**深 圳 大 学 实 验 报 告**

**课程名称： 信息检索**

**实验项目名称：布尔检索和倒排索引的实验**

**学院： 计算机与软件学院**

**专业： 计算机科学与技术**

**指导教师：潘微科**

**报告人： 沈晨玙 学号：2019092121 班级：19计科04**

**实验时间：2022年3月4日（周五）-2022年3月30日（周三）**

**实验报告提交时间： 2022年3月20日**

**教务部制**

|  |
| --- |
| **实验目的与要求：**  **实验目的：**掌握倒排索引（reverted index）的建立过程；掌握倒排记录表（postings lists）的合并算法；了解Lucene开源信息检索库（library）的简单调用。  **实验要求：**  (1). 阅读教材《Introduction to Information Retrieval》第一章（特别是第1页），谈谈你对“Information Retrieval”定义或含义的理解，特别是信息检索的主要特点。请对信息检索的主要特点逐个加以详细阐述（至少包含4个特点，每个特点写一段话，突出重点）。（20分）  (2). 阅读教材《Introduction to Information Retrieval》第8页Figure 1.4中所描述的倒排索引（reverted index）建立的详细过程，使用附件“HW1.txt”中的60个文档（每行表示一个document），用Java语言或其他常用语言实现倒排索引建立的详细过程。请在报告中附上代码截图（不要复制源代码，请用截图的方式）、运行结果截图和详细的文字说明。程序要有详细的注释。（40分）  (3). 阅读教材《Introduction to Information Retrieval》第11页Figure 1.6中所描述的倒排记录表（postings lists）的合并算法，使用第(2)题中的倒排索引，用Java语言或其他常用语言实现以下布尔检索：  (a) transfer AND learning  (b) transfer AND learning AND filtering  (c) recommendation AND filtering  (d) recommendation OR filtering  (d) transfer AND NOT (recommendation OR filtering)  请在报告中附上代码截图（不要复制源代码，请用截图的方式）、运行结果截图和详细的文字说明。程序要有详细的注释。（40分）  报告写作。要求：主要思路有明确的说明，重点代码有详细的注释，行文逻辑清晰、可读性强，报告整体写作较为专业。（20分）  **说明：**  （1）本次实验课作业满分为100分。  （2）本次实验课作业截至时间2021年3月30日（周三）22:00。  （3）报告正文：请在指定位置填写，本次实验需要单独提交源程序文件（源程序单独打包在Blackboard中上传，不要包含外部导入的包）。  （4）个人信息：WORD文件名中的“姓名”、“学号”，请改为你的姓名和学号；实验报告的首页，请准确填写“学院”、“专业”、“报告人”、“学号”、“班级”、“实验报告提交时间”等信息。  （5）提交方式：截至时间前，请在Blackboard平台中提交。  （6）发现抄袭（包括复制&粘贴整句话、整张图），**抄袭者和被抄袭者的成绩记零分。**  （7）延迟提交，不得分；如有特殊情况，请于截至日期之后的48小时内发邮件到panweike@szu.edu.cn，并在邮件中注明课程名称、作业名称、姓名、学号等信息，以及特殊情况的说明，我收到后会及时回复。  （8）期末考试阶段补交无效。 |

|  |
| --- |
| (1). 阅读教材《Introduction to Information Retrieval》第一章（特别是第1页），谈谈你对“Information Retrieval”定义或含义的理解，特别是信息检索的主要特点。请对信息检索的主要特点逐个加以详细阐述（至少包含4个特点，每个特点写一段话，突出重点）。（20分）  信息检索是指从大规模非结构化数据的集合中找出满足用户需求的资料的过程，通常是指对于文档的查询。我认为信息检索有以下几个特点：   1. 性能要求高   对于一个完整的信息检索系统，需要大量的数据量进行支撑。例如在互联网之中，对于同一个此条可能存在上亿甚至几十亿的数据，而且需要在较短时间内基于用户反馈。因此需要设计一个高效的查询算法，使得程序可以在较短的时间内给予用户反馈。   1. 扩展性强   由于信息检索系统中的文档种类众多，其中包括html，txt，doc等各种形式的文档，因此对于搜索引擎提出了要求，需要满足各种格式的文档检索，需要扩展到各种文件类型。   1. 数据多样性   对于需要查询的数据，具有极高的多样性。例如结构化数据，非结构化数据，半结构化数据。一个完整的信息检索系统需要满足各种数据的检索需求。   1. 数据规模不一   根据书中提到的信息，可以简单的讲数据规模分成三个级别：Web搜索，小规模（个人信息检索），中等规模（企业、机构内等）。因此对于不同数据规模的信息检索系统可能有着不同的信息检索技术，以免对不同的场景。  (2). 阅读教材《Introduction to Information Retrieval》第8页Figure 1.4中所描述的倒排索引（reverted index）建立的详细过程，使用附件“HW1.txt”中的60个文档（每行表示一个document），用Java语言或其他常用语言实现倒排索引建立的详细过程。请在报告中附上代码截图（不要复制源代码，请用截图的方式）、运行结果截图和详细的文字说明。程序要有详细的注释。（30分）  使用语言：Python3   1. 首先将文档读取到一个一维数据之中   代码：  结果：打印数组进行查看   1. 将每篇文档转换成一个个token的列表   步骤：   1. 将文本全部转换成小写 2. 根据“非字符”对文本使用正则表达式进行切割（注：当出现两个连续非字符，会切割出现空串，需要手工删除）   代码：  结果：打印token列表进行查看   1. 构建倒排索引   步骤：   1. 建立如下数据结构：   建立一个哈希表，key值为字符串，value值为列表。  其中key值中存储所有单词，并作为哈希表的索引；value值中第1位记录倒排索引长度，第2位开始记录每个单词出现文章的序号。  term  doc.freq  postings list   1. 遍历token列表：    1. 如果单词出现过，就将文章序号添加到列表尾部，并且长度加一。    2. 单词第一次出现时，将单词加入哈希表。   代码：  结果：  (3). 阅读教材《Introduction to Information Retrieval》第11页Figure 1.6中所描述的倒排记录表（postings lists）的合并算法，使用第(2)题中的倒排索引，用Java语言或其他常用语言实现以下布尔检索：  (a) transfer AND learning  (b) transfer AND learning AND filtering  (c) recommendation AND filtering  (d) recommendation OR filtering  (d) transfer AND NOT (recommendation OR filtering)  请在报告中附上代码截图（不要复制源代码，请用截图的方式）、运行结果截图和详细的文字说明。程序要有详细的注释。（30分）  在完成题目之前，首先完成AND、OR、AND NOT三个函数的编写   1. AND   思路：  参数为两个单词对应的倒排索引列表，返回值为完成AND操作后的结果列表。  需要完成的操作是将同时出现在list1，list2的index筛选出来。因为原先两个列表都是从小到大排序，因此，只需要不断地将指向较小数的指针不断向后移，遇到相同的index时，将index加入结果列表，直到一个指针走到底。   1. OR   思路：  参数为两个单词对应的倒排索引列表，返回值为完成OR操作后的结果列表。  需要完成的操作是将在list1，list2中出现的所有index合并筛选出来。思路与AND的解法大致类似，原先两个列表都是从小到大排序，因此，同样只需要不断地将指向较小数的指针不断向后移，区别是在index大小不相同时仍然需要将index加入结果列表，直到一个指针走到底。  因为OR操作是将两个列表合并，还需要将两个列表中剩余未遍历到的index加入结果列表之中。   1. AND NOT   思路：  参数为两个单词对应的倒排索引列表，返回值为完成AND NOT操作后的结果列表。  需要完成的操作是将出现在list1，但是未出现在list2的index筛选出来。原先两个列表都是从小到大排序，因此，同样需要不断地将指向较小数的指针不断向后移，并且当指向list1的index较小时，将index加入结果列表，直到一个指针走到底。  假设list1未遍历完，list2已经结束，那么list1剩余的index一定不会出现在list2中，所以还需要将剩余未遍历到的index加入结果列表之中。   1. 辅助函数：从哈希表中获取倒排索引列表，并删除第一个元素（用于记录元素个数）   (a) transfer AND learning  结果：transfer AND learning: [5, 7, 9, 10, 16, 17, 25, 32, 33, 49, 55, 56]  结果正确  (b) transfer AND learning AND filtering  先计算transfer AND learning，在计算AND filtering  结果：transfer AND learning AND filtering: [7, 25, 33, 55, 56]  结果正确  (c) recommendation AND filtering  结果：recommendation AND filtering: [13, 26, 38]  结果正确  (d) recommendation OR filtering  结果：recommendation OR filtering: [1, 2, 4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 15, 17, 18, 19, 20, 21, 22, 24, 25, 26, 27, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60]  结果正确  (e) transfer AND NOT (recommendation OR filtering)  先计算recommendation OR filtering，在计算AND NOT  结果：transfer AND NOT (recommendation OR filtering) [10, 16]  结果正确  ++++++++++++++++++++++++++++++++++++++++++++++++++++++  **其他（例如感想、建议等等）。**  此次实验帮我进一步了解了倒排索引以及它对应的合并算法，是信息检索的入门算法学习。因为我对于java 的掌握程度较低，因此此次实验也选择了自己更为熟悉的python语言进行操作，其中python自带的list和dict的数据结构减轻了许多工作量，帮我可以将更多的精力放在算法之上。 |

深圳大学学生实验报告用纸

|  |
| --- |
| 指导教师批阅意见：  成绩评定：  指导教师签字：  2022年 月 日 |
| 备注： |

注：1、报告内的项目或内容设置，可根据实际情况加以调整和补充。

2、教师批改学生实验报告时间应在学生提交实验报告时间后10日内。