IBM Data Science Capstone Project

Singapore Venues Cluster Analysis

by Agratama Arfiano

2. Data Acquisition and Cleaning

a. Data Description

Before start the analysis, we need to know the list of data that needed and where the data will be sourced:

- **List of neighborhood in Singapore**

  We can find the list of areas by scraping the data from Wikipedia

  https://en.wikipedia.org/wiki/Planning_Areas_of_Singapore#List_of_Planning_Areas

- **Coordinates of areas in Singapore**

  We will used the geopy modules to search each of the areas to find the coordinates

- **List of all venues on each area**

  List of all venues will be obtained from Foursquare API

To perform the analysis, we sourced the data from those multiple channels.

b. Data Exploration

Neighborhood in Singapore is called Planning Areas. Planning Areas data from Wikipedia shows 55 Planning Areas with additional information such as what region they are categorized, area size, population and density. Planning areas that have no venues will be dropped later to avoid interference with further analysis

After having list of all planning areas, we use geopy to get all the latitudes and longitudes coordinates

From Foursquare API we got location based data on different sort of venues. Foursquare indexes each location through the longitude and latitude of that location, and provides the data based on the endpoints selected. Regular endpoints are provided free and include basic venue data, category of venue, and venue ID. Premium endpoints require payment and included more depth of data such as user ratings, photo, tips, menu, hours of operation, etc.

c. Data Cleaning

After data scraped from different sources are combined together, we will do data cleaning by removing some information that irrelevant and will interfere our analysis.

d. Methodology

We will using clustering machine learning approach using K-means technique to create a cluster of venues with silhouette score for choosing the optimal number of clusters