IBM Coursera Applied Data Science Capstone Project

The Battle of Neighborhoods

Sridharan Sadagopan | IBM Coursera Applied Datascience | May 6, 2020

IBM Coursera Applied Data science Capstone Project

Overview

Introduction / Business Problem

Our client is a books and stationary seller based in Singapore. They would want expand their business to cities in nearby countries such as Malaysia, Indonesia, Thailand, Philippines and etc. Our client wants us to analyse and recommend top 5 locations for opening bookstores in those cities.

Selection of the right neighbourhood such as the ones with large number of schools, universities, shopping malls and etc is important for the bookstore business to be successful. It would also be important to analyse information on those neighbourhood about the existing bookstores to understand the competition.

Results of the recommendation would need to be presented for the cities Kuala lumpur, Jakarta, Bangkok and Manila with the top 5 recommended locations for the bookstores.

Data Understanding

To help up us in this process we would be using the different data about venues in the locations of interest. Most important data from Foursquare API that we would depend on is the Venu Categories.

Prior to using Foursquare API, we would be getting the information about the list Suburbs/Neighbourhood and their geolocation coordinates.

For example we would getting the list of suburbs from Wikipedia for Kuala Lumpur (one of the city of interest) from the page https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur

Using BeautifulSoup python library we extract the suburbs of Kuala lumpur as follows:

- 'Alam Damai',
- 'Ampang, Kuala Lumpur',
- 'Bandar Menjalara',
- 'Bandar Sri Permaisuri',
- 'Bandar Tasik Selatan',
- 'Bandar Tun Razak',
- 'Bangsar'
- ...

Then using geopy.geocoders 'Nominatim' we would get the geolocation of these suburbs:

Suburb	Latitude	Longitude
Alam Damai	3.06357	101.738974
Ampang	3.150256	101.760210
Bandar Menjalara	3.194136	101.633634
Bandar Sri Permaisuri	3.100205	101.718107
Bandar Tasik Selatan	3.076097	101.711447
Bandar Tun Razak	3.089695	101.712467
...	...	...

With the Name, Lattitude, Longitude we will proceed with the FourSquare API find the venues of interest (Educational Instituitions) as follows:

For example Fouresquare API returs the Venue categories such as schools, colleges and universities:

- School
    - Adult Education Center
    - Circus School
    - Cooking School
    - Driving School
    - Elementary School
    - Flight School
    - High School
    - Language School
    - Middle School
    - Music School
    - Nursery School
    - Preschool
    - ….
- College & University
    - Community College
    - Fraternity House
    - General College & University
    - Law School
    - Medical School
    - Sorority House
    - Student Center
    - Trade School
    - University

Using the category ids as defined by FourSquare API [https://developer.foursquare.com/docs/build-with-foursquare/categories] we can filter the venues of interest by specifying catergory of interest in the search/explore queries of FourSquare API. e.g Category Ids from FourSquare:

- College_iniversity='4d4b7105d754a06372d81259'
- Library='4bf58dd8d48988d12f941735'
- School='4bf58dd8d48988d13b941735'
- Shopping_mall='4bf58dd8d48988d1fd941735'
- Shopping_plaza='5744ccdfe4b0c0459246b4dc'

We would limit the venues that fall under the specified category as follows

index	name	categories	lat	lng
0	Metro Driving Academy College	Academic Building	3.063059	101.740452
1	Sek Rendah Agama Al Mukhlisin	College Classroom	3.062615	101.741722
2	Sekolah Rendah Agama Almukhlisin	Student Center	3.063011	101.740369
3	Sekolah menengah kebangsaan alam damai	College Administrative Building	3.063069	101.740458
4	Tadika Al-fath	Nursery School	3.064968	101.736886
...	...	...	...	...

Data Analysis

Now with the venues of interest, their category and geolocation we could count the unique categories of venues

	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
Suburb						
Alam Damai	5	5	5	5	5	5
Ampang, Kuala Lumpur	15	15	15	15	15	15
Bandar Menjalara	16	16	16	16	16	16
Bandar Sri Permaisuri	9	9	9	9	9	9
Bandar Tasik Selatan	6	6	6	6	6	6
Bandar Tun Razak	9	9	9	9	9	9
Bangsar	26	26	26	26	26	26
Bangsar Park	26	26	26	26	26	26
Bangsar South	27	27	27	27	27	27
Batu, Kuala Lumpur	18	18	18	18	18	18

Next we will proceed to summarize the top common venue category in the suburbs

Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Taman Tun Dr Ismail	Student Center	School	High School	Shopping Mall	College Classroom	Library	Music Venue	College Administrative Building	College Football Field	College Gym
Taman U-Thant	High School	College Academic Building	Medical School	General College & University	Office	University	College Gym	College History Building	College Lab	College Library
Taman Wahyu	Student Center	School	High School	College Classroom	Convention Center	Community College	College Engineering Building	College Football Field	College Gym	College History Building
Titiwangsa	College Classroom	General College & University	Medical School	College Library	College Science Building	College Administrative Building	Arcade	Nursery School	College Auditorium	College Academic Building
Wangsa Maju	Student Center	General College & University	Shopping Mall	Middle School	University	Preschool	Elementary School	College Library	High School	Law School

With top common venue category per suburb is available we will proceed with clustering (KMeans) to understand the similarity pattern among the suburbs

By studying further on the clustered suburbs further we would be able to recommend the top suburbs for the opening bookstores.
In addition to the basic venue details Foursquare API provide, we could potentially augment it with data FourSquare Places Location Data Can Offer.

For example we would be using place attributes like venue name, address, ratings, and reviews.

With problem to approach clearly defined and with these data that can be retrieved using Foursquare API, data requirements and correct sources of data for this project are clearly understood. The next steps of data science methodology Data Understanding, Data Preparation, Modeling, Evaluation and Potential Deployment.

Optionally an application can be built to help the business to continue to use the data from Foursquare API such as user ratings, reviews to continue improve the kinds of books and other items that can be stocked.