Information Retrieval CSF469

Lab Session - 4

Date - 12/02/2024

Marks: 10

Objective:

To read text from multiple files saved in a folder and then create an Information Retrieval System for efficient query processing.

Dataset: You are provided with a dataset containing multiple pdf files. The folder with pdf files has been uploaded on Canvas under “Praj Industries”.

Tasks:

1. Read the pdf files so that each pdf file is treated as one document. Please remember that only the text should be captured while ignoring the images, charts, etc. Moreover, the file reading process should not corrupt the text in the pdf file.

2. Create inverted index based on these conditions:
    a. The inverted index should have two kinds of tokens, one-word tokens or two-word tokens (including words with hyphen(-) in between. For example, words like bio-catalytic, pre-treatment, etc., are treated as one single token.
    b. Bi-word indices: The bi-word indices should be created by first checking the frequency of occurrence of bi-word terms in the text data.
    NOTE: First remove stopwords and then check frequency count.
    c. Select only the top 100 words in bi-word terms for indexing. For example, words such as order book, bio energy, bio products, etc., are part of indexing and are treated as one single token.

3. The next step is to create a simple boolean retrieval system for efficient query processing. (You can also use the code from previous labs.)