索引表的生成

19335025 陈禹翰 chenyh369@mail2.sysu.edu.cn 10月1日

摘要

本实验通过对输入文件中的单词进行读取，累计相同单词词频，生成索引表输出至输出文件中。

引言

要解决的问题：设计一个程序，输入是一个英文文本文件，输出是文件中所有词及其出现的频率，并且按照字母的 ASCII 码表顺序排列。

解决方法

运行形式：

```
.\count.exe .\inputfile .\outputfile
```

输入是一个英文文本文件，输出是文件中所有词及其出现的频率，并且按照字母的 ASCII 码表顺序排列。

使用的数据结构是数组，因为今天不太想用链表，算法主要是：

1. 进入readFile()函数，读取时过滤掉标点和换行符并将每个单词存储在string数组word中
2. 进入wordFreq()函数
    1. 先初始化cnt数组全部元素为1；
    2. 对word数组中的元素进行排序，按ASCII码表从小到大，使用冒泡排序的思想；
    3. 遍历word数组，当有一个string与flag的string相等时则说明重复，则在cnt数组中加一，同时record数组中后一个string的下标置1来屏蔽这个已经被合并掉的词，若没有则将这个不同的词设为flag继续进行词频统计；
3. 进入writeFile()函数，遍历将没有被屏蔽的词和它对应的词频写入文件中。

程序使用和测试说明

使用g++编译：

```
g++ .\count.cpp -o count.exe
```

运行：

```
.\count.exe .\inputfile .\outputfile
```

总结和讨论

我的实现特色在于我过滤了除了英文字母大小写之外的其他字符，并且我使用了静态数组进行实现。存在的问题便是使用数组实现不够直观，应该改进为采用链表进行实现。通过这次实验我学会了使用main函数的参数。

参考文献

命令行参数如何运行

C++文件和流