CSE 419L: LAB 08 Assignment: Compression Techniques

You have been given a dataset **(D1),** which contains 1000 documents, distributed into 10 folders.

https://www.kaggle.com/datasets/jensenbaxter/10dataset-text-document-classification

Task 01:

1. Read the Dataset.
2. Perform Text pre-processing steps.

Task 02:

1. Calculate TF for each term in each document.
2. Calculate IDF for each term in the vocabulary.
3. Create the TF-IDF matrix.

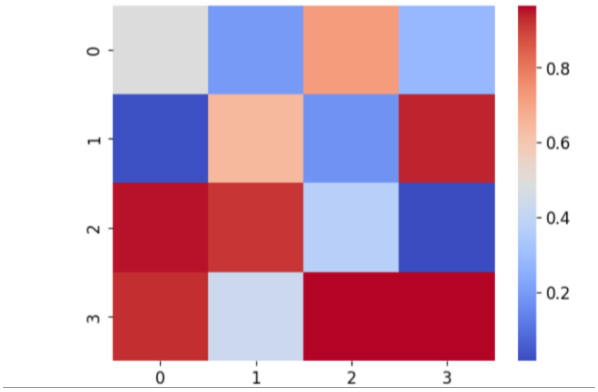
Task 03:

1. calculate cosine similarity of each document with all the other documents.
   Output will be a [1000 X 1000] matrix.

2. Generate the Heat map with help of similarity matrix calculated in the previous step.

Example heatmap for a given distance matrix:

0.49251769	0.19524299	0.72245212	0.28077236
0.02431597	0.6454723	0.17711068	0.94045858
0.95392858	0.91486439	0.3701587	0.01545662
0.92831856	0.42818415	0.96665482	0.96361998



Do not clear the outputs of the Notebook file. Output traces will be required for evaluation.