Comparison of London, Toronto and New York

Zhengjun Li

November 8, 2019

1 introduction

We have already explored the neighborhoods of New York and Toronto in the previous lab and assignment, one natural question would be how similar they are. Further, we could ask how similar some global cities are including Toronto and New York. We will explore this question in this reported project, and this report will be targeted to people who are curious about and eager to find out the similarities and differences between global cities.

In this project, we will try to determine how similar Toronto, New York and London are by comparing the neighborhoods of these three global cities. Specifically, our goal of this project is to explore the neighborhoods of London and then make comparison of these three cities.

Since we don't know much about the neighborhoods of London very well, We will explore them first before we make direct comparison of these cities. Because of too many neighborhoods in London, we will explore the neighborhoods of the most affluent areas of London, just as we did to Toronto and New York. Also, we are still interested in the mean frequency of occurrence of each venue category near each neighborhood. We will also cluster the neighborhood in London and try to determine the defining categories of each cluster.

After exploration of London, we would import the relevant data about Toronto and New York to make comparison of these three cities. We will first compare the cities based on mean frequency of occurrence of each venue category in each city, and then based on the proportion of each type of clusters in each city. In the end, how similar these cities are according to each comparison will be clearly expressed.

2 Data

Based on the goal of our project, The validity of comparison of cities will be highly affected by whether the neighborhoods we explore are representative. In this project, we decided to use the neighborhoods of the most affulent areas in these cities to make comparison. Thus we will explore the neighborhoods of City of London, Kensington and Chelsea(a borough's name), and Westminster in London and import data concerning the neighborhoods of Manhattan in New York and neighborhoods of Toronto boroughs with 'Toronto' in their names.

Following data source will be needed to extract and generate the required information:

- Every neighborhood of London along with its borough and OS grid reference will be scraped from `https://en.wikipedia.org/wiki/List_of_areas_of_London#Sub-districts`
- Geographical coordinates of London Neighborhoods will be obtained using OsGridConverter.
- Venues near the neighborhoods will be obtained using Foursquare API.
- Information about neighborhoods in New York will be imported from Neighborhood_Analysis_of_New_York, lab notebook of this course to explore the neighborhoods in New York.
- Information about neighborhoods in Toronto will be imported from Neighborhood_Clustering_of_Toronto, notebook of the week 3 assignment of this course to explore the neighborhoods in Toronto.

2.1 Data Acquisition and Processing

Libraries for web scraping and data processing were imported and table containing the information of neighborhoods in London was scraped from `https://en.wikipedia.org/wiki/List_of_areas_of_London#Sub-districts`, each neighborhoods' names, boroughs, and Os grid reference will be transformed into pandas dataframe. Then the latitude and longitude of each neighborhood was obtained based on its OS grid reference using package OSGridConverter and added to our dataframe.

2.2 Neighborhood Analysis of London

The neighborhoods of City of London, Kensington and Chelsea, and Westminster were extracted from our dataframe into a new dataframe, then these neighborhoods were shown on map of London to visualize them as shown in figure 1.

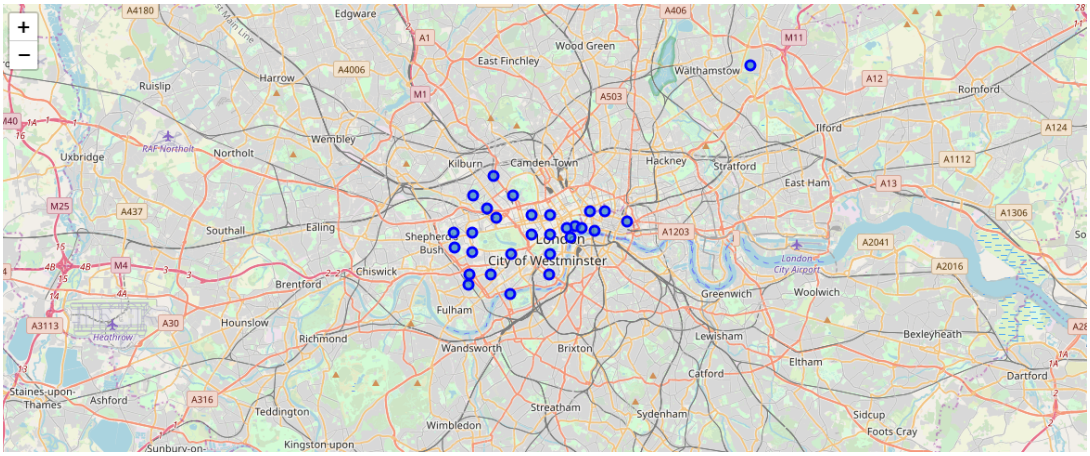

Figure 1: the neighborhoods of London

In the second step of neighborhoods exploration, nearby venues of these neighborhoods were gained through Foursquare API.

In the third step, one hot encoding was applied to these venues. On top of that, top 5 venue categories of each neighborhood were printed and each neighborhood with its top 10 venue categories were collected into a dataframe. The new data frame is shown in figure 2.

	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	Aldgate	Coffee Shop	Hotel	Pub	Salad Place	Cocktail Bar	Indian Restaurant	Thai Restaurant	Italian Restaurant	Japanese Restaurant	Sushi Restaurant
1	Aldwych	Theater	Coffee Shop	Hotel	Dessert Shop	Tea Room	Burger Joint	Restaurant	Pub	Bakery	History Museum
2	Barbican	Coffee Shop	Food Truck	Gym / Fitness Center	Hotel	Sushi Restaurant	Art Gallery	Turkish Restaurant	Vietnamese Restaurant	Indie Movie Theater	Concert Hall
3	Bayswater	Pub	Coffee Shop	Grocery Store	Greek Restaurant	Chinese Restaurant	Hotel	Yoga Studio	Thai Restaurant	Italian Restaurant	Gym / Fitness Center
4	Belgravia, Brompton, Knightsbridge	Boutique	Café	Italian Restaurant	Hotel	Coffee Shop	Clothing Store	Japanese Restaurant	Jewelry Store	Shoe Store	Tea Room

Figure 2: the top 10 venues of neighborhoods

In the next step, the neighborhoods were divided into 5 clusters using k means clustering. The clustered neighborhoods were shown in the map of London, as shown in figure 3.

In the end, each cluster of neighborhoods were examined and defining categories of it were determined. Defining categories of cluster 1 are coffee shop, pub and hotel. that of cluster 2 are hotel, cafe, pub, of cluster 3 are Pub, Garden, gym/ fitness center, of cluster 4 home service, construction landscaping, lounge, and theater for cluster 5. Some of the clusters are shown in figure 4.

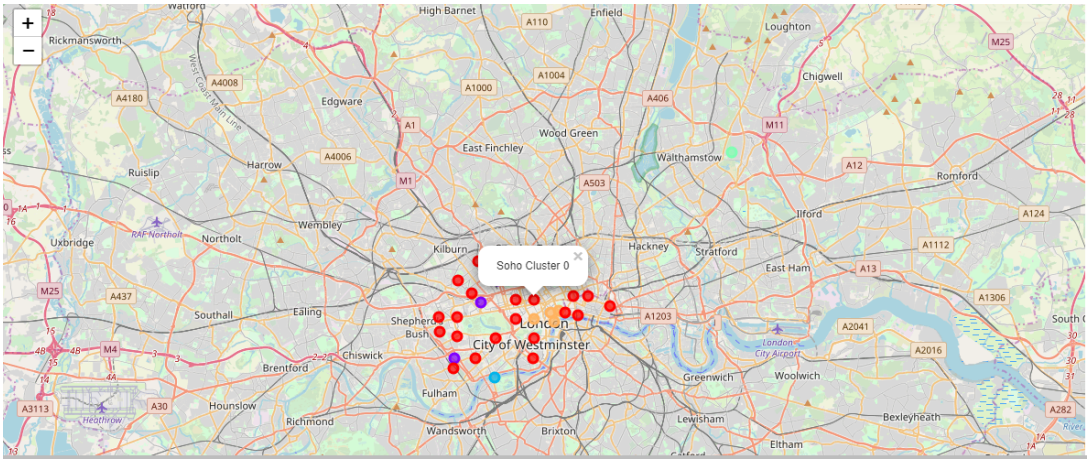

Figure 3: the labeled neighborhoods of London

Cluster 2

```
lon_merged.loc[lon_merged['Cluster Labels'] == 1, lon_merged.columns[[0] + list(range(5, lon_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	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
10	Earls Court	1	Hotel	Café	Pizza Place	Pub	Grocery Store	Garden	Burger Joint	Chinese Restaurant	Thai Restaurant	Lebanese Restaurant
14	Lisson Grove	1	Café	Hotel	Flower Shop	Pub	Hookah Bar	Coffee Shop	Thai Restaurant	Grocery Store	Lake	Snack Place
21	Paddington	1	Hotel	Café	Coffee Shop	Italian Restaurant	Pub	Sandwich Place	Indian Restaurant	Grocery Store	Beer Bar	Thai Restaurant

(a) Cluster 2

Cluster 3

```
lon_merged.loc[lon_merged['Cluster Labels'] == 2, lon_merged.columns[[0] + list(range(5, lon_merged.shape[1]))]]
```

	Neighborhood	Cluster Labels	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
7	Chelsea	2	Pub	Garden	Gym / Fitness Center	Harbor / Marina	French Restaurant	Monument / Landmark	Brazilian Restaurant	Farmers Market	English Restaurant	Grocery Store

(b) Cluster 3

Figure 4: Two of five neighborhood clusters

2.3 Import data for Toronto and New York

In the last stage of this section, we imported the relevant dataframes of Manhattan, New York and Toronto, which were generated in lab and previous assignment, to make comparison between cities. Necessary libraries were imported and then the relevant dataframes were imported and checked.

3 Methodology

In this project, we aim at determining how similar theses cities are by calculating the dissimilarity of either two cities.

In the first part of this project, we have explored the neighborhoods of London and import the data about neighborhoods in Toronto and New York.

In the second part, we will deploy the concept of Euclidian distance to represent the dissimilarity between cities. We will first calculate city dissimilarity based on venue category frequency, and we will use the square of Euclidian distance between the mean venue category frequency of city to represent the dissimilarity, which is:

$$DF = (vf1 - vf2) \cdot (vf1 - vf2)$$

Where DF represents dissimilarity based on frequency of venue category, vf1 and vf2 represent vector of venue frequency of city 1 and city 2. In this step, we will also visualize the obtained dissimilarities

between these cities.

In the second step of this part, we will calculate city dissimilarity based on neighborhood clusters, and we will use the square of Euclidian distance between the proportion of each type of cluster of city to represent the dissimilarity, which is:

$$DP = (vp1 - vp2) \cdot (vp1 - vp2)$$

Where DP represents dissimilarity based on proportion of each type of cluster, vp1 and vp2 represent proportion of each type of cluster in city 1 and city 2. In this step, we will also visualize the proportion of each type of neighborhood cluster in every city and the result of dissimilarity calculation.

4 Analysis

Some modifications were made in relevant dataframes. Information about city was added to dataframes of venues list of city, and the three dataframes of venues were combined. After that, one hot encoding was applied to the combined new dataframe.

After that, we moved into the next stage. The new dataframe was grouped by city and the mean frequency of occurrence of each venue category. Then the calculation of dissimilarity based on venue frequency was performed. The resulting dissimilarities between these cities are 0.00729 for London and New York, 0.00819 for London and Toronto, and 0.00676 for Toronto and New York. In the end, the dissimilarities was visualized, as shown below.

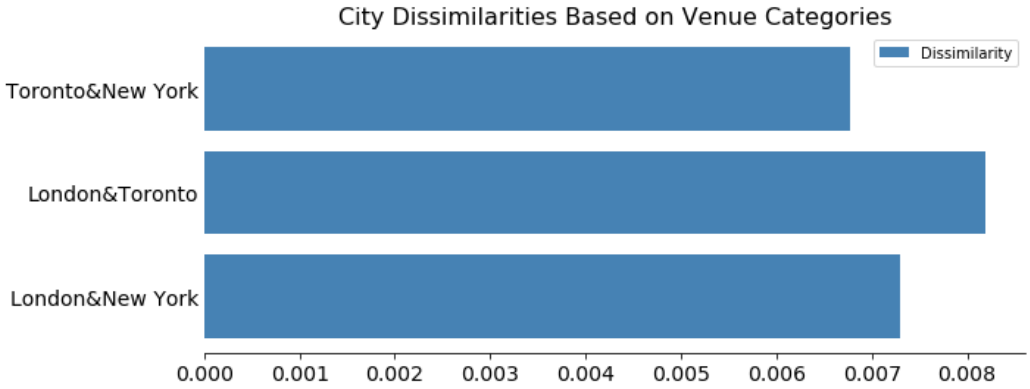

Figure 5: Dissimilarities based on frequency of venue categories

At the final step, the new dataframe was grouped by neighborhood and the grouped dataframe was divided into 5 clusters using k means clustering method. then the cluster label for each neighborhood was put back into a newly created dataframe which consists of only neighborhood and corresponding city. The new dataframe with cluster labels was grouped by city and cluster labels, and the number of each cluster label was replaced by its proportion, as shown below in figure 6. After that, the cluster distributions of these three cities were shown as a bar chart, as shown below in figure 7. From this bar chart, we could note that cluster distribution in Toronto and New York are similar, while that in London is much different. Finally, city dissimilarity based on proportion of different types of cluster were calculated, The resulting dissimilarities between these cities are 0.176 for London and New York, 0.211 for London and Toronto, and 0.011 for Toronto and New York. The calculated dissimilarities are also shown in figure 8.

5 Results and Discussion

In last section, We analyzed the dissimilarity between cities in terms of mean venue frequency of city and proportion of each type of neighborhood clusters.

City	Cluster Labels	Percentage
London	0	0.896552
	4	0.103448
New York	0	0.600000
	4	0.400000
Toronto	0	0.526316
	1	0.026316
	2	0.052632
	3	0.026316
	4	0.368421

Figure 6: Neighborhood cluster proportion in every city

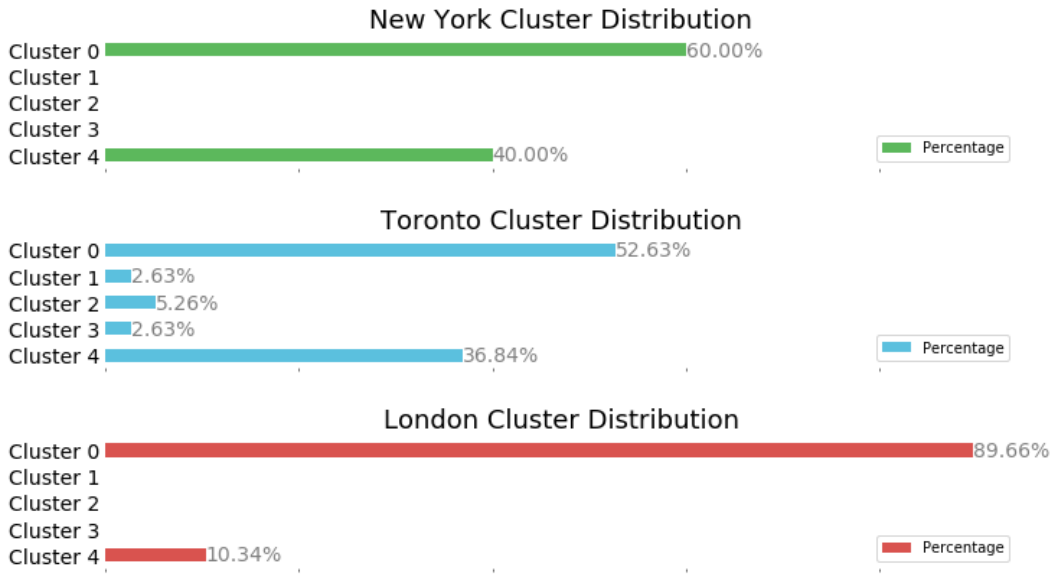

Figure 7: Dataframe of neighborhood cluster proportion in every city

For the former analysis, the resulting dissimilarities between these cities are 0.00729 for London and New York, 0.00819 for London and Toronto, and 0.00676 for Toronto and New York. Although the differences between the dissimilarities are small, we still could arrive a conclusion that Toronto and New York are most similar, followed by London and New York, and the least similar pair would be London and Toronto. We could also corroborate our point by the bar chart of the dissimilarities.

For the latter calculation, The resulting dissimilarities between these cities are 0.176 for London and New York, 0.211 for London and Toronto, and 0.011 for Toronto and New York. The difference between dissimilarities are prominent, especially the dissimilarity of Toronto and New York is much smaller than the other two dissimilarities. We could reach the same conclusion as the last one and we could also corroborate our point by the bar chart of the dissimilarities.

The two comparisons reach the same conclusion: Toronto and New York are most similar, followed by London and New York, and the least similar pair would be London and Toronto. Since the agreement of both outcomes, our conclusion would be valid.

6 Conclusion and Outlook

Our goal of this project is to explore the neighborhoods in London and compare it with other two global cities, which are Toronto and New York to determine how similar these cities are. We first extracted the neighborhoods in the most affluent areas of London and by clustering them, we identified the 5 type

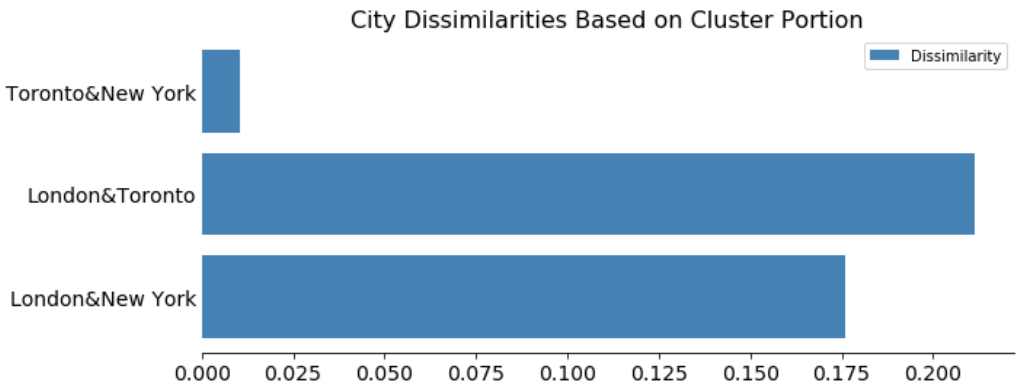

Figure 8: Dissimilarity based on cluster proportion in every city

of neighborhoods of theses areas in this city. After importing data about neighborhoods in the most affluent areas of the other two cities, we first grouped the nearby venues of neighborhoods by cities and calculated the dissimilarities of these cities based on the mean venue frequency of city and found that Toronto and New York are most similar, followed by London and New York, and the least similar pair would be London and Toronto. We then grouped the nearby venues by neighborhoods and cluster the neighborhoods afterwards. Next, we group the cluster labels of neighborhoods by cities and calculate the dissimilarities of them based on proportion of neighborhood cluster types. Again, We reached the same conclusion we got in the former dissimilarity calculation.

Although the conclusion we reached in this project is reasonable, we only focus on the most affluent areas of these three cities. More valid conclusion will be reached if research is conducted in the whole neighborhoods of these cities. And we only compare these cities in terms of venues of neighborhoods, other perspectives, such as criminal rate of neighborhoods, would make our comparison of cities more comprehensive.