

# 实验报告

## 一、实验目标 //说明：红色的为实验二的更新

实验目标为实现一个基于文本的音乐检索与推荐系统，本次实验的基本目标是使用。

## 二、实验环境

- 操作系统：Windows 10
- IDE：Visual Studio 2013
- 编程语言：C++

## 三、抽象数据结构说明

自定义的数据结构：

数据结构	描述
CharString 类	重实现 string 类的部分功能，进行字符串的存储和操作
DoHtml 类	存储音乐信息成员变量，实现搜索、推荐、GUI 等功能
CharStringLink 类	实现字符串链表，进行存储信息
Song 类	包含音乐作者等信息，进行解析网页，获取音乐信息，分词等基本操作
Stack 类	实现栈结构，网页解析时使用
Dicwords 类	载入词典，将词典存为 SBT（Simple Balanced Tree）
B-Tree 类	实现 B-tree 的添加，删除，遍历，搜索
Document 类	存储分词对应的倒排文档的节点
wordNode 类	分词的 ID，出现文件以及次数，其中指针指向文档链表
BTreeNode 类	B-树的节点

## 四、算法说明

- 1、网页解析：通过字符串截取，获得目标区域。进行遍历 s[i]，根据标签可获得音乐信息，并将歌词中的作词作曲提取出来。
- 2、中文分词：（正向最大匹配法）
- 3、建立倒排文档：分词之后，将每首音乐的分词加入 B 树中，之后遍历 B 树对 wordNode 指向的文档链表按照出现次数排序。
- 4、B-树：内部节点为 BNode 类，内部包含 WordNode 类，wordNode 指向已经建好的倒排链表。
- 5、搜索：在 B 树搜索到文档链表，每个词的搜索结果的链表进行或处理，得到一个链表进行输出
- 6、推荐算法：根据标题，搜索到 ID，根据 ID 在 m\_song 数组中取得歌曲信息，将分词链表加入 B-树，排序获得出现次数最多的 5 个词汇，调用搜索算法，输出这五个词汇的合并链表

7、GUI:使用 Socket，搭建一个简易的服务器，将自己写好的 html 挂上去，之后 html 中调用函数，返回 html 中的输入，调用相应算法，返回到网页，显示出来。

## 五、 流程概述

[读取网页]->[解析网页]->[提取音乐信息]->[输出.info 文件]->[载入词库]->[中文分词]->[获得处理好的歌曲信息以及分词]->[建立倒排文档、B 树]

搜索： [取得输入]->[在 B 树中搜索]->[获得文档链表]->[输出]

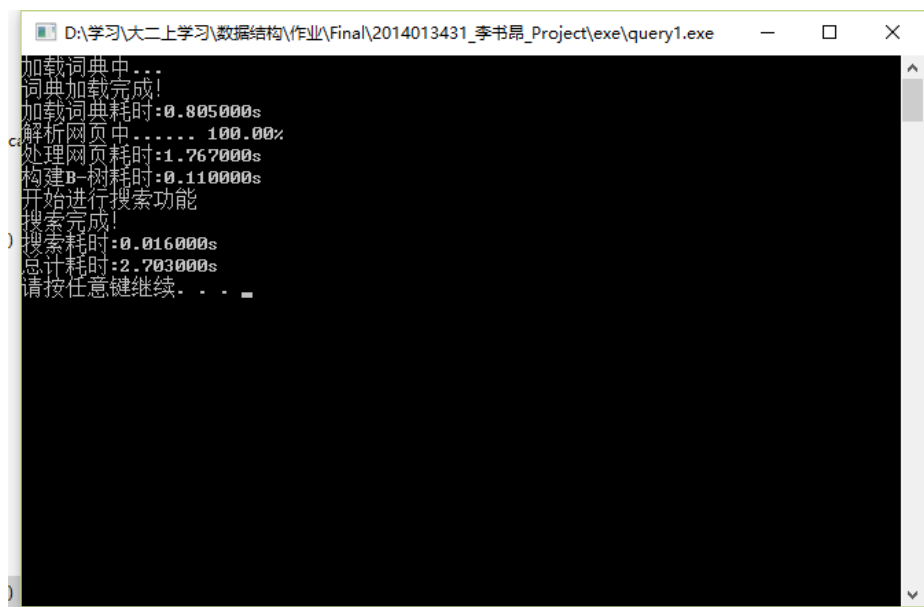
推荐： [取得输入] ->[获得对应歌曲 ID]->[获得分词链表]->[获得出现次数最多的五个词]->[在 B 树中搜索]->[获得文档链表]->[输出]

GUI： [搭建服务器]->[加载网页]-> [取得输入]-> [调用算法]->[输出]

## 六、 输入输出及操作相关说明

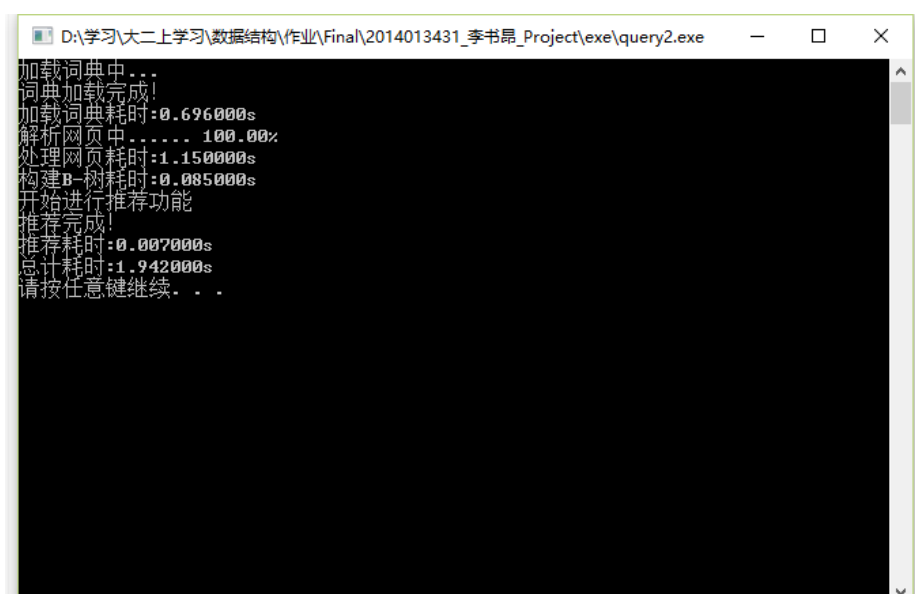
按照助教的说明，直接打开 exe 即可。gui.exe 需要输入，选择功能，点击搜索。

## 七、 实验结果

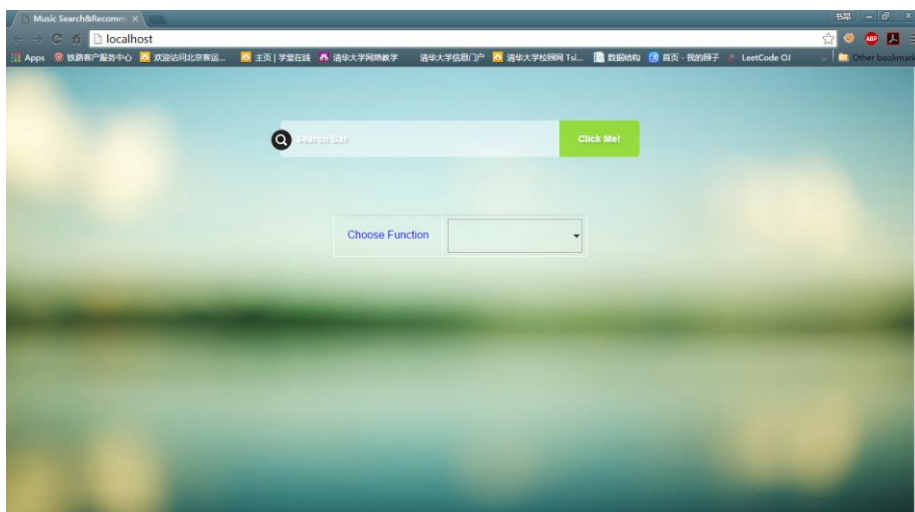


```
D:\学习\大二上学习\数据结构\作业\Final\2014013431_季书昂_Project\exe\query1.exe
加载词典中...
词典加载完成!
加载词典耗时:0.805000s
解析网页中..... 100.00%
处理网页耗时:1.767000s
构建B-树耗时:0.110000s
开始进行搜索功能
搜索完成!
搜索耗时:0.016000s
总计耗时:2.703000s
请按任意键继续. . .
```

Query1.exe



```
D:\学习\大二上学习\数据结构\作业\Final\2014013431_季书昂_Project\exe\query2.exe
加载词典中...
词典加载完成!
加载词典耗时:0.696000s
解析网页中..... 100.00%
处理网页耗时:1.150000s
构建B-树耗时:0.085000s
开始进行推荐功能
推荐完成!
推荐耗时:0.007000s
总计耗时:1.942000s
请按任意键继续. . .
```



## Gui.exe

### 八、 功能亮点说明

1. 可以处理任意数量，任意名称的网页，生成对应的.info，.txt
2. 使用栈结构处理标签，根据特定的 html 标签提取信息，解析网页
3. 使用自己实现的 SBT(Simple Balanced Tree)存储词典，时间复杂度  $O(\log n)$
4. 分词算法中实现数字匹配，对英文，以及一些字符进行处理，分割。
5. 分词算法中使用停用词表
6. 实现 B-树删除
7. 实现图形界面，用户使用性友好
8. 运行速度较快，1.5s 左右即可完成

### 九、 实验体会

感觉这次大作业还是很有效果的，学到了很多底层的结构设计，并且极大熟练了字符串操作，并且锻炼了栈，树，链表等等数据结构，感觉难度适中，不过思路比较清晰，比较容易设计类，进行封装，感觉算是从大一到现在，写的最舒服的一个大作业了。

希望明年能更难点->\_->，最好可以鼓励大家使用各种自己实现的树或结构去存储词典，进行算法时间，空间复杂度的比较，课堂上可以讨论一下。（辛苦助教！）

更新：

亲手写完 b 树的感觉非常赞，而且效率好高

还有感觉实验二交的时间比较急，比如自己想实现的很多网页功能比如播放音乐，加链表，都没有时间去写了，希望明年可以让大家有时间实现更多的东西