Comparing similarity between London and Toronto with similarity between London and Tokyo

1. Introduction:

In this project, the similarity between London and Toronto and the similarity between London and Tokyo are compared in terms of the categories of venues in each neighbourhood of these three major cities of respective countries. As we all known, London and Toronto are biggest cities of United Kingdom and Canada which are both Western countries. However, Tokyo is the capital city of Japan which is a Asian countries. Therefore, it is very interesting to investigate whether there is difference between the structure of Asian major city and Western major city. In this project, we choose London as the main object, and study the similarities between Canada and Japan with respect to London.

2. Data Description:

2.1 Spidering raw data from website:

In this report, the borough and neighbourhood of London, Tokyo and Toronto are extracted from Wikipedia websites.

borough and neighbourhood of London:
https://en.wikipedia.org/wiki/London_boroughs;
borough and neighbourhood of Tokyo:
https://en.wikipedia.org/wiki/Special_wards_of_Tokyo

borough and neighbourhood of Toronto:
https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The report uses BeautifulSoup package to parse html code and finally grab table in each website which is our raw data.

2.2 Data cleaning:

After creating three dataframes which including borough and neighbourhood, data need to be cleaned in terms of the structure of dataframe and the data format. For instance, in figure 1, neighbourhoods need to present as columns not in rows. Therefore, we need to clean the data to figure 2.

borough	neighbourhood_1	neighbourhood_2	neighbourhood_3	neighbourhood_4	neighbourhood_5
Greenwich	Greenwich (22a)	Woolwich (part) (22b)	NaN	NaN	NaN
Hackney	Hackney (9a)	Shoreditch (9b)	Stoke Newington (9c)	NaN	NaN
Hammersmith[notes 2]	Hammersmith (4a)	Fulham (4b)	NaN	NaN	NaN
Islington	Islington (10a)	Finsbury (10b)	NaN	NaN	NaN
Kensington and Chelsea	Kensington (3a)	Chelsea (3b)	NaN	NaN	NaN
Lambeth	Lambeth (6a)	Wandsworth (part) (6b)	NaN	NaN	NaN
Lewisham	Lewisham (21a)	Deptford (21b)	NaN	NaN	NaN
Southwark	Bermondsey (7b)	Camberwell (7c)	Southwark (7a)	NaN	NaN

Figure 1

	borough	neighbourhood
0	Greenwich	Greenwich
1	Greenwich	Woolwich
2	Hackney	Hackney
3	Hackney	Shoreditch
4	Hackney	Stoke Newington
5	Hammersmith	Hammersmith
6	Hammersmith	Fulham
7	Islington	Islington
8	Islington	Finsbury
9	Kensington and Chelsea	Kensington
10	Kensington and Chelsea	Chelsea

Figure 2

Moreover, compared to Figure 1, the name of borough and neighbourhood have been cleaned as well.

2.3 Obtaining location information (latitude and longitude):

For each neighbourhood, geopy library has been used to obtain latitude and longitude information.

Next, Foursquare dataset is used to get the venues information of each neighbourhood with limit 100 and radius 500 meters. For instance, venue name, venue latitude and longitude, venue category. Hence, we can group neighbourhoods in terms of the venue categories and frequencies of each category. Therefore, using the data visualisation techniques, we can obviously compare the similarity between London and Toronto with the similarity between London and Tokyo.

3. Methodology:

Firstly, the tables of venues of London, Tokyo and Toronto should be merged as an entire table. And then, we should get dummies in terms of the categories of venues. Finally, K-Means which is one of clustering methodologies has been used to cluster neighbourhood in terms of categories of the top 100 venues in each neighbourhood. In this report, K = 6, which means there are 6 clusters to be grouped. Consequently, a dataframe which including neighbourhood, latitude, longitude, and cluster labels should be created due to the need of visualisation.

	borough	neighbourhood	latitude	longitude	Cluster Labels
0	Greenwich	Greenwich	52.036732	1.168934	0.0
1	Greenwich	Woolwich	51.482670	0.062334	0.0
2	Hackney	Hackney	51.543240	-0.049362	1.0
3	Hackney	Shoreditch	51.526669	-0.079893	3.0
4	Hackney	Stoke Newington	51.557697	-0.077282	3.0

Figure 3

4. Result (data visualisation):

After using K-Means clustering methodology, there is a simple way to show the result which is data visualisation. The report uses green as cluster 3, blue as 2, purple as cluster 1, red as cluster 0. The result of neighbourhood of London is shown in Figure 4. The result of neighbourhood of Tokyo is shown in Figure 5. The result of neighbourhood of Toronto is shown in Figure 6.

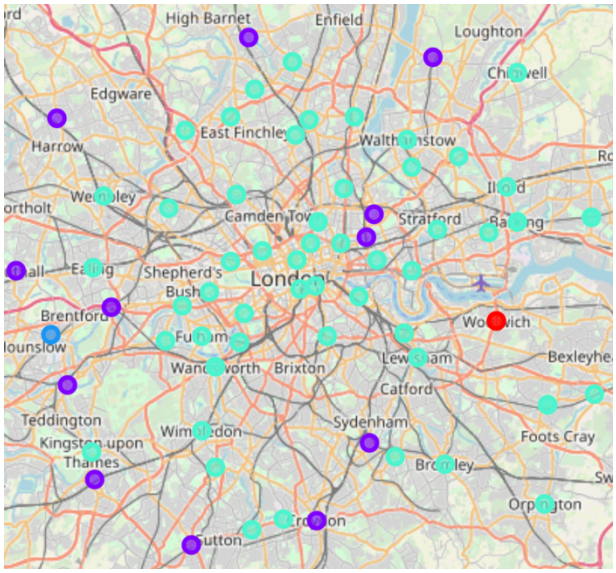

Figure 4

Figure 5

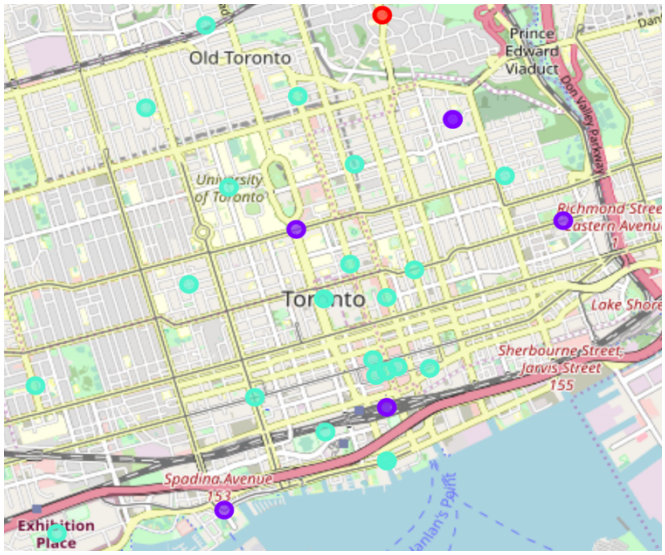

Figure 6

From the cluster visualisation map, there is an obvious result. We have grouped neighbourhoods from London, Tokyo and Toronto into 6 clusters. In the result, the neighbourhoods in London most likely belong to cluster 3 and cluster 1. The neighbourhoods in Toronto most likely belong to cluster 3 and cluster 1 as well. However, the neighbourhoods in Tokyo most likely belong to cluster 2 and cluster 3. Moreover, the most common cluster in both London and Toronto is cluster 3 and then cluster 1. There are least cluster 2 in those two cities. However, the most common cluster in Tokyo is cluster 2 and then cluster 3. Furthermore, there may be no cluster 1 in Tokyo. Therefore, we can conclude that the similarity between London and Toronto is apparently larger that the similarity between London and Tokyo.

5. Conclusion:

In terms of the result, it shows that London has more similarities with Western city than Asian city in terms of the categories of venues in each neighbourhood. Also, it is shown that there are differences of structure of city between Western cities and Asian cities.