ML Assignment: Clustering

Raktim Dey-MDS202132, Sucheta Jhunjhunwala-MDS202151

16th April, 2022

1 Introduction

Our target is to cluster the three datasets, NIPS blog entries, KOS blog entries and Enron Emails using KMeans Clustering Algorithm. We have worked on the datasets consisting of information about the word in each document, that is, docID, wordID and count. The data available might lie in a non-Euclidean space, our aim is to bring it to an Euclidean space and then apply KMeans Clustering on this.

2 Procedure

- We imported each of the dataset and created a matrix consisting of vectors of the form [docID,wordID,count].
- Using the array created above, we created a term document matrix consisting of entries as either 0 or 1 depending upon whether a particular word is present or not in the document. The columns of the matrix **term_d** represent each document of the dataset and each row basically indicates the presence or absence of a word from the vocabulary.
- We then compute the Jaccard matrix **jm** of each pair of document using the Jaccard Similarity Score as a metric. Note that we need to maximise the Jaccard Similarity score and hence maximize **I-jm**, which forms our distance matrix.
- We fitted the model for various values of k ranging from 2 to 10 and plotted Inertia against it. Since our points are in the Euclidean space, we choose inertia as a measure for distance to be minimised.
- The optimum value of k will be at the elbow point that is the point in the graph where there is a kink.
- We then try to visualize the clusters for each dataset which we do by plotting a 3D graph for the same.

- To verify our results we look at the Davies-Bouldin Index which measures the average similarity of each cluster with a similar cluster. So we need to minimise this index which indicates better clustering.

3 NIPS Blog Entries

- Apply all the steps as mentioned above, create the list $\mathbf{nnz}_vector_nips.Wefindtheoptimumvaluefork, whic$ 3.
-
- We then look at the 3D graph of the data points which indicates the occurrence of three clusters.

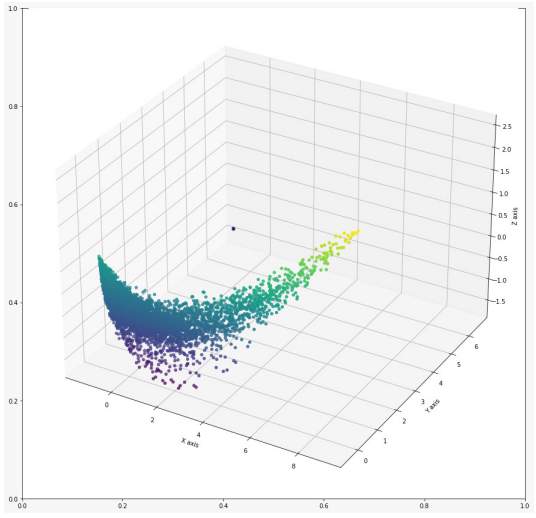

Figure 2: 3D graph of NIPS

- The Davies-Bouldin Index has value 1.4.

4 KOS Blog Entries

- Apply all the steps as mentioned above, create the list $\mathbf{nnz}_vector_kos.Wefindtheoptimumvaluefork, whic$ 3.
- We then look at the 3D graph of the data points which indicates the occurrence of three clusters.
- The Davies-Bouldin Index has value 1.19.

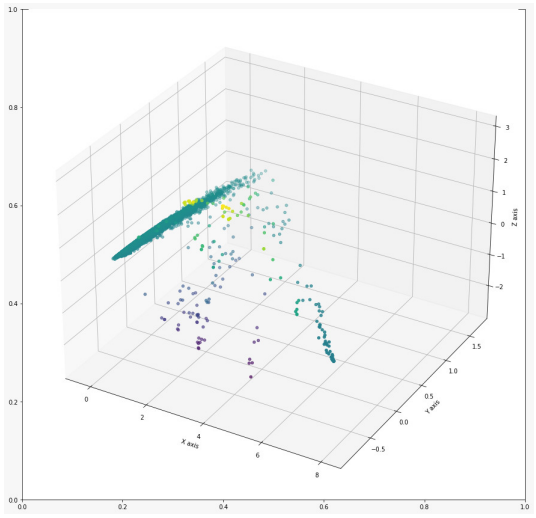

Figure 4: 3D graph of KOS

5 Enron Emails

- Apply all the steps as mentioned above, create the list $\mathbf{nnz}_v ector_e nron. Since the Enron Emails dataset is ve$
- We find the optimum value for k, which comes out as k=3.

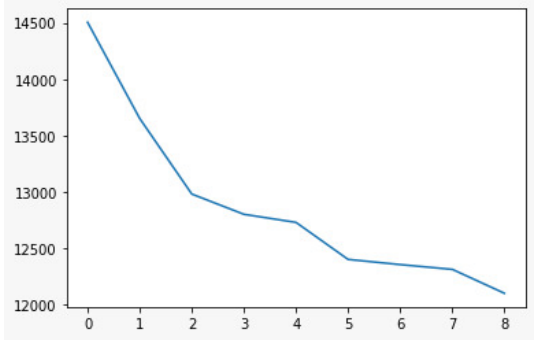

Figure 5: Elbow Point of KOS

- We then look at the 3D graph of the data points which indicates the occurrence of three clusters.
- The Davies-Bouldin Index has value 0.47.

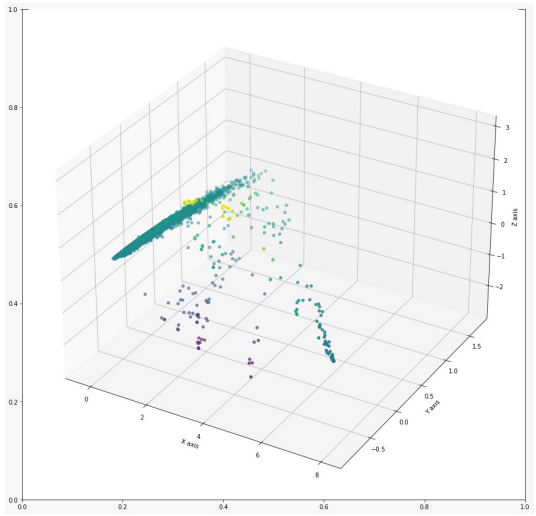

Figure 6: 3D graph of KOS

6 Comparative Evaluation

Dataset	k	Time	Davies-Bouldin Index	Memory
NIPS	3	36.3s	1.4	3631.52Mb
KOS	3	13.1s	1.19	1514.25Mb
Enron	3	68.6s	0.47	16594.32Mb

7 Conclusion

The documents in all the datasets, KOS dataset, NIPS dataset and Enron dataset formed three clusters separately. The Davies-Bouldin Index, which measures the goodness split by a K-Means clustering algorithm, was less than 1 for the NIPS dataset, slightly higher than 1 for the KOS dataset and very close to 0 for the Enron dataset. So, our clustering algorithm worked well for the three datasets. We tried implementing the same procedure on the Enron Emails dataset but due to the shortage of RAM the computation was infeasible and so we had to reduce the size. Even after reducing the sample size, the space and time needed for clustering Enron Emails dataset was very high.

8 Link

https://colab.research.google.com/drive/18T4bGzhlRoMZQbHAjEy$_N gQT3by5hmXA?usp = sharing$