
Yash Jain MDS202048

Yash Raj MDS202049

DMML Assignment 2

July 2021

OVERVIEW

We did K-Means clustering on the following three data sets: enron, kos and nips and found the optimum value of K for each data set.

APPROACH

We tried two different approaches; one for **kos** and **nips** and the other for **enron** because of the size of the datasets.

The number of documents in each dataset is:

Enron: 39861

Nips: 1500

Kos: 3430

Approach 1 for 'nips' and 'kos:'

First we calculate the **Jaccard matrix** which stores Jaccard distance (**1 - Jaccard Index**). The seeds (initial guess of centroids) are stored in a list and the clusters are stored in a dictionary. At each iteration we update the cluster using the current centroids. The Jaccard Matrix is then used to update the centroids. The same process repeats until we exhaust **max_iterations** or if the centroids stop updating.

The runtime is of the order n^2 since we are calculating the Jaccard distance for each document with respect to every other document in the dataset.

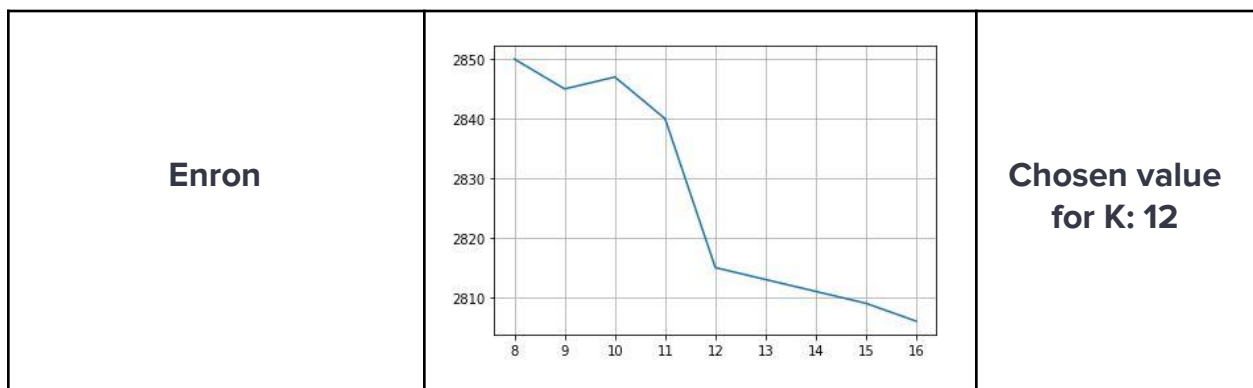
Approach 2 for 'enron:'

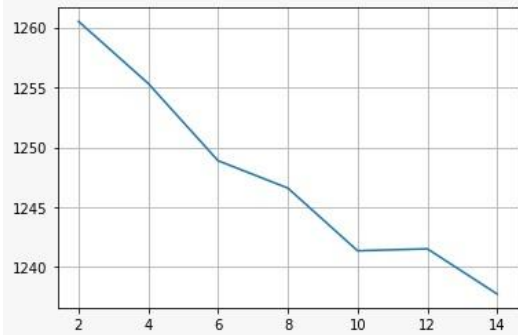
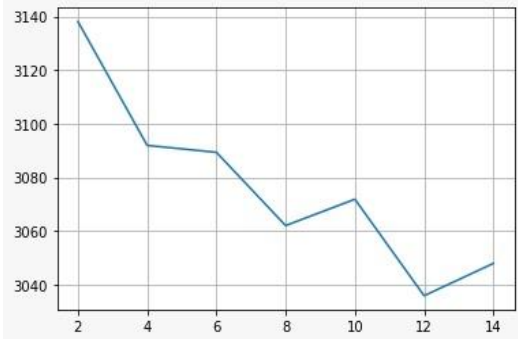
As storing the Jaccard matrix for a large dataset with many documents is quite expensive, we use a different approach to update centroids for the 'enron' dataset. Firstly, the seeds (initial guess of centroids) are stored in a list and the clusters are stored in a dictionary. To calculate the mean document (not necessarily in the dataset) we divide the frequency of each word in a cluster by the size of the cluster. A word is present in the mean document if the aggregated presence of the word in a cluster is greater than some threshold which in our case is 0.5. Then we find the nearest **docIDs** to each mean document in different clusters. And these **docIDs** are used to update the centroids on each iteration. Hence the clusters are also updated.

Selection of optimum K value:

We used the elbow method which is a heuristic used in determining the number of clusters in a data set. For different values of K we compared the total **Jaccard Distance** (sum of jaccard distance of each document with its corresponding centroid) to find the optimum K.

For the different datasets, we obtained the following graphs:



Nips	 <table><tr><th>K</th><th>Accuracy</th></tr><tr><td>2</td><td>1261</td></tr><tr><td>4</td><td>1255</td></tr><tr><td>6</td><td>1249</td></tr><tr><td>8</td><td>1247</td></tr><tr><td>10</td><td>1241</td></tr><tr><td>12</td><td>1241</td></tr><tr><td>14</td><td>1237</td></tr></table>	K	Accuracy	2	1261	4	1255	6	1249	8	1247	10	1241	12	1241	14	1237	Chosen value for K: 10
K	Accuracy																	
2	1261																	
4	1255																	
6	1249																	
8	1247																	
10	1241																	
12	1241																	
14	1237																	
Kos	 <table><tr><th>K</th><th>Accuracy</th></tr><tr><td>2</td><td>3138</td></tr><tr><td>4</td><td>3092</td></tr><tr><td>6</td><td>3089</td></tr><tr><td>8</td><td>3062</td></tr><tr><td>10</td><td>3072</td></tr><tr><td>12</td><td>3037</td></tr><tr><td>14</td><td>3050</td></tr></table>	K	Accuracy	2	3138	4	3092	6	3089	8	3062	10	3072	12	3037	14	3050	Chosen value for K: 12
K	Accuracy																	
2	3138																	
4	3092																	
6	3089																	
8	3062																	
10	3072																	
12	3037																	
14	3050																	