IBM Applied Data Science Capstone
Clustering Tokyo area and where we stay when Tokyo Olympic 2020

Introduction

In 2020, we have Tokyo Olympic and expect that many visitors come from over the world. They would see the Olympic game and also may want to go sightseeing in Tokyo and other district. Even if Tokyo is well known city, most of visitor may not know how to go to place where they would go, which area is continence to commute, where is the area we should go for lunch etc. Knowing the Tokyo helps tourist to their visit more comfortable and easy.

Business Problem

The Objective of this capstone project is to clustering the city of Tokyo, Japan to make the visitors easier to stay during Olympic period. Using the data science methodology and machine learning techniques to provide the useful data to foreigner and visitors in Tokyo.

Target Audience of this project

This project is particularly useful for tourist for Tokyo Olympic 2020.

Data

To get the insight, we will use the following data.

1. List of neighborhoods in Tokyo.
2. Latitude and Longitude coordinates of those neighborhoods.
3. Location data that would be useful for tourist.
    - Hotel, restaurant etc

Source of data and method of extract them

This page (https://qiita.com/butchi_y/items/3a6b70b38e13dc56ef13) contains list of Yamanote-Line station coordination. Yamanote-Line is one of Tokyo's busiest and most important lines connecting most of Tokyo key stations. I have downloaded into csv file and import into python using read_csv.

Also I use Foursquare API to get the venue data for these neighborhoods. Through the API, we can collect relative data which helps to give us the insight,

Methodology

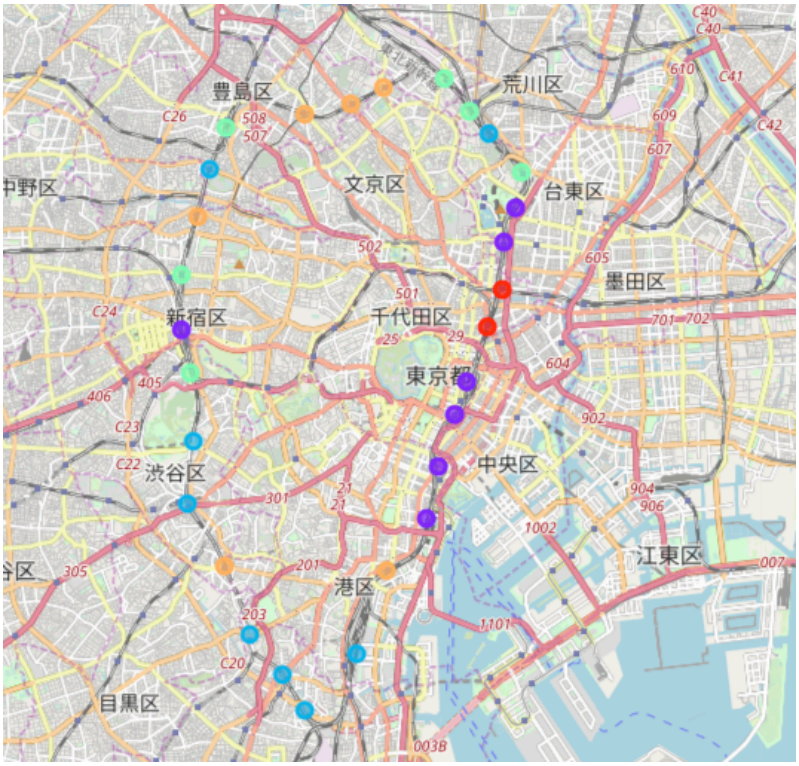
1. Read the Yamanote-Line station coordination and plot into the map to assure all station coordination is properly plotted.
2. Using Foursquare API using own client ID and client secret key, collect relative information.
3. Using One-Hot-Encoding, analyze the category nearby neighborhoods.
4. Clustering in K-Means clustering
5. Conjunction Stations data and Clustering data
6. Visualize in map.

Results

Cluster 0has more place to stay and Cluster 4 is less place to stay.

Cluster 0 (red)　　Cluster 1 (purple)　Cluster 2 (green)　　Cluster 3 (blue)

Cluster 4 (orange)



Discussion

As observations noted from the map, most of hotels are located next to major station. For example Kanda and Akihabara are nearby Tokyo station. I expect that major stations are business district and may have some hotel but not many since office space is more profitable than hotel use.

And less hotel cluster (No.4) are area known as resident area. Of course there are some hotels we can stay but not much since in japan co-existing both hotel and residential is not common and Japanese may not prefer living nearby hotel due to security and noise concern.

Conclusion

If you are planning to visit Tokyo 2020 Olympic, I would recommend to start booking the hotel since Tokyo area may not have sufficient hotel for all tourist. I have lived in Tokyo more than 15 years and I thought that we have more areas like cluster 1(Akihabara and Kanda).However, based on this study, my assumption was not correct.

Reference

Tokyo key station latitude and longitude

https://qiita.com/butchi_y/items/3a6b70b38e13dc56ef13

Foursquare developers documentation; Category code

https://developer.foursquare.com/docs/resources/categories