Assignment 2: Clustering

Aniket Santra (MDS202106)

Introduction:

The "Bag of Words" data set from the UCI Machine Learning Repository contains five text collections in the form of bags-of-words. We are given three collections from them: NIPS full papers, KOS blog entries, ENRON emails. Our task is to cluster the documents in these datasets via K-means clustering for different values of K and determine an optimum value of K.

Clustering algorithms are unsupervised machine learning approaches for grouping data based on similarity. Here we will use Jaccard Index as a similarity measure to measure similarity between two documents based on the overlap of words present in both documents. Then the K-means algorithm will be applied for clustering.

Procedure:

- For each of the datasets we first created a sparse matrix where the word IDs are in the rows and the doc IDs are in the columns.
- Next we calculated the jaccard similarity index using the inbuilt pairwise distance python function. After that we subtracted that from 1 we got a jaccard similarity matrix to get a square matrix with dimension the number of unique doc Ids.
- Then we applied the inbuilt K-means function from scikit-learn library. We ran a for loop and found out the inertia for a range of K (no.of clusters) values. After that we plotted the inertia for all these different K-means models and we found the best fit which is called the Elbow Method.
- Last we reduced the dimension of the Jaccard Matrix using PCA and then plotted the subsequent points which helped to visualize the clusters.

Results:

- NIPS DATASET : This dataset has approximately 1500 documents and 12419 unique words in the vocabulary. The dataset was not very huge and ran smoothly. The time taken to run the entire algorithm was 50 secs. According to the graph of the different values of inertia K=3 would give the best fit.
- KOS DATASET: This is a relatively small dataset with 3430 documents and 6906 unique words. The process to generate the optimal value of the number of clusters was the same as that of the NIPS dataset. The best value for K was 2 according to the Elbow Method. The time complexity for this program was 59 secs.

- ENRON DATASET: This dataset is quite large. It had approximately 39861 documents and 28102 unique words. It was difficult to use the earlier method on the dataset without optimizing it. So we reduced the dataset by taking stratified sampling. So here we have multiple stratas o based on the frequency of the word. On this reduced dataset we used the same procedure to get the jaccard similarity matrix and then the K-means clustering. The process to generate the optimal value of the number of clusters was the same. The best value for K was 3 according to the Elbow Method.

	NIPS	KOS	ENRON
No. of Clusters	3	2	3
Total Time Taken	50.40 sec	48.04 sec	2263.53 sec
Total Memory Used	341.00 Mib	388.82 Mib	6995.62 Mib