实验报告

2014012106 黄子懿

一、实验目标

用栈结构分析提取网页结构，提取关键信息，对正文进行分词操作

二、实验环境

VS2012 win7 操作系统（部分注释在 mac 下完成，已经重新编码…不排除有未知的 bug）

三、抽象数据结构说明

1.Stack

一个很普通的栈

主要为实现先入后出的结构所以内部存储使用指针，实现了普通的 pop\push\top 等必须的函数，总之没什么特别的，内存采用满了扩到 2 倍的方法来存

2.CharString

一个很普通的字符串类

实现了多种多样的构造和赋值函数

实现了子串、依 index 访问、连接、转换为 string，获取头指针，获取长度等普通的函数

尽量允许多种参数的调用

另外针对本实验实现了一个比较奇特的打印函数（跳过多余空格，忽略换行）

还有一些为了方便调用而存在的函数，比如查找下一个某字符，比如查找下一个非冗余字符等等

（详见注释）

3.CharStringNode

继承自 CharString

链表节点，没什么特别的

4.CharStringLink

实现了普通的获取节点个数，判断是否为空函数

实现了允许各种各样的参数的添加节点，删除节点，搜索节点函数（按节点，按数据，按 string，按指针等）

还有针对本实验的比较奇特的打印函数，（跳过中文标点符号和空行）

5.Tag

标签类，一个 Tag 对应一个 HTML 标签，存储了 HTML 标签应该有的东西（标签名，属性名，特别提取出的 class 名，内部文本等），并将存储孩子标签指针和父亲标签指针。

另有一些变量判断是否为重要的标签，是否为正文标签

成员函数就是分析整个字符串得到 Tag 各个成员变量的构造函数，以及针对本实验的 print 和 divide 函数（采用递归的方法，通过内部是否有其他标签来决定是否输出）

在判断重要性上依据标签名，类名来决定

6.TagChecker

网页标签提取类，一个 TagChecker 实例对应一个 HTML 文件，内部存储了源文件内容，以及顶端标签容器（对于标准 HTML 文档来说，应该只有 DOCTYPE，head 以及 body），实现的函数有构造函数（通过输入流获取 HTML 文件全内容并存储），solve 函数（用栈结构解析 HTML 文件并存储于容器），以及 print 以及 divide 的入口函数（详见 Tag），还

有预留的 errorreport 函数（用于判断标签正常关闭等问题）

7.Dictionary

字典类，存储一个 hash_map 用于记录词表

只有普通的通过输入流的构造函数和查找函数

8.SentenceSeparator

断句操作者类，用于断句，存储需要断句的字符串指针，结果字符串链表指针及依据词表指针，只有一个分词函数

他显然不局限于本题

9.App

预留接口类，提供了题目要求的三个接口，还有为了实现批量断句（毕竟正文很长）的另外一个接口

四、算法说明

网页解析：TagChecker::solve()函数

cursor 为当前光标位置，不断向后查找<以及</以及>，遇到<记录，遇到>查看是否已记录匹配的<，若是，压栈并设为当前父亲，否则报错，遇到</则直接去寻找配套的>并退栈并让出父亲位置，特殊处理对于一些<blabla/>标签，预留的 CORRECT_MODE 宏决定是否允许不关闭的标签。

中文分词算法：SentenceSeparator::separate()函数

cursor 为当前光标位置，每次做最大匹配，然后长度递减，匹配至 len<=3 或者找到字典中条目时加入链表，特殊处理英文字符以及数字，将连续的英文字符及数字也加入链表

五、流程概述

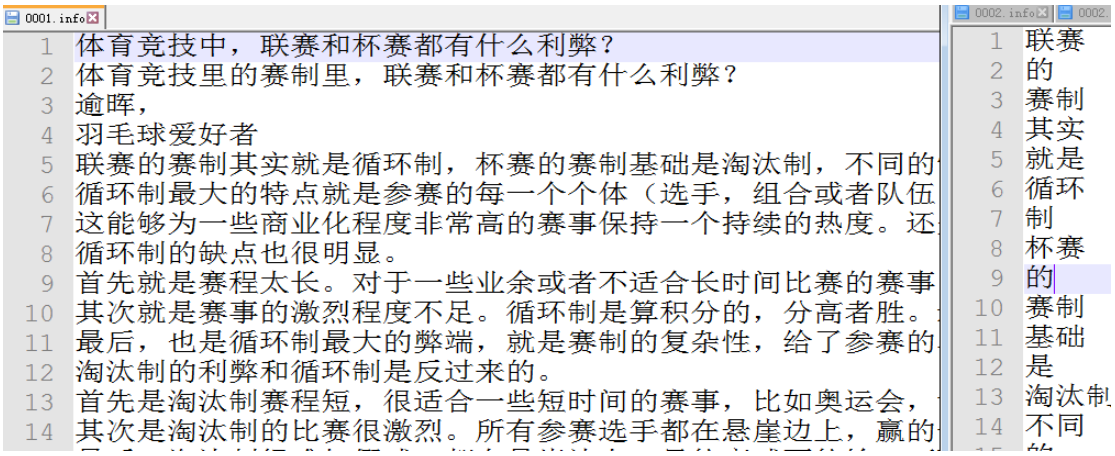
test.cpp 规定了 main 入口，然后构建 app 实例，app 实例初始化词典（initDictionary），由输入文件流初始化 TagChecker，TagChecker 执行 solve 提取重要信息，执行 extractInfo 接口打印重要（由 Tag 构造函数及 print()函数规定）条目，执行 doDivideWords 函数对正文条目执行分词算法（由 Tag 构造函数及 divide()函数规定），注:正文及重要两个 bool 变量采用父亲是则儿子一定是的规则。

六、输入输出及操作相关说明

双击即可，目前由 main 函数规定直接处理 0001.html~0010.html，如果变更需要修改源代码重新编译，如果出现预期之外的错误，会由 cerr 输出，比如出现标签之外的<或>符号

七、实验结果

基本符合预期



这是 0001.html 输出的部分内容

八、亮点

其实没什么亮点，非要说的话是自定义操作类尽可能多的提供了接口，对类成员变量和操作等基本进行了封装，程序拥有较好的鲁棒性，可以判断出输入文件的许多错误，提供了不同的标准