**深 圳 大 学 实 验 报 告**

**课程名称： 信息检索**

**实验项目名称：词典、倒排记录表和容错式检索的实验**

**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4月1日（周五）-2022年4月13日（周三）**

**实验报告提交时间： 2022年4月7日星期四**

**教务部制**

|  |
| --- |
| **实验目的与要求：**  **实验目的：**掌握跳表指针（skip pointers）技术、邻近搜索（proximity search）技术、编辑距离（edit distance）计算方法、基于发音的矫正技术（phonetic correction）等，了解PorterStemmer开源程序的简单调用。  **实验要求：**  (1). 考虑利用如下带有跳表指针的倒排记录表  和两个中间结果表（如下所示，不存在跳表指针）分别进行合并操作。  3 5 9 96 99 100 101  25 60 120 150  采用教材《Introduction to Information Retrieval》第37页Figure 2.10中所描述的基于跳表指针（skip pointers）的倒排记录表（postings lists）合并算法，请问：  a.跳表指针实际跳转的次数分别是多少（也就是说，指针p1的下一步将跳到skip（p1））？  b.当两个表进行合并时，倒排记录之间的比较次数分别是多少？  c.如果不使用跳表指针，那么倒排记录之间的比较次数分别是多少？  请在报告中附上详细的文字说明。（15分）  (2). 下面给出的是一个位置索引的一部分，格式为：  <position1, position2, …>; doc2: <position1, position2, …>; …  angels: 2: <36,174,252,651>; 4: <12,22,102,432>; 7: <17>;  fools: 2: <1,17,74,222>; 4: <8,78,108,458>; 7: <3,13,23,193>;  fear: 2: <87,704,722,901>; 4: <13,43,113,433>; 7: <18,328,528>;  in: 2: <3,37,76,444,851>; 4: <10,20,110,470,500>; 7: <5,15,25,195>;  rush: 2: <2,66,194,321,702>; 4: <9,69,149,429,569>; 7: <4,14,404>;  to: 2: <47,86,234,999>; 4: <14,24,774,944>; 7: <199,319,599,709>;  tread: 2: <57,94,333>; 4: <15,35,155>; 7: <20,320>;  where: 2: <67,124,393,1001>; 4: <11,41,101,421,431>; 7: <16,36,736>;  请问哪些文档和以下的查询匹配？其中引号内的每个表达式都是一个短语查询。  a.“angels fear”  b.“angels fear to tread”  c.“angels fear to tread”AND“fools rush in”  请在报告中附上详细的文字说明。（10分）  (3). 阅读教材《Introduction to Information Retrieval》第37页Figure 2.10中所描述的基于跳表指针（skip pointers）的倒排记录表（postings lists）合并算法，并用Java语言或其他常用语言实现该算法。要求在题(1)的例子上验证算法的正确性。  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。（10分）  (4). 阅读教材《Introduction to Information Retrieval》第42页Figure 2.12中所描述的邻近搜索（proximity search）中的两个倒排记录表（postings lists）的合并算法，并用Java语言或其他常用语言实现该算法，并按要求做适当改进。要求使用附件“HW2.txt”中的60个文档（每行表示一个document，按空格切词，文档中的单词全部转换为小写）建立positional index，两个词项之间的间距（注：相邻的两个词项的间距为1）的形式包括以下三种情形（x是一个正整数）：  “-x”、“+x”和“x”，  其中，“-x”表示第一个词项在第二个词项的左侧且间隔在x之内，“+x”表示第一个词项在第二个词项的右侧且间隔在x之内，“x”表示第一个词项与第二个词项的间隔（左侧和右侧均可）在x之内。要求在以下例子上验证算法的正确性：(ranking, filtering, -4), (ranking, filtering, -5), (ranking, filtering, -6), (ranking, filtering, -7), (heterogeneous, learning, +2), (recommendation, bias, 2)。  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）  (5). 阅读教材《Introduction to Information Retrieval》第59页Figure 3.5中所描述的基于动态规划（dynamic programming）来计算两个字符串的编辑距离（edit distance）的算法，并用Java语言或其他常用语言实现该算法。要求计算以下15组单词的编辑距离：  (business, buisness)  (committee, commitee)  (conscious, concious)  (definitely, definately)  (fluorescent, florescent)  (forty, fourty)  (government, goverment)  (idiosyncrasy, idiosyncracy)  (immediately, immediatly)  (millennium, millenium)  (noticeable, noticable)  (tendency, tendancy)  (truly, truely)  (weird, wierd)  (privilege, privledge)  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。（15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2022年4月13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). 考虑利用如下带有跳表指针的倒排记录表  和两个中间结果表（如下所示，不存在跳表指针）分别进行合并操作。  3 5 9 96 99 100 101  25 60 120 150  采用教材《Introduction to Information Retrieval》第37页Figure 2.10中所描述的基于跳表指针（skip pointers）的倒排记录表（postings lists）合并算法，请问：  a.跳表指针实际跳转的次数分别是多少（也就是说，指针p1的下一步将跳到skip（p1））？  b.当两个表进行合并时，倒排记录之间的比较次数分别是多少？  c.如果不使用跳表指针，那么倒排记录之间的比较次数分别是多少？  请在报告中附上详细的文字说明。（15分）  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **a.跳表指针实际跳转的次数分别是多少（也就是说，指针p1的下一步将跳到skip（p1））？**  跳表指针的跳转算法为：如果当前p1节点拥有跳表指针，当p1.val小于p2.val时不断向后跳转。（具体比较过程可以见第2题，或者看代码实现）  **1、和3 5 9 96 99 100 101进行合并操作：**  答：**2次**，<24, 75>, <75, 92>  比较到24与96时，24拥有skip指针且75 < 96，因此第1次跳转；  75拥有skip指针且92 < 96，因此第2次跳转；  92拥有skip指针但是115 > 96，结束跳转。  **2、和25 60 120 150进行合并操作：**  答：**3次**，<3, 24>, <75, 92>, <92, 115>  比较到3与25时，3拥有skip指针且24 < 25，因此第1次跳转；  24拥有skip指针且75 > 25，结束跳转。  比较到75与120时，75拥有skip指针且92 < 120，因此第2次跳转；  92拥有skip指针且115 < 120，因此第2次跳转;  115无skip指针，结束跳转。  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **b.当两个表进行合并时，倒排记录之间的比较次数分别是多少？**  当p1.val = p2.val时，加入答案列表，并且p1p2同时后移；当p1.val < p2.val时，如果p1无skip指针，则向后移，否则根据第1题的跳表指针算法向后移动p1指针；反之同理。（具体过程可以看代码实现）  **1、和3 5 9 96 99 100 101进行合并操作：**  答：**13次**，比较顺序如下：<p1.val, p2.val>  ['<3, 3>', '<5, 5>', '<9, 9>', '<15, 96>', '<24, 96>', '<75, 96>', '<92, 96>', '<115, 96>', '<96, 96>', '<97, 99>', '<100, 99>', '<100, 100>', '<115, 101>']  **2、和25 60 120 150进行合并操作：**  答：**10次**，比较顺序如下：<p1.val, p2.val>  ['<3, 25>', '<24, 25>', '<75, 25>', '<39, 25>', '<39, 60>', '<60, 60>', '<68, 120>', '<75, 120>', '<92, 120>', '<115, 120>']  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **c.如果不使用跳表指针，那么倒排记录之间的比较次数分别是多少？**  **1、和3 5 9 96 99 100 101进行合并操作：**  答：**18次**，比较顺序如下：<p1.val, p2.val>  ['<3, 3>', '<5, 5>', '<9, 9>', '<15, 96>', '<24, 96>', '<39, 96>', '<60, 96>', '<68, 96>', '<75, 96>', '<81, 96>', '<84, 96>', '<89, 96>', '<92, 96>', '<96, 96>', '<97, 99>', '<100, 99>', '<100, 100>', '<115, 101>']   1. **和25 60 120 150进行合并操作：**   答：**18次**，比较顺序如下：<p1.val, p2.val>  ['<3, 25>', '<5, 25>', '<9, 25>', '<15, 25>', '<24, 25>', '<39, 25>', '<39, 60>', '<60, 60>', '<68, 120>', '<75, 120>', '<81, 120>', '<84, 120>', '<89, 120>', '<92, 120>', '<96, 120>', '<97, 120>', '<100, 120>', '<115, 120>']  (2). 下面给出的是一个位置索引的一部分，格式为：  <position1, position2, …>; doc2: <position1, position2, …>; …  angels: 2: <36,174,252,651>; 4: <12,22,102,432>; 7: <17>;  fools: 2: <1,17,74,222>; 4: <8,78,108,458>; 7: <3,13,23,193>;  fear: 2: <87,704,722,901>; 4: <13,43,113,433>; 7: <18,328,528>;  in: 2: <3,37,76,444,851>; 4: <10,20,110,470,500>; 7: <5,15,25,195>;  rush: 2: <2,66,194,321,702>; 4: <9,69,149,429,569>; 7: <4,14,404>;  to: 2: <47,86,234,999>; 4: <14,24,774,944>; 7: <199,319,599,709>;  tread: 2: <57,94,333>; 4: <15,35,155>; 7: <20,320>;  where: 2: <67,124,393,1001>; 4: <11,41,101,421,431>; 7: <16,36,736>;  请问哪些文档和以下的查询匹配？其中引号内的每个表达式都是一个短语查询。  a.“angels fear”  b.“angels fear to tread”  c.“angels fear to tread”AND“fools rush in”  请在报告中附上详细的文字说明。（10分）  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **a.“angels fear”**  1. 符合查询的文档：**文档4**中，angles - 12 fear - 13或angles - 432 fear - 433  **文档7**中，angles - 17 fear - 18  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **b.“angels fear to tread”**  1. 符合查询的文档：**文档4**中，angles - 12 fear - 13 to - 14 tread - 15  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **c.“angels fear to tread”AND“fools rush in”**  1. “fools rush in”符合的文档：文档2中，fools - 1 rush - 2 in - 3  文档4中，fools - 8 rush - 9 in - 10  文档7中，fools - 3 rush - 4 in - 5或fools - 13 rush - 14 in - 15   1. 与第二组结果取AND，因此最终结果为**文档4**   (3). 阅读教材《Introduction to Information Retrieval》第37页Figure 2.10中所描述的基于跳表指针（skip pointers）的倒排记录表（postings lists）合并算法，并用Java语言或其他常用语言实现该算法。要求在题(1)的例子上验证算法的正确性。  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。（10分）  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **代码截图和详细的文字说明：**  首先定义一个数据结构：  链表节点Node：  值得注意的是，添加了一个skip指针。如果该节点可以跳转，则skip指针指向下一个节点，否则为None  定义一个含跳表指针的链表：  其中包括了三个方法：   1. 根据list添加节点 2. 添加单节点 3. 添加跳表指针   前两个方法为常见的链表算法。  第三个方法实现如下：  首先根据链表长度计算出skip指针的跳转长度为interval = sqrt(length)  然后从第一个节点开始，每隔interval个节点指向下一个跳转节点。  合并算法：（含跳表指针）  （不含跳表指针）  统计次数：    +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **运行结果截图和详细的文字说明（1）：**  带有跳表指针的倒排记录表  和3 5 9 96 99 100 101的合并操作  定义两个链表，并为list1加入skip指针，list2不加skip指针  调用函数，计算合并结果并输出统计结果  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **运行结果截图和详细的文字说明（2）：**  带有跳表指针的倒排记录表  和25 60 120 150的合并操作  定义list3链表，不加入skip指针  调用函数，计算合并结果并输出统计结果  (4). 阅读教材《Introduction to Information Retrieval》第42页Figure 2.12中所描述的邻近搜索（proximity search）中的两个倒排记录表（postings lists）的合并算法，并用Java语言或其他常用语言实现该算法，并按要求做适当改进。要求使用附件“HW2.txt”中的60个文档（每行表示一个document，按空格切词，文档中的单词全部转换为小写）建立positional index，两个词项之间的间距（注：相邻的两个词项的间距为1）的形式包括以下三种情形（x是一个正整数）：  “-x”、“+x”和“x”，  其中，“-x”表示第一个词项在第二个词项的左侧且间隔在x之内，“+x”表示第一个词项在第二个词项的右侧且间隔在x之内，“x”表示第一个词项与第二个词项的间隔（左侧和右侧均可）在x之内。要求在以下例子上验证算法的正确性：(ranking, filtering, -4), (ranking, filtering, -5), (ranking, filtering, -6), (ranking, filtering, -7), (heterogeneous, learning, +2), (recommendation, bias, 2)。  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）  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **代码截图和详细的文字说明：**   1. 读取文本并根据正则表达式对文本进行切割   结果：   1. 根据分解的词条，构建如下的数据结构：   外层hashtable是一个哈希表：key是term词项，value是一个哈希表  value的哈希表中，key是doc\_index文档序号，value是一个list（包含term出现的位置）  结果：   1. 编写含位置信息的倒排记录表合并算法def positionalIntersect(list1, list2, k)   有三个参数：list1对应第一个词项，list2对应第二个词项，k是一个’+k/-k/k’的字符串  输出为三元组列表<文档ID， 词项在p1中的位置， 词项在p2中的位置>  首先比较得出词项一和词项二共同出现的docID，方法类似AND方法。  最后利用二重循环比较两个term\_index列表，根据根据参数k的输入形式，判定取值范围。  举例：输入为+k，则当pos1在区间[pos2, pos2 + k]内，则为符合条件的查询  输入为-k，则当pos1在区间[pos2 - k, pos2]内，则为符合条件的查询  输入为k，则当pos1在区间[pos2 - k, pos2 + k]内，则为符合条件的查询  具体处理方式见以下完整代码：  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **例子1** **(ranking, filtering, -4)上的运行结果截图和详细的文字说明：**  27号文档中，ranking位置为4，filtering位置为8；  57号文档中，ranking位置为7，filtering位置为11  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **例子2** **(ranking, filtering, -5)上的运行结果截图和详细的文字说明：**  12号文档中，ranking位置为5，filtering位置为10；  27号文档中，ranking位置为4，filtering位置为8；  57号文档中，ranking位置为7，filtering位置为11  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **例子3** **(ranking, filtering, -6)上的运行结果截图和详细的文字说明：**  12号文档中，ranking位置为5，filtering位置为10；  27号文档中，ranking位置为4，filtering位置为8；  57号文档中，ranking位置为7，filtering位置为11  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **例子4** **(ranking, filtering, -7)上的运行结果截图和详细的文字说明：**  8号文档中，ranking位置为4，filtering位置为11；  12号文档中，ranking位置为5，filtering位置为10；  27号文档中，ranking位置为4，filtering位置为8；  57号文档中，ranking位置为7，filtering位置为11  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **例子5** **(heterogeneous, learning, +2)上的运行结果截图和详细的文字说明：**  7号文档中，heterogeneous位置为4，learning位置为2；  30号文档中，heterogeneous位置为5，learning位置为3；  33号文档中，heterogeneous位置为5，learning位置为3；  36号文档中，heterogeneous位置为6，learning位置为4；  56号文档中，heterogeneous位置为4，learning位置为2；  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **例子6** **(recommendation, bias, 2)上的运行结果截图和详细的文字说明：**  49号文档中，recommendation位置为5，bias位置为7；  50号文档中，recommendation位置为5，bias位置为3；  (5). 阅读教材《Introduction to Information Retrieval》第59页Figure 3.5中所描述的基于动态规划（dynamic programming）来计算两个字符串的编辑距离（edit distance）的算法，并用Java语言或其他常用语言实现该算法。要求计算以下15组单词的编辑距离：  (business, buisness)  (committee, commitee)  (conscious, concious)  (definitely, definately)  (fluorescent, florescent)  (forty, fourty)  (government, goverment)  (idiosyncrasy, idiosyncracy)  (immediately, immediatly)  (millennium, millenium)  (noticeable, noticable)  (tendency, tendancy)  (truly, truely)  (weird, wierd)  (privilege, privledge)  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。（15分）  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **代码截图和详细的文字说明：**  算法思想：动态规划  dp[i][j] 代表s1到i位置转换成s2到j位置需要最少步数  所以，当s1[i] == s2[j]，dp[i][j] = dp[i-1][j-1]；  当s1[i] != s2[j]，dp[i][j] = min(dp[i-1][j-1], dp[i-1][j], dp[i][j-1]) + 1  其中，dp[i-1][j-1] 表示替换操作，dp[i-1][j] 表示删除操作，dp[i][j-1] 表示插入操作。  注意，针对第一行，第一列要单独考虑，引入空行如下图所示：  第一行，是s1为空变成 word2 最少步数，就是插入操作  第一列，是s2为空，需要的最少步数，就是删除操作  实现代码如下所示：  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **计算得到的15组单词的编辑距离如下：**  (business, buisness)：2  (committee, commitee)：1  (conscious, concious)：1  (definitely, definately)：1  (fluorescent, florescent)：1  (forty, fourty)：1  (government, goverment)：1  (idiosyncrasy, idiosyncracy)：1  (immediately, immediatly)：1  (millennium, millenium)：1  (noticeable, noticable)：1  (tendency, tendancy)：1  (truly, truely)：1  (weird, wierd)：1  (privilege, privledge)：2  ++++++++++++++++++++++++++++++++++++++++++++++++++++++  **其他（例如感想、建议等等）。**  本次实验完成了跳表指针技术、邻近搜索技术、编辑距离计算方法的进一步理解与实现。其中在跳表指针技术的代码编写过程中遇到了一些问题。主要是由于书上的伪代码在判断过程中只用了and进行条件判断，简单的count+1会漏算比较次数，因此在不改变原算法的基础上进行了一些改动，花费了不少精力。不得不说，第二题的伪代码看起来简单，实现起来真的有点痛苦。 |

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。