DMML - Assignment 2: Clustering

Jayasooryan C S (MDS202119)

Aman Kumar (MDS202104)

The Task

The "Bag of Words" data set from the UCI Machine Learning Repository contains five text collections in t he form of bags-of-
words. The URL for the UCI repository is *https://archive.ics.uci.edu/ml/datasets/Bag+of+Words.*

Your 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.

As a similarity measure, use Jaccard index, that measures similarity between two documents based on t he overlap of words present in both documents. Note that this changes the underlying model from "bag of words" to "set of words".

The datasets are of different sizes. Report your results on the three smaller datasets (Enron emails, NIPS blog entries, KOS blog entries).

The Data Set

In each of the text collections, each document is summarized as a bag (multiset) of words. The individual documents are identified by document IDs and the words are identified by word IDs.

Information about the datasets in the repository (that you need to analyze)

Enron Emails: *orig source: www.cs.cmu.edu/~enron*
- D=39861
- W=28102
- N=6,400,000 (approx)

NIPS full papers: *orig source: books.nips.cc*
- D=1500
- W=12419
- N=1,900,000 (approx)

KOS blog entries:
orig source: *dailykos.com*
- D=3430
- W=6906
- N=467714

Approach:

K Means algorithm is implemented using Jaccard index as distance metric and centroid is defined as the point in the cluster whose cumulative Jaccard distance to rest of the points in that cluster is minimum. Clusters were evaluated using inertia to find the optimum value of k. Inertia is defined as the mean sum squared distance from centroids, where distance is measure in Jaccard metric.

Reading the Data

To read the data we are defining a txt2data function which reads the data form local drive and give a word-vector matrix as output

- Argument- filename: the name of file of interest
- output - sparse matrix of the interested file

Defining function:

- kmeans: Gives the centeroid and corresponding cluster after max_inter iterations
- kmean_inertia: finds inertia as mean sum squared distance from centroid using jaccard metric
- optimize: Gives interia values for Kmean for k upto m
- kmean_plot: render elbow plot to evaluate optimal value of number of clusters
- kmean_cluster: Evaluate clusters

kmeans

Input:

- data = doc-word vector(sparse matrix), output of txt2data
- jm = Jaccard distance matrix
- k = number of centroid
- max_iter = maximum number of iternation (shifting and clustering)

Output

- Type: Dict
- Dict[0] = clusters
- Dict[1] = centroids

kmean_inertia

Input:

- data = doc-word vector(sparse matrix), output of txt2data
- jm = Jaccard distance matrix
- k = number of centroid
- max_iter = maximum number of iternation (shifting and clustering)

Output

- Type: int
- Mean sum squared error

Optimize

Input:

- data = doc-word vector(sparse matrix), output of txt2data
- jm = Jaccard distance matrix
- m = maximum number of cluster for testing elbow point
- max_iter = maximum number of iternation (shifting and clustering)

Output

- Type: list
- Mean sum squared error for differnt values of clusters

Kmean_plot

Input:

- filename = doc-word file name
- k_range = maximum number of cluster for testing elbow point
- max_iter = maximum number of iternation (shifting and clustering)

Output

- Type: plot
- Mean sum squared error plot for differnt values of clusters

Kmean_cluster

Input:

- filename = doc-word file name
- k = optimal number of cluster
- max_iter = maximum number of iternation (shifting and clustering)

Output

- Type: dict
- Dict[0] = clusters
- Dict[1] = centroids

KOS blog entries

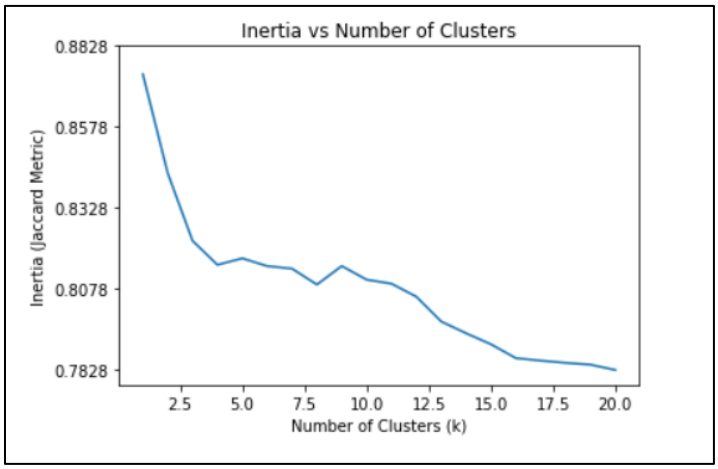

Fig1: Within Cluster Mean Squared Error For KOS

Total number of optimal cluster for KOS is 3

NIPS blog entries

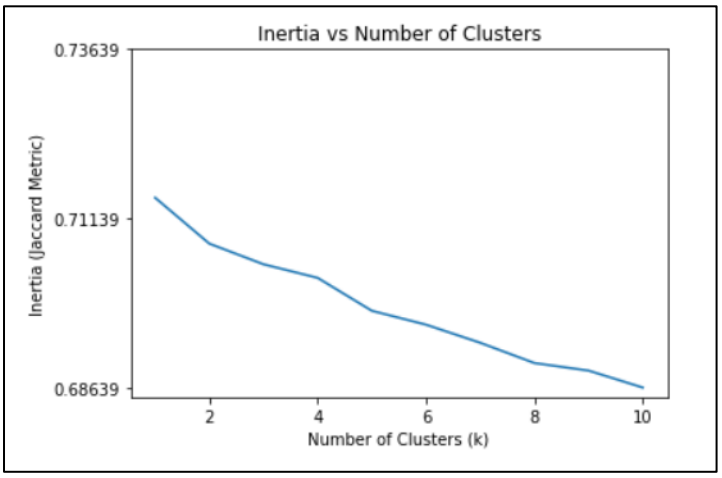

Fig2 : Within Cluster Mean Squared Error for NIPS

Here for the graphs for k upto 10 and k upto 20, the graph decreases in range of 0.025. Comparing it with the previous dataset of KOS, which showed more than 0.075 jump after the first three k values, the value for NIPS looks stabilized around k = 1. Hence based on the output, the optimum number of cluster for NIPS dataset is 1.

ENRON email

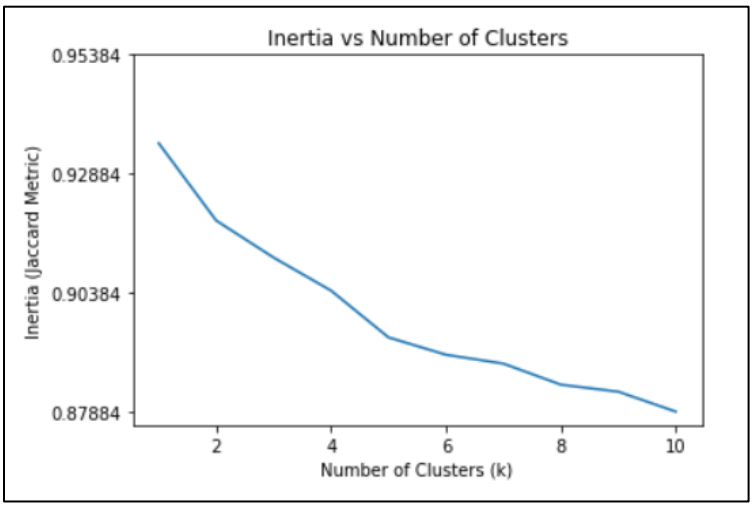

Fig3: Within Cluster Mean Squared Error for NIPS

The analysis was done with only 30% of the dataset, hence the results need not be repetitive of original dataset. Based on the results, the optimum number of clusters is 5.

Summary

Total number cluster/s:

- KOS: 3
- NIPS: 1
- ENORN: 5

Total Runtime:

- KOS: 5 mins
- NIPS: 6 mins
- ENRON: 24 mins