CSE 419L: LAB 12 Assignment:

Document Classification: Rocchio Classifier

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:

Select 70% documents from each Class [700 docs]. Keep rest of the 30% [300 docs] for Testing.

On 70% data: -

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: Implement Rocchio Classifier: -

3.1: - Calculate Centroids for each class as follows: **[beta=16, Gemma=4]**

Rochio classifier for a class c_p is computed as a centroid given by

$$\vec{c_p} = \frac{\beta}{n_p} \sum_{d_j \in c_p} \vec{d_j} - \frac{\gamma}{N_t - n_p} \sum_{d_l \notin c_p} \vec{d_l}$$

where

- n_p: number of documents in class c_p
- N_t: total number of documents in the training set
- terms of training docs in class c_p: positive weights
- terms of docs outside class c_p: negative weights

3.2: - Calculate cosine similarities of each Test doc with all centroids. Assign the class with highest similarity.

3.3: - Calculate the Accuracy, Precision, Recall, F-score. [Class wise]

[Do not clear the output before the submission.]