Text Clustering

Name: Yamini Aalla **Student Id:014331018**

Text clustering is used in various applications. It binds alike documents, (tweets, reviews which can be used for sentiment analysis). The unsupervised algorithms used are K-Means clustering and Hierarchical clustering.

Bisecting K-Means Clustering: The limitations of K-Means clustering can be overcome by using bisecting k-means clustering which can recognize clusters of various size and shape and solve issue of local minima.

Bisecting K-Means Clustering is used to solve the given problem.

Approach:

1.Convert the given text file into sparse matrix
2.Applying K-Means clustering where k=2
3.Implementing Bisecting K-Means by taking sse of two clusters and diving the cluster with higher sse value.
4.As per given requirements divide clusters till, they form 7 clusters.

Data Preprocessing: I have converted the given train.dat data file into CSR matrix format using csr_matrix. The matrix has 8580 rows and 126,356 columns and performed TFIDF on the matrix to get weightage for all the words used in the document. Where used normalization of l2 norm.

For dimensionality reduction used truncated singular value decomposition (SVD) with n_components = 300 and algorithm = ‘arpack’.

Clustering:

Finding initial centroids by shuffling the points and picking two points as centroids. clustering the points closest to these two initial centroids.

For stopping condition consider the number of iterations as 20. Recalculate the centroids by taking mean of the all the points in cluster and reform the clusters for each iteration.

Bisecting k-means:

For the clusters formed, calculate the sum of square error (that is from centroid to the points in cluster) for both clusters and add the cluster with minimum sse to the cluster list and divide the cluster with maximum sse. Stop bisecting when requirement is met the that is when number of clusters=7.

Pseudocode:

For size of clusters k<7:

1.Assign k=2 for k-means clustering.

2.Assign random initial centroids.

3.Generate two clusters.

4.Recalculate centroids by taking mean.

5.Form the clusters and repeat till number of iterations given.

6.Calculate sse values and For minimum sse value add the cluster to clusters list.

7. For maximum sse value cluster repeat steps 4 to

6.

Results:

Implementing Bisecting K-Means algorithm and plotting K values on xaxis and calinski_harabaz_score on y-axis.