Project: Design and implement a basic search engine for a collection of text documents.

Project Procedures:

1. **Data Collection: (1%)**

   - Gather a set of text documents to serve as your corpus. This could be a collection of articles, web pages, or any other textual content.
   - Ensure the documents are in a format that can be easily parsed and indexed

2. **Preprocessing: (1%)**

   - Tokenization: Split documents into individual words or tokens.
   - Lowercasing: Convert all text to lowercase for case insensitivity.
   - Stopword Removal: Eliminate common words (e.g., "and", "the", "is") that do not contribute much to the meaning of the document.
   - Stemming or Lemmatization: Reduce words to their base or root form (e.g., "running" to "run").

3. **Indexing: (1%)**

   - Build an inverted index
   - Create a data structure that maps each unique word (or term) to the documents that contain that word.
   - For each term, maintain a list of document IDs where the term appears along with the frequency of occurrence.

4. **Query Processing: (2%)**

   - Implement a simple query processing with expanded capabilities
   - Parse user queries (input text) and apply the same preprocessing steps (tokenization, lowercase, etc.) as used during indexing.
   - Identify relevant documents by leveraging the inverted index
   - Retrieve documents that contain all the terms from the query.
   - Rank the retrieved documents based ranking algorithm (TF-IDF).

5. **Query expansion: (3%)**

   - Apply relevance feedback by analyzing top-ranked documents for initial queries.
   - Incorporate synonyms or related terms using pre-built mappings or **Embeddings (ELMo and BERT)**.

6. **User Interface: (1%)**

- Develop a basic user interface to interact with the search engine
- Accept user queries.
- Display relevant search results.

7. **Evaluation: (1%)**

- Evaluate the performance of your search engine:
- Test with various queries to assess retrieval accuracy and speed.

Important information

Maximum points	10%
Due date	Week 14 (a complete schedule of discussions will be announced within week 13)

Project submission requirements:

1. Source Code:

Upload the complete source code of your search engine implementation via classroom before the final discussion session, with a deadline of within 1 complete day (**24 hours**) prior to the discussion. This deadline ensures that all submissions can be reviewed prior to our final discussion and allows time for any necessary preparations.

2. PDF Report:

Prepare a PDF report that contains a brief description of each stage of your search engine implementation. Each student will upload his/her PDF at the same time of uploading the source code.

Important precaution:

As a reminder, **the use of AI tools** to complete this project is strictly prohibited. Any student found violating this policy will receive **a zero** for the affected assignment and the final **course work**. This policy is in place to ensure that each student gains a comprehensive understanding of the subject matter through their own efforts and learning. Please adhere to this policy to maintain academic integrity and fair assessment.

Best regards
Dr. Mohamed Maher Ata