CSE 419L: LAB 02 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 Index to execute Boolean queries for accessing the required document.

Step 01:- Tokenize each file. Create Dictionary. Sort the data.

Step 02:- Create Posting-lists for each term.

Step 03:- Write a function to execute Boolean 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]