苏州大学实验报告

院、系	计算机学院	姓名	赵鹏	学号	2127405037
课程名称	信息检索课程设计				
指导教师	李正华		实验完成日期		2022/2/21

实验名称：分字

一. 实验目的

学习字符编码的基本知识。能够利用字符编码将中文字符用'/'分隔开来。

二. 实验内容

给定存有中文语句的文本文件 Sentences.txt,通过编写 C++程序运行程序后能够将文本中的字符字用 '/ '分开，例如输入"我爱中国"，输出"/我/爱/中/国/",并将修改后的文本输出到 out.txt.

三. 解决思路（流程图或伪代码）、遇到的问题和解决方法、运行结果

1.打开文件

2.依次读入每一行到字符数组 temp 中

3.对于每一行，定义一个从 0 开始的指针 i，通过定义的 getBytes 函数计算编码当前字符的字节长度 len;

4.将 temp[i]到 temp[i+len-1]存储到字符数组 c 中

5.输出字符数组 c 和一个/"并将指针 i 向后移动 len 位，若移动到末尾则输出空字符和换行符

6.循环以上过程直至处理完整个文件。

遇到的问题：1.不会读取和输出文本文件。

2.判断字符是由几字节编码

3.将 char 转为 int 时，由于编码最高位是 1，导致得到的 int 是负数，与预期不符

解决方法：将强制转换后得到的数字加 256 得到正确的编码数字。

解决方法：1.调用 fstream 库中的 ifstream 和 ofstream 函数以字符串的形式读取字符，输出字符。

2. 定义 getBytes 函数，通过位运算&和>>获得编码二进制下开头连续 1 的个数，即为结果。

3. 将强制转换后得到的数字加 256 得到正确的编码数字。

四. 实验总结

通过本次实验，深入理解了字符的编码规则，以及 C++中文件的读入和输出。能够利用 C++完成汉字的分字工作，在实验过程中练习了位运算的使用。较好的达到了实验的目的。