The Battle of Neighborhoods | Data Description

Data Description :

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

• List of neighbourhoods in Hyderabad. This defines the scope of this project which is confined to the city of Hyderabad, India.

• Latitude and longitude coordinates of those neighbourhoods. This is required in order to plot the map and also to get the venue data.

• Venue data, particularly data related to shopping malls. We will use this data to perform clustering on the neighbourhoods.

Sources of Data and Methods :

This Wikipedia page (https://commons.wikimedia.org/wiki/Category:Suburbs_of_Hyderabad,_India) contains a list of neighbourhoods in Hyderabad, with a total of 54 neighbourhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and beautiful-soup packages. Then we will get the geographical coordinates of the neighbourhoods using Python Geocoder package which will give the latitude and longitude coordinates of the neighbourhoods.

Since Foursquare has one of the largest database of 105+ million places and is used by over 125,000 developers, it could be used to get the venue data for the neighourhood.

Foursquare API will provide many categories of the venue data, we are particularly interested in the Restaraunt category in order to help us to solve the business problem. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium).

Foursquare API

We will need data about different venues in different neighborhoods of that specific city. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighborhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, we have chosen the radius to be 2000 meter.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

-Neighborhood

-Neighborhood Latitude

-Neighborhood Longitude

-Venue

-Venue Latitude

-Venue Longitude

-Venue Category