苏州大学实验报告

院、系	计算机学院	姓名	赵鹏	学号	2127405037
课程名称	信息检索综合实践				
指导教师	李正华		实验完成日期		2022/4/24

实验名称： 布尔查询

一. 实验目的

1.了解布尔查询并编写代码进行布尔查询

2.学习利用倒排索引文件提高单词查询效率

二. 实验内容

编写代码利用倒排索引进行查询，由用户输入数据，程序输出查询结果。

输入：作业五生成的倒排索引文件以及 1k-data 文件夹中已提取正文的网页文本文件

输出：用户所查询单词的结果，第一行输出符合查询结果的所有文件，随后以此输出每个文件以及其中符合查询要求的语句并将查询词用#标注，同时保存查询结果到 txt 文件中。

三. 解决思路（如流程图或伪代码）、遇到的问题和解决方法、运行结果

解决思路：1.读取./index 目录下的 index.txt 索引文件生成字典。

2.读取./data-1k 目录下的所有文件名并进行编号，利用两个字典建立文件名和编号的双向映射。

3.循环读入查询语句并调用 Query 函数，将查询语句以函数参数传入。

4.根据查询语句 split 后的列表长度判断查询模式，对于每次查询将单词映射为文件名编号列表并根据布尔操作对列表取交并集。

5.获得最终查询结果后将文件名编号列表映射为文件名列表。

6.获得最终结果后以此打开列表中每个文件，按行读入，若改行包含查询词则利用字符串替换进行标注。标注完成后打印结果并将保存到./TestExample 目录。

索引文件说明:

本程序使用作业五生成的索引文件，以 txt 文件进行存储，存储结构如下：

<word1> [filename1,filename2,filename3…]

<word2> [filename1,filename2,filename3…]

<word3> [filename1,filename2,filename3…]

…

其中

word1、word2、word3 为利用分词程序获得的词语

[filename1,filename2,filename3…]为文件内容包含 word 的文件名构成的列表

索引文件的部分内容如下：

```
涛 ['呵怒.txt', '黄鹤楼送孟浩然之广陵.txt']
心花 ['呵怒.txt']
众怒难任 ['呵怒.txt']
众怒难犯 ['呵怒.txt']
众怒 ['呵怒.txt']
直眉怒目 ['呵怒.txt', '暗目.txt', '眉案.txt']
震怒 ['呵怒.txt']
怨怒 ['呵怒.txt']
余怒 ['呵怒.txt']
麝 ['呵怒.txt', '孤拔.txt', '春盎.txt', '百壹.txt']
愠怒 ['呵怒.txt']
睚眦 ['呵怒.txt']
蓄 ['呵怒.txt', '暗示.txt', '蓁艾.txt', '艾芜.txt', '蒿艾.txt', '蓬艾.txt']
汹 ['呵怒.txt']
心花怒发 ['呵怒.txt']
城 ['呵怒.txt', '谨敖.txt', '险傲.txt']
泄 ['呵怒.txt', '奥渠.txt', '暗语.txt', '霑露.txt']
开心 ['呵怒.txt']
鲜衣怒马 ['呵怒.txt']
鲜车怒马 ['呵怒.txt']
羞恼 ['呵怒.txt', '敖恼.txt']
喜怒无常 ['呵怒.txt']
喜怒哀乐 ['呵怒.txt', '哀乐中节.txt']
喜怒不形 ['呵怒.txt']
嬉笑 ['呵怒.txt', '嘻嘻呵呵.txt', '敖嬉.txt']
```

问题： 1.将作业五的倒排索引转为字典

2.在将查询单词的两个编号列表合并后无法找回对应的文件名

3.如果输入分词中不存在的词进行查询，则会出现 KeyError 的报错

4.在文件中对查询词进行标注

5.对列表进行交集并集

6.进行例如 A AND B OR C AND D OR E 的多布尔查询

解决方法：1.从前往后遍历每一行，如果扫描到列表的左括号'['则停止，将前一部分作为字典的 key,将后面的部分用 eval 函数转为列表作为字典的 value

2.在给文件编号建立字典时同时建立一个逆向字典，使得能够找回编号对应的文件名。

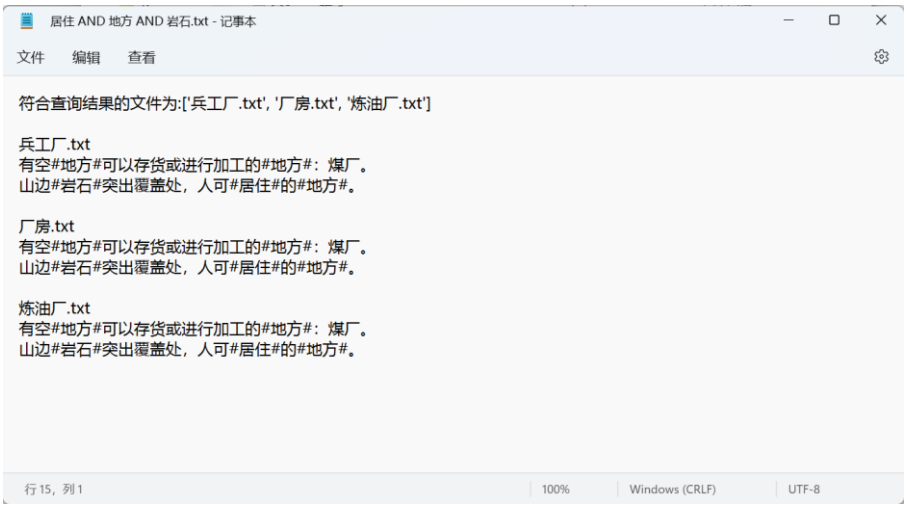
3.在分离出单词和对应的操作符后对单词进行一次特殊判断，如果两个单词中存在某个单词不在字典的 key 中则返回空列表

4.利用字符串自带的 replace 函数，将 word 替换为#+word+#

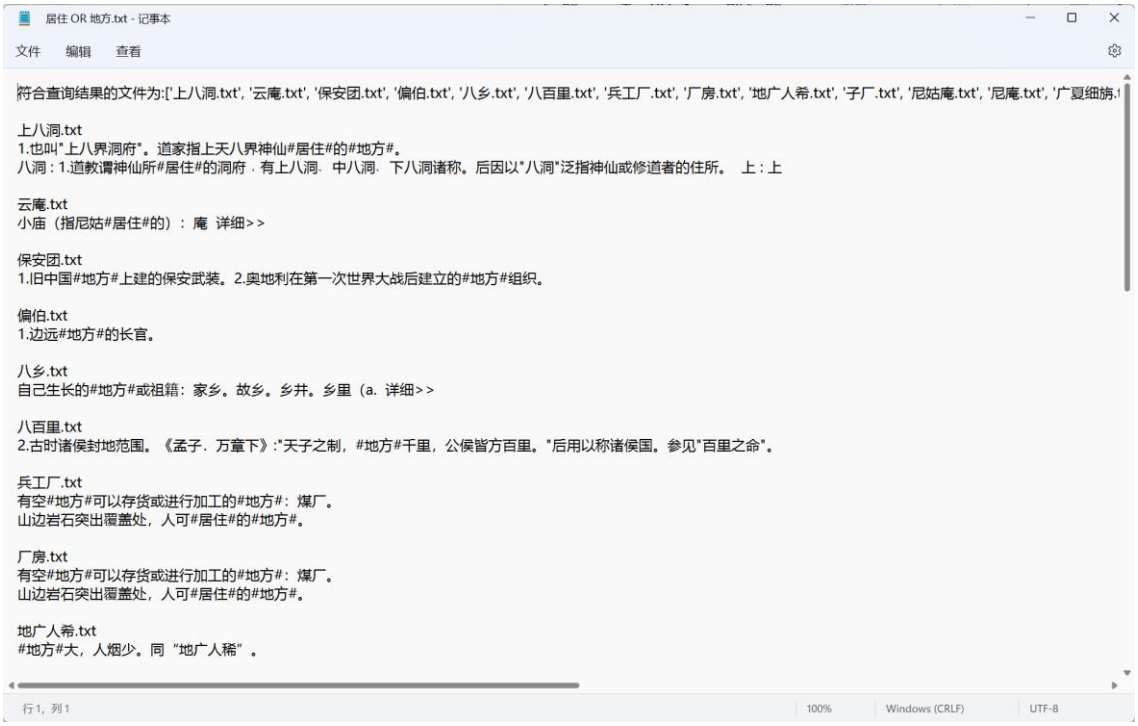
5.利用双指针的方法实现在 O(n)的时间复杂度内对两个列表取交并集，提高了程序的运行效率。

6.对于多布尔查询，先取出前三项(A AND/OR B)进行一次运算获得一个列表，随后用不断取出列表后的后两项与该列表进行运算。最终得到结果。

运行结果



教务处制



四. 实验总结

通过本次实验，我加强了对文件操作的使用，了解了布尔查询的方法，学习到了利用列表模拟集合交并操作的方法，并能够编写代码利用倒排索引进行布尔查询，提高查询效率，提升了自己的编程能力、发现问题、分析问题和解决问题的能力。