Tool Structure:

1. Indexing Process

The indexing reads documents from a directory that is specified, processes them (removes stopwords, cleans text), and creates an inverted index. The index is serialized to a file.

2. Querying Process

The querying loads created inverted index to search documents using Boolean queries. It processes the query, retrieves matching documents, and prints the result.

Compiling:

First compile the project.

For this project use this command:

javac -cp "lib/pdfbox-app-2.0.32.jar" -d bin src\main\java\com\example\invertedindex*.java

Usage Instructions:

1. Indexing Documents

After compiling index a collection of documents, use the Indexing class. The command requires two arguments:

- docs_path: The directory containing the documents to be indexed.
- data_path: The directory where the inverted index and dictionary will be stored.

  For this project use this command:

  java -cp "lib/pdfbox-app-2.0.32.jar;bin" com.example.invertedindex.Indexing docs/ data/

2. Querying Documents

Finally query the indexed data, use the Querying class. The command requires one argument:

- query: A Boolean query string.

  For this project use this command:

  java -cp "lib/pdfbox-app-2.0.32.jar;bin" com.example.invertedindex.Querying "string-to-find"

Requirements:

- **Java version**: Project uses Java version 11 or later.
- **PDFBox**: Make sure the pdfbox library is inside of lib/ for processing PDF documents.
- **DOCX Processor**: A custom DocxProcessor class handles the extraction of text from .docx files.