Data Mining and Machine Learning Assignment - 2

Srijit Saha

MDS202150

Project Report:

Aim:

The "Bag of Words" data set from the UCI Machine Learning Repository contains five text collections in the form of bags-of-words.

Our aim is to cluster the documents in these datasets via K-means clustering for different values of K and determine an optimum value of K.

Datasets used for the project:

- KOS dataset
- NIPS dataset
- ENRON dataset

Methodology

- Each dataset has a list of documents and a vocabulary of words from which each document is constructed (all stopwords such have been truncated).
- Each gzip file contains three columns consisting of the Document IDs, the Word IDS and their frequency counts (Non zero)
- Using the docword dataset, we constructed a matrix such that each entry records the usage of the particular word id in the corresponding doc ID.
- Next, the Jaccard distance is calculated using the pairwise distance function and the Jaccard similarity is calculated by subtracting 1 – each entry of the Jaccard distance matrix.
- Next we fit the Jaccard similarity matrix generated to the K-means model, and plot the inertia for each cluster.
- From the Elbow plot we identify the kink points on the curve and use those particular values of K for forming the clusters.
- K-means is run again with the optimal no of clusters and the data is separated into different clusters.
- Using the centroids for the clusters, the clusters are plotted on a graph, which shows how the data looks visually when clustered.
- As an accuracy metric we have used the Davies Bouldin score to analyse out performance.

Remarks:

- We have run the K-means clustering only on the KOS and Nips dataset, since the size of the dataset was feasible to run on a Colab or Jupyter Notebook.
- The Enron dataset consists of a large number of word and document entries and owing to limited computational power, running K-means clustering for the whole dataset is extremely expensive in terms of RAM usage.

- Filtering the dataset like trying to take a random sample of documents and words, or taking those documents featuring words only above a certain threshold, and applying K-means on that, the results are highly inconsistent with the actual dataset and may give us a faulty clustering.
- As an attempt to cluster the Enron dataset, stratified random sampling was used on the dataset and then the above mentioned methodology was carried out.
- An example of this faulty clustering is shown in the Colab notebook where, sampling of the dataset leads to a faulty representation of the K-means clustering.

Dataset\Attributes	No. of clusters	Time	Peak memory	Davies-Bouldin Score
KOS	3	7.4	824 MiB	2.09
Nips	5	2.8	448.3 MiB	2.56
Enron	6	7.96	11608.6 MiB	1.07