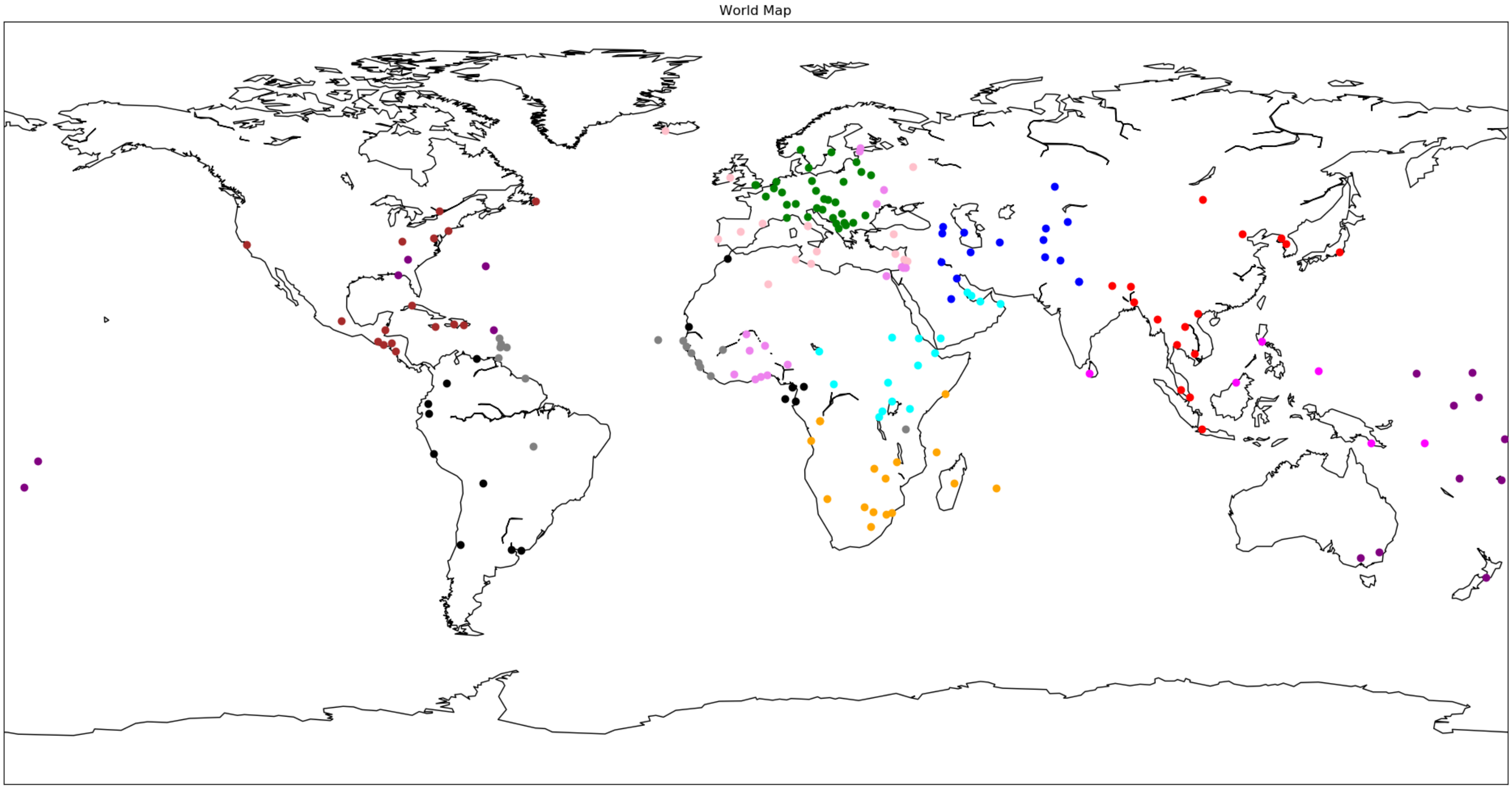
BIG DATA ANALYTICS (CSCI -720)

Homework-08 **Name: Rajkumar Lenin Pillai**

The program submitted has the functions to plot map and download latitude and longitude as comments.

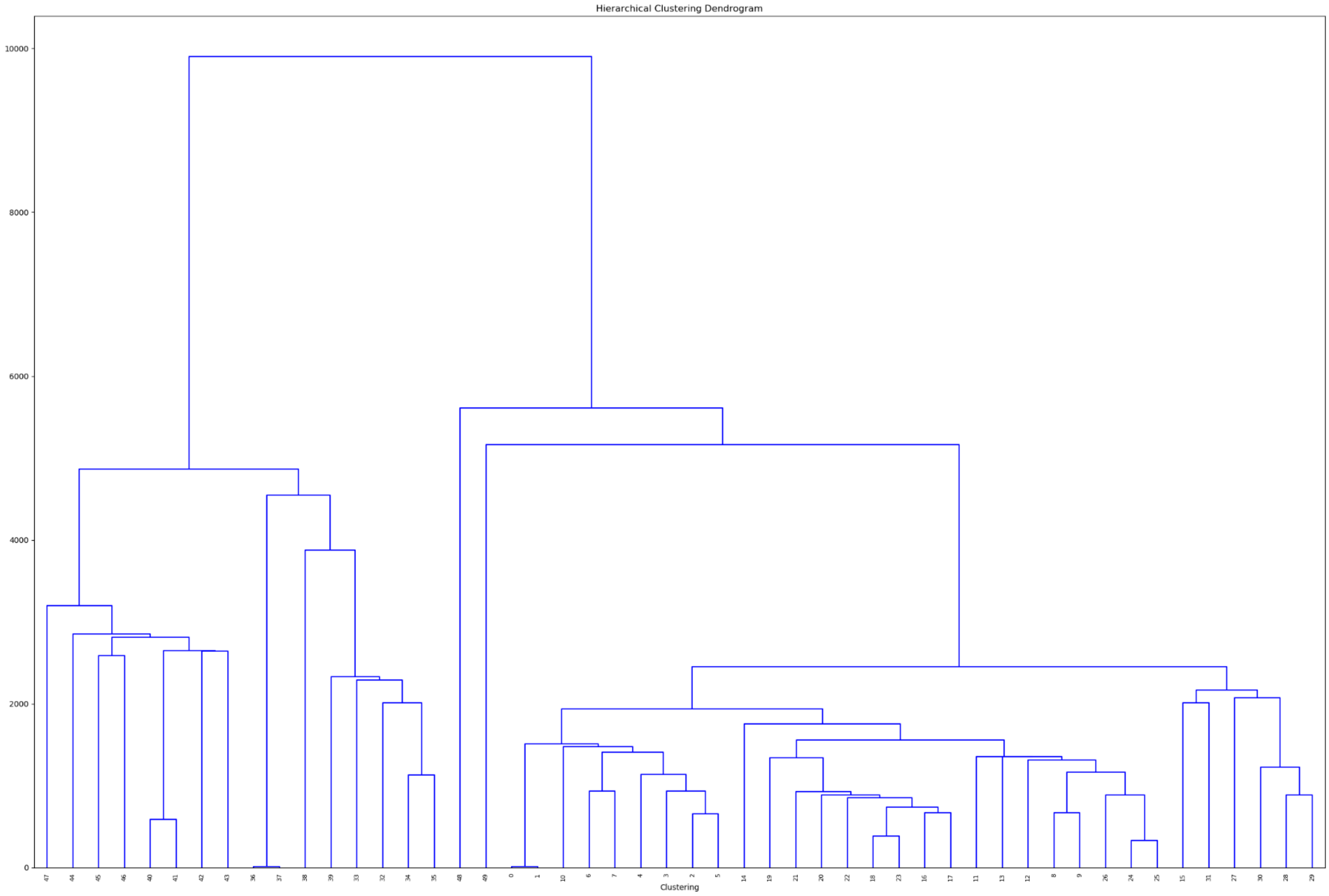
Question-a.)A map of the world, with the capital cities plotted on it. This means you need to figure out how to generate this. (See the python plotting documentation.) The cities should be clustered into 12 clusters. Each cluster should be plotted in a different namespace color (there are 12 of them). Use agglomeration, with the single linkage method, and the Haversine distance. (3/10)

Solution:



Question b.) A dendrogram showing the top 50 clusters. (3/10)

Solution:



Question c.) Who did What? How do what you did? Was anyone on quality assurance? How did you divide the tasks? What tasks where there?

Solution:

This assignment was done by me. I wrote the entire code for agglomerative single linkage clustering. The task was finished in a timely manner and it took 20 mins to write the code to download coordinates using 'Nominatim'. The entire Assignment took 8 hours to complete. Installation of packages took some 10 mins for each package. The tasks were first to download the co-ordinates. Then deciding the data structure to store the clusters which is a dictionary in this program. Then designing the algorithm which took 4 hours in which different cases like merging of clusters and deleting and iterating through cluster which is merged had to be taken care of in the python prgoram so that the algorithm doesn't consider it for next iteration. For the stopping condition which is no of clusters ,the length of keys in the tree is checked whether it is 12 which indicates there are 12 clusters. Plotting the dendrogram was done using the scipy package where the input to linkage function are the top 50 clusters. To understand and plot the dendrogram took 30 mins. Plotting the world map took one hour as it needed some packages and the points to be plotted were supposed to be in same color.

Question d.) Write a conclusion showing what you learned, and that you learned something. What issues did you face? What went wrong? What worked easily? Provide strong evidence of learning. Provide references as needed.

Solution:

I learned the agglomerative algorithm which just describes its first step as merging clusters with minimum distances actually requires four for loops in which every point must be considered and then the next points other than the one considered. The issues faced was in designing the algorithm. Since considering 2 clusters and merging them in one outer for loop causes problems in the inner for loop because the merged elements cannot be considered one more time in that iteration. To avoid this the condition were checked after the first for loop and second for loop which indicates if the clusters are merged or not.The downloading of the latitude and longitude worked easily with not much effort required other than understanding what are the values returned by the geocode function. Also I learned that the four for loops considered cannot be written in any fashion as the first outer for loop considerS the initial clusters then inner for loop for the next clusters and then inner for loop for the points of initial cluster and then innermost for loop for points of next points

Reference :

To downloading Latitude and Longitude: https://pypi.org/project/geopy/

To plot dendrogram:https://joernhees.de/blog/2015/08/26/scipy-hierarchical-clustering-and-dendrogram-tutorial/

Prof. Kinsman's Slides