IBM DATA SCIENCE CAPSTONE PROJECT

INTRODUCTION

The capstone project aims to give a taste of what data scientists do in real life when working with data. The project requires the knowledge of location data and different location data providers, such as Foursquare. One also needs the understanding of how to make RESTful API calls to the Foursquare API to retrieve data about venues in different neighborhoods around the world. It also encourages one to be creative in situations where data is not readily available by scraping web data and parsing HTML code. Python and its pandas library are utilized to manipulate data, which help to explore and analyze data. Folium library is also used to great maps of geospatial data and to communicate results and findings.

BUSINESS PROBLEM

An entrepreneur would like to open a café that specializes in cheesecakes in Dubai. Being a city with diverse population and a tourist hub, Dubai is considered great choice for café and restaurant owners to begin or expand their business. The entrepreneur would like for the café to be affordable and easily accessible to the public.

Taking into account the price level and easy access to public, the café needs to be located within a community in Dubai that has a good amount of footfall of the general public as well as the tourists.

Being a resident of Dubai for 20 years now, the intention behind is to derive optimal communities within Dubai for the location of the restaurant using unsupervised machine learning in addition to application of knowledge of the environment of the city.

Although this business problem is specific for a particular café owner, this model can also be extrapolated to the audience of any potential entrepreneur looking to open a new restaurant or café.

DATA

To build this model, 3 different data sources will be used:

1) List of Communities in Dubai
   https://en.wikipedia.org/wiki/List_of_communities_in_Dubai

This data will be retrieved from the URL using Web Scraping. The *pandas* package on Python will be used to retrieve this data.

2) Geospatial data of the Communities in Dubai from the above list

The latitude and longitude of the communities in Dubai will be retrieved by using the *geocoder* package on Python.

This data will then be merged with the data obtained from Wikipedia to create the base data.

3) Top Venues per Community

The top venues per community will be retrieved by using Foursquare through an API by using the data collected in points 1&2 as base data.