Suneha Sen, MDS202351
Sayantani Saha, MDS202345
DMML Assignment 2

Aim:

The bag of words dataset from the UCI Machine Learning repository consists of a collection of five corpus of texts. The objective is to cluster the documents from KOS, NIPS and ENRON using the K-means clustering algorithm.

Methodology:

1. The dataset comes in a tabular format where each row contains a word (denoted by wordID), the corresponding document (denoted by docID) and the frequency of the word in the document. The format of this raw data is unsuitable for clustering.
2. The clusters would be formed based on the similarity of the word-contents of the documents so we process the data using appropriate libraries and functions such that the processed data represents a bag of words. After processing, the data consists of tuples for each document. The tuples are composed of the word-IDs.
3. The current dataset represents a set of indices, for which the Euclidean distance is not defined. So, the clustering has to be done with an appropriate measure of distance. An appropriate choice would be the Jaccard distance which is defined for two sets A and B as follows:
$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$
4. To the best of our knowledge, there are no off-the-shelf implementations of the K-Means algorithm which offer the functionality to define the distance measure between the clusters. So we choose to perform K-means clustering with Euclidean distance on the pairwise Jaccard similarities of the data-points. The Jaccard similarity is defined as follows:
$$\text{Similarity(A, B)} = 1 - \frac{|A \cap B|}{|A \cup B|}$$
5. Justification of the above methodology: Let us assume that the data points reside on some space V, the distance metric on this space is given by f (jaccard-similarity in our case). We project the data on the Euclidean space as follows:
$$h : V \to \mathbb{R}^{|X|}$$
where for $x \in V$ and X denoting the set of all given data
$$h(x) = \sum_{i=1}^{|X|} f(x, X[i]) e_i$$
where $|X|$ is the number of data-points and e_i is the standard basis of $\mathbb{R}^N$. $h()$ takes a

data-point from X and converts it into a vector of jaccard similarity (f) from all points in V. So in a way, this preserves the global structure of V in the transformed Euclidean space. The benefit of this preservation is that the clustering that is performed on this transformed space is a good representative of the cluster in the original space V.

6. Calculating the entirety of the Jaccard-similarity matrix is computationally expensive, since the matrix is of order $(|X| \times |X|)$. However the matrix is symmetric in nature, so we calculate the jaccard similarity for the upper triangular part of the matrix and fill the rest accordingly.
7. In order to determine the number of optimal clusters we calculate the inertia for the “elbow-plot”. However the elbow-plot was ambiguous for all the text-corpora, so we use an additional diagnostic known as the Silhouette Score.
8. After clustering we visualise the data, labelling them according to the clusters. This acts as a good sanity-check for the number of optimal clusters. The matrix plot shows us a 2-dimensional view of the clusters formed. A better visualisation alternative is to project the data on a 3-dimensional space using dimensionality reduction techniques. We use PCA and t-SNE to project the data in a 3-dimensional space and verify that our choice of the number of clusters was indeed optimal.

Evaluation:

Apart from a visual sanity-check we evaluate the performance on three evaluation metrics, namely: Davies-Bouldin Score, Calinski-Harabasz Score and Silhouette Score. The scores for the different tasks are reported as follows:

Dataset	**No. of clusters**	**Time(Clustering)**	**Peak memory (Clustering)**	**Time(Jaccard distance)**	**Peak memory (Jaccard Distance)**	**Davies Bouldin score**	**Silhouette score**
KOS	2	4.28s	580.07MiB	1 min 18 s	462.26 MiB	0.55	0.68
Nips	3	4.65 s	1665.5 MiB	1 min 27s	635.09 MiB	1.74	0.14
Enron	2	11.4s	1903.2 MiB	5 mins 25 s	1902.3 MiB	0.61	0.6

Note: ENRON is a large corpus with 3710420 words in total, clustering on such a large dataset requires more computational resources. So we take a sample from the dataset (around 10%) and learn the centroids of the cluster. Once the centroids are learnt, we allocate the rest of the data to these centroids. Due to the large size of the dataset, a simple random sample without replacement cannot ensure that the sample drawn is in fact representative of the entire dataset.

So we draw samples using a Stratified Sampling scheme where each stratum is based on the length of the document, this ensures that we get documents of all sizes.