There are four datasets from different sources. The datasets are from different points of view such as

1. geometry (centroid_edges.mat)
2. density (points.mat),
3. image (binaryalphadigs.mat)
4. text (20news_w100.mat).

Even though the datasets have ground truth labels. The task is to perform clustering operations using the datasets. There is some kind of uniqueness in all the datasets. It is up to the user to identify the uniqueness which is common among all the datasets.

Evaluation Criteria:

1. Data visualization
2. Using principal component analysis (PCA) for real datasets namely binaryalphadigs.mat and 20news_w100.mat
3. Implementing different clustering techniques such as k-means, KNN, k-medoids, min-max cut clustering, DBSCAN or any other hierarchical or Partitional clustering.
4. Identification of number of clusters in each dataset (simple – but has to identify automatically through clustering metric).
5. Using clustering metric to perform computation.
6. Evaluation using ground truth labels.

Note: These are the key pointers against which we will evaluate your solution. Your solution may cover the above mentioned points but do not restrict yourself just to these.

Recommended Time: 4hrs