Final Project: Mini Search Engine for www.muhlenberg.edu

Grading and Deadlines:

- **Program**: 50 points, due Dec 5th.
- **Paper**: 30 points, due Dec 9th.
- **Analysis**: 20 points, due May 9th.
- **Extra Credit**: Up to 5 points for a high-quality GUI.

Team Composition: Option for individual work or collaboration in pairs.

Project Overview: Create a specialized search engine for the muhlenberg.edu website. Your engine will rank web pages based on user queries, leveraging the Cranfield Collections for testing and evaluation. The project involves several key phases:

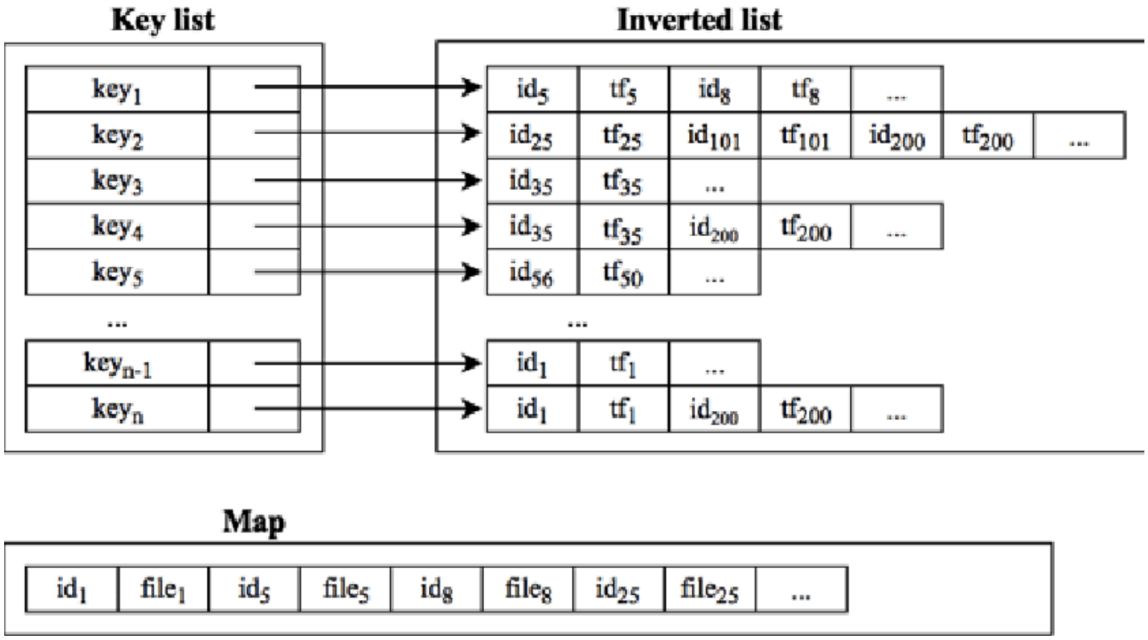
Step 1: Web Crawler

Design and implement a web crawler to extract content from up to 10,000 documents within the muhlenberg.edu domain. Make sure to store content in uniquely named files (e.g., doc1, doc2, etc.), so you don't have to crawl the sites again. Make sure to put the URL of the in the first line of the file.

Implement polite crawling practices and efficient data retrieval with robust error handling.

Step 2: Inverted Index

Build an inverted index for efficient data organization. Include preprocessing features for StopWord removal and Stemming, with toggle options. Use advanced data structures for the index and incorporate text normalization techniques. Save the index in a quickly accessible format. The format from the image is a recommended format.



Step 3: Retrieval

- Develop a retrieval system using TF-IDF and cosine similarity and implement vector space models and efficient similarity calculations for optimized retrieval. Don't forget to apply consistent preprocessing to both documents and search queries.

Step 4: Display of Results

- Design a user interface to display the top ten search results with functional hyperlinks.

Step 5: Analysis Using Cranfield **Collections**

- Use the Cranfield Collections to evaluate your search engine in four scenarios:
    - StopWords ON, Stemming ON
    - StopWords OFF, Stemming ON
    - StopWords ON, Stemming OFF
    - StopWords OFF, Stemming OFF

Cranfield Collections Structure: Consists of test queries and a collection of documents with manual relevance judgments.

Calculating Precision and Recall:

- Run each Cranfield query through your search engine in all four scenarios.
- Retrieve and compare the top N documents against Cranfield's judgments.
- Calculate precision and recall for each query, then average these metrics across all queries for each scenario.