CSE 419L: LAB 03 Assignment

You have been given a dataset, which contains 1000 documents, distributed in to 10 folders.

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

[Pick 10 documents randomly from each folder:- 100 documents for your lab assignment]

Prepare an **Inverted Positional Index** to execute **phrasal queries** for accessing the required document.

Step 01: Load the data and perform tokenization.

Step 02: Create **Inverted Positional Index by** [For each Term: storing the DOC_ID and locations of the term in each DOC.]

Step 03: Write a function to execute **phrasal queries**.

[OPTIONAL: Repeat your assignment on all 1000 documents]

[Lecture PPT is shared for the reference].

[Note:- Use meaningful identifier names. Create sections, sub-sections as per requirement]