I implemented Lloyds Method to calculate the k-means value and the complete linkage algorithm. After I read in the input, I begin by finding whether the maximum number of possible configurations is less than 100. In the event that n choose k is less than 100 then that is the maximum number of possible initial configurations. Otherwise, I begin running it with 100 possible initial configurations, chosen randomly. I have a cluster point object that stores some property data of each individual data value that was read in and the cluster object which is used as a parent container for the cluster points. After initializing the points and clusters, I begin by moving nodes into the clusters and recalculating the centroids. This continues until convergence or in other words no turns are made consecutively. Lloyd’s method has continued calculation of the distance to the new centroid which can slow down its run time. The complete linkage data initializes the cluster and cluster point objects. This algorithm precomputes the distances to each other point. The algorithm then runs until the required k is met. It begins finding the maximum distance of each point in a cluster and takes the minimum one. Then it combines them and continues until there are k clusters. In these data sets, data sets 1 and 2 were creating the same clusterings and the data set 3 was creating a clustering pair that had a hamming distance of roughly 12. The times for each of these runs are listed in the chart below.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Lloyds Method Time (ms) | Complete Linkage Time (ms) | Hamming Distance |
| Data Set 1 | 140.61 | 125.00 | 0 |
| Data Set 2 | 149.62 | 31.23 | 0 |
| Data Set 3 | 500.05 | 367.94 | 12 |

My program was created in Python using version 3.4. The README file will indicate to run the program on linprog using the run command “python3.4 ass4.py”. The program will then prompt the user for an input file. This is where data1.txt, data2.txt, or any other test file can be entered. The output will show the Lloyd’s method and the Complete Linkage runs with their times. Choosing which one you would like to use depends on how you want to view the data. The user should be evaluating how they would like to cluster the data then choose which algorithm they would like to use to do so.