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 **nnz_vector_nips** consisting of vectors of the form [docID,wordID,count].

- We saw for NIPS blog entries the dataset had 1500 docIDs and 12419 words in total.

- Using **nnz_vector_nips**, 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 distance 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 minimise **I-jm**.

- 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.

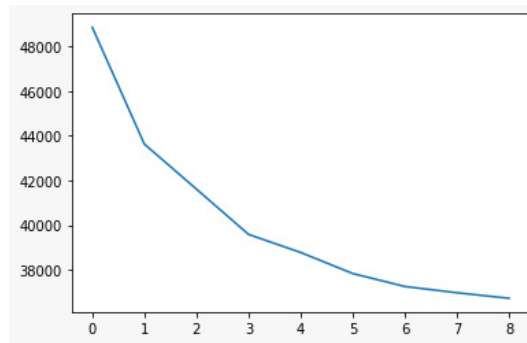- We find the optimum value for k, which comes out as k=3.

Figure 1: Elbow Point of NIPS

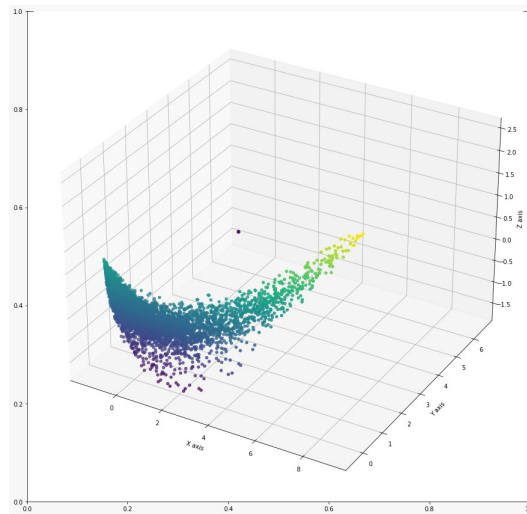- We then look at the 3D graph of the data points which indicates the occurrence of three clusters.

Figure 2: 3D graph of NIPS

2

- The Davies-Bouldin Index has value 1.4.

4 KOS Blog Entries

- Apply all the steps as mentioned above.

- We find the optimum value for k, which comes out as k=3.
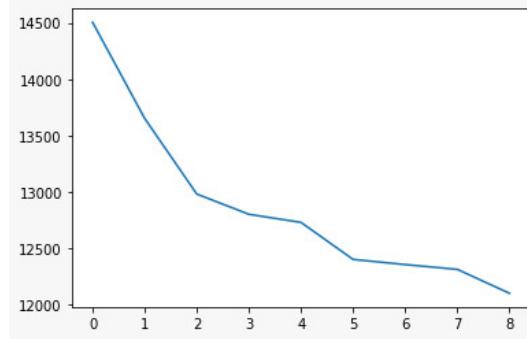

Figure 3: Elbow Point of KOS

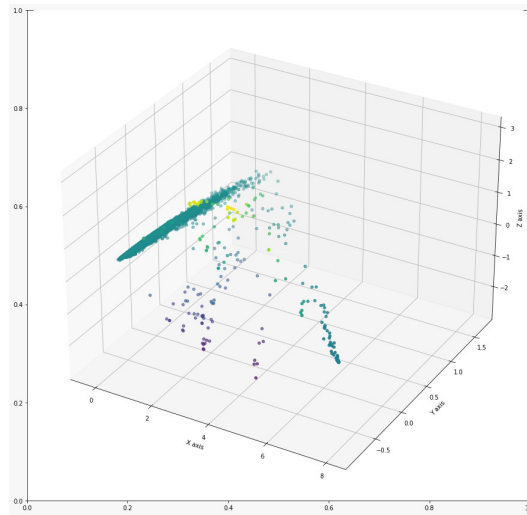- We then look at the 3D graph of the data points which indicates the occurrence of three clusters.

Figure 4: 3D graph of KOS

- The Davies-Bouldin Index has value 1.19.

5   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

6   Conclusion

The documents in both the KOS dataset and NIPS 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 and slightly higher than 1 for the KOS dataset. So, our clustering algorithm worked well for both the datasets. We tried implementing the same procedure on the Enron Emails dataset but due to the shortage of RAM the computation was infeasible. We even tried taking a random sample of size 3000 to implement the clustering but it didn't work as well.

7   Link

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