.

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 ：　CS1708

学　　号 ：　U201714823

姓　　名 ：　张鸿飞

指导教师 ：　纪俊文

2020 年 **4** 月 **23** 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

自行对题目要求进行细化、补充， 例如发生异常的条件。

功能一：

二、系统设计

1. 概要设计

介绍设计思路、原理。将一个复杂系统按功能进行模块划分、建立模块的层次结构及调用关系、确定模块间的接口及人机界面等。

要有总体结构、总体流程（图）。

2. 详细设计

设计每个模块的实现算法（处理流程）、所需的局部数据结构。具体介绍每个模块/子程序的功能、入口参数、出口参数、流程（图）等。

三、软件开发

简单介绍采用什么开发环境，如何编译、连接生成可执行文件。使用了什么调试工具。篇幅不要长。

四、软件测试

对照题目要求，构造测试例，给出程序界面截图，举证题目要求的功能（以及自行补充的功能）已实现。

分析测试效果。

注意：已实现但未在报告中主动举证的功能可能被当作没有实现。

五、特点与不足

1. 技术特点

创新和得意之处

2. 不足和改进的建议

不足和改进的建议

六、过程和体会

1. 遇到的主要问题和解决方法

课程设计中所遇到的主要问题和解决方法

2. 课程设计的体会

课程设计的体会

七、源码和说明

1. 文件清单及其功能说明

提交程序资料的构成，各文件作用是什么。哪个是执行文件，哪个是源码……

2. 用户使用说明书

简要介绍如何安装、使用你的程序。

3. 源代码

打印源码清单。源码关键位置要有注释。