Document Content Search using Elasticsearch

Students:

Milconu Adina-Maria

Vlășceanu Daniela-Claudia

Contents

Project Description

This project implements a web-based document management and search system that allows users to upload documents (PDF and Word), extract their textual content, store it in Elasticsearch, and perform full-text and partial searches with highlighted results.

The system provides a REST API and a simple web interface where users can:

- Upload documents
- View list uploaded documents
- Delete documents
- Search inside document content using single or multiple terms
- View highlighted matches inside documents

Elasticsearch is used as the search engine, while Apache Tika is used for text extraction from documents. The backend is implemented using Spring Boot (Java).

List of Use Cases

UC1 – Upload Document

- Description: The user uploads a PDF or Word document.
- Result: The document content is extracted and indexed in Elasticsearch.

UC2 – List Documents

- Description: The user requests a list of all uploaded documents.
- Result: The system returns document IDs and filenames.

UC3 – View Document Content

- Description: The user selects a document.
- Result: The full extracted content of the document is displayed.

UC4 – Search in Document Content

- Description: The user searches using one or multiple keywords.
- Result: Matching document fragments are returned with highlighted terms.

UC5 – Delete Document

- Description: The user deletes a document.

- Result: The document is removed from Elasticsearch.

Elasticsearch Mapping

Mapping used in the project

Index: documents

PUT documents

```
{
  "mappings": {
    "properties": {
      "filename": { "type": "keyword" },
      "content": { "type": "text" },
      "file_type": { "type": "keyword" },
      "word_count": { "type": "integer" },
      "upload_date": { "type": "date" }
    }
  }
}
```

Implementation description

Backend

- Implemented using Spring Boot
- REST APIs handle: upload document, get list of all documents, view content of document, search in the content of a document, delete document

Text Extraction

- Apache Tika extracts content from PDF and Word documents
- Extracted text is indexed into Elasticsearch

Search

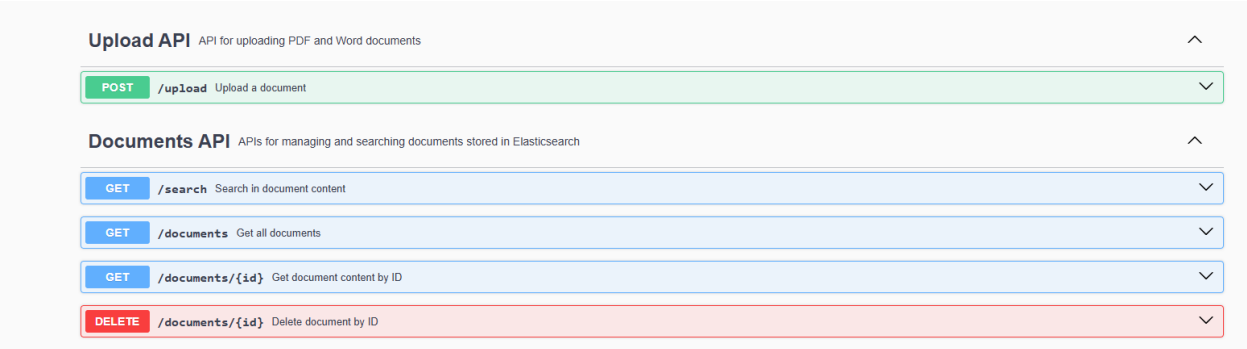
- Supports: Partial search, Multi-term search
- Elasticsearch highlights matched terms

Frontend

- HTML + JavaScript UI
- Sidebar displays documents
- Clicking search results opens the correct document with highlight

API Documentation – Swagger

APIs used are documented in swagger:



Postman Testing

Collection of Postman was created to test the APIs: