華中科技大學

课 程 实 验 报 告

课程名称：Java 语言程序设计

实验名称：基于内存的搜索引擎设计和实现

院　　系 ：计算机科学与技术

专业班级 ： 计算机科学与技术 2305 班

学　　号 ： U202315573

姓　　名 ： 叶润莹

指导教师 ： 纪俊文

2025 年 4 月 29 日

一、需求分析

1. 题目要求

实现一个基于内存的英文全文检索搜索引擎，需要完成以下功能：

功能 1：将指定目录下的一批.txt 格式的文本文件扫描并在内存里建立倒排索引，这里面包含必须的子功能包括：

（1）读取文本文件的内容；

（2）将内容切分成一个个的单词；

（3）过滤掉其中一些不需要的单词,例如数字、停用词（the, is and 这样的单词）、过短或过长的单词（例如长度小于 3 或长度大于 20 的单词）；

（4）利用 Java 的集合类在内存里建立过滤后剩下单词的倒排索引；

（5）内存里建立好的索引对象可以序列化到文件，同时可以从文件里反序列化成内存里的索引对象；

（6）可以在控制台输出索引的内容。

功能 2：基于构建好的索引，实现单个搜索关键词的全文检索，包含的子功能包括：

（1）根据搜索关键词得到命中的结果集合；

（2）可以计算每个命中的文档的得分，并根据文档得分对结果集排序；

（3）在控制台显示命中的文档的详细信息，如文档的路径、文档内容、命中的关键词信息（如在文档里出现次数）、文档得分；

功能 3：基于构建好的索引，实现二个搜索关键词的全文检索。包含的子功能包括：

（1）支持这二个关键词的与或查询。与关系必须返回同时包含这二个单词的文档集合，或关系返回包含这二个单词中的任何一个的文档集合；

（2）可以计算每个命中的文档的得分，并根据文档得分对结果集排序；

（3）在控制台显示命中的文档的详细信息，如文档的路径、文档内容、命中的关键词信息（如在文档里出现次数）、文档得分；

功能 4：基于构建好的索引，实现包含二个单词的短语检索，即这二个单词必须在作为短语文档里出现，它们的位置必须是相邻的。**这个功能为进阶功能**。

除了以上功能上的要求外，其他要求包括：

（1）针对搜索引擎的倒排索引结构，已经定义好了创建索引和全文检索所需要的抽象类和接口。**学生必须继承这些预定义的抽象类和和实现预定义接口来完成实验的功能，不能修改抽象类和接口里规定好的数据成员、抽象方法；也不能在预定义抽象类和接口里添加自己新的数据成员和方法**。但是实现自己的子类和接口实现类则不作任何限定。

（2）自己实现的抽象类子类和接口实现类里的关键代码必须加上注释，其中每个类、每个类里的公有方法要加上 Javadoc 注释，并自动生成 Java API 文档作为实验报告附件提交。

（3）使用统一的测试文档集合、统一的搜索测试案例对代码进行功能测试，构建好的索引和基于统一的搜索测试案例的检索结果最后输出到文本文件里作为实验报告附件提交。

（4）本实验只需要基于控制台实现，实验报告里需要提供运行时控制台输出截屏。

关于搜索引擎的倒排索引结构、相关的抽象类、接口定义、还有相关已经实现好的工具类会在单独的 PPT 文档里详细说明。同时也为学生提供了预定义抽象类和接口的 Java API 文档和 UML 模型图。

2. 需求分析

对题目要求进行细化如下：

功能一：将指定目录下的一批.txt 格式的文本文件扫描并在内存里建立倒排索引

（1）读取文本文件，并将它们的相关信息存储在 document 结构里，包括文档 id、文档绝对路径、文档包含的内容（用三元组描述）。

（2）处理内容，即分词：将每个文本文档的内容切分成一个个单词，再针对每一个单词构建三元组：单词本身、出现次数、出现的当前位置，以此来储存每个单词的相关信息。

（3）分词后的处理：包括大小写转换和过滤。将所有大写英文字符转换成小写；过滤掉其中一些不需要的单词：停用词表里的词、过短或过长的单词、正则表达式不要的词（如数字）。

（4）建立索引：索引中包含两个要素，一是词语的索引，将文本文档中的每个三元组与 PostingList 关联起来；二是文件的索引，即将文档 id 和它的路径关联起来。

（5）通过 map 遍历每一个成员，调用序列化/反序列化方法，来实现索引对象与文件的相互转化。

（6）实现 toString 方法，以便在控制台输出索引的内容。

功能二：基于构建好的索引，实现单个搜索关键词的全文检索

（1）根据搜索关键词得到命中的结果集合：输入单词后，程序根据索引查找包含该关键词的所有文档。

（2）根据词频计算每个命中的文档的得分，并根据文档得分对结果集排序。

（3）在控制台显示命中的文档的详细信息，包括文档的路径、文档内容、命中的关键词信息（包括在文档里出现次数和出现的位置）、文档得分。

功能三：基于构建好的索引，实现二个搜索关键词的全文检索

（1）判断这两个关键词的关系：若为与关系，则要返回同时包含这二个单词的文档集合；若是或关系，则返回包含其中一个单词的文档集合。

（2）通过将两个关键词的得分相加，计算命中文档得分，并依此给文档排序。

（3）输出文档具体信息，这部分与功能二中的相同。

功能四：基于构建好的索引，实现包含二个单词的短语检索

（1）在索引中查找包含这两个单词且位置相邻的文档集合。

（2）根据短语出现的频率计算每个命中的文档的得分，并根据文档得分对结果集排序。

（3）在控制台显示命中的文档的详细信息，包括文档的路径、文档内容、命中的关键词信息（包括在文档里出现次数和出现的位置）、文档得分。

二、系统设计

1. 概要设计

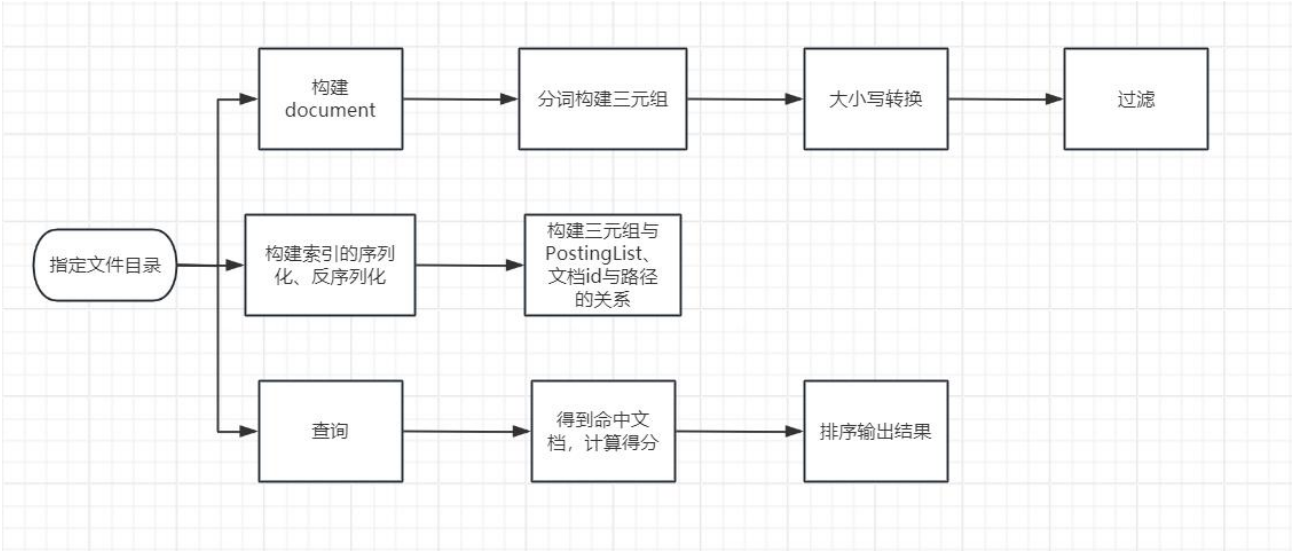
整个倒排索引搜索引擎包含三个模块：建立索引模块、过滤模块和搜索模块。

建立索引模块：主要功能是读取文档信息，将其内容拆分成单词；为每个单词建立 Posting 对象，并将它添加到对应的 PostingList 中；最后建立倒排索引。

过滤模块：判断读取的单词是否符合过滤条件：是数字、是停用词、过短或过长；摘除符合条件的词语。

搜索模块：根据用户的搜索词，查找 PostingList，将命中的文档生成 Hit 对象；对每个 Hit 对象进行评分并排序；最后输出搜索结果。

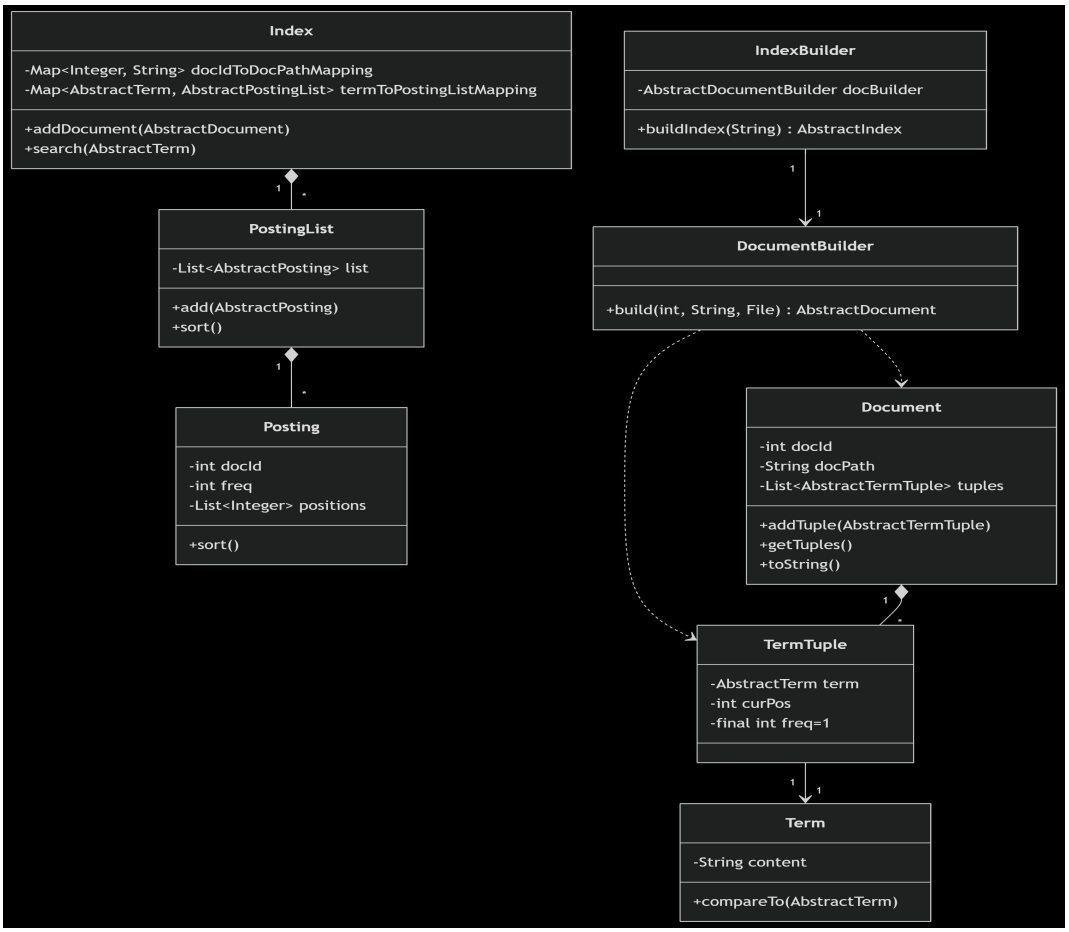
索引构建流程：



2. 详细设计

（1）package hust.cs.javacourse.search.index

UML 图：



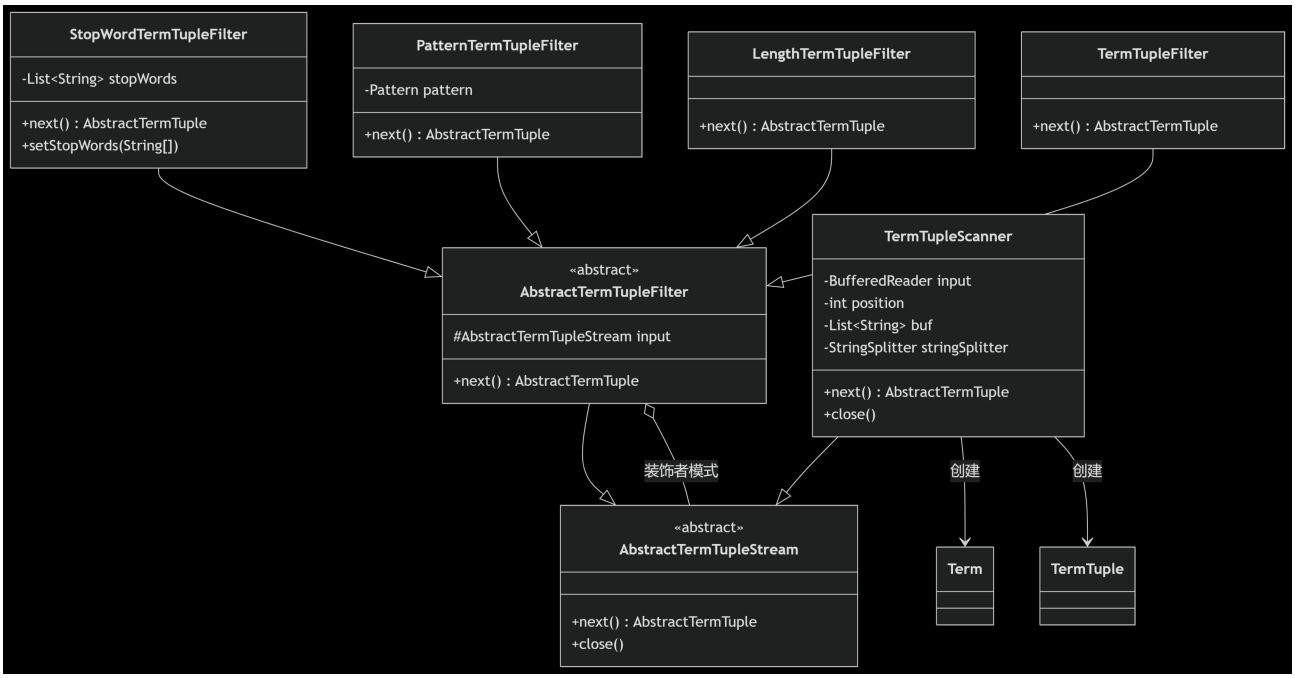
类分析：

类名	作用	重要方法
Term	表示文档中的一个单词（term），实现内容比较和序列化	equals, toString, compareTo 方法的重写，以及输入输出流函数 writeObject 和 readObject
TermTuple	表示 term 的三元组：(term, freq=1, curPos)，记录单词及其位置	equals 和 toString 方法的重写
Document	代表一个文档，包含 docId、路径和所有 TermTuple 的集合	重写了针对 List 的一些函数，如 addTuple 插入三元组、getTuple 查找三元组、contains 是否包含三元组、getTupleSize 返回 List 大小
DocumentBuilder	解析文档文件： 1. 通过 TermTupleScanner 生成词	包含了 build 方法：根据文档 id、文档路径等信息构建 Document 对象

	流 2. 应用过滤器（停用词、长度等） 3. 构建 Document 对象	
Posting	倒排索引项，记录 term 在某个文档中的(docId, freq, positions)	equals, toString, compareTo, sort 方法的重写，以及输入输出流函数 writeObject 和 readObject
PostingList	管理一个 term 对应的所有 Posting（按 docId 排序）	包含了 List 相关的函数，如 add, get, remove, clear, indexOf, contains, size, isEmpty, sort 等；还有输入输出流函数 writeObject 和 readObject
Index	完整的倒排索引，维护两个映射： - term→PostingList - docId→文档路径	包含文件相关的方法，如 load, save；有根据单词搜索 PostingList 的方法 search；有返回文档中所有词语集合的函数 getDictionary；还有排序函数 optimize：排序 Map 中的 PostingList 和 PostingList 中的 Posting
IndexBuilder	遍历目录下的所有文档，调用 DocumentBuilder 构建 Document 并添加到 Index 中	buildIndex 方法：通过根目录信息读取文件序列，再将每个文件的索引加入到 Index 中

（2）package hust.cs.javacourse.search.parse

UML 图：



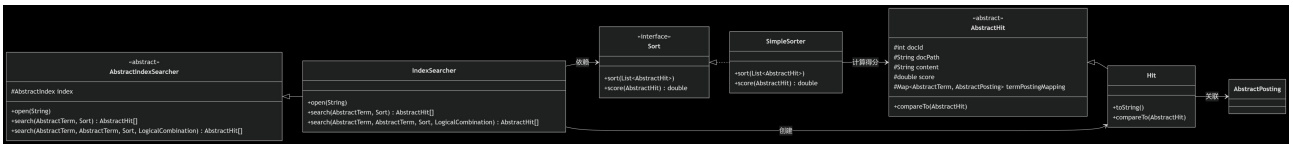
类分析：

类名	作用	过滤规则
AbstractTermTupleStream	定义三元组流的抽象接口	
AbstractTermTupleFilter	过滤器基类，采用装饰者模式包装其他流对象	
TermTupleScanner	从文本文件读取内容，按正则分词生成 TermTuple 流	无过滤
StopWordTermTupleFilter	过滤停用词	term.content ∈ StopWords.STOP_WORDS
PatternTermTupleFilter	通过正则表达式过滤无效词	!term.content.matches(Config.TERM_FILTER_PATTERN)

	（如数字、符号）	
LengthTermTupleFilter	过滤过长或过短的单词	length ∉ [minLen, maxLen]
TermTupleFilter	综合过滤器，整合长度、正则、停用词三种过滤逻辑	依次应用所有规则

（3）package hust.cs.javacourse.search.query

UML 图：

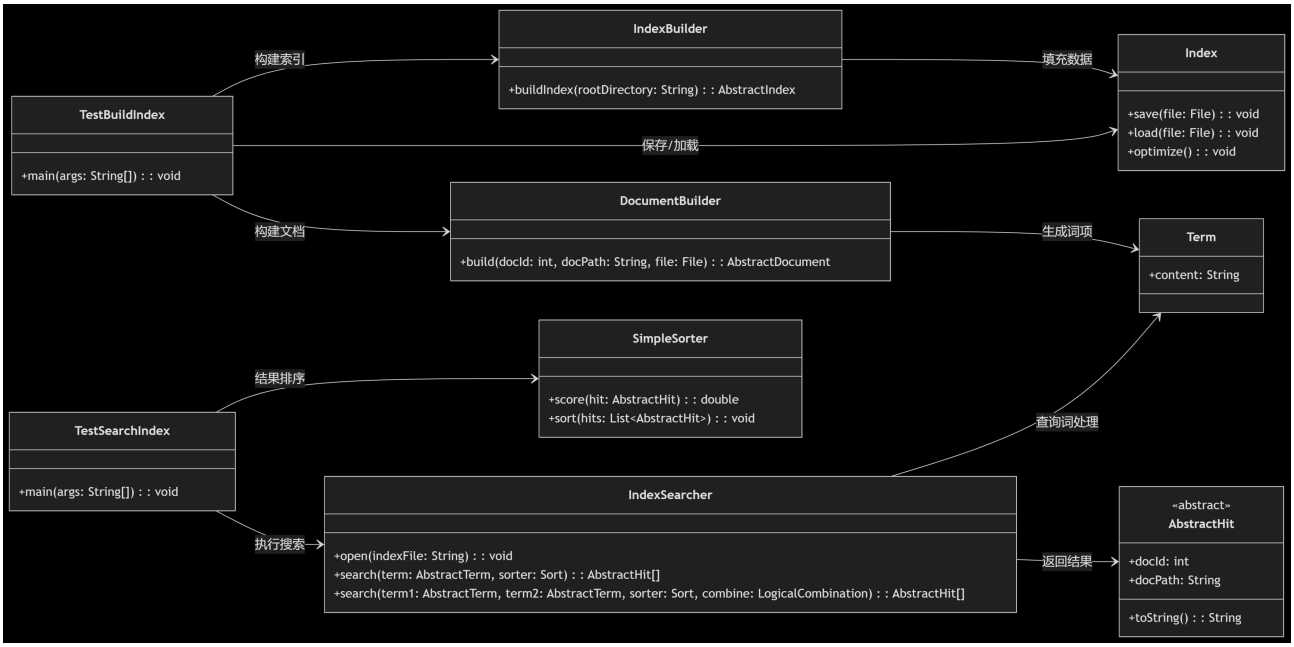


类分析：

类名	作用	关键点
Hit	表示单个命中结果	- 存储文档 ID、路径、内容、得分 - 实现 compareTo 按得分排序 - toString()高亮显示命中词
SimpleSorter	基于词频的排序策略	- score(): 累加所有命中词的频率（取负值实现降序） - sort(): 调用 Collections.sort
IndexSearcher	执行搜索并返回结果	- 单词搜索：直接获取 PostingList 并包装为 Hit - 双词搜索：支持 AND/OR 逻辑合并结果 - 依赖 Sort 实现结果排序

（4）package hust.cs.javacourse.search.run

核心类与 TestBuildIndex 和 TestSearchIndex 两个测试类的关系图：



功能总结：

类名	用途	输入	输出
TestBuildIndex	测试索引构建流程	文档目录路径	序列化的索引文件
TestSearchIndex	测试搜索功能	用户查询词	命中文档的格式化结果

（5）package hust.cs.javacourse.search.util

各工具类的作用：

Config（配置类）：

属性	作用
PROJECT_HOME_DIR	项目根目录
INDEX_DIR	索引文件存储路径
DOC_DIR	文档文件存储路径
IGNORE_CASE	是否忽略大小写
STRING_SPLITTER_REGEX	分词正则表达式
TERM_FILTER_PATTERN	单词过滤正则（保留字母）
TERM_FILTER_MINLENGTH	单词最小长度
TERM_FILTER_MAXLENGTH	单词最大长度

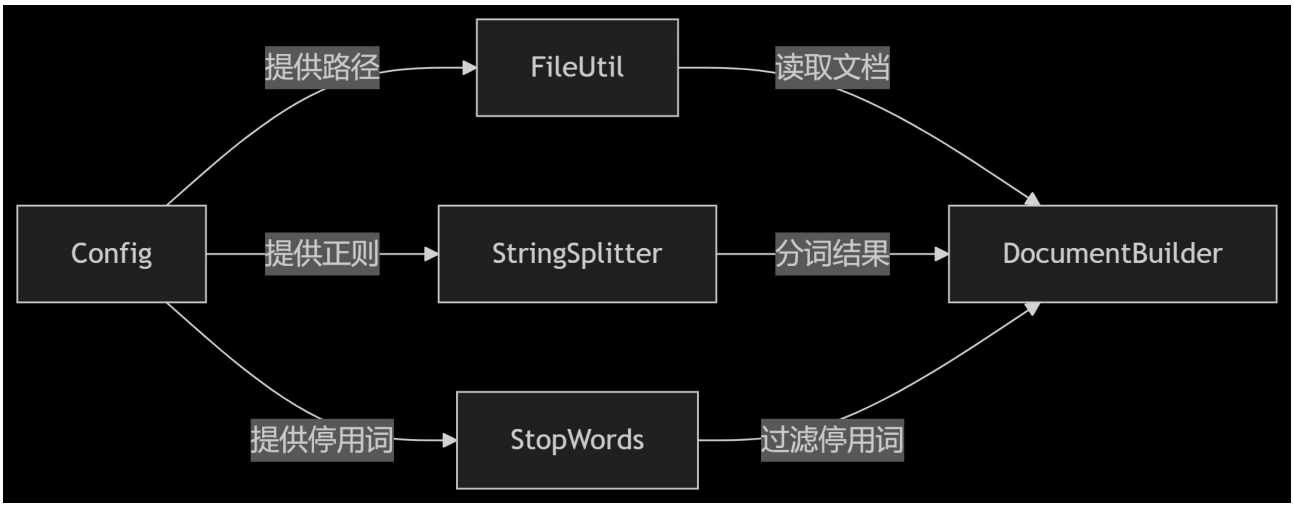
FileUtil（文件工具类）

方法	作用
read(String filePath)	读取文件内容
write(String content, String filePath)	写入内容到文件
list(String dirPath)	列出目录下所有文件路径
list(String dirPath, String suffix)	列出目录下指定后缀的文件路径

StringSplitter（字符串分割类）

方法	作用
setSplitRegex(String regex)	设置分词正则表达式
splitByRegex(String input)	按正则分割字符串

关键协作流程：



三、软件开发

操作系统：Windows 11

开发语言：java version "17.0.10" 2024-01-16 LTS

开发平台：Visual Studio Code

四、软件测试

利用实验自动测试程序进行测试，测试用例 106 个，通过 106 个，跳过 0 个。

```
================================================
All Test Suite
Total tests run: 106, Failures: 0, Skips: 0
================================================
```

Test	# Passed	# Skipped	# Failed	Time (ms)	Included Groups	Excluded Groups
All Test Suite						
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/index/TermTest.java	9	0	0	40		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/index/TermTupleTest.java	4	0	0	10		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/index/PostingTest.java	16	0	0	17		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/index/PostingListTest.java	17	0	0	13		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/index/DocumentTest.java	12	0	0	9		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/parse/TermTupleScannerTest.java	2	0	0	7		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/parse/StopWordTermTupleFilterTest.java	3	0	0	9		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/parse/PatternTermTupleFilterTest.java	2	0	0	5		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/parse/LengthTermTupleFilterTest.java	2	0	0	3		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/parse/ScannerFilterAllInOneTest.java	2	0	0	4		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/index/DocumentBuilderTest.java	7	0	0	446		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/index/IndexTest.java	10	0	0	186		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/index/IndexBuilderTest.java	1	0	0	12		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/query/HitTest.java	16	0	0	27		
D:/IdeaWorkspace/SeachEngine/test/hust/cs/javacourse/search/query/IndexSearcherTest.java	3	0	0	35		
Total	**106**	**0**	**0**	**823**		

使用给出的真实数据测试集，进行查询操作时，程序运行结果如下图：

```
country

docPath:C:\Users\y-one\Desktop\FinalJava\SearchEngineForStudent\text\1.txt
content:The novel coronavirus death toll has reached 21 as of Saturday in Britain as the num
ber of confirmed cases totalled 1,140, according to the latest figures released by the Briti
sh Department of Health and Social Care.

The new figures showed an increase of 342 confirmed COVID-19 cases in Britain, the largest r
ise on a single day since the start of the outbreak in the country. Ten more patients who co
ntracted coronavirus died, bringing the death toll in Britain to 21.

All the 10 patients who died were aged over 60 and had underlying health conditions, said Ch
ris Whitty, chief medical officer for England.

According to health authorities, most of the cases are in England. There have been 121 confi
rmed cases in Scotland, 60 in Wales and 34 in Northern Ireland.

The British government said on Friday that it estimated the true number of infected cases in
 Britain to be around 5,000 to 10,000. People who are self-isolating with mild symptoms are
```

```
The British government said on Friday that it estimated the true number of infected cases in
 Britain to be around 5,000 to 10,000. People who are self-isolating with mild symptoms are
no longer being tested for the virus.
post:[docId:0, freq:1, positions:[65]]

docPath:C:\Users\y-one\Desktop\FinalJava\SearchEngineForStudent\text\6.txt
content:With a population of 602,000, Luxembourg is one of Europe's smallest countries -- ye
t it suffers from major traffic jams.

But that could be about to change. As of March 1, 2020 all public transport -- trains, trams
 and buses -- in the country is now free.

The government hopes the move will alleviate heavy congestion and bring environmental benefi
tsThe government hopes the move will alleviate heavy congestion and bring environmental bene
fiThe government hopes the move will alleviate heavy congestion and bring environmental bene
TheThe government hopes the move will alleviate heavy congestion and bring environmental ben
The government hopes the move will alleviate heavy congestion and bring environmental benefi
ts, according to Dany Frank, a spokesperson for the Ministry of Mobility and Public Works.
post:[docId:11, freq:1, positions:[44]]

docPath:C:\Users\y-one\Desktop\FinalJava\SearchEngineForStudent\text\8.txt
```

```
docPath:C:\Users\y-one\Desktop\FinalJava\SearchEngineForStudent\text\8.txt
content:Mandarin could be taught in Welsh secondary schools in a bid to raise Wales' interna
tional profile.
The Welsh Government wants to establish the country as a visitor destination for Chinese peo
ple as part of its new strategy to boost trade.
It is also looking to have a "Wales in Germany" themed year in 2021, and have an increased p
resence in EU countries.
post:[docId:13, freq:1, positions:[23]]
```

五、特点与不足

1. 技术特点

清晰的模块化设计：从文档解析→索引构建→搜索处理，分层明确，各模块职责单一；运用装饰着模式，通过过滤器链（如 StopWordFilter→LengthFilter）灵活扩展处理逻辑。

高效的数据结构：如 Posting 类存储文档 ID、词频、位置列表，支持快速合并和排序；PostingList 类自动按 docId 排序，支持二分查找；Index 类使用 Map<AbstractTerm, PostingList> 实现快速词典检索。

可扩展性好：运用策略模式，Sort 接口允许自定义排序算法（如 SimpleSorter 按词频排序）。

序列化支持：所有核心类实现 writeObject/readObject，支持索引的持久化存储（如 Index.save()）。

2. 不足和改进的建议

异常处理还不完善，如文件操作未处理 IOException 的恢复。可以添加重试机制或友好错误提示。

六、过程和体会

1. 遇到的主要问题和解决方法

开始进行实验时，我遇到的第一个难题是如何解构需求，将搜索引擎这个整体拆分成一个个类，来分别实现不同的功能。通过仔细阅读实验要求、认真查看 javadoc 文件，我逐渐理清了类与类之间的关系，建构起自己的思路。

在实现索引的建立时，我遇到的问题是如何高效地将文档 id 和文档路径、每个单词和它对应的 PostingList 一一对应起来。在深入了解了 HashMap 提供的键值对关系后，我成功用 Map 实现的对应关系的连接和存储，建立起索引。

2. 课程设计的体会

通过本次课程设计，我深刻体会到了理论知识与实践结合的重要性。在实现倒排索引搜索引擎的过程中，我不仅巩固了 Java 编程技能，还对搜索引擎的核心原理有了更深入的理解。从最初的文档解析、分词处理，到构建倒排索引，再到最终的搜索功能实现，每一个环节都让我认识到系统化思维和模块化设计的重要性。尤其是装饰者模式在过滤器链中的应用，让我直观感受到设计模式如何提升代码的灵活性和可扩展性。

在开发过程中，我也遇到了不少挑战。例如，最初设计索引结构时，没有充分考虑到性能优化问题，导致处理大规模文档时效率较低。通过反复调试和优化，我学会了如何使用更高效的数据结构（如 Map 和排序列表）来提升查询速度。此外，异常处理的不足也让我意识到健壮性对于软件系统的重要性，这促使我在后续开发中更加注重错误处理和日志记录。

最后，这次课程设计让我认识到自身技术能力的不足之处，比如对算法复杂度的分析还不够熟练，对设计模式的理解有待加深。未来，我将继续学习分布式系统和性能优化相关的知识，进一步提升工程实践能力。总体而言，本次设计不仅让我掌握了搜索引擎的核心技术，更培养了我解决复杂问题的能力和严谨的工程思维。

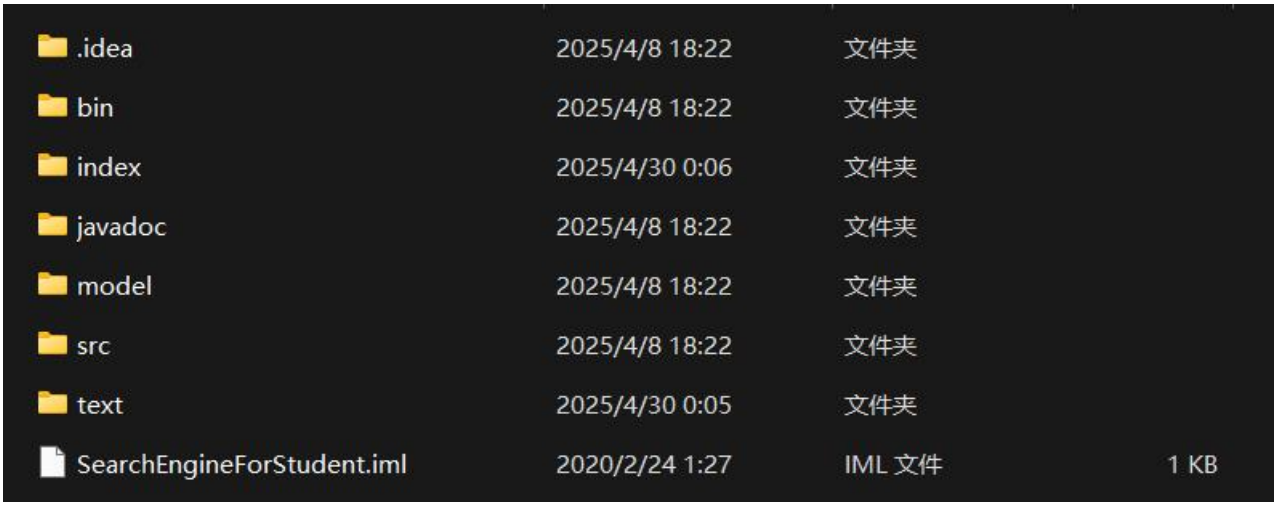
七、源码和说明

1. 文件清单及其功能说明

提交的文件目录如下所示：

其中 Experiment1Test(JDK17)是自动测试文件，所有.class 文件已经放入 betest 文件夹中。

SearchEngineForStudent 是实验的源码工程，它的文件目录如下：



bin 文件夹中包含所有代码的.class 文件；javadoc 文件夹里包含 javadoc 文件，对代码中的类和方法做简要说明；src 文件夹中包含所有代码的.java 文件；text 文件夹中包含测试集文件。

Java 程序设计实验报告文件夹中有实验报告的 word 版本和 pdf 版本。

测试结果截图文件夹中包含了使用自动测试程序和真实测试数据集的测试结果截图。

README 中记录了实验开发环境说明，包括操作系统和 IDE。

2. 用户使用说明书

自动测试：打开 Experiment1Test(JDK17)，在 betest 目录下放入项目的.class 文件（这里已提前放入），在 Windows 环境下，在终端运行 test.bat 脚本文件；在 Linux 环境下，运行 test.sh，运行后在 test-output 中的 index.html 可以得到自动测试的结果。

手动测试：可以将要检索的文档加到项目工程 SearchEngineForStudent 的 text 目录下，先

后对 hust.cs.javacourse.search.run 包下的 TestBuildIndex.java 和 TestSearchIndex.java 编译运行。

3. 源代码

见压缩包中的 SearchEngineForStudent\src 文件。