Rex Technologies™

Digitally Yours!

Task: Information Retrieval System Development

Task Objective:

Your mission is to design and implement an Information Retrieval (IR) system capable of efficiently retrieving relevant documents from a given dataset. Below are the key steps and details for this project:

1. Dataset Selection and Preparation:

- Your first task is to select a dataset of text documents. This dataset should represent a collection of documents from a specific domain (e.g., news articles, research papers, or product descriptions).

- Ensure that the dataset is appropriately labeled or categorized.

2. Data Preprocessing:

- The next step involves cleaning and preprocessing the text data. This includes tasks such as:

  - Tokenization.

  - Lowercasing.

  - Removing punctuation.

  - Handling missing values, if any.

- You should also create an inverted index, a data structure that maps terms (words) to their corresponding documents.

3. User Query Interface:

- Create a user-friendly query interface that allows users to input search queries.

- Ensure that the interface can handle natural language queries.

4. Retrieval Algorithm:

- Implement the Vector Space Model (VSM) or Term Frequency-Inverse Document Frequency (TF-IDF) as your retrieval algorithm. These are beginner-friendly approaches.

- Your system should rank documents based on their relevance to user queries.

5. Query Processing:

- Preprocess user queries similarly to how you processed documents.

- The system should compare the user query to the indexed documents and return a ranked list of relevant documents.

6. Evaluation:

- Define a set of test queries and relevant documents to evaluate the system's performance.

- Use common IR evaluation metrics like Precision, Recall, and F1-score to assess how well the system retrieves relevant documents.

7. User Feedback (Optional):

- If time allows, you can incorporate user feedback mechanisms to improve the system's performance over time.

8. Documentation and Presentation:

- Compile all your work into a comprehensive document that includes:
  - Cleaned and preprocessed dataset.
  - Code and scripts used for preprocessing, indexing, and retrieval.
  - Detailed explanations of each step, including the rationale behind the chosen methods.
  - Evaluation results, including retrieval metrics and any visualizations.
  - Recommendations and insights, if any.

Results and Reporting:

- When reporting the results, include:
  - A summary of the system's performance based on evaluation metrics.
  - Any challenges encountered during system development and how they were addressed.
  - Suggestions for future improvements or extensions to the system.