CSE 419L: LAB 07 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

T1: Dictionary Compression Task:

1. Read the Dataset.
2. Perform Text pre-processing steps.
3. Create the Term-Postings list (Inverted Indexing).
4. Store the Inverted Index in a Data Frame as follows:

Freq.	Term	Posting List
34	a	[……..]
5	call	[……..]
23	python	[……..]

5. Compress the Dictionary as follows:

   Term_string = “acallpython”

Freq.	Term_start_index	Posting List
34	0	[……..]
5	1	[……..]
23	5	[……..]

6. Calculate the size of Dictionary before and after the compression.

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