第一次实验报告

软 73 沈冠霖 2017013569

一．实验目标

本次实验有四个目标：首先，自己实现字符串，栈，链表，哈希表四个基本类，并且尽量做成模板类来复用。其次，能够读取几十万长度的词库，并且进行保存，散列变换。之后，能够对符合要求的网页进行解析，提取其标题，时间，来源，正文。最后，对正文进行中文分词。这一切都是为了构建之后的新闻检索与推荐系统。

二．实验环境

Win10，VS2017 下进行开发。

三．抽象数据定义

定义了一个字符串类 NCharString，可以进行空字符串构造（空构造函数），通过 char 来构造（重载构造函数），赋值(m_SetValue)，复制(m_Duplicate),截取（m_CutString），拼接(f_Concat，m_PushBackChar，m_PushBackString),查找子串(m_FindSubString),比较（m_Compare）等功能。

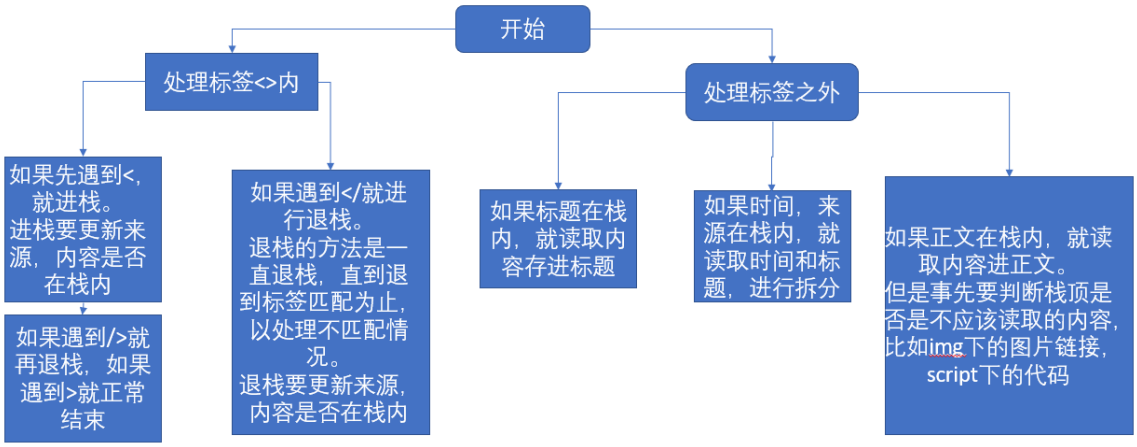
定义了一个链表类 NChainLink，是模板类。定义了查找（查找第 i 个位置的 m_SearchAtPlace,查找第一个和给定数据相等的 m_SearchEqual),在给定位置后插入(m_InsertAtPlace),删除给定位置(m_DeleteAtPlace)，构造和析构功能。还继承了一个字符串链表 NCharStringLink,有一个 m_PushBack 函数。

定义了一个栈 NStack，基于链表，里面存储标签和标签内的内容，便于进行解析。可以实现入栈（m_PushBack）,出栈(m_PopBack),判断是否为空(m_JudgeEmpty).

定义了一个 29 万长度的哈希表 NHash，可以把一个字符串链表转换成哈希（构造函数），也可以读入一个我自定义的字符串来搜索其位置（m_SearchList）。此处有引用代码[1]。

四．算法说明

解析算法使用栈实现，在 NNews 类的构造函数中。具体流程如下：



分词算法和课上讲的大体一致，需要用哈希表。首先，建立词库就要建立对应的哈希表。之后，在分词的时候要先忽视一切 ASCII 码，然后按照从长到短的顺序遍历子串的长度，并且截取子串，放到哈希表里进行比较。如果能匹配成功，就存储，去比较之后的文章。否则就减少子串长度再比较。

五．流程概述

【读取词库，建立哈希表】→【读取目录下全部 html 文件】→【逐个文件进行解析】→【逐个文件进行分词】→【逐个文件进行输出】

六．输入输出及相关操作说明

如果要执行我的代码，请把输入文件放到当前目录下的 input 文件夹内，全都命名为 xxx.html，长度不要超过 1000 个字节。词库请使用我提交的 exe/WordList/WordList.txt，并放到本目录下的 WordList 文件夹中，取名为 WordList.txt。必须在本目录下新建文件夹 output。之后打开工程，编译运行即可，详情见代码目录下的 readme。

如果要执行可执行文件，请不要动我的词库 WordList,否则会出现错误。请把要添加的测试网页放到 input 文件夹里，命名为 xxx.html，不要超过 1000 个字节。请不要删除 output 文件夹，在那里输出每个文件的.txt 和.info。直接点击.exe 就可以运行，详情见可执行文件目录下的 readme。

七．实验结果

解析结果大体符合预期，大部分网页都能正常解析，也能规避标签不匹配现象，也能解析图集的文字。但是我没有读取图片和其他链接，部分网页的编程不符合规范，无法正常解析时间，来源，正文也会多一些杂质。

分词结果大体符合预期，解析分词 20 个网页大概需要几秒到几十秒不等。

八．功能亮点

自己实现了哈希表 NHash，算法来自附录 1，能够把分词速度从卡死优化到 20 个网页只需要几秒到几十秒不等。

九．实验体会

网页的质量参差不齐，或许这就是真实的工程，需要大量除杂等，很麻烦。这也让我大大怀疑网易工程师的道德素质和业务水平，并且坚定了绝对不去网易就业的想法。

个人觉得大作业本身比较简单，但是词表的长度过长，网页数据杂质太多则大大增加了完成难度，大部分时间都花在除杂，实现哈希表了。

这次有很多不足之处，之前写小作业没有把类模板化，导致需要重写。开发的时候没有及时进行单元测试，后期很多错误来自最基本的类。没有及时测试 release 模式，导致最后 release 也出过问题。没有时间进行更加精致的提取和分词。

个人觉得大作业布置时间不合理，按照软院的课表，4-8 周是课程最密集，压力最大的时候，而这个时候完成大作业难度很大，我大作业一发下来就开始写，到了 ddl 前一天才写完，而且这段时间几乎没进行任何休息，还经常熬夜到晚上 4-5 点，第二天八九点又要起来继续上课，我想不少同学和我差不多。而且这次的基本内容在十一前就能实现了。希望以后大作业和第一次难题作业在十一前发布。

十．我的编程规范

我的注释主要写在.hpp 和.h 的函数之前，详细说明了函数的参数，输出，功能，注意事项。每个文件前也有简要的注释。在核心函数的部分代码上，我有简单的注释。

我的命名规范如下：局部变量用小写字母命名，用下划线隔开。类属性和函数用匈牙利命名法，前缀加上 m_,友元函数加上 f_，常量前加 c_,全局函数加 g_。

十一．引用的代码和感谢

【1】https://www.cnblogs.com/youngerchina/p/5624453.html 哈希表的算法

【2】https://blog.csdn.net/qq_15947787/article/details/51384258 获取指定目录下全部文件

【3】http://www.cnblogs.com/hnrainll/archive/2011/07/27/2118812.html malloc 和 realloc 用法

感谢王世杰同学和我探讨哈希表，感谢吴海隽同学帮我测试可执行文件，感谢

黄翔同学提供的静态编译方法。