

浙江大学计算机学院

Java 程序设计课程报告

2024—2025 学年秋冬学期

题目	HW3：搜索引擎
学号	3220104147
学生姓名	余卓耘
所在专业	软件工程
所在班级	2202

目录

1 引言.....	1
1.1 设计目的.....	1
1.2 设计说明.....	1
2 总体设计.....	2
2.1 功能模块设计.....	2
2.2 流程图设计.....	3
3 详细设计.....	5
3.1 核心模块设计.....	5
3.2 Indexer 类的设计.....	7
3.3 Parser 类的设计.....	8
3.4 Searcher 类的设计.....	8
3.5 WebCrawler 类的设计.....	9
4 测试与运行.....	10
4.1 程序测试.....	10
4.2 程序运行.....	11
5. 总结与参考文献.....	17

1 引言

本次开发的是一个搜索引擎,使用 JAVA 语言编写,结合 Jsoup, apache tika, lucene 等来搭建搜索引擎,支持网络爬虫搜索和利用本地文档资料搜索,支持解析 txt、doc、pdf、html 等文件格式,在命令行界面直接交互。

1.1 设计目的

本项目旨在开发一个功能较完善的文件搜索引擎系统,实现对本地文件和网页内容的智能检索。具体功能如下:

(1) 支持多源索引功能,用户可以选择添加本地文件目录(\documents)或网页 URL 进行内容索引。系统能够实时监控本地文件变化,自动更新索引内容。

(2) 提供灵活的搜索范围选择,用户可以选择在所有内容、仅本地文件或仅网页内容中进行搜索,提高搜索精确度。

(3) 实现智能网页爬取搜索功能,系统可以自动分析网页内容,提取关键信息并建立索引。爬虫具有深度限制和访问控制机制,可自行输入爬取深度(1-3,3 最深),避免过度爬取

(4) 实现本地文件搜索功能,搜索结果展示丰富。本地文件显示文件相对路径、类型和内容预览,对于网页内容则显示标题、URL 和相关内容片段,帮助用户快速定位所需信息。

(5) 采用面向对象设计思想,将索引器、解析器、搜索器等功能模块解耦,便于系统的维护和扩展。系统支持多种文件格式的解析,包括 txt、docx、pdf、html 等常见格式。

(6) 提供友好的命令行交互界面,用户可以通过简单的菜单操作完成所有功能,操作直观,使用方便。系统具有完善的异常处理机制,确保稳定运行。

1.2 设计说明

本程序采用 Java 程序设计语言,在 VSCode 平台下编辑、编译与调试。具体程序由 3220104147-余卓耘独立完成。运行方式详见压缩包中的 ReadMe.txt 文

件。

2 总体设计

2.1 功能模块设计

本程序需实现的主要功能有：

- (1) 支持多源索引功能：用户可以选择添加本地文件目录或网页 URL 进行内容索引，如代码所示：

```
switch (choice) {
case 1: // 添加本地目录
    handleLocalDirectory(scanner);
    break;
case 2: // 添加网页索引
    handleWebCrawling(scanner);
    break;
```

- (2) 提供灵活的搜索范围：用户可以选择在所有内容、仅本地文件或仅网页内容中进行搜索，提高检索精确度。
- (3) 实现智能网页爬取：系统可以自动分析网页内容，提取关键信息并建立索引。爬虫具有深度限制和访问控制机制。
- (4) 提供丰富的搜索结果展示：包括文件路径、类型、内容预览等信息，帮助用户快速定位。

程序的总体功能如图 1 所示：

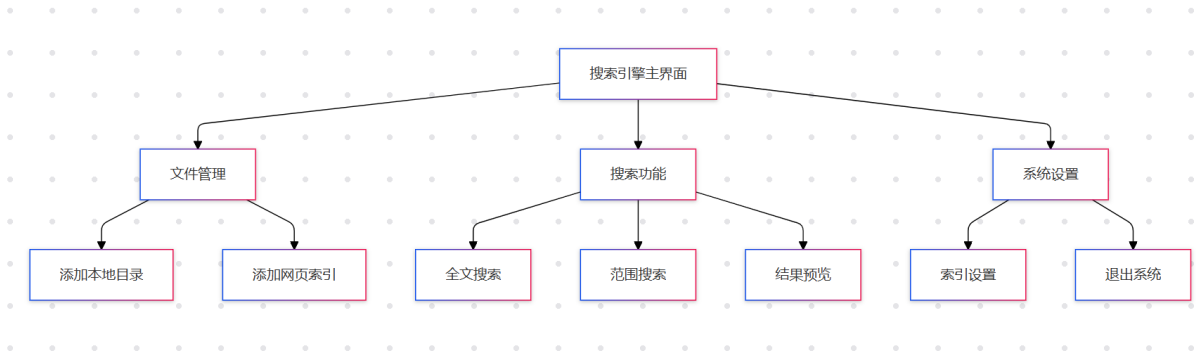


图 1 总体功能图

2. 2 流程图设计

程序总体流程如图 2 所示：

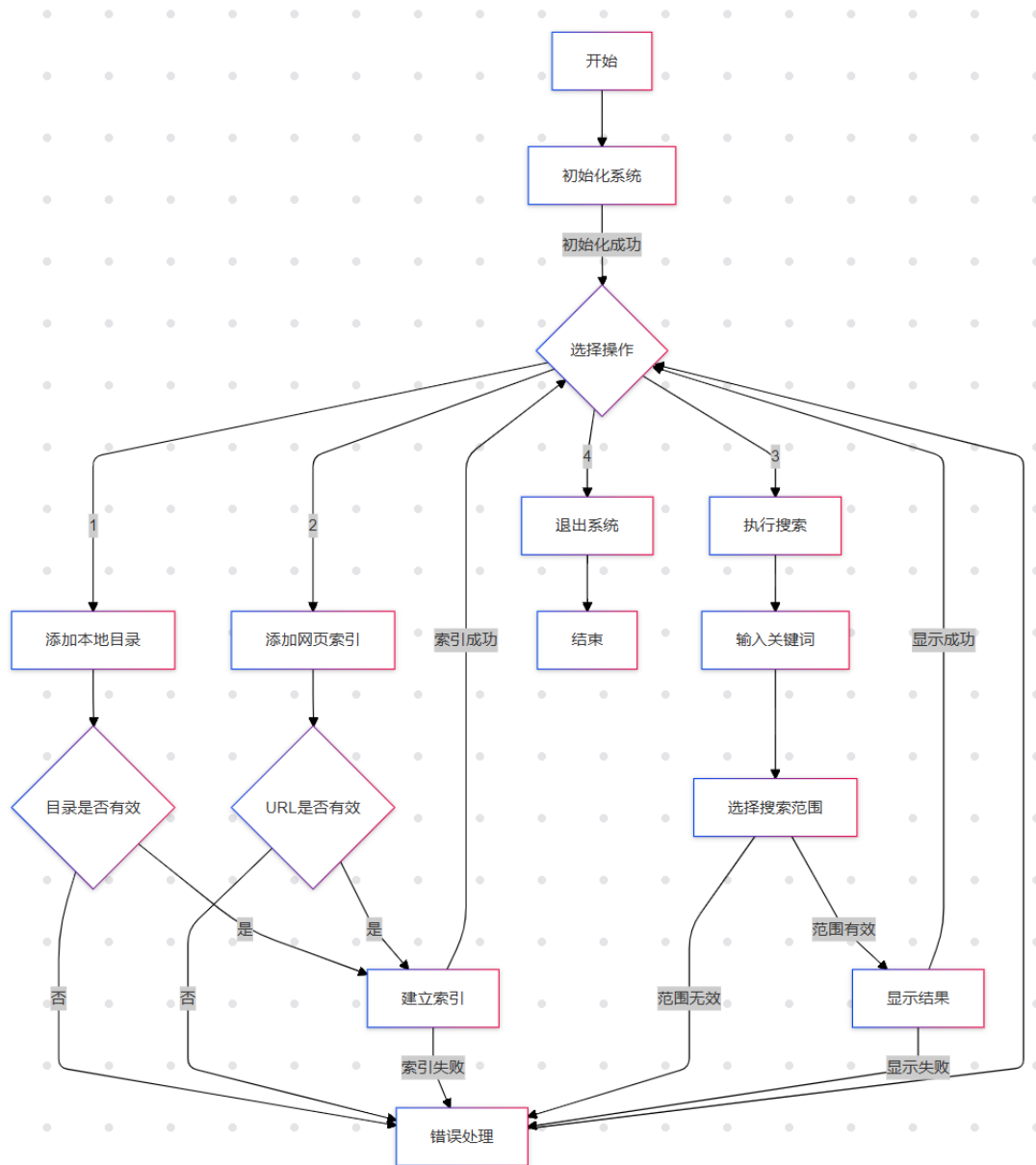


图 2 总体流程图

主功能——搜索流程如下图所示：

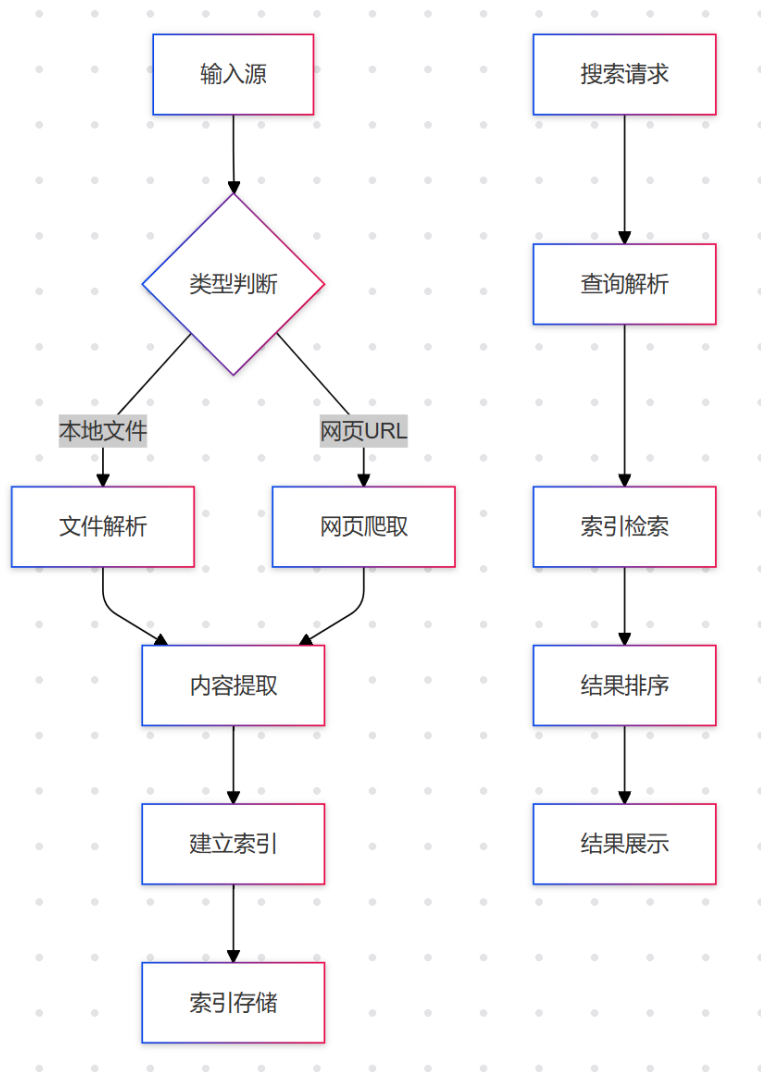


图 3 搜索功能流程图

3 详细设计

3.1 核心模块设计

Main 类是整个搜索引擎的核心类, 实现了文件索引和搜索功能。主要包含以下几个关键类:

索引器(Indexer): 负责建立和维护文档索引

解析器(Parser): 负责解析不同格式的文档内容

搜索器(Searcher): 提供搜索功能

网页爬虫(WebCrawler): 负责网页内容的获取和分析

UML 图如下:

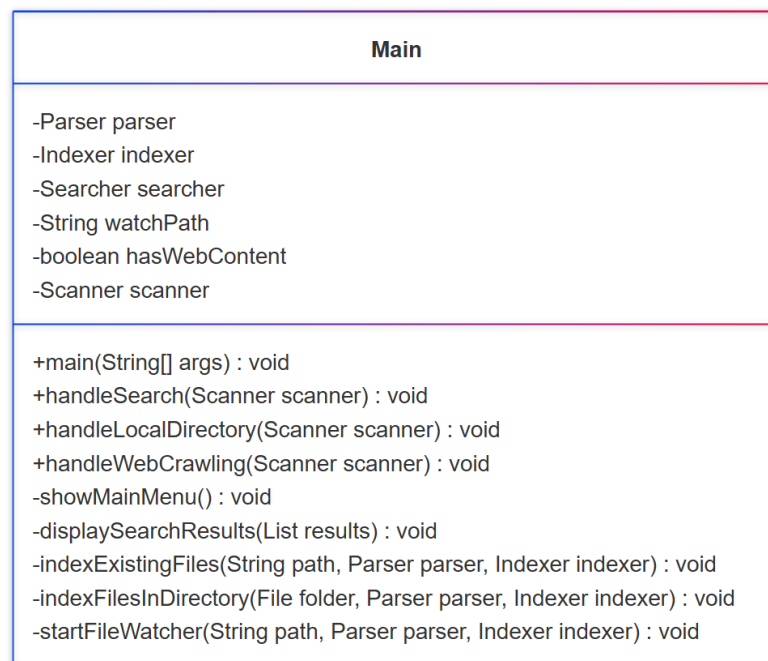


图 4 Main 类 UML 图

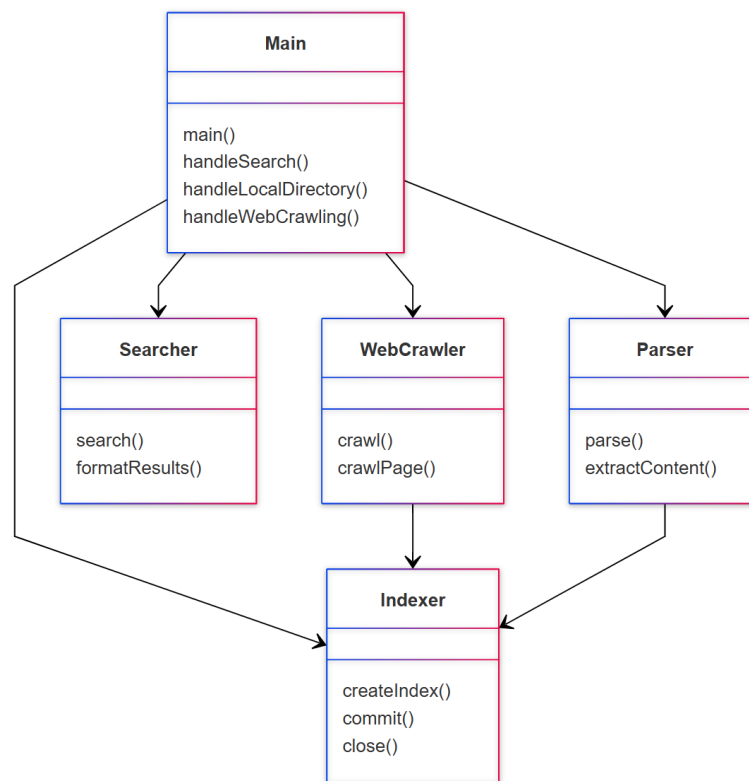


图 5 模块关系 UML 图

以下是 UML 图中有关数据和方法的详细说明：

(1) 成员变量

- ① parser 是 Parser 类型的对象,用于解析不同格式的文档内容；
- ② indexer 是 Indexer 类型的对象,负责创建和维护文档索引；
- ③ searcher 是 Searcher 类型的对象,提供搜索功能；
- ④ watchPath 是 String 类型的对象，负责监控本地文件目录路径；
- ⑤ hasWebContent 是 boolean 类型的对象，负责判断是否包含网页内容。

(2) 方法

- ① `main()`: 程序入口,初始化系统组件
- ② `handleLocalDirectory()`: 处理本地文件索引；
- ③ `handleWebCrawling()`: 处理网页内容爬取；

- ④ `handleSearch()`: 处理搜索请求;
- ⑤ `displaySearchResults()`: 格式化显示搜索结果。

3.2 Indexer 类的设计

Indexer 类是负责创建和管理索引的核心类。标明 Indexer 类的主要成员变量、方法以及和其他类之间组合关系的 UML 图如下所示:

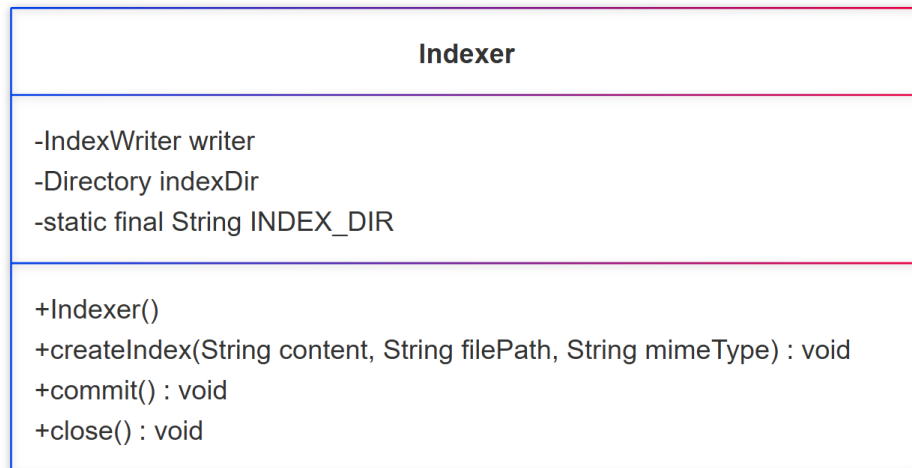


图 6 Indexer 类 UML 图

以下是 UML 图中有关数据和方法的详细说明:

(1) 成员变量

- ① `writer` 是 `IndexWriter` 类型的对象, 用于写入索引
- ② `indexDir` 是 `Directory` 类型的对象, 表示索引存储目录
- ③ `INDEX_DIR` 是静态常量, 定义索引目录路径

(2) 方法

- ① `Indexer()` 构造方法初始化索引写入器
- ② `createIndex()` 方法创建文档索引
- ③ `commit()` 方法提交索引更改
- ④ `close()` 方法关闭索引写入器

3.3 Parser 类的设计

Parser 类是 javax.swing 包中的文档解析类,实现了对多种格式文档的内容提取。所创建的对象 parser 是系统中重要的解析组件之一,负责将不同格式的文档转换为可索引的文本内容。标明 Parser 类的主要成员变量、方法以及和其他类之间组合关系的 UML 图如下所示:



图 7 Parser 类的 UML 图

以下是 UML 图中有关数据和方法的详细说明:

(1) 成员变量

① tika 是 Apache Tika 解析器实例,用于解析不同格式的文档

(2) 方法

① parseFile() 方法负责解析文件内容,支持多种文档格式

② extractContent() 方法从解析结果中提取纯文本内容

3.4 Searcher 类的设计

Searcher 类是搜索功能的核心实现类,提供了强大的全文检索能力。所创建的对象 searcher 负责执行搜索查询并返回相关结果。标明 Searcher 类的主要成员变量、方法以及和其他类之间组合关系的 UML 图如下所示:



图 8 Searcher 类的 UML 图

以下是 UML 图中有关数据和方法的详细说明：

(1) 成员变量

① searcher 是 IndexSearcher 类型的对象, 用于执行搜索

② analyzer 是 Analyzer 类型的对象, 负责文本分析

(2) 方法

① search() 方法执行搜索查询并返回匹配结果

② formatResults() 方法格式化搜索结果以便显示

3.5 WebCrawler 类的设计

WebCrawler 类是网页内容爬取的核心类, 实现了网页的自动获取和分析。所创建的对象 webCrawler 负责从指定 URL 开始爬取网页内容。标明 WebCrawler 类的主要成员变量、方法以及和其他类之间组合关系的 UML 图如下所示：

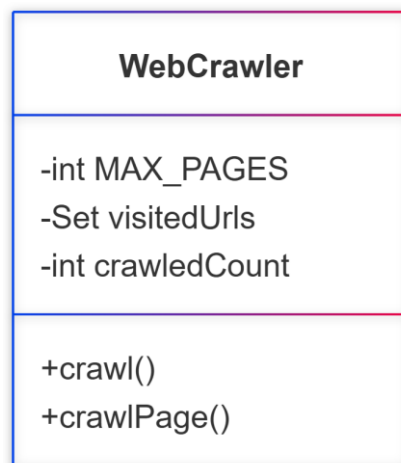


图 9 Searcher 类的 UML 图

以下是 UML 图中有关数据和方法的详细说明：

(1) 成员变量

① MAX_PAGES 是整型常量, 限制最大爬取页面数

② visitedUrls 是 Set 类型的集合, 记录已访问的 URL

③ crawledCount 是整型变量, 统计已爬取的页面数

(2) 方法

① crawl() 方法开始网页爬取过程

② crawlPage() 方法爬取单个网页的内容

4 测试与运行

4.1 程序测试

在程序代码基本完成后，经过不断的调试与修改，最后测试本次所设计的搜索能够正常运行，功能满足实验要求。详情见 4.2 程序运行处截图。

测试数据如下：

我在本地准备了一些 txt、docx、pdf、html 文件，存放于根目录\documents 下，准备了网址 <https://www.runoob.com/java/java-tutorial.html> 供爬取信息。

测试时输入“java”关键词进行搜索。

根目录 documents 文件组织结构如下：

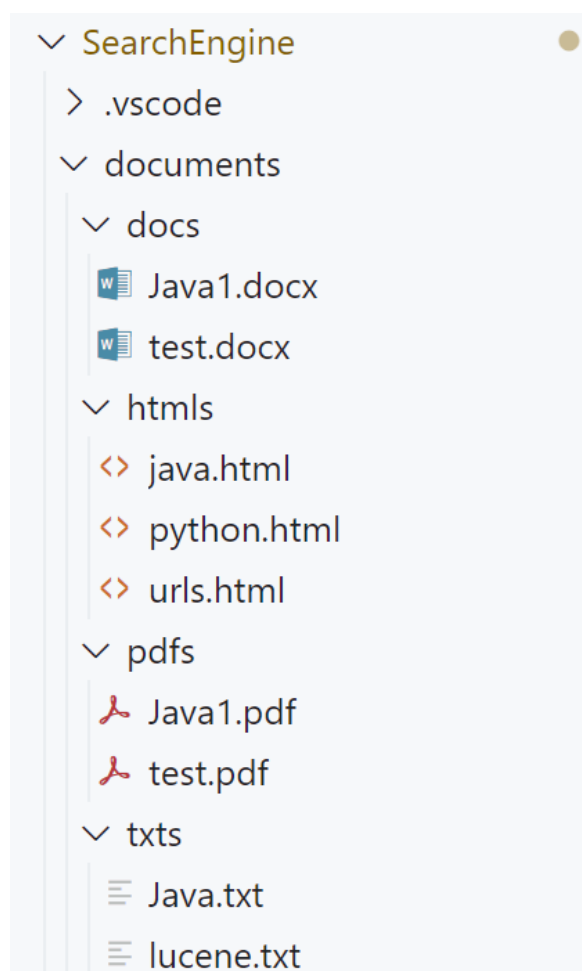


图 10 本地测试文件组织

具体测试样例文件内容可在根目录\documents 下查看，这里限于篇幅只给出.txt 文件的样例：

Java是一种广泛使用的编程语言，具有跨平台、面向对象、泛型编程的特性，广泛应用于企业级Web应用开发、移动应用开发、桌面应用、嵌入式系统和云计算等领域。以下是Java的一些关键特点：

1. 简单性：Java设计时尽量减少了复杂性，去掉了指针直接操作等容易引起错误的部分。
2. 面向对象：Java是一种面向对象的语言，这意味着它支持类和对象的概念，继承、封装和多态等特性。
3. 平台无关性：Java程序是“一次编写，到处运行”（Write Once, Run Anywhere, WORA）。Java程序在Java虚拟机（JVM）上运行，JVM可以在多种操作系统上实现。
4. 健壮性：Java的强类型机制、异常处理和垃圾自动收集等特性使得Java程序非常健壮。
5. 安全性：Java提供了强大的安全机制，包括在运行时进行字节码验证和实施安全策略。

行 1, 列 1 716 个字符

100%

Windows (CRLF)

UTF-8

图 11 txt 文件测试样例

4. 2 程序运行

程序运行主界面、主菜单如图 12 所示：

```
PS D:\User\Desktop\JAVA\2024hw\ZJU_JAVA_2024\hw3\SearchEngine>
欢迎使用文件搜索引擎！
本程序支持本地文件和网页内容的索引与搜索。

请选择操作：
1. 添加本地目录
2. 添加网页索引
3. 开始搜索
4. 退出
```

图 12 程序主界面

添加本地目录如图 13 所示：

```
请选择操作：
1. 添加本地目录
2. 添加网页索引
3. 开始搜索
4. 退出
1
请输入要索引的目录路径（例如：documents）：
documents
开始索引现文件...
已索引：documents\docs\Java1.docx
已索引：documents\docs\test.docx
已索引：documents\htmls\java.html
已索引：documents\htmls\python.html
已索引：documents\htmls\urls.html
已索引：documents\pdfs\Java1.pdf
已索引：documents\pdfs\test.pdf
已索引：documents\txts\Java.txt
已索引：documents\txts\lucene.txt
索引完成！
```

图 13 添加本地目录

添加网络索引如图 14 所示：

```
请选择操作：
1. 添加本地目录
2. 添加网页索引
3. 开始搜索
4. 退出
2
请输入要索引的网页URL（例如：https://www.baidu.com）：
https://www.runoob.com/java/java-tutorial.html

开始爬取网页内容...
已成功索引第 1 个网页
网址：https://www.runoob.com/java/java-tutorial.html
标题：Java 教程 | 菜鸟教程
-----

爬取完成！共爬取 1 个页面
网页内容爬取完成！
```

图 14 添加网络索引

开始搜索后输入关键词如图 15 所示：

```
请选择操作：
1. 添加本地目录
2. 添加网页索引
3. 开始搜索
4. 退出
3
请输入搜索关键词：
java
```

图 15 开始搜索

各种搜索选项的搜索结果如图 16~19 所示：

```
搜索范围：
1. 所有内容
2. 仅本地文件
3. 仅网页内容
1

找到 5 个结果：

结果 #1:
文件路径: webpage:https://www.runoob.com/java/java-tutorial.html
文件类型: webpage
相关度: 0.7414384
内容预览: 网页标题:
网页地址: :https://www.runoob.com/java/java-tutorial.html

-----

结果 #2:
文件路径: documents\docs\Java1.docx
文件类型: doc
相关度: 0.71907103
内容预览: Java是一门高级的、面向对象的编程语言，由James Gosling等人在1995年于Sun Microsystems公司开发。它被设计为具有跨平台能力和高度的可移植性，这使得Java应用程序能够在任何安装了适当Java虚拟机（JVM）的设备上运行。以下是Java的一些独特特点和优势：

1. **跨平台能力**：Java的核心优势之一是其“编写一次，到处运行”（WORA）的理念，这意味着Jav...

-----

结果 #3:
文件路径: documents\pdfs\Java1.pdf
文件类型: pdf
相关度: 0.71907103
内容预览:
Java 是一门高级的、面向对象的编程语言，由 James Gosling 等人在 1995 年于 Sun Microsystems 公司开发。它被设计为具有跨平台能力和高度的可移植性，这使得 Java 应用程序能够在任何安装了适当 Java 虚拟机（JVM）的设备上运行。以下是 Java 的一些独特特点和优势：

1. **跨平台能力**：Java 的核心优势之一是其“编写一...

-----

结果 #4:
文件路径: documents\txts\Java.txt
文件类型: txt
相关度: 0.7185646
内容预览: Java是一种广泛使用的编程语言，具有跨平台、面向对象、泛型编程的特性，广泛应用于企业级Web应用开发、移动应用开发、桌面应用、嵌入式系统和云计算等领域。以下是Java的一些关键特点：

1. 简单性：Java设计时尽量减少了复杂性，去掉了指针直接操作等容易引起错误的部分。

2. 面向对象：Java是一种面向对象的语言，这意味着它支持类和对象的概念，继承、封装和多态等特性。

3. ...

-----

结果 #5:
文件路径: documents\htmls\java.html
文件类型: html
相关度: 0.6971112
内容预览: html> head> Java编程介绍 /head> body> Java编程语言 Java是一种广泛使用的计算机编程语言，拥有跨平台、面向对象、泛型编程的特性。 Java可以开发： 桌面应用程序 Web应用程序 Android应用程序 分布式系统 /body> /html>
```

图 16、17 在所有内容(本地+网页索引)搜索

```

请输入搜索关键词：
java

搜索范围：
1. 所有内容
2. 仅本地文件
3. 仅网页内容
2

找到 4 个结果：

结果 #1:
文件路径：documents\docs\Java1.docx
文件类型：doc
相关度：0.71921813
内容预览：Java是一门高级的、面向对象的编程语言，由James Gosling等人在1995年于Sun Microsystems公司开发。
它被设计为具有跨平台能力和高度的可移植性，这使得Java应用程序能够在任何安装了适当Java虚拟机（JVM）的设备
上运行。以下是Java的一些独特特点和优势：

1. **跨平台能力**：Java的核心优势之一是其“编写一次，到处运行”（WORA）的理念，这意味着Jav...
-----

结果 #2:
文件路径：documents\pdfs\Java1.pdf
文件类型：pdf
相关度：0.71921813
内容预览：
Java 是一门高级的、面向对象的编程语言，由 James Gosling 等人在 1995 年于 Sun
Microsystems 公司开发。它被设计为具有跨平台能力和高度的可移植性，这使得 Java
应用程序能够在任何安装了适当 Java 虚拟机（JVM）的设备上运行。以下是 Java  的一些独特特点和优势：

1. **跨平台能力**：Java 的核心优势之一是其“编写一...
-----

结果 #3:
文件路径：documents\txts\Java.txt
文件类型：txt
相关度：0.71870923
内容预览：Java是一种广泛使用的编程语言，具有跨平台、面向对象、泛型编程的特性，广泛应用于企业级Web应用开
发、移动应用开发、桌面应用、嵌入式系统和云计算等领域。以下是Java的一些关键特点：

1. 简单性：Java设计时尽量减少了复杂性，去掉了指针直接操作等容易引起错误的部分。

2. 面向对象：Java是一种面向对象的语言，这意味着它支持类和对象的概念，继承、封装和多态等特性。

3. ...
-----

结果 #4:
文件路径：documents\htmls\java.html
文件类型：html
相关度：0.6971994

```

图 18 仅本地文件搜索


```

请选择操作：
1. 添加本地目录
2. 添加网页索引
3. 开始搜索
4. 退出
3
请输入搜索关键词：
java

搜索范围：
1. 所有内容
2. 仅本地文件
3. 仅网页内容
3

找到 1 个结果：

结果 #1:
文件路径：webpage: https://www.runoob.com/java/java-tutorial.html
文件类型：webpage
相关度：0.7419481
内容预览：网页标题：Java 教程 | 菜鸟教程
网页地址： : https://www.runoob.com/java/java-tutorial.html
-----

```

图 19 仅网页搜索

错误信息处理如图 20~22 所示：

```

请选择操作：
1. 添加本地目录
2. 添加网页索引
3. 开始搜索
4. 退出
1
请输入要索引的目录路径（例如：documents）：
aaaa
错误：无效的目录路径！请确保目录存在且路径正确。

```

```

请选择操作：
1. 添加本地目录
2. 添加网页索引
3. 开始搜索
4. 退出
5
无效的选项，请输入1-4之间的数字

```

图 20、21 不合法输入的报错

```
请选择操作：
1. 添加本地目录
2. 添加网页索引
3. 开始搜索
4. 退出
3
请输入搜索关键词：
jaav

搜索范围：
1. 所有内容
2. 仅本地文件
3. 仅网页内容
1
未找到匹配结果
```

图 22 未找到结果的报错

退出界面如图 23 所示：

```
请选择操作：
1. 添加本地目录
2. 添加网页索引
3. 开始搜索
4. 退出
4
感谢使用搜索引擎，再见！
PS D:\User\Desktop\JAVA\2024hw\ZJU_JAVA_2024\hw3\SearchEngine>
```

图 23 退出界面

5. 总结

本次开发的搜索引擎项目，旨在结合 Jsoup, apache tika, lucene 等来搭建搜索引擎。在开发过程中，我面临了众多挑战，这些问题既有基础性的，也有较为复杂的系统性问题。尽管困难重重，但通过不懈的调试和优化，搜索引擎终于达到了预期的功能。

在项目开发中我遇到最大的困难是爬虫内容不显示，我增加了诸多打印语句，插在程序执行的各个位置，一方面能帮助我查看程序运行情况排查错误，另一方面在主菜单上显示信息对于用户也能起到提示功能，改善使用体验。经过调试，我发现是索引建立时目录定位错误的问题，后续我便顺利修改完了这个问题，实现了搜索引擎功能的正常使用。

通过这个项目，我深刻体会到了细节在编程中的重要性。每一个小问题都可能成为提升编程技能的契机，它们不仅锻炼了我的编程能力，也培养了我严谨的编程态度。同时，这次经历为我未来的编程工作积累了宝贵的经验。

在完成搜索引擎的开发后，我意识到自己还有许多需要提高的地方。单独完成这样一个项目对我来说是一项巨大的挑战，因为它涉及到众多复杂的算法和数据结构。这次经历让我深刻认识到扎实的基础知识对于编程的重要性。我也明白了，只要我们愿意投入努力去学习，就能够创造出优秀的作品。原创性是我们最宝贵的资产，它推动我们不依赖于他人的成果，而是通过自己的努力实现创新。我期待将这些经验应用到未来的学习和工作中，继续在编程的道路上不断前进。

参考文献

- [1] 耿祥义. Java 大学实用教程[M]. 北京：清华大学出版社，2009.
- [2] 耿祥义. Java 课程设计[M]. 北京：清华大学出版社，2008.
- [3] 王鹏. Java Swing 图形界面开发与案例详解[M]. 北京：清华大学出版社，2008.
- [4] 丁振凡. Java 语言实验教程[M]. 北京：北京邮电大学出版社，2005.
- [5] 郑莉. Java 语言程序设计[M]. 北京：清华大学出版社，2006.