PR1: Text Clustering

Anuhya Gankidi,
015897323

Objective of this programming homework is implementing bisecting k-means algorithm to cluster text documents as provided in the format of sparse vector matrix.

Approach:

1. The input data containing 8580 text records are read and these documents are converted into sparse matrix.

2. This CSR matrix is then scaled by sklearn's TF/IDF and further normalized by its L2-norm.

3.Then SVD Matrix factorization is applied on the data matrix with 500 components: After scaling the matrix by IDF and L2 Normalization, SVD is applied on the data matrix for dimensionality reduction.

4.Over this bisecting k means is to get data into desired number of clusters (7). The clusters with the lowest similarity will be returned from k means method. K means method randomly selects two centroids to form a cluster. After 7 clusters, i stored the records to the output file.

Pseudo code for bisecting k means:

1. Setup initial cluster

2. implement K-means method to get 2 clusters

3. Calculate mean similarity of clusters

4. Add the scores to an array of scores

5. calculate minimum scores in array

6. Call k-means on the cluster that has the minimum similarity score (stop after 7 clusters are obtained for optimal)

7. Repeat step 3-6

This silhouette metric has been plotted on the y-axis against the values for k on the x-axis: