Capstone Project Report

IBM Data Science Specialisation

Introduction

The proposed software will have the following main functionality. Provided an input location and a destination, it will investigate various characteristics of the input location, and suggest the most similar locations within a predetermined area of the destination.

This functionality is expected to address the following two use cases:

A.) Suppose a user would like to move from his/her current location to a new city, perhaps due to professional reasons, climate change or religious beliefs. It is difficult to find an area that has similar characteristics to the location the user is moving away from, and this software could help

B.) Suppose a user has a “dream” location in mind and would like to see if something similar is available closer to his/her current location.

The Minimal Viable Product (MVP) of this software will do the following:

- Accept 2 input coordinates
- Retrieve information on various characteristics of around the 2 input coordinates
- Perform hierarchical clustering
- Display the top 10 areas within a predefined range of input coordinate 2 most similar to input coordinate 1

Optional features:

- Allow users to set the range parameter for the 2 input coordinates
    - E.g. Find the top 10 most similar locations within a 10 miles range of input coordinate 2
- Allow users to set the limit for displaying the top *n* most similar locations
- Allow users to specify characteristics which should be considered
    - E.g. There needs to be a school in the are and a Mexican restaurant
- Allow users to provide city and country names as input instead of coordinates

Data

This software will use publicly accessible location data from Foursquare - As prescribed by the capstone project assignment.

For the MVP, no additional data source is required. The user will provide 2 sets of coordinates and the software will retrieve location characteristics data for both locations using the Foursquare API. The data will be parsed, cleaned, trimmed and combined into a data frame. Hierarchical clustering algorithms will be deployed on the data frame, centred around the input coordinate. The output will be a map, using the folium package, that will highlight the top 10 locations around the second input coordinate which are highly similar to the first input coordinate.

The proposed workflow is the following:

1. The user provides 2 input parameters:
    a. @coordinate_1: (latitude: float, longitude: float)
    b. @coordinate_2: (latitude: float, longitude: float)
2. The software calls the Foursquare API and gets data within a 1-mile range of coordinate_1 and 10 miles range grid of coordinate_2
    a. This will yield 101 boxes, each of them 1-by-1 mile
3. The data will be cleaned, selecting useful features and discarding others
4. The 2 data frames will be combined, i.e. 101 rows and *n* number of columns
5. Hierarchical clustering will be performed, centred around the 1-by-1 box of the input coordinate_1
6. The results will be sorted by similarity to the reference row
7. The top 10 similar rows will be plotted on a map of the area around coordinate_2