**深 圳 大 学 实 验 报 告**

**课程名称： 信息检索**

**实验项目名称：索引构建和压缩的实验**

**学院： 计算机科学与技术**

**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15日（周五）-2022年4月27日（周三）**

**实验报告提交时间： 2022年4月16日星期六**

**教务部制**

|  |
| --- |
| **实验目的与要求：**  **实验目的：**掌握文本数据集的统计分析方法，以及索引的构建和压缩技术。  **实验要求：**  (1). 针对附件“HW3.txt”中的600个文档（每行表示一个document，文档ID为1至600）：(i)使用jieba中文分词（https://pypi.org/project/jieba/）或其他中文分词工具进行分词；(ii)统计600个文档中的token的总数和term的总数；(iii)构建倒排索引，并输出以下七组查询的文档ID：“迁移”，“迁移学习”，“推荐”，“深度学习”，“隐私”，“跨领域”，“跨域”。  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）  (2). 阅读教材《Introduction to Information Retrieval》第97页Figure 5.8中所描述的采用VB的编码和解码过程（VB encoding and decoding），并用Java语言或其他常用语言实现该算法。要求在三个数字（113，309，720）上验证实现的算法的正确性。  注意使用合理的数据结构。  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。（20）  (3). 阅读教材《Introduction to Information Retrieval》第98-99页（Section 5.3.2）中所描述的采用Gamma codes的编码和解码过程（encoding and decoding of Gamma codes），并用Java语言或其他常用语言实现该算法。要求在三个数字（113，309，720）上验证算法的正确性。  注意使用合理的数据结构。  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。（20）  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27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。  （8）期末考试阶段补交无效。 |

|  |
| --- |
| (1). 针对附件“HW3.txt”中的600个文档（每行表示一个document，文档ID为1至600）：(i)使用jieba中文分词（https://pypi.org/project/jieba/）或其他中文分词工具进行分词；(ii)统计600个文档中的token的总数和term的总数；(iii)构建倒排索引，并输出以下七组查询的文档ID：“迁移”，“迁移学习”，“推荐”，“深度学习”，“隐私”，“跨领域”，“跨域”。  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）  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **代码截图和详细的文字说明：**   1. 读取文档  1. 计算token，term  1. 建立倒排索引  1. question3展示   +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **(i)使用jieba中文分词（https://pypi.org/project/jieba/）或其他中文分词工具进行分词**  首先读取文档内容  然后调用jieba.cut函数进行分词  jieba分词0.4版本以上支持四种分词模式：   1. 精确模式：试图将句子最精确地切开，适合文本分析； 2. 全模式：把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义 3. 搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词 4. paddle模式：利用PaddlePaddle深度学习框架，训练序列标注（双向GRU）网络模型实现分词。同时支持词性标注。   精确模式：  全模式：  搜索引擎模式：  paddle模式：（m1 macbook暂不支持paddlepaddle安装，故没有测试）  可以看出，全模式的切词是最粗糙的，将所有的词汇都进行了返回。主要有以下几个问题：   1. 未结合语境，容易有歧义：协同过滤----> 协同+同过+过滤 2. 不了解词汇：鲁棒----> 鲁+棒   精确模式与搜索引擎模式可以结合具体需求进行选择。  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **(ii)统计600个文档中的token的总数和term的总数**  经过上方的测试，此处选择精确模式进行分析。  将分解出的词汇分别存入list和set。set会自动进行去重，达到计算term的目的  token数量：6190  term数量：1128  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **(iii)构建倒排索引，并输出以下七组查询的文档ID：“迁移”，“迁移学习”，“推荐”，“深度学习”，“隐私”，“跨领域”，“跨域”**   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。   最终得到了如下的倒排索引表：  同时也对需要进行搜索的单词进行了筛选：  可以看到jieba对于计算机专业词汇的分割，并不是十分了解。因此我又尝试了许多其他分词库，包括thulac，pkuseg，FoolNLTK等，发现结果大同小异。最终，我觉得手动添加一些计算机专业词汇进入jieba。  在网站：https://github.com/ylfeng250/cs-dict中，有许多已经整理好的计算机专业词汇。  使用jieba.load\_userdic()添加自定义词库  最终，倒排索引搜索结果如下：  (2). 阅读教材《Introduction to Information Retrieval》第97页Figure 5.8中所描述的采用VB的编码和解码过程（VB encoding and decoding），并用Java语言或其他常用语言实现该算法。要求在三个数字（113，309，720）上验证实现的算法的正确性。  注意使用合理的数据结构。  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。（20）  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **代码截图（不要复制源代码，请用截图的方式）和详细的文字说明：**  编码过程：将数据循环除以128取余并将十进制结果记录，将延续位变为1并将8位二进制字符串结果记录。  解码过程：将二进制字符串变成十进制，将延续位改为0。   1. VB编码单数字（输入数字，返回VB编码）  1. VB编码数字流（输入数字流，对数字一次编码，并返回字节流）  1. VB解码字节流（输入数字流，返回解码后的数字流）  1. 注：在python中，无法对于数字的数据类型进行指定，因此在代码中不再进行特别指定。如果使用C/C++，可以使用unsigned char来存储（0-255），刚好一个字节的大小，可以尽量节省存储空间。   +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **第1个数字113上的运行结果截图和详细的文字说明：**  原始: num = 113  十进制: VBEncodeNumber10(num) = [241]  二进制: VBEncodeNumber2(num) = ['11110001']  解码: VBDecode(VBEncodeNumber2(num)) = [113]  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **第2个数字309上的运行结果截图和详细的文字说明：**  原始: num = 309  十进制: VBEncodeNumber10(num) = [2, 181]  二进制: VBEncodeNumber2(num) = ['00000010', '10110101']  解码: VBDecode(VBEncodeNumber2(num)) = [309]  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **第3个数字720上的运行结果截图和详细的文字说明：**  原始: num = 720  十进制: VBEncodeNumber10(num) = [5, 208]  二进制: VBEncodeNumber2(num) = ['00000101', '11010000']  解码: VBDecode(VBEncodeNumber2(num)) = [720]  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **在数字流[113,309,720]上的运行结果截图和详细的文字说明：**  原始: nums = [113, 309, 720]  十进制: VBEncode(nums) = [241, 2, 181, 5, 208]  二进制: VBEncode(nums) = ['11110001', '00000010', '10110101', '00000101', '11010000']  解码: VBDecode(VBEncode(nums)) = [113, 309, 720]  (3). 阅读教材《Introduction to Information Retrieval》第98-99页（Section 5.3.2）中所描述的采用Gamma codes的编码和解码过程（encoding and decoding of Gamma codes），并用Java语言或其他常用语言实现该算法。要求在三个数字（113，309，720）上验证算法的正确性。  注意使用合理的数据结构。  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。（20）  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **代码截图（不要复制源代码，请用截图的方式）和详细的文字说明：**  编码过程：将数据转化为二进制，去除首位得到偏移量。对偏移量的长度进行一元编码，gammar编码最终为两者相连。  解码过程：获取第一次出现0的位置，获取偏移部分，首位添一，转换为十进制。   1. gammar编码数字流  1. gammar编码单数字  1. gammar编码解码（根据数字流，返回解码后的字节流）   +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **第1个数字113上的运行结果截图和详细的文字说明：**  num = 113  gammarEncodeNumber(num) = 1111110110001  gammardecode(gammarEncodeNumber(num)) = [113]  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **第2个数字309上的运行结果截图和详细的文字说明：**  num = 309  gammarEncodeNumber(num) = 11111111000110101  gammardecode(gammarEncodeNumber(num)) = [309]  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **第3个数字720上的运行结果截图和详细的文字说明：**  num = 720  gammarEncodeNumber(num) = 1111111110011010000  gammardecode(gammarEncodeNumber(num)) = [720]  +++++++++++++++++++++++++++++++++++++++++++++++++++++++  **在数字流[113,309,720]上的运行结果截图和详细的文字说明：**  nums = [113, 309, 720]  gammarEncode(num) = 1111110110001111111110001101011111111110011010000  gammardecode(GammarEncode(nums)) = [113, 309, 720]  ++++++++++++++++++++++++++++++++++++++++++++++++++++++  **其他（例如感想、建议等等）。**  通过本次实验了解并掌握文本数据集的统计分析方法，以及索引的构建和压缩技术。第一题中，尝试使用了多重的分词库，一开始都没能很好的解决计算机专业词汇的问题，后来不得已只能自行添加词典，完成分词任务。第二三题中，由于python的语言特性，无法选择字节类型的数据结构，无法测试实际的压缩比等数据，如果有机会可以尝试使用C语言进行一下测试。 |

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。