David Strickland

CS5402 Intro to Data Mining

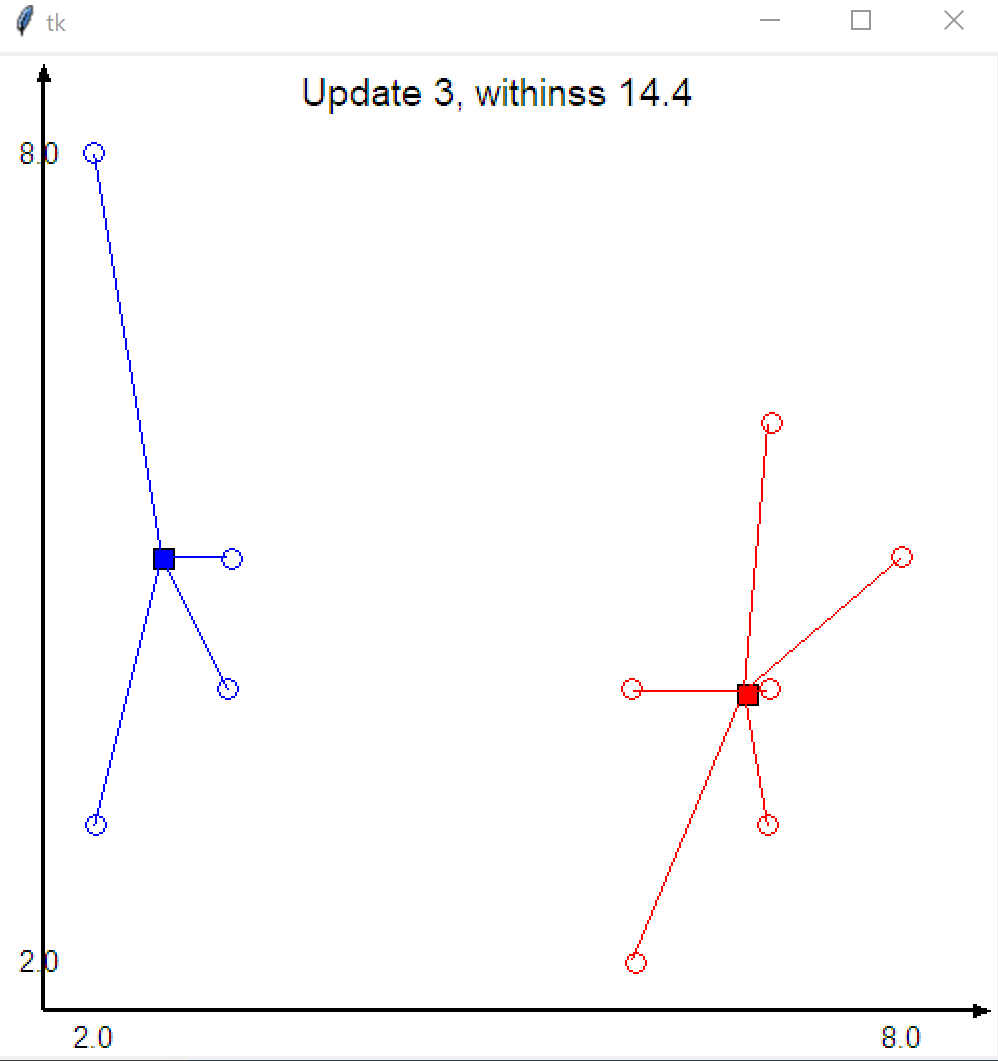
Homework 4

Python file located at the following github repository: https://github.com/ds5b4/CS5402-IntroToDataMining-Fall2018/tree/master/HW5

1.1.

Centroid0 = (2.50, 5.00)

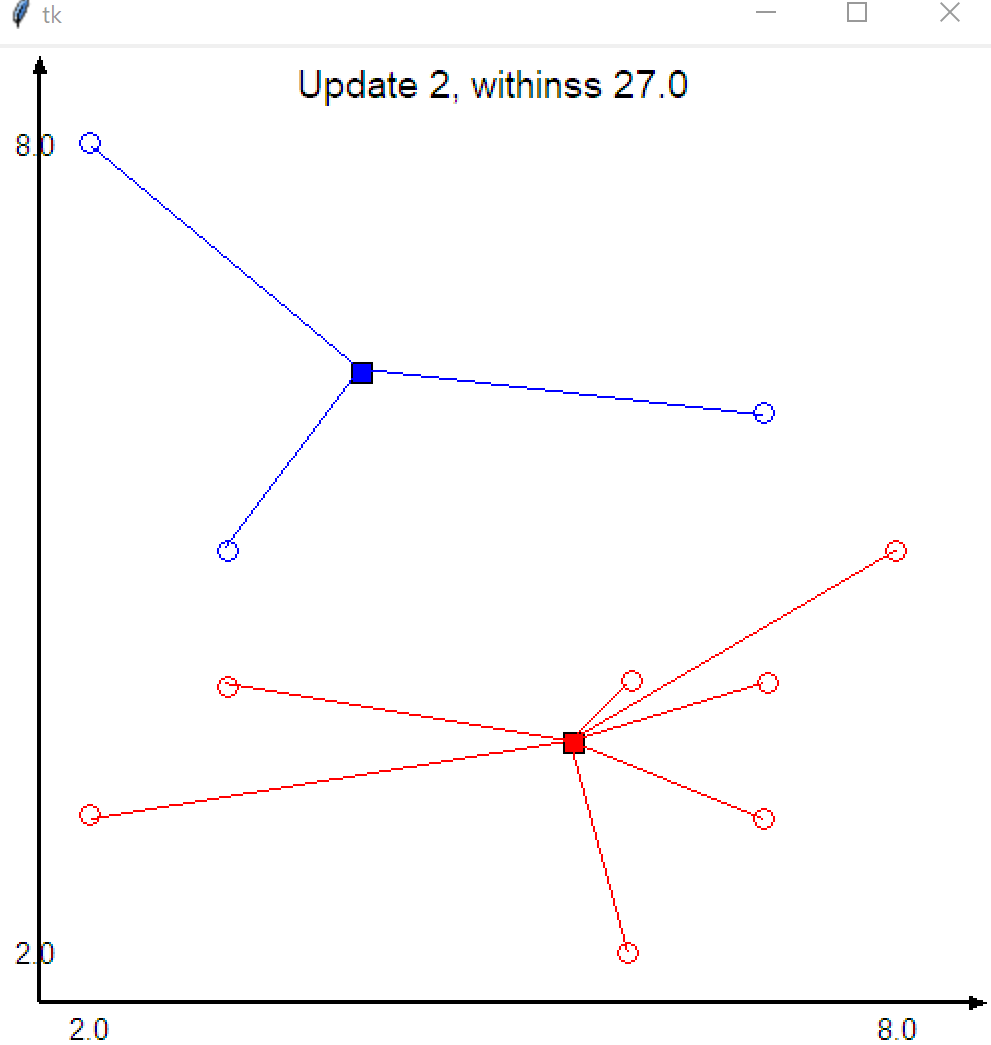
Centroid1 = (6.83, 4.00)



2.

Centroid0 = (4.00, 6.33)

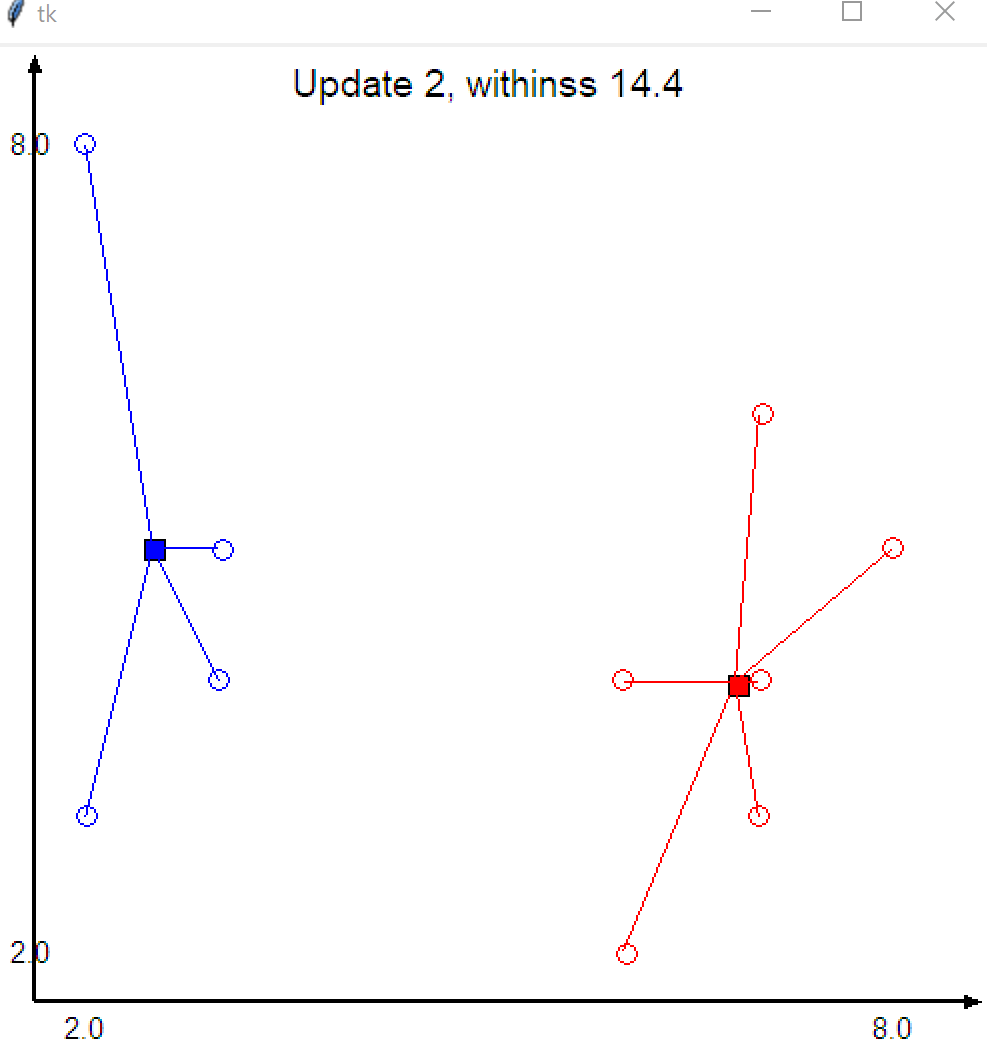
Centroid1 = (5.57, 3.57)



3.

Centroid0 = (2.50, 5.00)

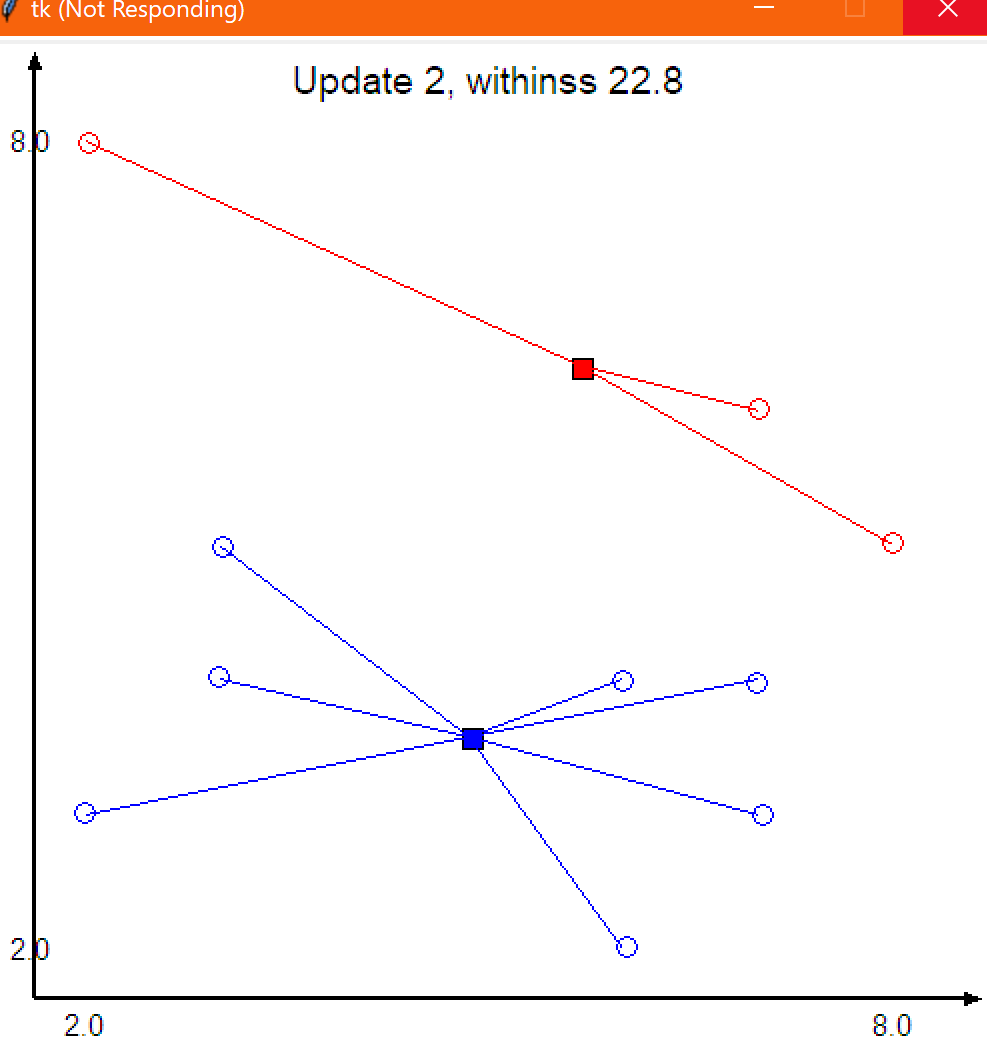
Centroid1 = (6.83, 4.00)



4.

Centroid0 = (4.86, 3.57)

Centroid1 = (5.67, 6.33)



Using Euclidean distances can lead to local maxima which may result in the same centroids being found even when different initial centroids are given. As these are local centroids, it is still possible to get different final centroids depending on the initial centroids as seen in 1.4. Manhattan distance tends to not find the best maxima compared to euclidean distance as seen with 1.1 and 1.2.

2. 1.

Euclidean SSE, Cluster 0: 19.938723404255317

Euclidean SSE, Cluster 1: 18.864905660377364

Euclidean SSE, Cluster 2: 1.42

Euclidean SSE, Cluster 3: 3.4114285714285715

Cosine SSE, Cluster 0: 5.619941837855495e-05

Cosine SSE, Cluster 1: 7.190338788685357e-05

Cosine SSE, Cluster 2: 2.2281090202011664e-05

Cosine SSE, Cluster 3: 0.000348044466785213

Jaccard SSE, Cluster 0: 0.21867083042872582

Jaccard SSE, Cluster 1: 0.5446471048234596

Jaccard SSE, Cluster 2: 0.1625300989777855

Jaccard SSE, Cluster 3: 0.18060062548094147

Jaccard distance seems to be the best of the methods as it has errors consistently under 1 where the other two also have at least one cluster with an error over 1.

2.

3. Cosine requires the most iterations and thus time as the changes made using cosine distance are smaller and thus to reach a maximum it takes more iterations.

4.

3. K-means is the attempt to group like items based on their properties. These properties must be able to be expressed in terms of either numbers as distance metrics are used to measure their proximity from one to another. The center of the clusters are called centroids and this is the point which all of the data points are measured off of. A data point is assigned to a cluster when its distance to either the centroid, nearest point of a cluster, or furthest point of a cluster is less than the same distance to every other cluster. Clusters can merge when they overlap. K-means is easy to implement and generally faster than hierarchical clustering when k is small. It is hard to predict the correct number of clusters and initial values can widely sway the results as seen in 1.1 and 1.4. This can be helped by visually selecting good centroids rather than having a randomly assigned initial centroid.