Battle of Neighborhoods of -Port Elizabeth & Durban in South Africa

Capstone Project

2. Data Requirements

Required data can be gathered from:

- **Port Elizabeth** and **Durban City** information, including districts and neighborhoods, can be obtained from Wikipedia:
  - (Source - https://en.wikipedia.org/wiki/Port_Elizabeth) &
  - (Source - https://en.wikipedia.org/wiki/Durban)
- Photos and Picture of **Port Elizabeth** and **Durban** City used for Presentation from (Source - https://afrotourism.com/travelogue/)
- The data used for this project will be acquired from http://www.sapostalcodes.info. The datasets consists of the postal codes and suburb names of each city.
- In order to obtain venues and their categories we will use Foursquare API search feature [FOURSQUARE] (https://foursquare.com/) will be used to collect neighborhood venue information as well as the longitude and latitude details of each suburb. Details about local venues and locality will provide insight into the qualities of a neighborhood.
- In addition to Foursquare, various python packages will be used to create maps and machine learning models to gather further insights and provide efficient recommendations and results into our neighborhood battle project.

  These packages includes:

  1. Pandas - Library for Data Analysis
  2. NumPy – Library to handle data in a vectorized manner
  3. JSON – Library to handle JSON files
  4. Geopy – To retrieve Location Data
  5. Geocorder - For geolocation of neighborhoods
  6. Requests – Library to handle http requests
  7. Matplotlib – Python Plotting Module
  8. Sklearn – Python machine learning Library
  9. Folium – Map rendering Library

Basic Work Flow followed as:

- HTTP requests would be made to this Foursquare API server using postal codes of Port Elizabeth Suburbs and Durban Suburbs to pull out the latitude and longitude which will be used for creation of the map as well data analysis.
- Using credentials Foursquare API search feature would be enabled to collect the nearby places of the suburbs. Due to http request limitations, the number of places per suburb parameter would be set to 100 and the radius parameter would be set to 700.
- Folium- Python visualization library would be used to visualize the suburbs cluster distribution of Port Elizabeth and Durban over an interactive leaflet map.
- Extensive comparative analysis of two suburbs world be carried out to derive the desirable insights from the outcomes using python's scientific libraries Pandas, NumPy and Scikit-learn.
- Unsupervised machine learning algorithm K-mean clustering would be applied to form the clusters of different categories of places residing in and around the neighborhoods. These clusters from each of those two chosen suburbs would be analyzed individually collectively and comparatively to derive the conclusions.