Coursera IBM Data Science Professional Certification
Coursera Capstone Project

Instructor: Alex Aklson

Submitted by: Luis Lopez

Table of Content

Contents

Capstone Project – Week 1 Part 2

Assignment Requirements.

Part 1. Submit a description of the problem intended to solve and a discussion of the background information.

Part 2. Clearly define a problem or an idea of your choice, where you would need to leverage the Foursquare location data as main resource to solve the problem.

The development of this capstone project will proceed along the guidelines and procedures in the methodology presented in course 3 “Data Science Methodology”of “Coursera IBM Data Science Professional Certification” specialization.

This methodology is driven by 10 methodology questions that provide context to the 10 development phases that a typical data science project should adhere in the path toward successful completion.

Assignment Part 2: Data Section.

This part of the assignment deals with the following crucial questions of the data science methodology:

Question 3: What data do you need to answer the question?

Question 4: Where is the data coming from and how will you get it?

Question 5: Is the data that you collected representative of the problem to be solved?

Question 6: What additional work is required to manipulate and work with the data?

2.1 Data Requirements Phase.

Question 3: What data do you need to answer the question?

By definition of the assignment we will data mine Foursquare location data.

However there are ancillary data requirements; such as:

A. Sydney_Suburb.csv. A file containing suburb_name, postcode, latitude, longitude. One row per suburb.

B. Sydney_Restaurants.csv. A file containing restaurant_name, restaurant_id, latitude, longitude.

C. Sydney_Restaurant_Clusters.csv. A file containing restaurant_name, restaurant_id, postcode, cuisine_cluster_id, demografic_cluster_id, atmosphere_cluster_id.

2.2 Data Collection Phase.

Question 4: Where is the data coming from and how will you get it?

A. Foursquare location data will be access via API program call.
B. Sydney_Suburb.csv file. This ancillary file will be obtained from data mining the web. This file will include very few rows.
C. Sydney_Restaurants.csv. This ancillary file is generated by looping over the Sydney_Suburb.csv rows and accessing Foursquare using, suburb's postcode and searching by "Restaurant".
D. Sydney_Restaurant_Clusters.csv. This ancillary file is generated by clustering algorithms. There are 3 clustering process: one for Cuisine, one for Demographics, and one for Atmosphere.

Important note on the 3 clustering processes: A metric, or measure similarity, needs to be define for clustering by Cuisine, by Demographics, and by Atmosphere (or Ambience).

2.3 Data Understanding Phase

Question 5: Is the data that you collected representative of the problem to be solved?

According to the Data Science Methodology, the Data Understanding Phase encompasses all activities related to constructing the datasets and/or ancillary files.

The clustering by Location was used in one of the labs of this course, where the city of New York and the city of Toronto were partition using the postcodes and finding their longitude and longitude by using the Nominatin module of the geopy.geocoders package.

For clustering by Cuisine it is proposed to use the number "tips" to calculate the "distance between two restaurants". That is, two restaurants are close thgether if their number of "tips" are similar numerically.

Same metric can be used for clustering by Atmosphere.

The metric for clustering by Demographic could be defined after investigating further the 'gender' key of the Foursquare user's information.

This phase is still early stage in the project. An important objective of this phase is also identify early feasibility issues. Por instance, it's possible that some of the client requirements can not be met or can be postpone for a future developments.