Applied_Data_Science_Capstone_Part1Q2

August 22, 2019

Peer-graded Assignment: Capstone Project - The Battle of Neighborhoods (Week 1, Question 2)

2.) A description of the data and how it will be used to solve the problem. (15 marks) (revised requirement)

The original requirement was changed from 2.) A full report consisting of all of the following components (15 marks)

The exemplar heading was 2. Data acquisition and cleaning. The first 2 sections seemed to correspond to this week's requirements although the subsequrnt parts did not directly match those specified for the next phase.e

0.0.1 2.1 Data Sources

A JSON file containing New York data (i.e. Borough, Neighborhood, Latitude, and Longitude) will be loaded into a Pandas DataFrame to act as the primary source of data and help facilitate any further processing required. Moreover, a POI csv file will be manually constructed so it can read into a Dataframe. Furthermore, the FourSquare API will be used to get the hotels and other nearby venues which will help analyses of the retrieved data and provide the ability to "drill-down" to help aid the quality of, or yield further, insights.

External websites like LatLong.net (https://www.latlong.net/convert-address-to-lat-long.html) to get the Latitude and Longitude of POIs. And Movable Type Scripts (https://www.movable-type.co.uk/scripts/latlong.html) to help verify that the calculations using Python to compute distances given a start and end set of coordinates (in Latitude and Longitude) and correctly implementing the Haversine formula.

All these will help inform a predictive model.

0.0.2 2.2 Data Cleaning

To be safe, all DataFrames will first be checked for redundancies and any duplicates found will be eliminated. Seeing as the data is relatively "clean" (which was a huge factor for it being chosen), minimal, or hardly any, effort for data cleaning will be necessary (unlike what was done in the previous assignment). That said, columns with categorical data can be transformed using one-hot encoding to enable the use of Machine Learning functions built into Python (since I'm not a statistician). Particular attention needs to be paid to the Neighborhood column because of the sheer volume of different values.

0.0.3 2.3 Feature Selection

Folium maps (or other "meaningful" Data Visualization methods) will be used to represent the contents of the New York city DataFrames and help evaluate candidate accommodations (and their respective Boroughs and Neighbohoods). Moreover, they can be used to help clarify if the clusters "roughly" conform to the 5 major boroughs of New York city (namely Manhattan, Brooklyn, Queens, the Bronx, and Staten Island). This will help determine the relevant columns, refine the area to search, and reduce the number of rows of the dataset (when possible).

The final number of columns comprise the feature set and help determine the dependent, as well as, independent variable(s) . Selection will be influenced by the contents of the FourSquare database. Of interest are which columns correlate with the Haversine formula so that a predictive model can be tried to be built.

0.0.4 2.4 Other Tools

Although a best effort will be made to appropriately incorporate the Data Science technologies covered in the course videos (or used in the labs), if time permits other previous courses will be investigated, if any other tools can help enhance the functionality or analyses that were inadvertently omitted as changing requirements may arise that are not always possible to consider in advance. Particularly, the other ML techniques may be useful.

Note: 1. Feasibility Check. Partial Python code was written to help validate the requirements. 2. External "Eyes". A "reputable" 3rd -party was utilised for input and any suggested modifications were incorporated for clarity. 3. Final Format. A best effort was made to try to combine my personal style, professional exposure, the course requirements, and the exemplar provided (despite being designed for a different type of problem) to produce the text. 4. Specification Differences. The 'Instructions' and 'Submission' tabs of the assignment were slightly different causing me to rereview my hard work prior to submission.

[ ]: