《数据结构实验报告》

2015080121 软件工程52班 李在弦

一、实验目标

-实现一个基于文本内容的知乎日报文章检索。使用栈分析网页结构，提取文章关键信息，并对正文进行分词操作。

二、实验环境

-开发环境（建议）

操作系统：Windows10

IDE：Visual Studio 2015

编程语言：C++

三、抽象数据结构说明

-字符串

有char和wchar_t形式的两个字符串。

-字符串链表数组

普通的链表数组（自定义的哈希图）。

-栈

普通的栈。

四、算法说明以及流程概述

-网页解析

1. 先对标签进行处理，提取有效标签。（同时有‘<’和‘</’的标签）

2. 查找“input”文件夹里面的“*.html”文件。

3. 开始做网页解析（遇到有效标签的‘<’部分就进展，然后开始保存内容。遇到‘</’退栈。）

-中文分词

1. 查找“词库”文件夹里面的“词库.dic”文件。

2. 把文件里面的单词都保存在字符串链表。保存方法是这样的，先提取单词的长度，然后一直加（一个字的数字是 * (index+1)）值，index值为长度时停止。然后做 %1000 就得到字符串链表的 “index”值，再把单词保存在字符串链表[index]里面。

(如果对“联赛”进行处理，（设‘联’= 24635，‘赛’= 36452，字符串链表是‘CSindex’）

所以，CSindex = （(24635 * (0+1)) + (36452 * (1+1))）%1000 = 539；

3. 取单词同时取文件里面中单词的最大长度，然后打开“*.info”文件读入一行一行。读入一行之后把文章保存在某个字符串里，然后对它进行分词。做分词用逆向最大匹配。

五、输入输出及操作相关说明

-把输入文件（如 “*.html”）放在“input”文件夹里。

-点击“exe”文件夹里面的“大作业.exe”可执行文件。（不要改变这文件的位置）

-输出文件会生成在“output”文件夹里。

六、实验结果

-可以自动查找“input”文件夹里面的“*.html”文件，然后开始做网页解析。

-做完网页解析之后自动开始查找“词库”文件夹里面的“词库.dic”文件，然后把“词库.dic”里面的所有单词都保存到字符串链表的数组里。

-保存好之后自动做分词算法。

八、功能亮点说明

做了有跟哈希图相似结构的字符串链表数组。（其实不确定是不是跟哈希图的结构相似。。哈哈）

七、实验体会

做这次实验之前，我不知道“html”文件的结构，“UTF-8编码的特性”，“哈希图”和“wchar_t”的存在。通过这次实验我提高了我的写代码能力！！^_^