Programming Assignment 6 - Hierarchical Clustering

Data Visualization (CSE 578)

As in your previous assignments, the administrators of the Dino Fun World theme park, you must have the data analysis in order to help them administer the park. In this case, your task is the same. In a priori, you were asked to find the distance between a set of visitor trajectories using a simple edit distance algorithm and report the distances. For this task, you must construct and display a dendrogram of those distances. Again, the administrators of the park.

Provided Database

The database provided by the park administration is formatted to be readable by any SQL database library. The course staff recommends the sqlite3 library. The database contains three tables, named 'checkin', 'attractions', and 'sequences'. The information contained in each of these tables is listed below:

`checkin`:

- Description: check-in data for all visitors for the day in the park. The data includes two types of check-ins, inferred and actual checkins.
- Fields: visitorID, timestamp, attraction, duration, type

`attraction`:

- The attractions in the park by their respective AttractionID, Name, Region, Category, and type. Regions are from the VAST Challenge map such as Coaster Alley, Tundra Land, etc. Categories include Thrill rides, Kiddie Rides, etc. Type is broken into Outdoor Coaster, Other Ride, Carussel, etc.
- Fields: AttractionID, Name, Region, Category, type

`sequences`:

- The check-in sequences of visitors. These sequences are the list of the positions of each visitor to the park. If the visitor has not entered the part yet, the sequence has a value of 0 for that time interval. If the visitor is in the park, the sequence lists are the most visited.
- Fields: visitorID, sequence

The database is named 'dinofunworld.db' and is available at read only / dinofunworld.db.

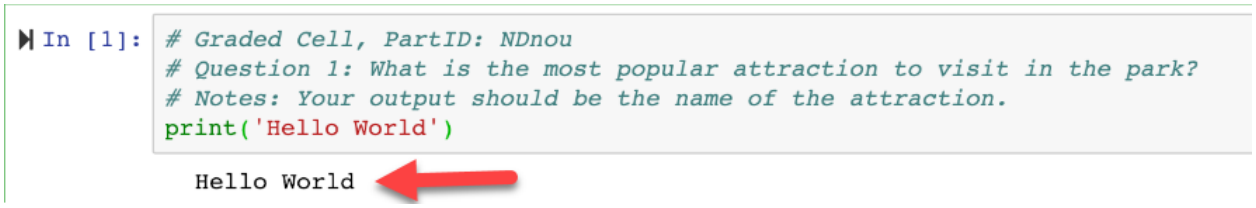
Assignment

This task consists of only one question, which will require you to generate a dendrogram graph. Create this dendrogram using the trajectories of the visitors with the IDs: 165316, 1835254, 296394, 404385, and 448990. If you are unsure about how to create a dendrogram, please refer to the Jupyter Notebook example which is creating a dendrogram. When performing clustering over the trajectories to inform the dendrogram, use an average distance over all points in the cluster.

Instructions

Coursera's grading system. In order for your answers to the question, you must have the code for each question. In addition, you should submit the assignment in the cell's display area. Contain no extraneous information. Each cell that is going to be graded has a set of comment lines at the beginning of the cell. These lines are extremely important and must not be modified or removed.

Please execute each cell in Jupyter Notebook before submitting.



Feedback

There is one part in the grading with a total of 1.0 point. If some part of your data is incorrect, you will get a score of 0.0. If the submission failed, we will return the corresponding error messages. If the submission is correct, you will see "Correct" with 1.0 point for the part.