Document Clustering using Kmeans with Jaccard distance

Utpalraj Kemprai(DS1), Sayan Bose(CS1)

March 2024

1 Introduction

Our task is to cluster the three documents in the Bag of Words datasets via K-means clustering for different values of K and determine an optimum value of K. The three datasets are Enron emails, NIPS blog entries and KOS blog entries. As a similarity measure we use the Jaccard Index.

2 Data storage for clustering

For each collection, we store the documents as a (D,W) sparse matrix (csr_matrix from scipy.sparse) where (i,j) th entry is 1 if word with wordID (j+1) is present in document with documentID (i+1) and 0 otherwise.

The sparsity (no. of zeros/total no. of entries) were 0.985,0.960 and 0.997 for the KOS, NIPS and ENRON datasets respectively. Hence we used csr_matrix for storing them for efficient storage and speed.

3 Kmeans clustering with the Jaccard distance

The Jaccard distance is defined as 1 - (Jaccard Similarity). We use the Jaccard distance as our distance metric in our implementation of Kmeans instead of the usual l_2 norm.

3.1 Algorithm

Our algorithm for clustering is as follows:

1. Select K points from the data set uniformly at random and set them as centers of the clusters.

2. Assign each point of the data set to the cluster, to whose center it is the closest in Jaccard distance.

3. Calculate new centers for each cluster by taking the means of each of the clusters (same as Kmeans with l_2 norm).

4. Repeat 1,2 and 3 till a fixed number of times or if we get the same centers after performing 3.

5. output the points with their assigned clusters and the centers of the clusters.

4 Implementation

For finding the optimal value of K for Kmeans clustering, we plot the sum of squared distances (SSD) between points in a cluster to their centers, for different values of K and using the elbow method find the optimal value of K for clustering.

4.1 KOS dataset

For the KOS dataset we plot SSD for k = 1 to 20 , with max iteration as 100 for each of the kmeans for k = 1 to 20.

4.2 NIPS dataset

For the NIPS dataset we plot SSD for k = 1 to 20 , with max iteration as 100 for each of the kmeans for k = 1 to 20.

4.3 ENRON dataset

For the ENRON dataset we plot SSD for k = 1 to 12 , with max iteration as 50 for each of the kmeans for k = 1 to 12.

5 Results

We find the following optimal values of K and the time taken for implementation (doing Kmeans for different values and plotting SSD) in seconds , for each of the datasets using the elbow method:

Dataset	Optimal K	Time needed (sec)
KOS	14	349
NIPS	13	109
ENRON	10	3402