CMI/2022/Assignment2 Deadline: April 17,2022

Data Mining & Machine Learning

***Group Members*:**

1. Avik Das - MDS202112
2. Meghna Mondal – MDS202123

K-means clustering Models on Bag of Words dataset:

In this assignment we have built K-means Clustering for 3 data sets:

1) KOS,

2) NIPS

3) ENRON.

Then we used suitable evaluation metric to compare the performance of the three classifiers.

Table for Comparing the classifier models:

Performance Measure	KOS	NIPS	ENRON (On 1% Sample)
Time Taken(s)	**128.02**	**51.86**	**3962.09**
Space required(MiB)	**726.61**	**582.61**	**5115.92**

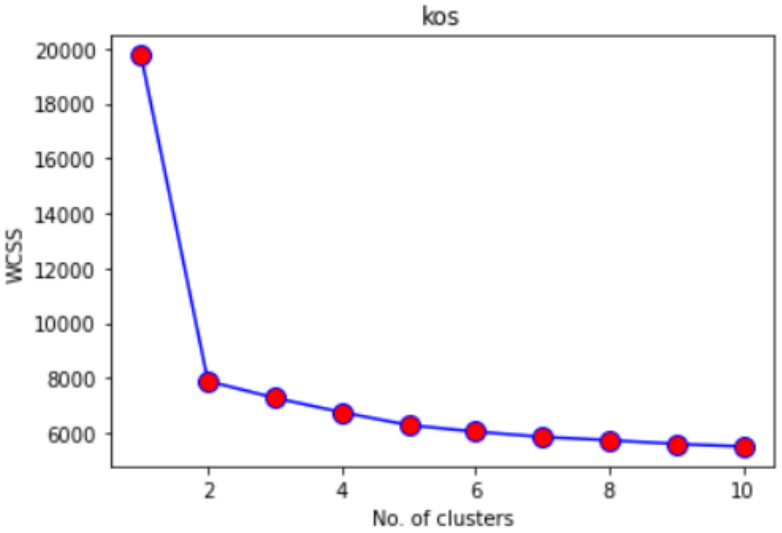
Procedure of fitting the Models:

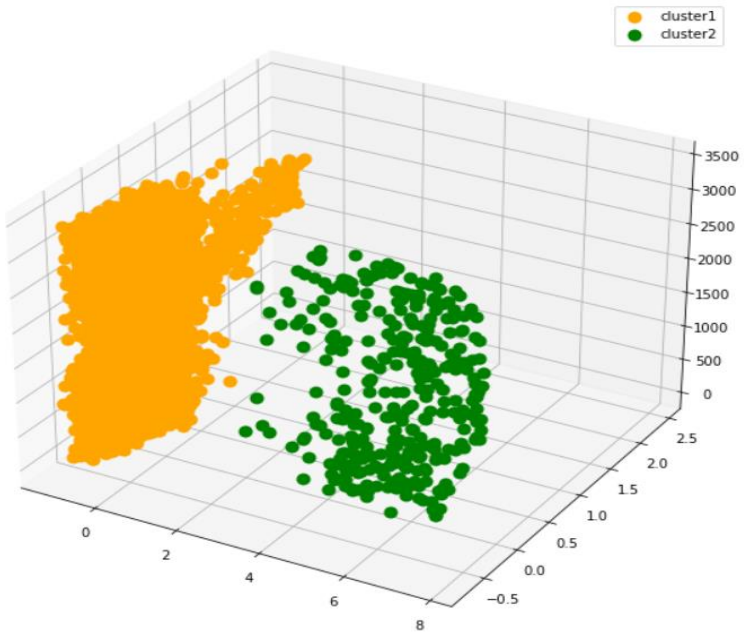
- At first, we read the dataset and skip the 1st 3 lines of each data. And then we take the matrix containing Doc_ID, Word_ID, Word_Count and made the data frames.
- Then we create the sparse matrix with Doc_IDs in the column and Word_IDs in the rows. And it contains 1 at ijth position if a ith Word_ID has appeared in a jth Doc_ID.
- Then we calculate the Jaccard Index for each of the document and made a symmetric matrix containing Jaccard index between ith and jth document at ijth position.
- After that we pass that symmetric matrix to build Kmeans clustering model with different number of clusters and then by plotting the

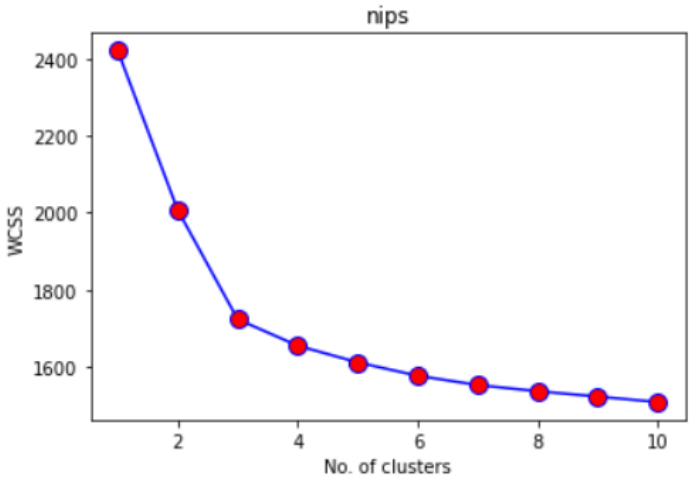
graph of inertia vs # clusters, we find the optimum number of clusters by elbow method.

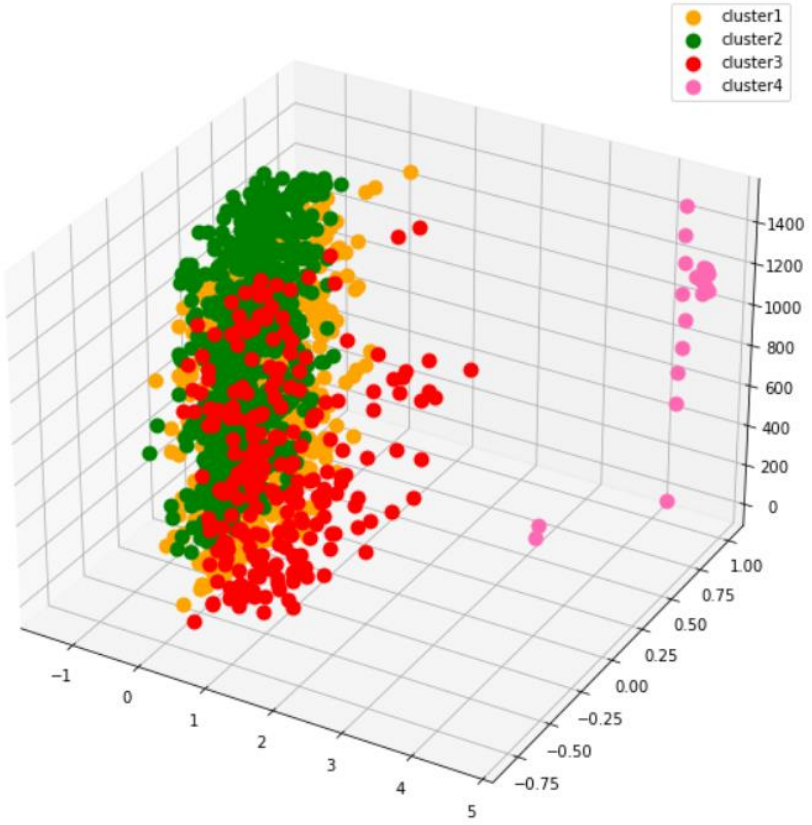
- Then we finally fit the Kmeans clustering model with the optimum number of clusters for KOS and NIPS dataset.
  - In KOS dataset the optimum number of clusters is 2.
  - In NIPS dataset the optimum number of clusters is 4.
- After that we perform dimension reduction for visualizing the clusters in 3-dimensional space.
- And at last, we observed the Doc_IDs belonging to different clusters and give them as the final output in a dictionary format.
- In case of ENRON dataset, we at first did stratified sampling and take 1% of the data as sample grouping by Word_counts(i.e., keeping the ratio of word frequencies same as the original data).
- Then we go through the same procedure, but here we didn't get a good elbow, so we can't able to fit the model in a good way.

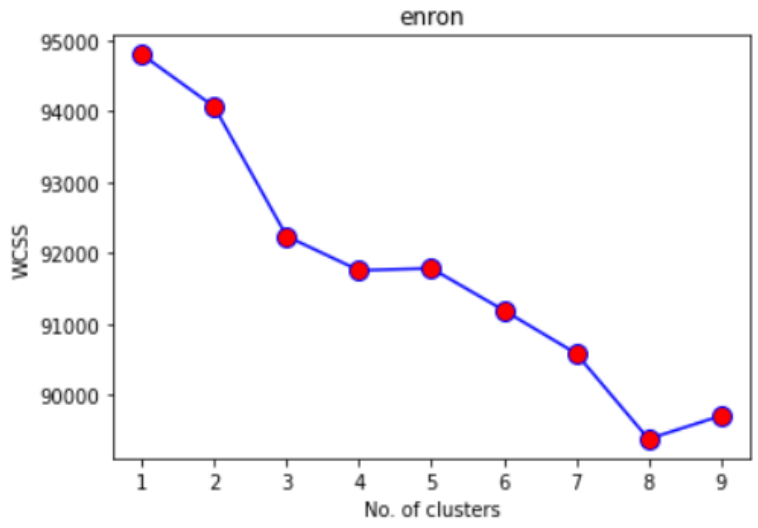
Visualizing the Clusters:

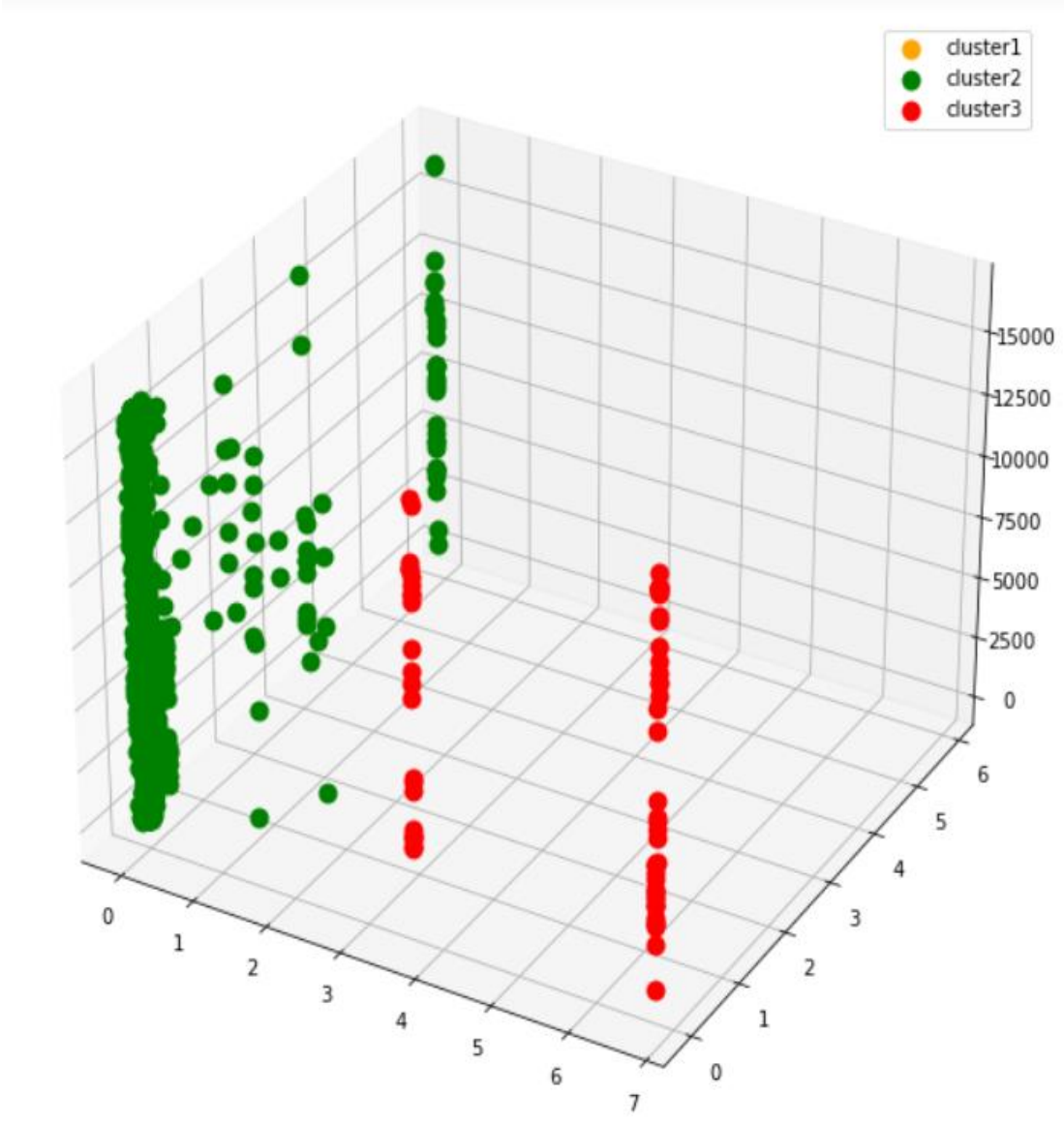





nips
WCSS
No. of clusters
2400
2200
2000
1800
1600
2
4
6
8
10


cluster1
cluster2
cluster3
cluster4
1400
1200
1000
800
600
400
200
0
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1
0
1
2
3
4
5


enron
WCSS
No. of clusters
95000
94000
93000
92000
91000
90000
1
2
3
4
5
6
7
8
9


cluster1
cluster2
cluster3
15000
12500
10000
7500
5000
2500
0
0
1
2
3
4
5
6
7
0
1
2
3
4
5
6