CSE 419L: LAB 11 Assignment:

Okapi BM 25 Ranking Functions

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:

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: Implementation of Okapi BM 25 Ranking Function.

Given a query Q, containing keywords $q_1, \ldots, q_n$, the BM25 score of a document D is:

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot \left(1 - b + b \cdot \frac{|D|}{\text{avgdl}}\right)}$$

where $f(q_i, D)$ is the number of times that the keyword q_i occurs in the document D, $|D|$ is the length of the document D in words, and avgdl is the average document length in the text collection from which documents are drawn. k_1 and b are free parameters, usually chosen, in absence of an advanced optimization, as $k_1 \in [1.2, 2.0]$ and $b = 0.75$.[3] $\text{IDF}(q_i)$ is the IDF (inverse document frequency)

Calculate ***score(D, Q)*** for each Document ***(D)*** in the collection for the given query **(Q).** Provide the name of top 10 documents at the output. Execute your program for 4 different input queries.

[Do not clear the output before the submission]

https://en.wikipedia.org/wiki/Okapi_BM25