Zeyu Xiong

City Similarity based on Venues

Introduction and Data Description

As we all know, Toronto, New York, London and Hong Kong, they are all regional or worldwide financial capitals, so they should be similar in the central area in terms of lots of financial companies' skyscrapers. However, Toronto, New York and London are well-known western cities with, to some extent, similar culture, does it necessary mean that their geographical venues in the central area should be similar as well while Hong Kong is completely different with them? Alternatively, UK has colonized Hong Kong for more than 100 years, from this perspective, it is not unreasonable to think that current Hong Kong should still be similar to UK. Therefore, I will determine which two or three are more similar so that they would be clustered into the same group while the other two or one will be clustered into another group based on the venue category in the central area. This question may be interesting to other city governors who wants to develop their cities, and historian may also be interested at it as they could research deeper in how colonization will shape the building style in the colony.

The data I will use is venue datasets of each city queried from **Foursquare API**. Simply speaking, I will explore all venues around each city, and base on the venue category, I am able to determine which two are more similar. In addition, I will also explore the major venue components in Downtown Toronto, so I will also need the borough coordinate data from Wikipedia and public coordinate data. After cleaning, the venue data of each city is as follows

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
1	Totonto, ON	43.653963	-79.387207	Art Gallery of Ontario	43.654003	-79.392922	Art Gallery
101	New York City, NY	40.712728	-74.006015	The Beekman - A Thompson Hotel	40.711173	-74.006702	Hotel
201	London, UK	51.489334	-0.144055	Apollo Victoria Theatre	51.495622	-0.142689	Theater
301	Hong Kong, HK	22.279328	114.162813	Hong Kong Park Aviary (香港公園觀鳥園)	22.277140	114.161399	Zoo

And venue data of each neighborhood in Downtown Toronto is as follows

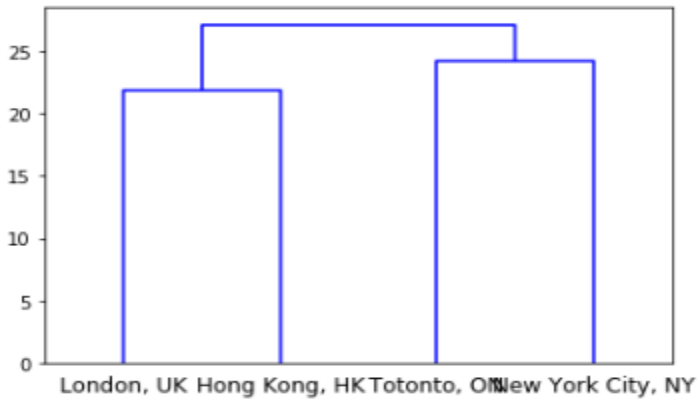
	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Harbourfront	43.654260	-79.360636	Roselle Desserts	43.653447	-79.362017	Bakery
1	Harbourfront	43.654260	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
51	Regent Park	43.654260	-79.360636	Tandem Coffee	43.653559	-79.361809	Coffee Shop
101	Ryerson	43.657162	-79.378937	Burrito Boyz	43.656265	-79.378343	Burritos
200	St. James Town	43.651494	-79.375418	Gyu-Kaku Japanese BBQ	43.651422	-79.375047	Japanese

Methodology and Result

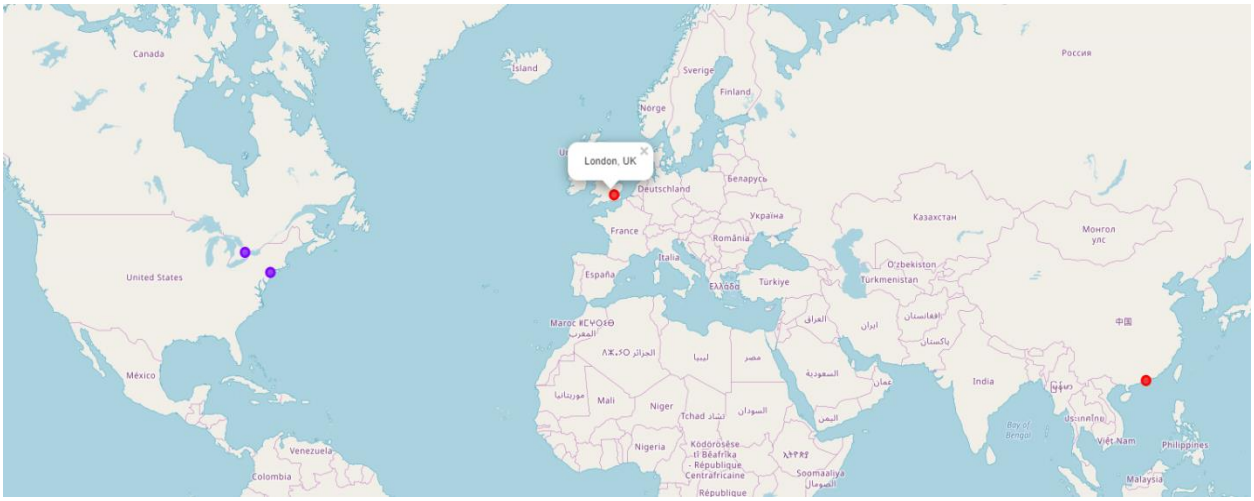
Using Foursquare API, I collected 100 venues around each city's coordinate, I encoded the category of these venues into lots of binary variables and group them again according to the city they belong to for the clustering, so final data will be in this shape

City	American	Apparel	Aquarium	Argentinian	Art Gallery	Art Museum	Arts & Crafts	Asian	Athletics & Sports	Australian	...	Train Station	Udon	Vegetarian / Vegan	Vietnamese	Volleyball Court	V
Hong Kong, HK	0	2	0	1	2	0	0	1	1	0	...	0	0	1	1	0	
London, UK	1	1	0	0	2	3	0	1	0	0	...	0	0	0	0	0	
New York City, NY	0	0	0	0	1	0	1	2	1	1	...	0	1	1	0	1	
Totonto, ON	1	1	1	0	2	0	1	0	0	0	...	1	0	1	0	0	

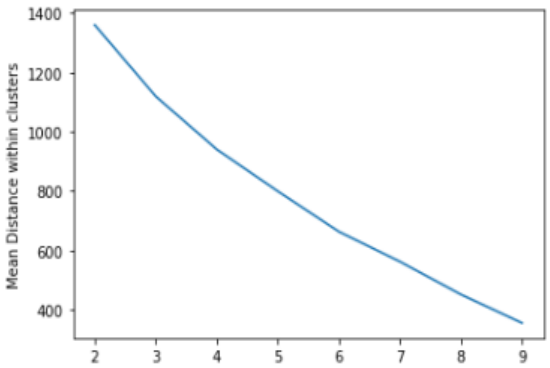
And I choose **Agglomerative Clustering** method as it can give me a hierarchical relation graph between each pair called dendrogram. As we can see from the diagram below, Hong Kong is more similar to London while Toronto and New York are clustered together.



I used folium library to visualize geographic details of four cities, and I added different color to different cluster.



Afterwards, following similar trick, I included more venues in neighborhoods in Downtown Toronto so that I could cluster neighborhoods into several major classes. At this time, I used K-Mean algorithm due to its popularity and simplicity. Unfortunately, I couldn't find a suitable 'clusters' parameter based on elbow method, so I directly choose to cluster them into 3 clusters.



Based on the clustering result, I am able to group these neighborhoods into 3 clusters and summarize the most common venues in each cluster.

	class	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	...
0	0	Venue Category_Coffee Shop	Venue Category_Café	Venue Category_Bakery	Venue Category_Park	Venue Category_Bar	Venue Category_Italian	Venue Category_Restaurant	Venue Category_Pizza	Venue Category_Beer Bar	..
1	1	Venue Category_Coffee Shop	Venue Category_Café	Venue Category_Hotel	Venue Category_Restaurant	Venue Category_Gastropub	Venue Category_Deli / Bodega	Venue Category_Steakhouse	Venue Category_American	Venue Category_Bar	..
2	2	Venue Category_Airport Service	Venue Category_Terminal	Venue Category_Lounge	Venue Category_Airport	Venue Category_Harbor / Marina	Venue Category_Plane	Venue Category_Bar	Venue Category_Boutique	Venue Category_Coffee Shop	..

Discussion and Conclusion

From the dendrogram, we find that even though Toronto, New York and London are western cities sharing lots of cultural similarities, London is more similar to Hong Kong than to his western peers. As I raised at the beginning, it's reasonable because Hong Kong has been the colony of UK for a long period in the history, most of its venues may still keep the features of UK. On the other side, Toronto and New York are much similar to each other than to the others due to the cultural and geographical similarity.

As I dive deeper into the venues of neighborhoods in Downtown Toronto, I clustered neighborhoods into 3 clusters according to the venues they have. However, it seems that those 3 classes are pretty similar as cafe, coffee and restaurants are very common in all classes, after all, food-related venue is always the most popular one in all neighborhoods, and it's easier to open a restaurant then to open a museum or gym. We can still extract some differences:

1. As the park is 4th popular in the first class, so the first class should be 'Entertainment and food' neighborhood;

2. As Hotel is the 3rd most common venue, so the second class should be 'Vacation and Leisure' neighborhood;

3.There are several airport-related venues in the third class, so it should be 'Transportation Service' neighborhood.