P3: Unsupervised Cluster

Purpose:

The purpose of this project is to familiarize ourselves with unsupervised learning from data. For that we are asked to implement K-mean clustering Algorithm, Bisecting K-mean Algorithm and use DBSCAN. We are asked to plot the SSE vs no. of clusters. The purpose of this project was to see the change of SSE in different algorithm.

Dataset:

The 20-newsgroups dataset was used in this project. It was imported from sklearn.dataset and fetched using the function fetch_20newsgroups. The arguments passed for this were (subset = "train", remove =remove, shuffle=True). Total of 11314 data were read.

Preprocessing:

For preprocessing, 11314 read data was filtered by top 10 news groups. After filtration 5954 data was left. Bag of word approach was used for training the clustering. First all the messages were taken one by one, then symbols, stop words were removed. It was then lemmatized. Then a list of words was generated for each message. These lists were then extended to a super list called all_words. From this list the frequency of each words was processed, and a feature was extracted with no of features 5000. It was done to reduce the no of computation. To extract the features, TD-idf weighting system was used.

$$W_{t,d} = (1 + \log tf_{t,d}) * \log N/df_t$$

The generated matrix was then normalized so that cosine similarity can be applied

Kmean algorithm:

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

For Kmean first K was assigned. Then k centroids were assigned randomly. For this the random extracted features were used. Once the centroids were generated, then all the data were labeled closest to the centroid. For this cosine similarity was used, where

$$\cos(A, B) = \sum_{i=0}^{n} \mathrm{A[i]} * \mathrm{B[i]}$$

After first clustering, the centroid was recomputed to the mean of the values in that cluster. The process of clustering is repeated by checking the cosine similarity of the value and data values. Once there is no change in labeling of clusters, then the final cluster is generated. The value of K was selected to be 4 after experimenting with other numbers.

SSE was used as an internal measure to check the clustering. It was observed that after the final cluster is generated the SSE decreased compare to the SSE of 1st iteration of clusters.