IBM Applied Data Science Capstone Project

By Liu Yuanxiu

July 2019

PART 2

Opening a shopping mall in Kuala Lumpur, Malaysia

Data

To solve the problem, we will need the following data:

1. List of neighbourhoods in Kuala Lumpur – to define the scope of this project which is targeted for Kuala Lumpur, the capital city of Malaysia.
2. Latitude and longitude coordinates of these neighbourhoods – to plot the map and obtain the venue data.
3. Venue data related to shopping malls – to perform clustering on the neighbourhoods.

Sources of data and methods for extraction:

First, we pick a list of neighbourhoods in Kuala Lumpur from Wikipedia (https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur), which contains 71 neighbourhoods.

We will use web scraping techniques to extract the data from Wikipedia via Python requests and Beautifulsoup packages. Next, we will obtain the geo coordinates of the neighbourhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighbourhoods.

We will then use Foursquare API to obtain the venue data for these neighbourhoods. Foursquare API will provide categories of the venue data, especially categories on the shopping malls to help us solve the business problem mentioned in Part 1.

This project will make use of several data science skills from web scraping, Foursquare API, data cleaning, data wrangling, machine learning, and map visualisation.

End of Part 2

Thank you.