# Assignment 4 Report: Data Clustering

Annelise Wittenberg & Cody Stammer

([wittenberg.18@osu.edu](mailto:wittenberg.18@osu.edu); [stammer.3@osu.edu](mailto:stammer.3@osu.edu))

## Approach

For this assignment, we chose to import the Orange Library in python to write a program that clusters the dataset. We chose k-means and hierarchical clustering as the algorithms, and as our distance metrics, Euclidean and Manhattan. Initially, we attempted to use the Orange GUI but were limited by it as it couldn’t give us the definite values that we needed. We also compared results to Weka, from which we were able to get good visual results.

Using the functions in the Orange library made it possible to run the entire data set using less RAM than in the classification assignment.

Furthermore, we removed any articles with a topic label that occurred 3 or less times in order to have a better clustering efficiency and fewer outliers. It especially helped in determining the number of clusters for the dendrogram.

## Results

The table below contains data for our runs on the entire dataset.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Hierarchical**  **Euclidean** | **Hierarchical**  **Manhattan** | **K Means** |
| **Time** |  |  |  |
| **# Clusters** |  |  |  |

## Issues

We first ran into issues when deciding what methods to use for the clustering algorithms. Weka has a good interface for showing timing of each run, and the distributions of the clusters, but does not show plots and dendrograms well. Orange has a very easy GUI to use, but it does not provide any kind of visual representation for k-means clustering results. When we implemented the lab using the Orange library for python, we were able to get better visual results. Weka, however, provided the best information about the clusters themselves, providing in table format the actual breakdown of each cluster. Weka also provided information about timing of the algorithm, which Orange did not. To solve the issue of timing, we decided to use the CPU clock timer to calculate the time it took for each of our coded clustering algorithms to run.

## Assumptions

An assumption we made for the purposes of this assignment was that everything we have done up to this point was done correctly. For example, when we wrote code to create the distance matrix and do the clustering, we assumed that the way we implemented it, based on the results we printed out, was correct. We also assumed that the Orange and Weka implementations of hierarchical and k-means clustering are relatively accurate. Our assumption was backed by the relatively consistent results we found across multiple test runs.

## Work Distribution

Clustering Algorithms - Annelise

K Means Plot - Cody

Dendrogram - Annelise

Report – Annelise and Cody

Testing - Cody