Capstone Project - The Battle of Neighborhoods

Introduction

Due to the recent political events that happened in Hong Kong, our client, Mr.X, starts looking for information related to investment immigration in three cities. These cities are

- Montreal, Canada
- London, UK
- Tokyo, Japan

(The cities are suggested by https://www.leeabbamonte.com/travel-blog/30-best-cities-in-the-world.html, which Toronto, Canada is changed to Montreal, Canada because Toronto has already been analyzed in the previous weeks' exercises.)

To prevent a strong feeling of nostalgia, Mr.X would like to understand which city is most similar to where he currently lives, so he can save time for deep research of one city.

Data

We will use the following data for analysis

1. Area data of the three cities

By using Beautiful Soup library, we will extract the information from

- https://en.wikipedia.org/wiki/London_boroughs - List of areas of London
- https://en.wikipedia.org/wiki/Boroughs_of_Montreal - List of boroughs of Montreal
- https://en.wikipedia.org/wiki/Special_wards_of_Tokyo - List of special wards of Tokyo

2. Foursquare Using the Foursquare API, we will extract the information related to the neighborhoods.
    - https://foursquare.com/city-guide

Methodology

The project will perform analysis by:

1. Define Mr.X's neighborhood
    - Mr.X's neighborhood is located in Tsim Sha Tsui, Hong Kong
    - We use the foursquare API to collect the top 50 venues within the radius of 100,000 meters, or 100 kilometers.
2. Define neighborhoods in the previously listed three cities
    - We collect the data from the Wikipedia pages and extract the name of the districts / boroughs
    - Using the geopy library to retrieve the coordinates of the boroughs
    - Demonstrate the location of the boroughs in a map
    - Using the same parameters when retrieving Mr.X's neighborhood's venues.
3. Perform clustering on all the neighborhoods, including Mr.X's neighborhood

4. Calculate the similarity of the cities and provide suggestion

Results

After clustering, we group the data first on the name of the city, and then the cluster labels. Which generates the following results.

City	Cluster Lable	Borough (Count)
London	0	29
	3	3
Montreal	0	6
	2	1
	3	11
	4	1
Mr.X	3	1
Tokyo	1	8
	3	15

Mr.X's neighborhood is distributed to cluster 3, then we will calculate the proportion of boroughs belongs to cluster 3 in each city.

```
London: 9.375 %
Montreal: 57.8947 %
Tokyo: 65.2174 %
```

Therefore, Mr.X should first look for Tokyo, then Montreal, and finally London.

Reflection

According to the results generated by the clustering analysis, we can clearly see the boroughs in the same city are mostly allocated to the same cluster. This can be due to the following reasons

1. The cultural difference between the cities made the boroughs in the same city easier to be grouped into the same group
2. The number of venues collected and the radius selected for venues data collection can be improved
3. The community contribution is not enough in some of the cities /boroughs; this might be due to the difference in culture, borough size, the scale of economic activity, and the popularity in each city.

```
Number of venues in Mr.X's neighborhood: 50

Number of boroughs in London: 32
Number of venues in London: 1246
Average number of venues in each borough in London:  38.9375

Number of boroughs in Montreal: 19
Number of venues in Montreal: 529
Average number of venues in each borough in Montreal:  27.842105263157894

Number of boroughs in Tokyo: 23
Number of venues in Tokyo: 1150
Average number of venues in each borough in Tokyo:  50.0
```

4. Some labels of venues are in detail, but some of them are in general form, this can be explained by using the result.

	Cluster Labels	borough	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	1	Adachi	Convenience Store	Supermarket	Café	Ramen Restaurant	Fast Food Restaurant	Restaurant	Japanese Restaurant	Steakhouse	Donburi Restaurant	Discount Store
1	3	Ahuntsic-Cartierville	Café	Pizza Place	Sandwich Place	Hockey Arena	Train Station	Park	Middle Eastern Restaurant	Breakfast Spot	Liquor Store	Ice Cream Shop
2	3	Anjou	Restaurant	Italian Restaurant	Coffee Shop	Hockey Field	Recreation Center	Sushi Restaurant	Food	Bowling Alley	Paper / Office Supplies Store	Thai Restaurant
3	1	Arakawa	Convenience Store	Italian Restaurant	Park	Sake Bar	Fast Food Restaurant	Grocery Store	Café	Sushi Restaurant	Train Station	Supermarket
4	0	Barking	Hotel	Grocery Store	Supermarket	Park	Coffee Shop	Gas Station	Gym	Business Service	Discount Store	Breakfast Spot

In row index 2 (Anjou), the most common venue is “Restaurant,” which is a general description of a venue, but the existence of its sub-categories (Italian Restaurant, Sushi Restaurant, Thai Restaurant) can disturb the accuracy of clustering. To improve the result, generalizing the labels / category of a venue might be needed.