Assignment 1

1. Collect any dataset without decision attribute. The dataset may be collected from UCI machine learning repository. If the dataset contains the decision attribute then remove it as you will perform clustering of objects in the dataset.

2. Let the dataset has *m* rows and *n* columns where, each row is an object and each column is an attribute or feature of the object in the dataset. So you can consider the dataset as an $m \times n$ matrix.

3. Normalize the attribute values within the range [0,1] using any normalization technique to give all the attributes an equal importance.

4. Create a similarity matrix*S*of size $m \times m$ where, each (*i*, *j*)-th entry in the matrix gives the dissimilarity measurement between *i*-th and *j*-th objects. Use Eucledian distance to measure the dissimilarity

5. The *i*-th row indicates similarity of *i*-th object with all other objects. Find the average dissimilarity of *i*-th object with other objects and form a cluster C_i with *i*-th object and objects having dissimilarity less than the average similarity. Repeat this process for all rows of the similarity matrix. Thus, you have now *m* clusters.

6.Remove the clusters (if any) which are subset of some other clusters. As a result you have now say, *p* (<*m*) clusters.

7. Create a similarity matrix *C* of size $p \times p$ where, each (*i*, *j*)-th entry in the matrix gives the similarity measurement between *i*-th cluster C_i and *j*-th cluster $C_j$using following similarity measure.

$$c_{ij} = \frac{|C_i \cap C_j|}{|C_i \cup C_j|}$$

8. Out of all p^2entries in matrix *C*, find out the maximum value. If multiple maximum values occur, choose any one randomly. Let, c_{kl} is the maximum value selected, that implies clusters $C_k$and C_l are the most similar clusters among all *p* clusters. Merge these two clusters $C_k$and C_l to get a new cluster C_{kl}, i.e. $C_{kl} = C_k \cup C_l$.

9. Repeat steps 6 to 8 until desire number (say, at most K) of clusters are obtained.

Note: You will get the overlapping clusters. Next find the probability of an object to be in all the clusters. For this, you may compute the similarity of it to the mean of the cluster in which it lies. Set similarity of it to a cluster, in which it does not lie, as zero. Also put the object into a single cluster to which the similarity is maximum. Break the tie arbitrarily.