Applied_Data_Science_Capstone_Part1

August 22, 2019

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

0.1 1.) A description of the problem and a discussion of the background. (15 marks)

0.1.1 1.1 Background

New York city is a global destination, cosmopolitan “wonderland”, world icon and, as of late, hosts more than 60 million visitors a year. This is, probably, exclusive of those on “short” business trips. It is, often, difficult for travellers to determine where to stay if they are free to choose. The sheer number of “attractions” may pose logistical nightmares. Given travellers have different tastes and preferences “generic: guides (such as books or the Internet) may be overwhelming (and at the same time,”insufficient“) in suggesting accommodation to try and make the most of, sometimes”limited“, time. It is, therefore, advantageous to have something that will recommend venues that attempts to minimise physical distance to expressed”Points of Interest" (POI).

0.1.2 1.2 Problem

It may prove difficult for travellers to determine where to stay given the “ironic” psychological concept of the “tyranny of too much choice”. Visitors may want to “maximise” their time by selecting locations that are geographically proximate to venues they would like to visit. The primary aim of this project is to help the user plan to stay where the total physical distance of all POIs are considered. This is specifically broken down in the next section.

Aims, Objectives, and Benefits:

1. Provide the neighborhoods and borough of hotels (or their equivalents) with the least total geographic distances. This may help reduce costs by making POIs either walkable or “close by” if transportation need to be used.,
2. Identify the number of nearby venues. This may make a location more “desirable”.

Aside from travellers, this may be of some interest to the New York city tourism board (https://www.viator.com/New-York-City-tourism/New-York-Top-Tours-and-Activities/d687-t30053?semLander=true&m=28353&supag=6517262291&supsc=kwd-355514978618&supai=273084059420&supap=1t3&supdv=c&supnt=nt:g|clk:EAIaIQobChMIlKKwzfWS5AIVmg
cvD_BwE&suplp=9070825&supli=&supti=kwd-355514978618&tsem=true&supci=kwd-355514978618&supap1=&supap2=&gclid=EAIaIQobChMIlKKwzfWS5AIVmg4rCh0nnAnvEAAYAyAAEgL-cvD_BwE).

There may be other benefits that accrue once the final deliverable is submitted but it is essential that certain project objectives be made explicit so it would be possible to gauge the efficacy of the outcome. I have observed a tendency to overpromise and underdeliver to boost the chances of securing a contract. Since the financial pressure does not technically exist, it seems more prudent to offer a realistic proposal rather than have an overoptimistic schedule or an impractical set of deliverables.

0.1.3 1.3 Other Pertinent Factors to Consider

Aside from the benefits, limitations of the project are also enumerated to be pragmatic.

Justifications: Since New York city was used in a previous lab, the investigator can use (or reuse) various informational resources and code to try to build on (and leverage) previous assets. Moreover, my family and I visited New York and it would have been "good" to have had a resource such as this that would have been instrumental in our planning process.

Assumptions: A "heavy cluster" of hotels in an area can be considered a "good" location as this is can be taken as a proxy for proximity to "popular" sights or amenities (as chains often survey the area prior to commencing building). Moreover, an "ideal" place to stay is mainly determined by physical distance.

Constraints: Different categories for accommodation were purely dictated by the classifications in the FourSquare database. Under Hotel are Bed & Breakfast, Boarding House, Hostel, Hotel Pool, Inn, Motel, Resort, and Vacation Rental.

Being a person with a disability, I wanted to include accessibility. However, I noticed the differences in nomenclature in my travels to America and Japan. Despite also being developed economies, the facilities were not as accommodating as Australia and the terminology used often exhibited differences. While a "good" idea, it was impractical and difficult to incorporate this aspect.

While the sharing economy apps (like AirBnB) are not currently supported (as far as I'm aware), future iterations may also consider these non-traditional forms of dwelling.

The outcomes for next week's set of requirements also need to be considered so the alignment of outputs can be consciously taken into account. This not only serves to guide the validity of the development but can act as a potential roadmap to possible future directions or entirely separate branches.

Predefined (later on allowing users to specify their own interests) POIs will allow for the amalgamation of distances to various lodgings (i.e. the total distance from the shelter to place they want to visit). While not a full-blown recommender system, it can suggest various accommodations (factoring in the "nearest" Boroughs and Neighborhoods of New York city) given their proximity to certain venues. Regardless of whether they are potential tourists or are on a short stay business trips, it endeavours to provide users with places to stay to try to get the most out of their visit.

Software Development Issues A Jupyter notebook stored in a Skills Network lab was purposely chosen. This was explicitly done for several reasons: 1.) This was the format used in the course and familiar to peer reviewers. Moreover, it has the benefit of being "submit ready" and could easily be exported to a text format., 2.) This allowed for the flexibility to use Python code to test concepts (as it is more important to determine feasibility and specify requirements)., and 3.) IBM's backup regime is probably more comprehensive than what I personally have set-up. Despite the requirements being mainly text-based (and formats like MS word, .txt, .rtf, or .pdf would have sufficed), further validation and the importance of feasibility became more crucial given the “shorter” time frame. Based on my prior, personal experience, 3 months is often a reasonable period for software development. Given I have only 2 weeks to complete the capstone project (aside from my other commitments) a Design-to-Schedule methodology (that is, given a fixed deadline do your best to deliver what is possible) will be used. The result is a “Proof of Concept” prototype - typically the practice is be prepared to throw the first one away.

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

0.1.4 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.

0.1.5 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.

0.1.6 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) . Of interest are which columns correlate with the Haversine formula so that a predictive model can be tried to be built.

0.1.7 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 to produce the text.