Java Web Crawler Project Report

1. Web Crawler Design

The web crawler is designed to take a list of seed URLs and download the text content of those pages. It uses the Jsoup library to connect to each URL, parse the HTML, and extract the plain text for indexing. The crawler also extracts hyperlinks from the seed pages to allow optional expansion if needed. Special care is taken to handle links containing special characters like "#".

Key Features:

- Fetches content from given seed URLs.
- Extracts text while ignoring irrelevant HTML tags.
- Collects valid hyperlinks for potential further crawling.
- Normalizes URLs to handle special cases like "#".

2. Inverted Index Building

An InvertedIndex class is used to map each stemmed term to:

- A list of documents where the term appears.
- The frequency of the term within each document.

When adding a document:

- The text is tokenized into words.
- Stop words (common, irrelevant words) are filtered out.
- Remaining words are stemmed using a Stemmer Class with a Method "Stem".
- The term-document-frequency mappings are stored.

3. TF-IDF and Cosine Similarity Computation

- **TF-IDF Calculation:**

For a term t in document d:
TF (Term Frequency): 1 + log10(frequency of term in document)
IDF (Inverse Document Frequency): log10(total number of documents / number of documents containing the term)
Thus:
TF-IDF(t, d) = TF * IDF

Each document is represented by a vector of TF-IDF values, one per term.

- **Query Vector:**

Queries are processed similarly:
- Tokenize the query.
- Remove stop words.
- Stem the remaining terms.
- Build a query vector with TF-IDF scores.

- **Cosine Similarity:**

To rank documents, cosine similarity between query vector q and document vector d is calculated:
Cosine(q, d) = (q · d) / (||q|| * ||d||)

Where:
q · d is the dot product.
||q|| and ||d|| are vector magnitudes (norms).

Higher cosine values indicate more relevant documents.

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

How to Input a Query

When prompted Enter your search query:

type any phrase for example: “pharaohs”.

The program will display the top 10 matching documents ranked by relevance.
Type exit to quit the program.