Capstone Project – Battle of Neighborhoods (Part 1): Determining the top Japanese Restaurants in Toronto using K-means clustering method

By Hoo Chee Hau

DATA

In this section, the description on the data used in this study, with the source of data from where it is obtained and how the data will be utilized is detailed as below:-

1. Toronto's **FSA (Forward Sortation Area)** 3-letter codes can be found from Wikipedia's webpage titled - '**List of postal codes of Canada: M**' with the URL given below:- https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M. According to Canada Post, Toronto's FSA code begins with '**M**'.
   A process known as **webscraping** is performed to scrape the relevant data related to Toronto's FSA postal codes and the corresponding names of **boroughs** and **neighborhoods** from this Wikipedia webpage and convert those to a Pandas dataframe in this project.
2. The **geographical coordinates** in Longitude and Latitude corresponding to Toronto's postal code **FSA** are obtained from : http://cocl.us/Geospatial_data in comma delimited text format. This comma delimited file will be imported and converted into a Pandas dataframe in the project.
3. The two aforementioned dataframes will eventually be merged to create a new expanded dataframe that contains columns as the following : **Postal Code**, **Borough**, **Neighborhood**, **Longitude** & **Latitude**.
4. This study will utilize **Foursquare API** (https://api.foursquare.com/v2/) extensively to perform **venues search** and to retrieve the **name** and **id** of the respective **Japanese restaurants** in each neighborhood based on its geographical coordinates in **longitude** and **latitude**. Foursquare is a popular social location service that requires login authentication.
5. Next, by utilizing **Foursquare API premium** search, **venue details** search will be performed to retrieve **venue details** information, such as **ratings**, number of **likes**, number of **tips** and **price tiers** based on the **id** of each **Japanese restaurant**, retrieved in the previous **venues search**.
6. Following that, **k-means clustering**, a machine-learning technique, with the objective to rank the Japanese restaurants based on how these restaurants are grouped in the clusters, is performed on a dataset that consists of the retrieved **venues details** information, namely **ratings**, number of **likes** and number of **tips** of these Japanese restaurants.
7. Based on the analysis of k-means clustering result, these Japanese restaurants are ranked and the top Japanese restaurants on the list can be determined. **Foursquare API premium** search will be then utilized to retrieve **venue details** information on the **address**, **phone numbers**, **opening times** and **actual location coordinates** in **longitude** and **latitude** for the purpose of map visualization.