《数据结构与算法》课程实验报告

基于文本内容的音乐检索与推荐(第二部分)

实验目标

- 倒排文档及查询系统的构建
- 推荐系统的简单实现

实验环境

- 操作系统：Windows10
- IDE：Visual Studio 2012
- 编程语言：C++
- 内存：16GB

抽象数据结构说明

B 树：
1.每个结点 x 有如下属性： a. x.n，表示结点 x 中的关键字个数 b. x.n 个关键字以非降序排列 c. x.leaf，表示 x 是否为叶结点
2.每个内部结点包含 x.n+1 个指针指向孩子们 c[i]
3.x 中的关键字 x.key[i] 对子树中的关键字 k[i] 进行分割，n 个关键字，n+1 个子树
4.每个叶结点有相同的深度，即树的高度 h
5.每个结点的关键字数由 最小度数（minimum degree）t>=2 控制： a. 除根节点外，每个结点至少含 t-1 个关键字 b. 每个结点至多含 2t-1 个关键字
本次实验实现了 B 树的**构建、查找、插入**

文档链表：
本次实验实现了文档链表的**构建、插入、查询、释放**

算法说明

倒排文档的构建：

遍历已有的分词结果，对每一个词，先在 B 树种查找一次，如果查找到了，则根据已有的文档链表，添加相应信息，否则，如果没查找到，则在 B 树中插入节点，并添加相应信息。

推荐算法：

在构建倒排文档的同时，将每个文档中出现最多的词记录下来，用作推荐因素之一。对于每个歌曲，有一个推荐的权值，权值的计算综合了歌手 4，作词 3，作曲 3，专辑 5，及出现最多的词 2，数字分别为所占权重。

之所以这么分配权重是由于只能根据内容来推荐，个人认为同一专辑的歌相似度相对比较高、歌手次之、作词作曲再次之，在加上一个出现最多的单词，综合比较选出权值最高的十首歌。

实验流程

程序开始 >> 加载词库 >> 加载停用词库 >> 文件解析 >> 分词 >> 利用 B 树构建倒排文档 >> 批量查询或推荐或 GUI >> 程序结束

操作说明

输入、输出文件均与可执行文件同目录，网页文件位于 pages_300 中。分别双击执行 query1.exe，query2.exe，gui.exe 即可。

对于交互界面输入的说明：

请按照提示一步一步完成即可。

实验结果

实验结果保存在 exe 同级目录下

```
1 (30,45) (131,1) (222,2) (258,1)
2 (2,9) (192,2) (5,6) (6,6) (8,2)
3 (1,1) (2,2) (9,1) (10,4) (11,3)
```

result1.txt 的格式是每一行为对应的查询结果，使用(文件名， 多个关键字出现的总次数)，出现多个关键词文档的排序在前。

```
根据您的输入"教学"，我们找到2首对应的歌曲！
歌曲"免费教学录像带(内地版)"的推荐结果如下：
(258,免费教学录影带),(199,跨时代),(23,好久不见),(3,烟花易冷),(
歌曲"免费教学录影带"的推荐结果如下：
(257,免费教学录像带(内地版)),(3,烟花易冷),(258,免费教学录影带)

根据您的输入"七里"，我们找到1首对应的歌曲！
歌曲"七里香"的推荐结果如下：
(107,止战之殇),(240,乱舞春秋),(295,园游会),(68,我的地盘),(296,

根据您的输入"夜空中最亮的星"，我们找到1首对应的歌曲！
歌曲"夜空中最亮的星(微电影《摘星的你》主题曲)"的推荐结果如下：
(105,梦想启动),(1,蒲公英的约定),(2,手写的从前),(3,烟花易冷),(4
```

result2.txt 的格式说明：如果输入为准确歌名，则返回准确结果，如果输入为歌名的一部分，返回可能的歌名及其推荐结果，权值得分高的歌排在前。

GUI 的主菜单界面如下：按照提示进行即可

功能亮点

1：实现英文分词及其查询
2：停用词表
3：输入友好性
当用户查询歌曲输入模糊时，会根据用户输入返回可能的结果，如：

```
根据您的输入"教学"，我们找到2首对应的歌曲！
歌曲"免费教学录像带(内地版)"的推荐结果如下：
(258,免费教学录影带),(199,跨时代),(23,好久不见),(3,烟花易冷),(
歌曲"免费教学录影带"的推荐结果如下：
(257,免费教学录像带(内地版)),(3,烟花易冷),(258,免费教学录影带)

根据您的输入"七里"，我们找到1首对应的歌曲！
歌曲"七里香"的推荐结果如下：
(107,止战之殇),(240,乱舞春秋),(295,园游会),(68,我的地盘),(296,

根据您的输入"夜空中最亮的星"，我们找到1首对应的歌曲！
歌曲"夜空中最亮的星(微电影《摘星的你》主题曲)"的推荐结果如下：
(105,梦想启动),(1,蒲公英的约定),(2,手写的从前),(3,烟花易冷),(4
```

实验体会

首先，本次实验学习了利用 B 树构建倒排文档、并根据歌曲的信息设计了一个比较平凡的推荐算法，让我学习到了数据结构的实际应用，而不仅仅是在做一些练习题。由于数据量比较大，所以这次实验 2 很大一部分的时间花在了填实验 1 的坑上，比如内存泄露之类的问题，这让我意识到了自己的代码鲁棒性不够强，而且在写的时候没有考虑到代码的扩展性，导致有很多冗余的代码量。

同时，我还学习了如何在控制台界面写一个用户输入比较友好的界面，这对我来说还是挺新鲜的。

另外，这次 B 树的建立是照着《算法导论》这本书上写的，所以我觉得我学到的很重要的一个能力就是利用工具书，因为《算法导论》上的代码全是伪代码，从其翻译过来的这个过程我遇到了许许多多的错误和困难，这也让我意识到了前人的强大之处，每一个下标、判断条件都是精心设计好的，即使自己能够写出来，恐怕也没法写的那么清楚，还有许多需要学习的地方。

还有一点，我觉得通过大作业来完成一个小工程的成就感远远不是做些题能够比得上的，虽然这个过程中有很多郁闷、艰辛的时候，但也让我找到了一个学写代码的方法，虽然和很多大神的作业比起来，自己的程序还是显得很简陋，但是我也收获颇丰、慢慢进步着！

最后，谢谢老师和助教的精心设计！