: PCA plot of neighborhoods based on the prevalence of different types of venues (with the neighborhoods colored based on the k-means cluster tags). The three clusters are Green Cluster 1: Neighborhoods that are similar across both; Red Cluster 2: Toronto dominated; Blue Cluster 3: New York dominated

The three groups (Cluster1_Neighborhoods, Cluster2_Neighborhoods, Cluster3_Neighborhoods) will be helpful to those shifting from one location to another as well as interested in setting up businesses in similar areas (Figure 2). Please refer to the notebook for the detailed list.

CONCLUSION

The above analysis gives a rough overview of how cities can be compared among each other, to identify:

a. Neighborhoods that are similar (important especially for travelers and people trying to look into similar neighborhoods for moving in).

b. What venues are highly prevalent or less prevalent across cities and venues (important especially for individuals looking to set up business).