浙江大学计算机科学与技术学院

Java课程设计课程报告
2024-2025学年秋冬学期

题目：简易搜索引擎
姓名：姜雨童
学号：3220103450
所在专业：计算机科学与技术
所在班级：计科2202

目录

1 引言

本次项目开发的是一个在终端实现交互的简易搜索引擎。这是一个综合性较强的题目，不仅用到了Java编程知识，还需要学会使用已有的Java库，正确编写依赖关系并构建项目。因此该项目不仅提高了我的编程水平，也让我对Java语言中的各项功能有了更深的理解和掌握，为以后的工作打下一定基础。

1.1 设计目的

本项目使用Java语言开发一个能够处理多种格式文档（如PDF、Word、HTML等）的搜索引擎，并允许用户根据自己的需求进行搜索。具体功能如下：

1. 用户能自行输入被索引文件所在的文件夹路径（文件夹目录下所有文件均被纳入索引）；
2. 用户输入被搜索内容后，搜索引擎输出被搜索内容所在文件名和出现该内容的相关行；
3. 本项目不需要实现GUI界面，输入与输出均在终端进行。

1.2 设计说明

本项目采用Java程序设计语言，使用Maven构建项目依赖，由本人独立完成。其中用到了jsoup、apache tika、lucene、poi四个库：

- Jsoup：用Java HTML Parser来抽取HTML文件中的内容； https://jsoup.org/
- Apache Tika：从多种格式（如pdf）的文档中提取文本内容； https://tika.apache.org/
- Apache Poi：从word文档（.doc, .docx）中提取文本内容； https://poi.apache.org/
- Apache Lucene：构建高效、可扩展的文档索引。 https://lucene.apache.org/

2 总体设计

2.1 功能模块设计

本项目主要分为三个功能模块，外加一个集成模块（SearchEngineApp），功能模块简述如下：

1. DocumentIndexer：负责文档的索引和存储；
2. SearchEngine：负责接收用户的搜索请求，并查询索引；
3. SearchResult：负责展示搜索结果。

项目总体功能如下图所示：

图1 总体功能图

2.2 流程图设计

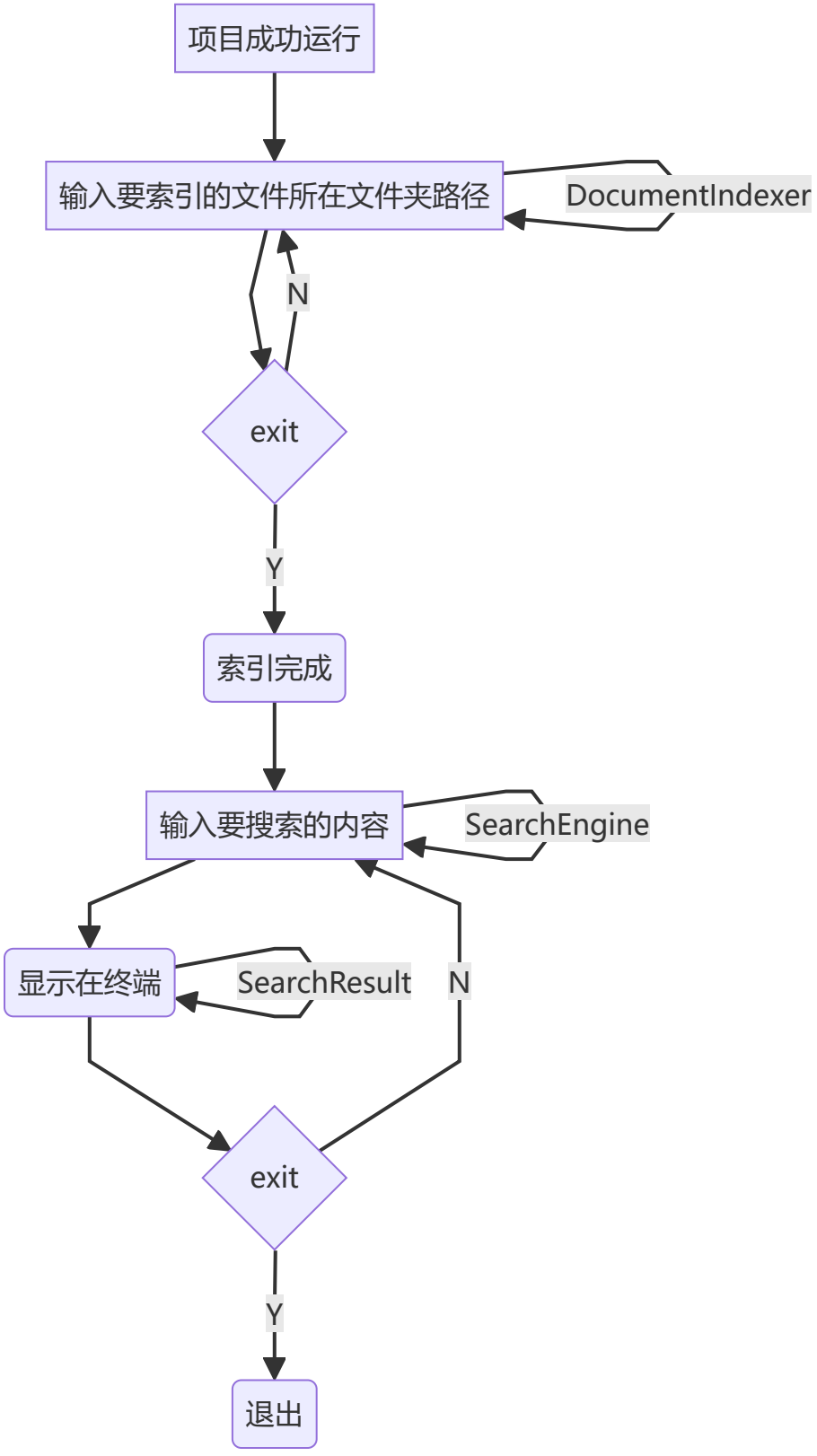


图2 总体流程图

3 详细设计

3.1 SearchEngineApp

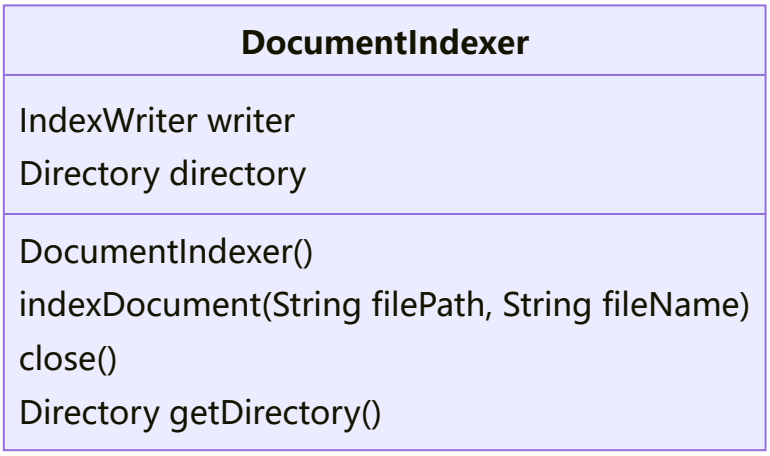
该程序为本项目主程序，实现两个功能：扫描用户输入文件夹路径（用户不输入时默认为src\main\resource）中的所有文件，并对其进行索引；在已索引的文件中搜索用户输入关键词，返回相关结果。

异常处理模块可能捕捉的异常如下：IOException（处理文件访问相关错误）、TikaException（处理文档解析相关错误）、ParseException（处理搜索查询解析错误）。

数据结构等内容在其他模块进行说明。

3.2 DocumentIndexer

DocumentIndexer 是一个用于索引文档的类，使用了 Apache Lucene 和 Apache Tika 来解析和索引不同格式的文件（如 PDF、DOC、DOCX 和 HTML）。以下是该类的UML图：



以下是UML图中有关数据和方法的详细说明：

3.2.1 成员变量：

- `writer`：`IndexWriter`，用于写入索引。
- `directory`：`Directory`，表示索引的存储位置，使用 `RAMDirectory` 在内存中存储索引。

3.2.2 构造方法：

- `DocumentIndexer()`：初始化 `directory` 和 `writer`，设置 Lucene 的索引配置。

3.2.3 方法：

1. **indexDocument(String filePath, String fileName)**：
   - 功能：根据给定的文件路径和文件名来索引文档。
   - 参数：
     - `filePath`：文档的路径。
     - `fileName`：文档的名称。
   - 异常：

  - `IOException`: 读取或写入索引时发生错误。
  - `SAXException`: 解析文档时发生错误。
  - `TikaException`: 使用 Tika 解析文档时发生错误。
- 处理：根据文件后缀名判断文件类型，使用相应的解析器提取文本内容，并通过 `IndexWriter` 将文档内容和文件名添加到索引中。

2. **close()**:

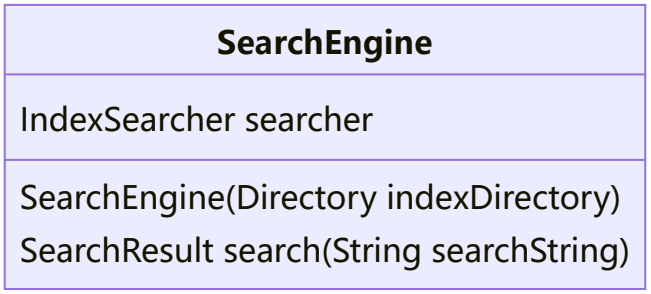
- 功能：关闭索引写入器，释放资源。
- 异常：`IOException`: 关闭索引写入器时发生错误。

3. **getDirectory()**:

- 功能：获取索引存储的目录。

3.3 SearchEngine

SearchEngine 类负责在生成的 Lucene 索引中执行搜索操作。它提供了构造函数以接收索引目录，并包含一个搜索方法，用于根据搜索字符串查找匹配的文档。以下是该类的UML图：



以下是UML图中有关数据和方法的详细说明：

3.3.1 成员变量：

- `searcher`: `IndexSearcher`，用于执行搜索操作。

3.3.2 构造方法：

- **SearchEngine(Directory indexDirectory)**
  - 功能：初始化 `IndexSearcher`，接收一个 `Directory` 类型的参数，该目录包含已创建的 Lucene 索引。
  - 异常：`IOException`: 开启索引时发生错误。

3.3.3 方法：

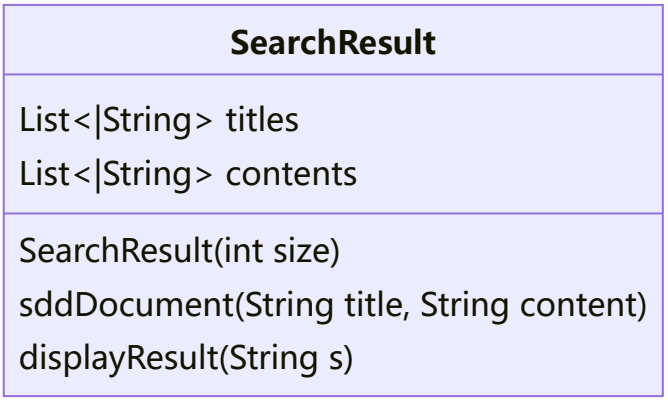
1. **search(String searchString)**

- 功能：根据给定的搜索字符串在索引中查找匹配的文档。
- 参数：`searchString`: 用于查询的搜索字符串。
- 返回：`SearchResult`: 包含搜索结果的对象。
- 异常：

  - `IOException`: 搜索过程中出现错误。
  - `ParseException`: 搜索字符串无效时抛出。
- 处理：使用 `QueryParser` 对搜索字符串进行解析并生成查询，调用 `IndexSearcher` 执行查询并返回匹配的文档。

3.4 SearchResult

SearchResult 类用于表示搜索的结果，包含多个文档的标题和内容。该类提供方法用于添加文档信息以及显示搜索结果。以下是该类的UML图：



以下是UML图中有关数据和方法的详细说明：

3.4.1 成员变量：

- `titles`: `List<String>`，保存搜索到的文档标题的列表。
- `contents`: `List<String>`，保存搜索到的文档内容的列表。

3.4.2 构造方法：

- **SearchResult(int size)**
  - 功能：根据指定的大小初始化标题和内容的列表。
  - 参数：`size`: 预计要存储的文档数量，可以优化列表的初始容量。

3.4.3 方法：

1. **addDocument(String title, String content)**:
   - 功能：将文档的标题和内容添加到搜索结果中。
   - 参数：
     - `title`: 文档的标题。
     - `content`: 文档的内容。
2. **displayResults(String s)**:
   - 功能：显示与搜索字符串相关的搜索结果。
   - 参数：`s`: 进行搜索的字符串。

- 处理：打印出搜索结果的标题和对应的内容，突出显示与搜索字符串匹配的部分

4 测试与运行

4.1 测试环境

- Apache Maven 3.9.9
- Java version: 1.8.0_431, vendor: Oracle Corporation
- 依赖包：
  - Tika：1.28.5
  - Lucene：8.11.0
  - Poi：5.2.3
  - Jsoup：5.15.3

4.2 程序测试

（注：本项目文件均采用GBK编码模式。）

构建项目：

```
mvn clean install
```

```
D:\PointMe\HW\Java\hw3>mvn clean install
[INFO] Scanning for projects...
[INFO]
[INFO] ----------------------< com.example:search-engine >-----------------------
[INFO] Building search-engine 1.0-SNAPSHOT
[INFO]   from pom.xml
[INFO] --------------------------------[ jar ]---------------------------------
[INFO]
[INFO] --- clean:3.2.0:clean (default-clean) @ search-engine ---
[INFO] Deleting D:\PointMe\HW\Java\hw3\target
```

```
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  3.124 s
[INFO] Finished at: 2024-12-10T12:20:31+08:00
[INFO] ------------------------------------------------------------------------

D:\PointMe\HW\Java\hw3>
```

运行：

```
mvn exec:java -Dexec.mainClass="com.example.SearchEngineApp"
```

```
D:\PointMe\HW\Java\hw3>mvn exec:java -Dexec.mainClass="com.example.SearchEngineApp"
[INFO] Scanning for projects...
[INFO]
[INFO] ----------------------< com.example:search-engine >----------------------
[INFO] Building search-engine 1.0-SNAPSHOT
[INFO]   from pom.xml
[INFO] --------------------------------[ jar ]---------------------------------
[INFO]
[INFO] --- exec:3.5.0:java (default-cli) @ search-engine ---
请输入要索引的文件夹路径('exit'退出，'enter'使用默认路径resource):
```

输入回车和exit对默认路径的文件夹（src\main\resource）下的文件进行索引：

```
文档已索引: D:\PointMe\HW\Java\hw3\src\main\resource\06 - Methods(4).pdf
文档已索引: D:\PointMe\HW\Java\hw3\src\main\resource\07 - Single-Dimensional Arrays(5).pdf
文档已索引: D:\PointMe\HW\Java\hw3\src\main\resource\08 - Multidimensional Arrays(5).pdf
文档已索引: D:\PointMe\HW\Java\hw3\src\main\resource\cpszd.html
文档已索引: D:\PointMe\HW\Java\hw3\src\main\resource\hw3.pdf
文档已索引: D:\PointMe\HW\Java\hw3\src\main\resource\Java_3220103450_hw1_report.pdf
文档已索引: D:\PointMe\HW\Java\hw3\src\main\resource\Java_3220103450_hw2_report.pdf
文档已索引: D:\PointMe\HW\Java\hw3\src\main\resource\test1.docx
文档已索引: D:\PointMe\HW\Java\hw3\src\main\resource\test2.html
文档已索引: D:\PointMe\HW\Java\hw3\src\main\resource\一个word文档.docx
请输入要索引的文件夹路径('exit'退出，'enter'使用默认路径resource):
exit
索引完成。
请输入搜索关键词（输入 'exit' 退出）：
```

输入内容进行搜索：

- loop（测试了docx和pdf）：

```
exit
索引完成。
请输入搜索关键词（输入 'exit' 退出）：
loop
----------RESULT FOR: loop-----------------
Title: 05 - Loops(4).pdf
|_Content: ? To write programs for executing statements repeatedly using a while loop
|_Content: ? To follow the loop design strategy to develop loops (§§5.2.1–5.2.3).
|_Content: ? To control a loop with a sentinel value (§5.2.4).
|_Content: ? To write loops using do-while statements (§5.3).
|_Content: ? To write loops using for statements (§5.4).
|_Content: ? To discover the similarities and differences of three types of loop statements
|_Content: ? To write nested loops (§5.6).
```

```
---
Title: 一个word文档.docx
|_Content: Loop，loop，loop！会被测试到吧。
---
Title: 07 - Single-Dimensional Arrays(5).pdf
|_Content: ? To simplify programming using the foreach loops (§7.2.7).
|_Content: i ( =5) < 5 is false. Exit the loop
|_Content: Enhanced for Loop (for-each loop)
|_Content: JDK 1.5 introduced a new for loop that enables you to traverse the complete array
|_Content: Using a loop:
|_Content: the loop.
|_Content:      Enhanced for Loop (for-each loop)
---
Title: 06 - Methods(4).pdf
|_Content: A variable declared in the initial action part of a for loop
|_Content: header has its scope in the entire loop. But a variable
|_Content: declared inside a for loop body has its scope limited in the
|_Content: loop body from its declaration and to the end of the block
--------------------------
请输入搜索关键词（输入 'exit' 退出）：
```

- 测试（测试了docx和pdf）：

```
----------------------------
请输入搜索关键词（输入 'exit' 退出）：
测试
----------RESULT FOR: 测试------------------
Title: 一个word文档.docx
|_Content: 这是第二个word测试文档。
|_Content: Loop，loop，loop！会被测试到吧。
|_Content: 测试文档（x）测试样例一览（?）
---
Title: Java_3220103450_hw2_report.pdf
|_Content: 撰写相关?档并进?测试说明和性能分析。
|_Content: 在测试结果中可以看到，对可变类的修改成功执?，?对不可变类则不会进?修改：
|_Content: 测试结果如下：
|_Content: 3 完整测试文档&结果
|_Content:      完整测试文档&结果
---
Title: test1.docx
|_Content: 这是一个.docx的测试文档。
---
Title: hw3.pdf
|_Content: 详细设计、测试与运行、总结
---
Title: Java_3220103450_hw1_report.pdf
|_Content: 尽管两个?任务在实现上有共通之处，为了测试和使??便，我将两个任务分别写在 SudoGen.java 和
|_Content: Chapter 3: 结果测试
|_Content:      Chapter 3: 结果测试
```

- 产品（测试了html和docx）：

```
----------------------------
请输入搜索关键词（输入 'exit' 退出）：
产品
----------RESULT FOR: 产品------------------
Title: cpszd.html
|_Content: Welcome to CPSZD 产品就是产品啊，家产真的特别美味555。饭呢我要吃饭，我的产品饭谁给我补上
---
Title: 一个word文档.docx
|_Content: 我产品好真。
----------------------------
请输入搜索关键词（输入 'exit' 退出）：
```

- Lucene（测试了pdf）：

```
请输入搜索关键词（输入 'exit' 退出）：
Lucene
----------RESULT FOR: Lucene------------------
Title: hw3.pdf
|_Content: - Lucene库：构建高效、可扩展的文档索引。
|_Content: Apache LuceneTM is a high-performance, full-featured text search engine
----------------------------
请输入搜索关键词（输入 'exit' 退出）：
```

输入exit退出程序：

```
----------------------------
请输入搜索关键词（输入 'exit' 退出）：
exit
[INFO] ------------------------------------------------------------------------
[INFO] BUILD SUCCESS
[INFO] ------------------------------------------------------------------------
[INFO] Total time:  11:23 min
[INFO] Finished at: 2024-12-10T12:33:18+08:00
[INFO] ------------------------------------------------------------------------

D:\PointMe\HW\Java\hw3>
```

_

添加多个文件夹路径对其下文件进行索引：

```
文档已索引：D:\PointMe\HW\Java\hw3\src\main\resource\Java_3220103450_hw2_report.pdf
文档已索引：D:\PointMe\HW\Java\hw3\src\main\resource\test1.docx
文档已索引：D:\PointMe\HW\Java\hw3\src\main\resource\test2.html
文档已索引：D:\PointMe\HW\Java\hw3\src\main\resource\一个word文档.docx
请输入要索引的文件夹路径('exit'退出，'enter'使用默认路径resource):
./src/main/added
文档已索引：D:\PointMe\HW\Java\hw3\.\src\main\added\added.html
文档已索引：D:\PointMe\HW\Java\hw3\.\src\main\added\新建 Microsoft Word 文档.docx
请输入要索引的文件夹路径('exit'退出，'enter'使用默认路径resource):
exit
索引完成。
请输入搜索关键词（输入 'exit' 退出）：
```

输入搜索内容进行测试：

- added（仅含新增路径文件夹下的文件）：

```
索引完成。
请输入搜索关键词（输入 'exit' 退出）：
added
----------RESULT FOR: added------------------
Title: added.html
|_Content: Added Page This is the added page.
---
Title: 新建 Microsoft Word 文档.docx
|_Content: 只有添加了该文档所在文件夹路径才会索引added文档。
----------------------------
请输入搜索关键词（输入 'exit' 退出）：
```

- 文档（含有默认文件夹和新增文件夹下的文档）：

```
请输入搜索关键词（输入 'exit' 退出）：
文档
----------RESULT FOR: 文档-----------------
Title: hw3.pdf
|_Content: 的文档（如PDF、Word、HTML等），并允许用
|_Content: - Apache Tika库：用于从多种格式的文档中提取
|_Content: - Lucene库：构建高效、可扩展的文档索引。
|_Content: 解析HTML文档中的信息（如标题、段落、链接等）。
|_Content: 类型文件的元数据。它支持多种文件格式，包括文档、图片、音频和
|_Content: ? 文档要求：
|_Content: ? 作业包括：java文件 + 文档 + 数据
---
Title: 新建 Microsoft Word 文档.docx
|_Content: 默认情况下，不会索引该文档。
|_Content: 只有添加了该文档所在文件夹路径才会索引added文档。
---
Title: 一个word文档.docx
|_Content: 这是第二个word测试文档。
|_Content: 测试文档（x）测试样例一览（?）
```

–

输入错误文件夹路径程序输出错误提示：

```
[INFO] --- exec:3.5.0:java (default-cli) @ search-engine ---
请输入要索引的文件夹路径('exit'退出，'enter'使用默认路径resource):
aaa
指定的路径不存在或不是一个文件夹。请重新输入：
请输入要索引的文件夹路径('exit'退出，'enter'使用默认路径resource):
```

–

可以看到，程序运行均符合预期，测试通过。

5 总结

该项目实现了一个简易的终端交互的搜索引擎，需要用到Tika等其他的库。整个过程对我来说难度较大，开始构建项目时因为版本不匹配的原因出过很多问题，比如Java版本和使用的库不匹配、使用的库在不同版本时对某些类进行了修改导致名称/用法改变等。另外有一个印象很深的内容是在进行对doc/docx文档内容的提取时，使用Tika库内的方法程序会报错显示文档已加密`[Solved]org.apache.tika.exception.EncryptedDocumentException: Unable to process: document is encrypted`，但是我并没有对word文档进行加密，最后使用了Poi的库解决了该问题。

同时，这也是Java课程中第一个独立完成完整项目（并对代码和文档做出要求）的任务，不仅提高了我的编程能力，也让我对程序开发流程有了更清晰的认识。写下项目的版权声明还有各个函数的注释时，第一次体会到由自己创造的，从零到一的喜悦。

本项目实现的功能并不复杂，程序也还有很大的改进空间，但是项目编写时收获的经历和体验弥足珍贵。