CSE 419L: LAB 05 Assignment: SPIMI Algorithm

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

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

SPIMI Algorithm implementation is given at:

https://github.com/vartanbeno/SPIMI

Default implementation uses Reuters dataset (~21500 documents). Perform task 01 on the same dataset. Do not download dataset manually. Dataset collection is built in.

Task 01:

a. Set up the code on Google Collab. [install required library and perform a test run]
b. Default value for #document is 500 in each block.

    i. Run your algorithm from #document = 20 for the first run. In the next run double the count and so on.
    ii. Continue this process until whole data fits in one block.
    iii. For each experiment log the time taken by the algorithm. Draw a line plot to show the relation b/w block count vs time.
    [Add your RollNo. In the title of the plot].

Task 02:

Execute the SPIMI algorithm on **D1** dataset. [#Document 1000]

i. Run your algorithm from #document = 10 for the first run. In the next run double the count and so on.
ii. Continue this process until whole data fits in one block.
iii. For each experiment log the time taken by the algorithm. Draw a line plot to show the relation b/w block count vs time.
[Add your RollNo. In the title of the plot].

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