Finding the most similar Toronto neighborhood with South Korea

**Applied Data Science Capstone course by IBM/Coursera**

Eunkyung Gu

December 10, 2020

# Introduction: Business Problem

In this project I will try to find the best place for Koreans to live in Toronto, Canada. Specifically, this report will be targeted to stakeholders interested in immigrating from Seoul, South Korea to Toronto, Canada. In order to narrow the target down and clarify the project's performance, the project will set the scope of people migrating from Seongdong-gu, Seoul to Downtown Toronto, Toronto.

Nowadays, there are quiet lots of Korean who try to immigrate to Canada for a variety of personal reasons, including career and academic issues. I will try to find the locations that have similar facilities with Seoul. I would like to find a place that meets this condition.

This project could help Koreans who want to immigrate to Canada by representing the most adaptable neighborhood in Toronto.

# Data acquisition and cleaning

Based on definition of my problem, factors that will influence our decision are:

* category and frequency of facilities in Seongdong-gu and Downtown Toronto.
* similarity of existing facilities.

Following data sources will be needed to extract/generate the required information:

* The information of Neighborhood in Toronto, Canada including postal codes will be obtained from Wikipedia page. (<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>)
* The geographical coordinates of each postal code of Canada will be obtained from a csv file (<http://cocl.us/Geospatial_data>)
* similarity of existing facilities in the neighborhood of Toronto and South Korea will be obtained using Foursquare API
  1. Data of Downtown Toronto, Toronto, Canada

First, I scraped data from Wikipedia page, merged it with csv file, and made into a dataframe. It includes postal code, borough, neighbourhood, and coordinate(latitude, longitude).

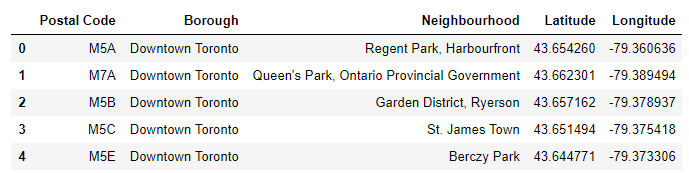


Figure 1 Dataframe of Downtown Toronto including Postal Code, Geographical coordinates etc.

* 1. Geographical coordinate of Seongdong-gu, Seoul, South Korea

Using geolocator, I got the coordinate of Seongdong-gu. It is:

* Latitude: 37.5635
* Longitude: 127.0365
  1. Foursquare API

To bring information on facilities in Seoul and Toronto, I used Foursquare API. I got the 100 venues within a radius of 1000 meters of each region. These data were made into dataframes.

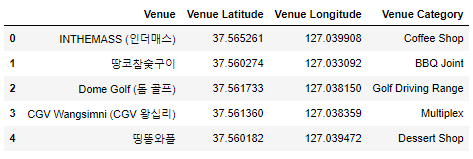


Figure 2 Dataframe of venues in Seongdong-gu

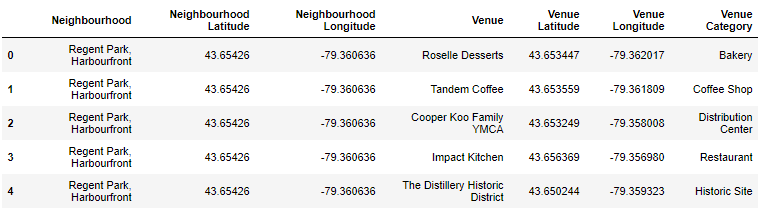


Figure 3 Dataframe of venues in Toronto.

# Methodology

In this project I directed my efforts on detecting best areas of Downtown Toronto, Toronto that have similarity of the presence and frequency of facilities to Seongdong-gu, Seoul.

In first step, I had collected the required data: location and top 100 Venues and its Category for both borough (Dowontown Toronto, Seongdong-gu).

Second step in my analysis was clustering Neighborhoods of Downtown Toronto using their Venue category and its frequency. I used K-Means clustering model and one-hot encoding method for categorical variables clustering.

In third step, I focused on categorizing the Seongdong-gu. With cluster index of Downtown Toronto, using it as categorizing index, I classified which cluster Seongdong-gu belongs to. I used decision tree model and one-hot encoding method for categorical variables categorizing.

# Analysis

* 1. Explanatory Data analysis

For generating explanatory data analysis and deriving additional info, I picked top 10 categories that appears most. I figured the frequency out using one-hot encoding.

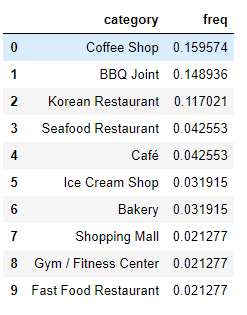


Figure 4 One-hot encoding matrix of Seongdong-gu

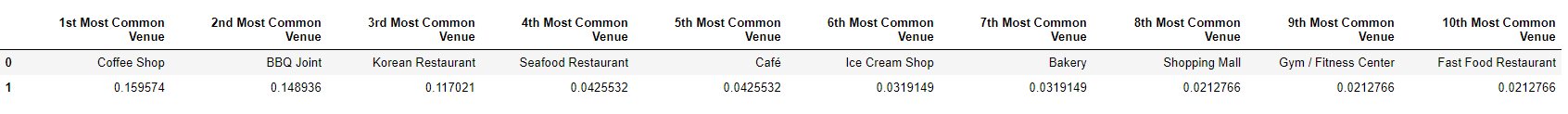


Figure 5 Top 10 Common Venue Categories of Seongdong-gu

I successfully extracted top 10 Common Venues Categories and their frequency of Seongdong-gu.

1. Coffee Shop: 0.16
2. Hotel: 0.09
3. Korean Restaurant: 0.07
4. Chinese Restaurant: 0.05
5. Lounge: 0.04
6. Bookstore: 0.04
7. Japanese Restaurant: 0.04
8. Cafe: 0.04
9. Historic Site: 0.03
10. Plaza: 0.03

Next, I analyzed as same way to get top 10 common venues of each neighborhood in Toronto.

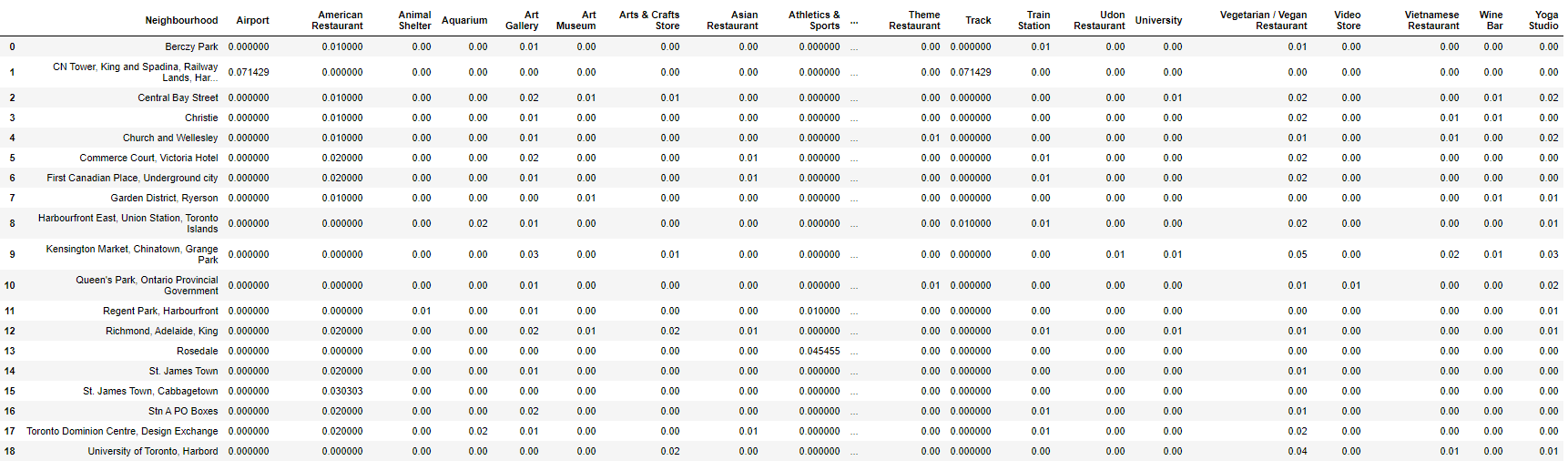


Figure 6 One-hot encoding matrix of Downtown Toronto

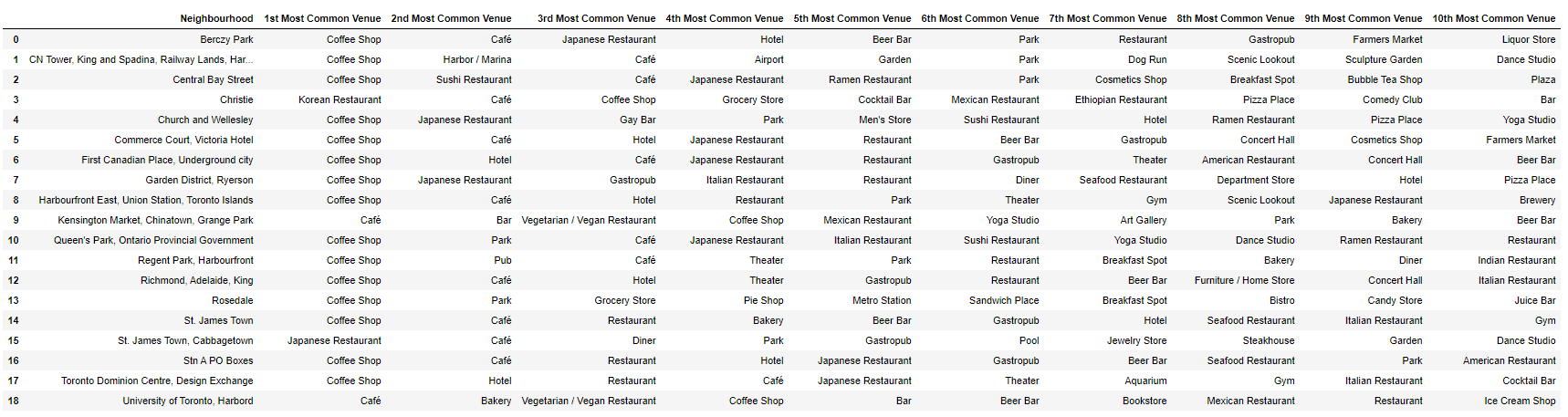


Figure 7 Top 10 Common Venue Categories of Downtown Toronto

* 1. Clustering Neighborhoods in Downtown Toronto

I clustered the neighborhoods in Downtown Toronto. Because variables for cluster are 10 most common venue categories were categorical values, so I used onehot encoding dataframe to cluster. I used K-Means Clustering model, and the number of clusters will be 5. Clustered data was shown on the map using Folium (Fig. 7).

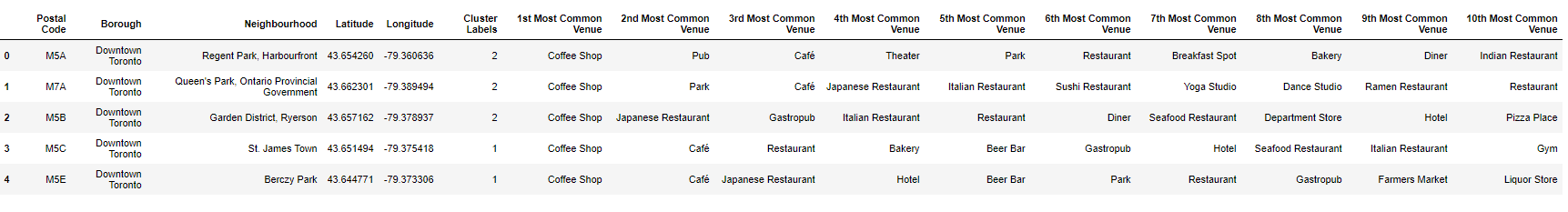


Figure 8 Cluster Labeled

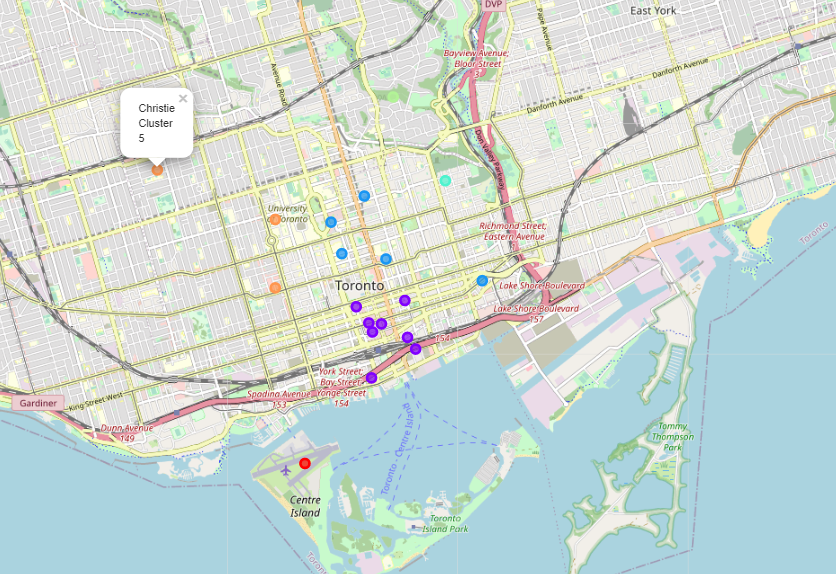


Figure 9 Clustered data on Map using Folium

* 1. Classifying which cluster Seongdong-gu belongs to.

Using decision Trees model, I classified which cluster Seongdong-gu belongs to among cluster 1-5. I will use onehot encoding dataframe for classification. Decision Tree classifier criterion was 'entropy' and max depth was 4. For training the classification model, I splited the data into trainset and testset. 19 dataset was splited into 16 trainset and 3 testset(test size is 0.15). The accuracy of trained model was 0.67. From the prediction result, Seongdong-gu belongs to cluster 2.

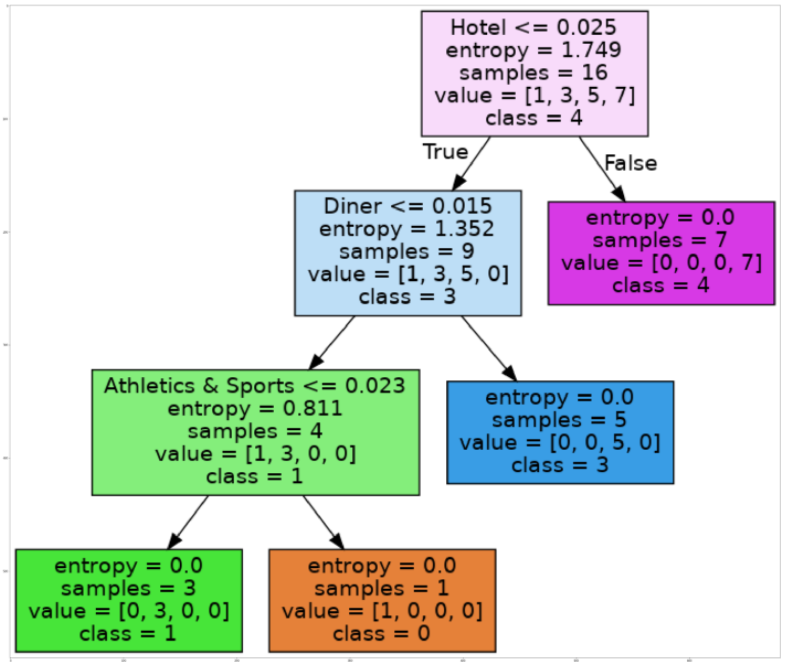


Figure 8 Visualization of Decision-making Tree

# Results

Analysis showed that Seongdong-gu was predicted to belong to cluster 2. Cluster 2 is including the neighborhoods such as:

* Christie
* University of Toronto, Harbord
* Kensington Market, Chinatown, Grange Park

In other words, there is a similarity of facilities between in Seongdong-gu and these areas. From this point on, Koreans from Seongdong-gu will be able to adapt quickly and well when they move to these areas mentioned above.

# Discussion

However, the project following limitations:

1. The number of dataset used for clustering and classification was 19. And this was not enough (actually, lack) to train. The accuracy of Decision tree in this project showed quiet low at 0.67. There is room for further improving the model by increasing the amount of data.
2. Clustering and classification index were created simply in proportion to the number of facilities. In other words, it is calculated without proper consideration of the overall ratio, which may give ambiguity to analysis. For example, coffee shops or restaurants are a lot in any area, and historical sites are a few in any area, but this project did not take that into account.

# Conclusion

Purpose of this project was to reveal which neighborhood in Downtown Toronto is most similar to Sengdong-gu, in order to aid stakeholders the search for optimal location for the best immigration destination. Based on data from Foursquare API, I clustered and classified the areas. And it showed calculated best place.

The final decision on choosing the optimal location will be made by stakeholders based on the characteristics (included in this project or not) of the region. This project will be a good reference for them.