数据结构与程序设计专题实验报告

姓名	学号	班级
袁--	216-	
韩--	216-	
匡--	216-	
王--	216-	

实验指导教师：李峰、顿玉洁

实验地点：西一楼一层计算机中心机房。

实验结束日期：2017-12-10

联系方式:

cantjie@cantjie.com

一、实验任务：

信源编解码是通信系统的重要组成部分。本实验旨在通过程序设计实现基于哈夫曼编码的信源编解码算法。程序具备以下功能：

对于给定的源文档 SourceDoc.txt，

1) 统计其中所有字符的频度（某字符的频度等于其出现的总次数除以总字符数），

包括字母（区分大小写）、标点符号及格式控制符（空格、回车等）。

2) 按频度统计结果生成哈夫曼编码码表。

3) 基于哈夫曼码表进行编码，生成对应的二进制码流，并输出到文件 Encode.dat。

4) 对二进制码流进行哈夫曼解码，把结果输出到文件 DecodeDoc.txt。

5) 判断 DecodeDoc.txt 与 SourceDoc.txt 内容是否一致，以验证编解码系统的正确性。

二、实验内容：

1）需求分析以及模块划分。

2）C 中队列和栈的实现以及使用。

3）链表和二叉树数据结构的应用

4）文件的正常读写和二进制读写。

5）对数据按位操作。

6）排序方法的比较和选择，归并排序。

三、程序的算法描述：

具体请见附录流程图

1) 主控流程图 A——main 函数流程控制部分

2) 主控流程图 B——main 函数流程控制部分
3) 主要功能流程图 A——从文件统计词频、建立哈夫曼树、打印树形
4) 主要功能流程图 B——得到某个字符的编码、打印码本、空间优先压缩、时间优先压缩
5) 主要功能流程图 C——解压缩文件、比较两文本文件是否相同

四、本组优点：

1、开发过程：

a) 开发前先进行了需求分析，并按需求实现将任务和程序分为若干模块，尽可能地实现了高内聚、低耦合。

b) 开发前完成了代码规范文档，因此无论是变量和函数的命名还是注释的格式都十分一致，debug 或阅读他人代码都很方便。

c) 在开发过程中使用坚果云完成版本控制和多人协作。相比 QQ 文件或 u 盘实现协作，效率大幅提升，且每个人都可以通过 header.h 了解别人提供的接口和数据结构。

d) 开发过程中有开发日志（随笔），便于记录每一个功能实现的思路以及完成进度。

2、程序特性：

a）程序鲁棒性较高。

在任何用户输入的地方，都有合法性检验。例如输入文件名的地方，不仅会检验文件名是否合法（filenameVaild 函数），还可以检验文件名是否以特定字符为后缀（filenameEndsWith 函数）；在要求选择功能的地方，如果输入不符合要求，也会提示重新输入。

在任何调用 malloc 的地方，都会检验是否开辟成功。此外，

由于程序中统计词频和建立哈夫曼树是依次进行的，为了保证事件的原子性，如果哈夫曼树建立失败，不仅会把已经开辟的空间全部释放，还会把词频信息链表也释放。由此，可以避免出现因为某一步失败而造成之前申请了的空间还没有被释放掉，而指向这些空间的指针就已经丢失了的情况。

b) 程序适用于多种场景。

压缩算法有两种选择，时间优先和空间优先。在小文件时，空间优先具有较高的速度和空间优势；对于大文本文件，时间优先又可以保障较快的速度。此外，对于非 GB2312 编码的文本文件，理论上虽然没有办法统计词频，但仍能实现压缩与解压缩功能。

c) 输出格式统一、美观。

在输出词频统计结果时，对于\t \n \r 等字符，若直接输出会引起格式的混乱。而本程序在输出结果时，对于 ASCII 码<32 的字符进行了特殊处理(具体可见程序运行结果)，因此对于不同文件，本程序都可以格式统一、十分美观地输出统计结果和码本信息。

d) 功能创新。

在完成全部需求的基础上，还增加了打印树形的功能，这对于了解哈夫曼树的构造也有一定帮助。

e) 具有文件名默认值。

在要求输入文件名的地方，都用尖括号提供了默认值，方便用户了，也方便了程序调试。同时，当用户输入了与默认值不同的文件名，在本次运行中，此默认值将会改成用户输入的文件名，避免用户多次输入相同文件名的麻烦。

五、程序运行结果：

现以一个较小的文本文件 source.txt 展示程序运行界面。

source.txt 中包含中文、英文和特殊字符（如'\n','\t',空格），具体内容如图 1。

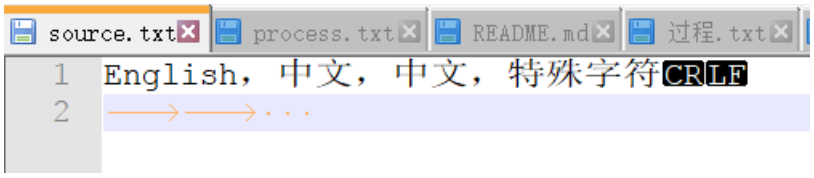


图 1　source.txt 文件内容

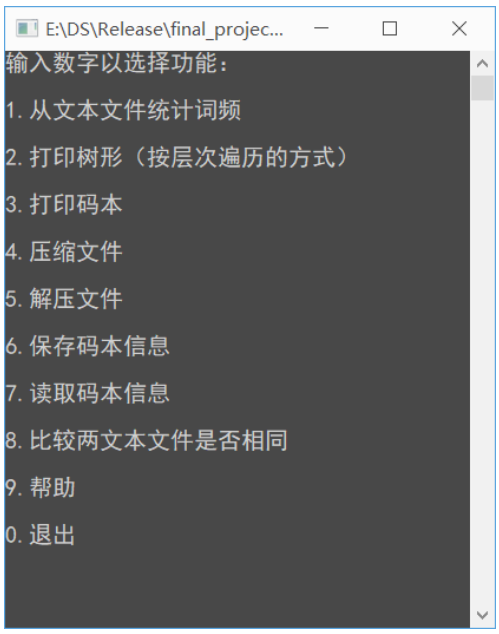


图 2　程序主界面

图 3　对 source.txt 进行词频统计

主界面内容如图 2 所示，输入“1”后，输入回车，可以对 source.txt 进行词频统计，程序运行结果如图 3。进行词频统计后，便可以打印树形和码本，分别如图 4 和图 5 所示。除了打印到屏幕，还可以将词频和码本信息保存为“可视化”的 txt 文件（如图 6），或者可以被该程序读取的.tf（term frequency）文件。

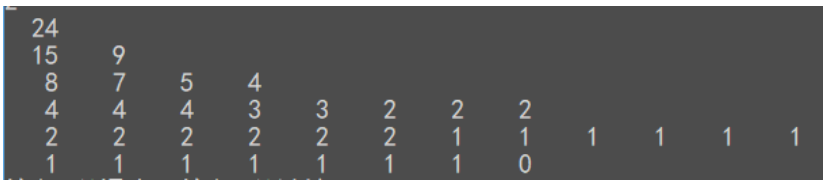
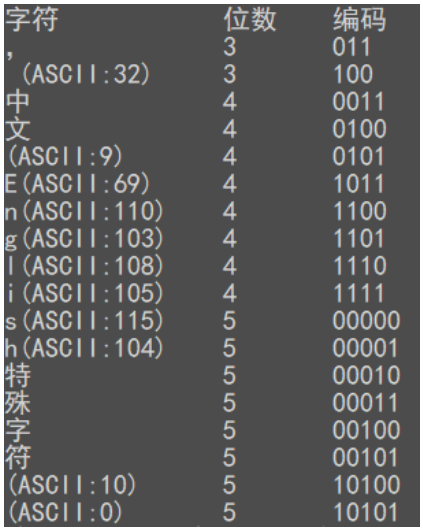
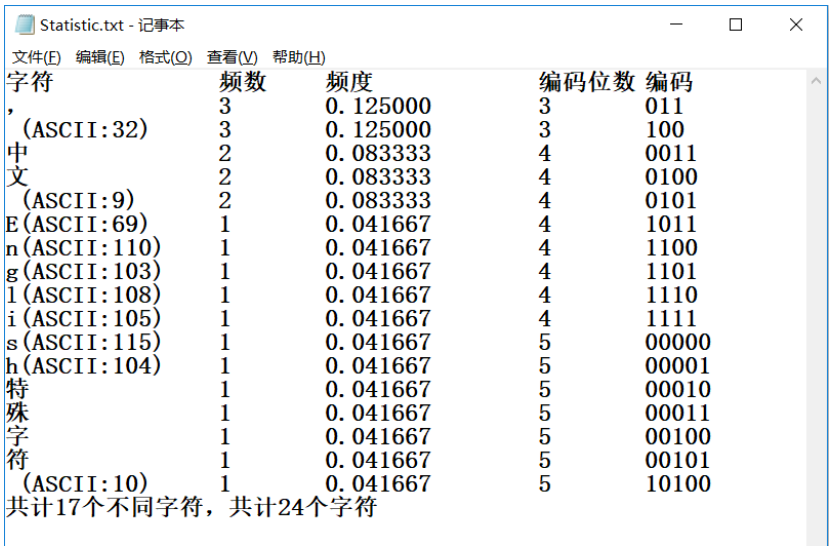


图 4　打印树形（层次遍历）



字符	位数	编码
，	3	011
(ASCII:32)	3	100
中	4	0011
文	4	0100
(ASCII:9)	4	0101
E(ASCII:69)	4	1011
n(ASCII:110)	4	1100
g(ASCII:103)	4	1101
l(ASCII:108)	4	1110
i(ASCII:105)	4	1111
s(ASCII:115)	5	00000
h(ASCII:104)	5	00001
特	5	00010
殊	5	00011
字	5	00100
符	5	00101
(ASCII:10)	5	10100
(ASCII:0)	5	10101

图 5　打印码本



字符	频数	频度	编码位数	编码
，	3	0.125000	3	011
(ASCII:32)	3	0.125000	3	100
中	2	0.083333	4	0011
文	2	0.083333	4	0100
(ASCII:9)	2	0.083333	4	0101
E(ASCII:69)	1	0.041667	4	1011
n(ASCII:110)	1	0.041667	4	1100
g(ASCII:103)	1	0.041667	4	1101
l(ASCII:108)	1	0.041667	4	1110
i(ASCII:105)	1	0.041667	4	1111
s(ASCII:115)	1	0.041667	5	00000
h(ASCII:104)	1	0.041667	5	00001
特	1	0.041667	5	00010
殊	1	0.041667	5	00011
字	1	0.041667	5	00100
符	1	0.041667	5	00101
(ASCII:10)	1	0.041667	5	10100

共计17个不同字符，共计24个字符

图 6　将码本和词频信息保存为可视化的 txt 文件

选择压缩功能后，会提示选择时间优先模式还是空间优先模式，选择完成后会要求输入源文件名和目标文件名，如图 7 所示。压缩成功后有提示，利用 HxD 软件可读取该文件，文件内容如图 8。

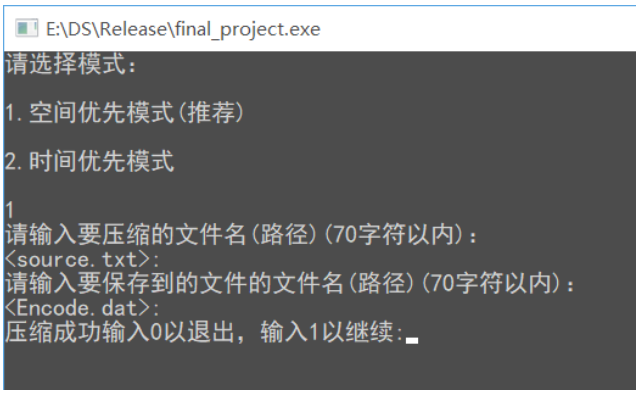


图 7　文件压缩界面

图 8　按 16 进制显示 Encode.dat

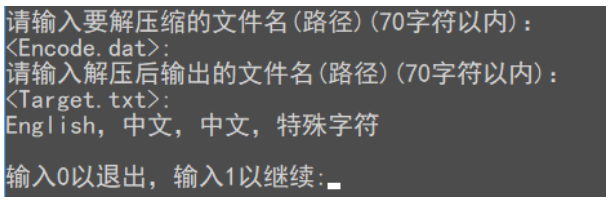


图 9　解压文件

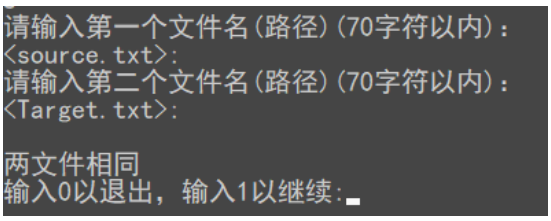


图 10　比较两文件是否相同

解压文件的界面如图 9，解压完成后还可比较两文件是否相同，如图 10。

此外，我们以功能 7——读取频率信息为例，检验程序鲁棒性。如图 12，我们可以见三次输入中，分别出现了 a)未以.tf 结尾，b)文件名中出现了非法字符，c)文件名不存在的错误，程序提示错误后依然正常运行。

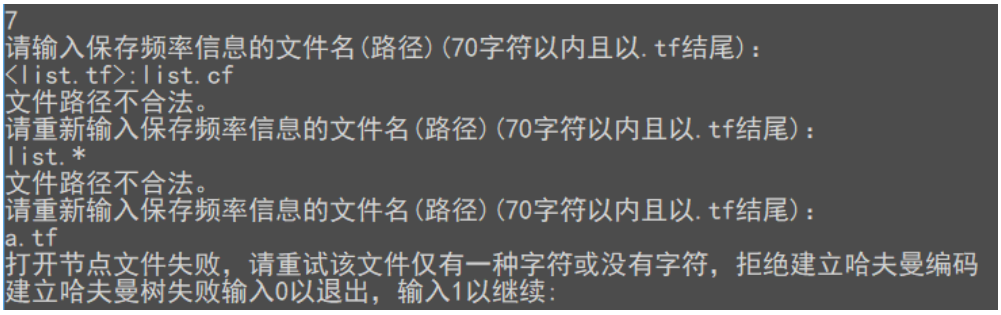


图 12 鲁棒性检验

现在我们以取了两个大小不同的、包含中英文的文本文件为 source.txt 检查该程序的压缩效果。

	文本文件大小	压缩后大小	压缩比
Source1	16.4 MB	11.1 MB	67.68%
Source2	3.98KB	2.13 KB	53.52%

六、实验心得：

这次实验中，我们实践了项目开发的很多流程，从分析需求，模块划分，到协作开发出最终的产品，每个人都感到新鲜又有参与感和成就感。

在实验中，我们又进一步巩固了上学期学习的 C 语言基础，同时锻炼了搜索技巧，学到了很多新的函数、新的项目开发技巧。同时在多人的协作中，整个项目在大家的各种思路的交锋下不断优化，不断完善，还有些时候我们将不同的思路都实现，将不同的 branch 合并，反而产生了适用于不同场景的功能。

这门实验课程很是培养我们的能力，不止编程能力，还有团队

合作分工能力，从一开始的小组讨论到逐渐明确思路，完成分工，然后知道编程中的困难再到解决困难，其中我们牺牲了平时自习的时间，也熬过夜，一次次实验，一次次失败然后互相讨论又或是向老师询问，最后到问题解决，其中的汗水以及乐趣恐怕只有小组一起讨论，亲身体会了才会明白。

就考核方式而言，我们觉得汇报的形式比笔试要好得多，汇报更可以看出能力；有些内容较难，我觉得还需要根据同学自己本身的能力自己去选择，或者是两个人或者三个人甚至一个组共同去攻克去完成。

再者，这节课不能光从网上复制写成的代码，一定要自己弄懂算法的内涵，理解算法，不然完成下来也只是一无所获。

总的来说我们十分赞同开设这样的实验课程，虽然课程并不是特别久，但的确是对程序设计能力，小组合作能力有极大的帮助与提高。

七、致谢词：

值此实验完成之际，最要感谢的是我们的指导教师李峰、顿玉洁老师。老师从一开始上课的时候就非常耐心的对我们进行项目的解读，一些基本问题的讲解。并且在编码过程中给我们组提供了大量建议，告诉应该注意的细节问题，细心的指出错误，并和我们一起找出错误所在。他们对数据结构程序设计的见解，使我们组可以顺利地完此项任务，并受益良多。李峰、顿玉洁老师严肃认真的工作作风，一丝不苟的工作态度，诲人不倦的教学风格给我们留下深刻的印象。在此，谨向李峰、顿玉洁老师表达衷心的感谢！其次，也要感谢小组内其他组员，我们一起解决困难、共同努力，让我们感受到了合作的力量，也让我意识到小组中的每一部分对小组成功的重要性，谢谢你们!