Java Web Crawler

Name	ID	Group
Omar Mohamed Abdelmelek	20220232	S4
Ahmed Saber Ahmed	20220021	S4
Belal Mohamed	20220086	S8
Abdelrahman Ahmed Lotfy	20220474	All
Mohab Mohamed	20220472	All
Anas Mohamed Monir	20220072	S6

1. Crawler Design

The crawler is designed to initiate from a set of given seed URLs, Its main steps are:

Extracting Links:
It downloads the HTML content of each page.

Normalizing Links:

It scans for all anchor (<a>) tags and filters valid Wikipedia links (those starting with /wiki/).

Filtering Links:
Skips non-article pages such as special pages, help pages, and external references.

Storing Data:
It saves the page title and cleaned plain text content.

Breadth-First Crawling:
New links are crawled in a breadth-first manner with a limit on the number of pages to prevent infinite crawling.

2. Inverted Index

An inverted index maps each term to the list of documents it appears in, along with the term's frequency.

Text Processing

Tokenization: Split text into words.

Stop Words Removal: Eliminate common English words like "the," "and," etc.

Stemming: Reduce words to their root forms.

Index Construction

For each remaining term, document IDs and term frequencies are recorded.

Terms already present in the index are updated with new postings.

This structure allows fast and efficient document retrieval during search.

3. TF-IDF and Cosine Similarity

TF-IDF Calculation:

Each document and query is converted into a TF-IDF weighted vector.

- **Term Frequency (TF):** 1 + log10(tf) where tf is the frequency of the term in the document.
- **Inverse Document Frequency (IDF):** idf(t)=log10(N/ df(t)).
- **TF-IDF Weight:** (TF * IDF).

Cosine Similarity

- Similarity between a document and the query is computed as:
  - *Dot Product of Document Vector and Query Vector / (Norm of Document * Norm of Query).*
- Higher scores indicate higher relevance.

4. User Manual

How to Run

Compile the Code:

- Ensure you have Java 17 or higher installed.
  Compile using:
  ```
  javac -cp .;jsoup 1.14.3.jar *.java
  ```

Run the program:

-
  ```
  java -cp .;jsoup-1.14.3.jar Main
  ```

Query Input

When prompted Enter your search query:

type any phrase.
Example:
pharaohs

The program will display the top matching documents ranked by relevance.
Type exit to quit the program.

Thank You