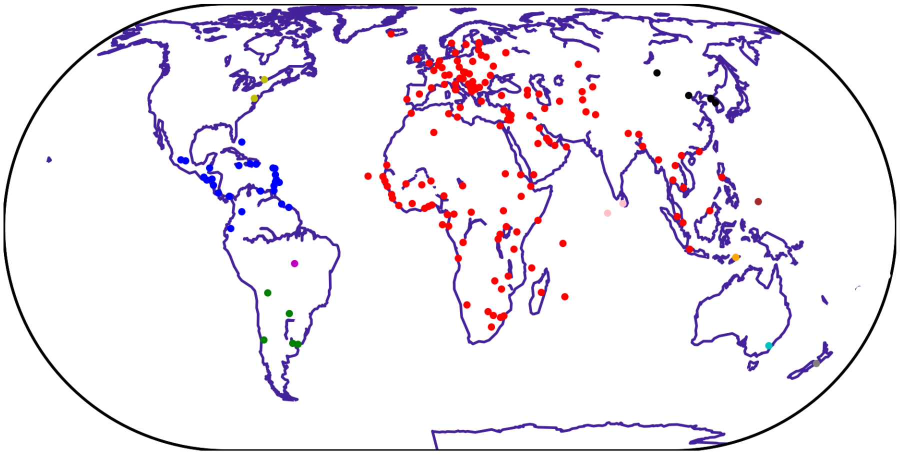
Homework 08

Erikson Sodergren & Justin Lad

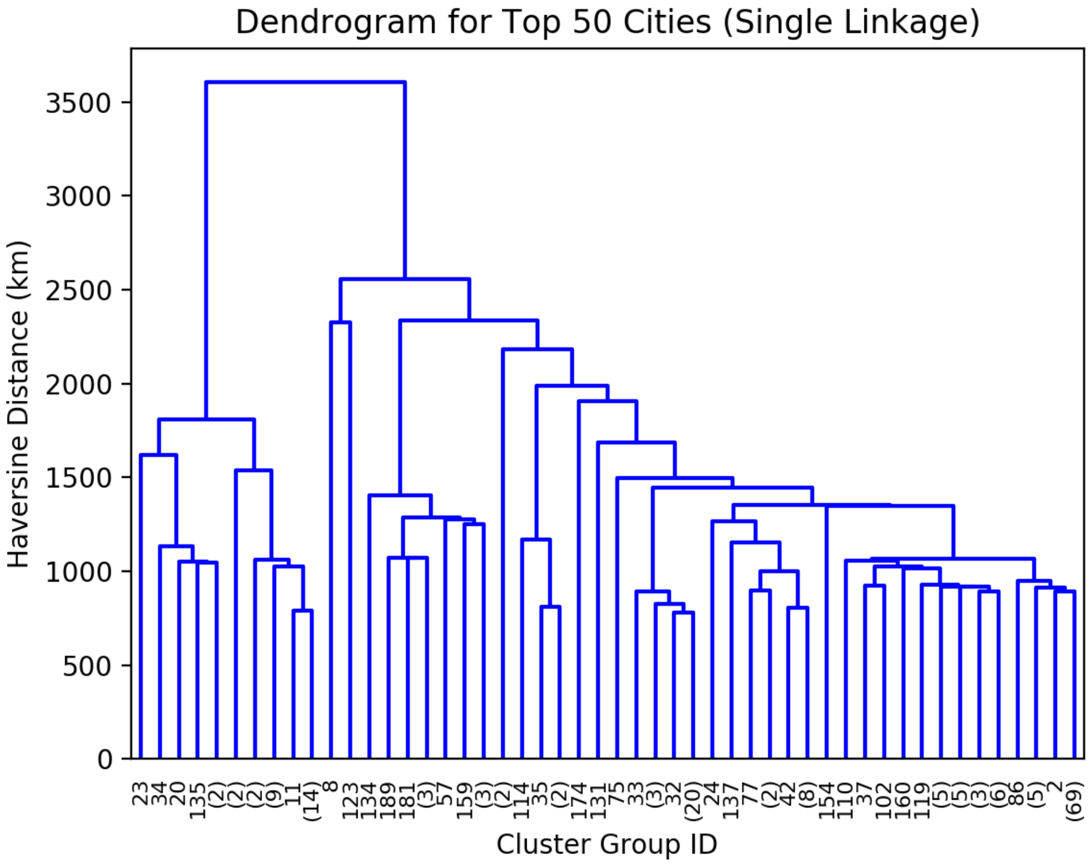
CSCI 720 — Big Data Analytics

March 31, 2019

A. Map



B. Dendrogram



C. Work Breakdown

We split the work by breaking the assignment into components and individually completing components. Required tasks break down into:

- Convert city,contry to lat,long (Justin)
- Agglomeration function (Erikson)
- Plot clusters onto worldmap (Justin & Erikson)
- Create Dendrogram (Justin)
- Verify results (Justin & Erikson)
- Write report (Justin & Erikson)

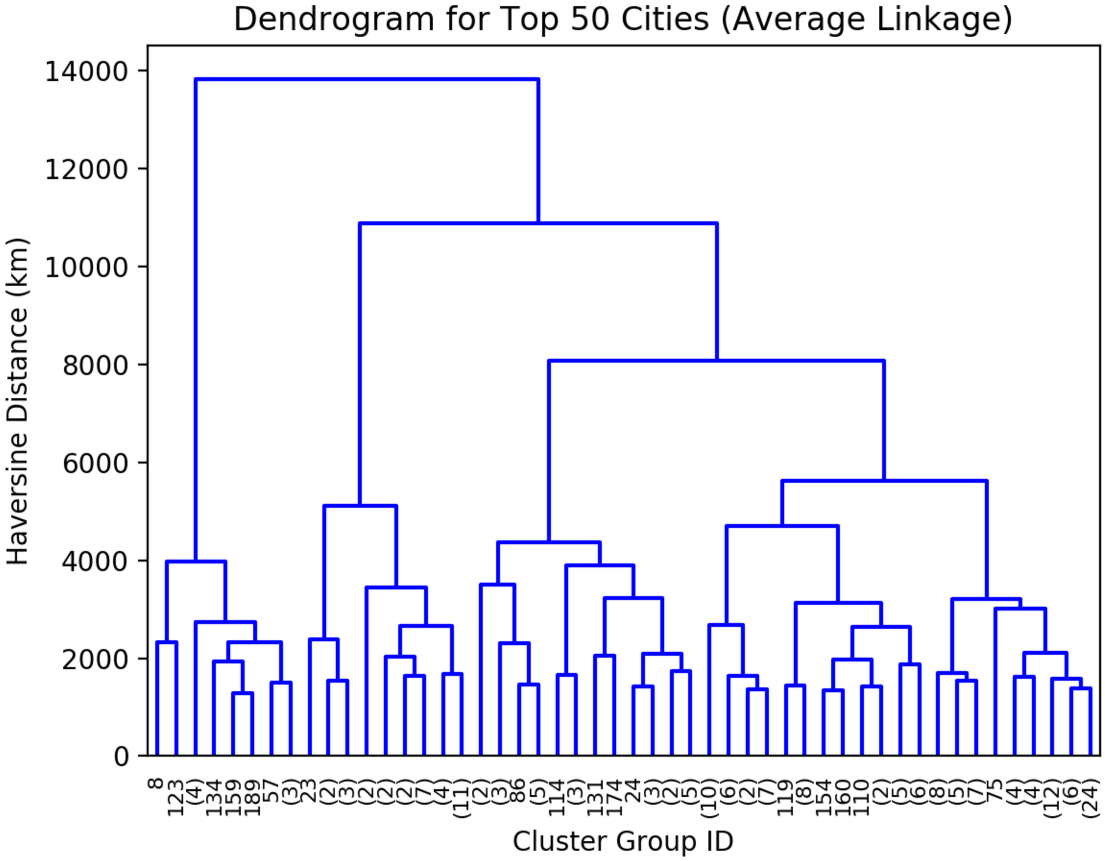
Justin wrote the function to create a new csv document with lat/long instead of city/country. He was assigned the task of plotting the output of the agglomeration algorithm on the map.

He did research on this task and found the basemap package and code to achieve the desired functionality, but had an error running the basemap package on his computer. Erikson was able to get it installed, so he wrote the map drawing function. Erikson wrote the agglomeration function, plotted the color-coded points on map, and helped with issues Justin had.

Erikson primarily acted as quality assurance, making sure that Justin delivered what he said he would. We independently compared results from the dendrogram and map, and concluded that the agglomeration function works properly. The dendrogram plot has a very large cluster on the right part of the chart, and merges into a very large cluster when the dendrogram gets to 12 clusters, which is reflected by the large cluster. The results weren't what we were expecting, but after verifying with the dendrogram, we felt confident it worked correctly. We acknowledged that the large cluster is due to the linkage method. In the conclusion, we show average linkage, and how the cluster sizes are more uniform.

D.Conclusion

Through this assignment, we learned and saw how the agglomeration procedure works. We note that the selection of linkage type is very important when using agglomeration. Single linkage resulted in poor performance, shown in the map above. There is a very large single cluster in Africa/Europe/Asia, while there are 6 singleton cluster points. We would hope to see slightly more Using average linkage, clusters are much more evenly distributed by size.



Learning the scipy vs sklearn libraries to produce the dendrogram was more frustrating than I anticipated. For example, to plot the top 50 cities, I was using the ‘level’ command in truncate mode. It made sense at the time, but when I revisited the documentation the next day, it was obvious that ‘lastp’ is the truncate mode selection to get p starting clusters. Additionally, using haversine distance as the distance metric for the dendrogram was a little tricky at first. The linkage object accepts distance.pdist distance metrics. The distance.pdist library has 22 standard distance functions, not including haversine. However, Erikson pointed out that you can pass a function as the pdist metric.

Getting basemap working correctly took a little fiddling, notably the clusters were originally being plotted underneath the continent outlines before i found the ’zorder’ argument for plotting. Additionally, it was initially worrying to me that the final 12 clusters had such a lopsided form compared to the example on the assignment description, but after extensive debugging and comparison with the dendrogram, we concluded that the fault lies with using single linkage.