Hierarchical Clustering

Project #6

Due: Wednesday, April 5th

For this assignment, we will be using the **UrbanGB** dataset located in the University of California at Irvine's Machine Learning Repository located at:

https://archive.ics.uci.edu/ml/index.php

This website hosts a large collection of datasets used by machine learning researchers. You should use the **UrbanGB.cntr** text file as your input. Treat this file as a list of Cartesian coordinates.

You are tasked with designing a program that creates a bottom-up/agglomerate clustering of **UrbanGB.cntr**. Your program must use group average linkage. Your program should output (at every step of building the hierarchy) the numbers of items in the first cluster, the number of items in the second cluster and the distance between these two clusters before they are joined.

You must author your own code using either the C++, Java or Python programming language and without the use of *external* software packages. Your project submission will be run through a source code verifier against previous submissions to check for plagiarism. Any evidence of cheating (CODE NOT WRITTEN SOLEY BY YOU) will result in a failing grade for this course. Treat this seriously. I do!

When submitting your assignment include:

- the output file
- a copy or your source code
- a README file containing any information required to run your program.
- A file containing your interpretation of the clustering. How many clusters are there and why? Are there outliers?

Feel free to consult any relevant literature on clustering algorithms.