IBM Coursera Capstone Project

The Battle of Neighborhood

Hanyu Qi

Introduction: Business Problem

In this project we will try to investigate what venues have higher frequency in **New York** and **Toronto**. And we also will compare the distribution of venues in two cities to find out what venues are more common in one than the other.

Since there are lots of venues we will try to detect **the top 10 popular venues in each city**. We are also particularly interested in **venues with low frequency**.

This project is aiming to help new business to choose their venue types and avoid venues with high density in certain area. Eventually, it can create higher profit for the new business starter.

Data

We will use three data sources to complete this project. For the data of New York Neighborhood, we obtain a JSON file from the following link:

https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-DS0701EN-SkillsNetwork/labs/newyork_data.json

Next, we get the latitude and longitude data by geopy library.

For the data of Toronto Neighborhood, we get the postal code and neighborhood information from Wikipedia and merge latitude and longitude data into the dataset. The file contains latitude and longitude data can be downloaded via the following link"

http://cocl.us/Geospatial_data

Finally, we use Foursquare API to collect the information about venues corresponding with neighborhoods to derive a integral dataset containing Neighborhood, Neighborhood Latitude, Neighborhood Longitude, Venue, Venue Latitude, Venue Longitude, and Venue Category for the further analysis.

For each city, we merge the latitude and longitude of the neighborhoods into the neighborhood table.

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Wakefield	40.894705	-73.847201	Lollipops Gelato	40.894123	-73.845892	Dessert Shop
1	Wakefield	40.894705	-73.847201	Rite Aid	40.896649	-73.844846	Pharmacy
2	Wakefield	40.894705	-73.847201	Carvel Ice Cream	40.890487	-73.848568	Ice Cream Shop
3	Wakefield	40.894705	-73.847201	Walgreens	40.896528	-73.844700	Pharmacy
4	Wakefield	40.894705	-73.847201	Dunkin'	40.890459	-73.849089	Donut Shop

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Malvern, Rouge	43.806686	-79.194353	Wendy's	43.807448	-79.199056	Fast Food Restaurant
1	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	Royal Canadian Legion	43.782533	-79.163085	Bar
2	Rouge Hill, Port Union, Highland Creek	43.784535	-79.160497	SEBS Engineering Inc. (Sustainable Energy and ...	43.782371	-79.156820	Construction & Landscaping
3	Guildwood, Morningside, West Hill	43.763573	-79.188711	RBC Royal Bank	43.766790	-79.191151	Bank
4	Guildwood, Morningside, West Hill	43.763573	-79.188711	G & G Electronics	43.765309	-79.191537	Electronics Store

Methodology

In this project we will work on detecting top 10 popular venues in New York and Toronto to help new business finding their direction.

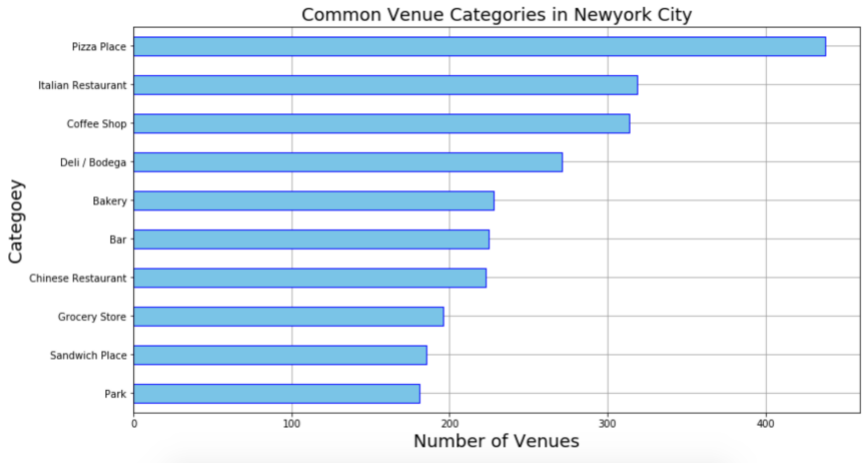
In first step we have collected the required data: Neighborhood, Venue, Latitude and Longtitude. We have also plot the location of each neighborhood in the map.

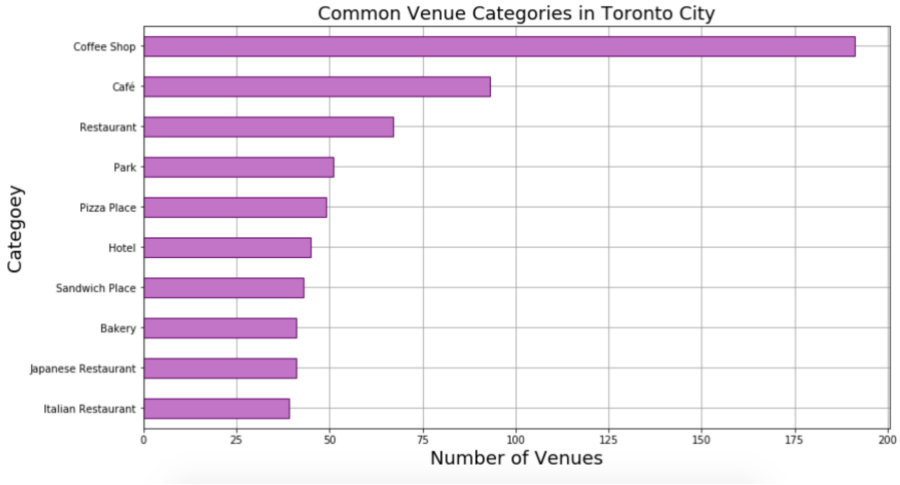
Second step in our analysis will be finding the top 10 popular venues in each neighborhood and located them in the map.

In third and final step we will focus on cluster the locations of the venues in two cities to identify similar venue types in different areas.

Analysis

We find the top 10 common venues in New York and Toronto. And we plot the bar charts to present the number of top 10 common venue category from two cities.





From the above charts, we can observe that the coffee shop takes account as the top 3 popular venue in both New York and Toronto. And in New York, people prefer fast food like pizza and fine food like Italian restaurant. In Toronto, Japanese restaurant has slightly more venues than Italian restaurant. So restaurant and café would be a great choice for new business starter but in the meanwhile, this also means high competitive market.

Also, we combine the data of New York and Toronto and use KNN to find the clusters among these neighborhoods.

Cluster 1

	% of venues
Pizza Place	13.0587
Pharmacy	5.34619
Deli / Bodega	5.08326
Grocery Store	4.73269
Chinese Restaurant	4.38212
Bank	3.5057
Sandwich Place	2.62927

Cluster 2

	% of venues
Pool	57.1429
Baseball Field	28.5714
River	14.2857

Cluster 3

	% of venues
Construction & Landscaping	40
Baseball Field	40
Bar	20

Cluster 4

	% of venues
Park	51.7241
Convenience Store	10.3448
Pool	6.89655
Bus Stop	3.44828
Intersection	3.44828
Playground	3.44828
Boat or Ferry	3.44828

Cluster 5

	% of venues
Coffee Shop	4.37376
Italian Restaurant	3.09958
Pizza Place	3.0544
Café	2.3405
Bar	2.26821
Bakery	2.23206
Deli / Bodega	2.05133

Result and Discussion

The differences between the clusters can be seen from the figure; each cluster has a different distribution of common venue categories. The first cluster consist of venues related with food and restaurants such as Pizza Place, Deli/Bodega and Chinese Restaurant. The second cluster consists of River, Pool, Baseball Field which are mostly the outdoor sports venues. The third cluster consists of landscaping, bar and baseball field again which may be a good area for tourism. In the fourth cluster, the park has 51.7% as the most common venues and there are also some transportation such as bus stop, boat or Ferry and Intersection. In the fifth cluster, the food and drink venues are the major venues again. So we can find the cluster 1 is quite similar to the cluster 5.

Conclusion

In this project, the areas of New York City and Toronto were clustered into various groups dependent on the classifications (kinds) of the venues in these areas. The outcomes demonstrated that there are venue categories that are more common in 1st and 5th clusters; Also these most common venue categories contrast from one cluster to the next. So using this information one can make decisions and be able to find similar neighbourhoods in the new City. It will also help business starters to find the best types of business to start and they may also can use the development of the other clusters as a reference. In the future if a more profound investigation is performed considering more viewpoints, it may bring about finding various styles in each cluster dependent on the most common categories in the cluster.