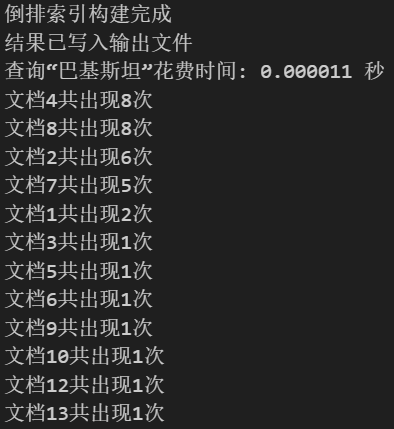
自然语言处理第三次实验报告

唐屹 10235501451

1.代码实现功能

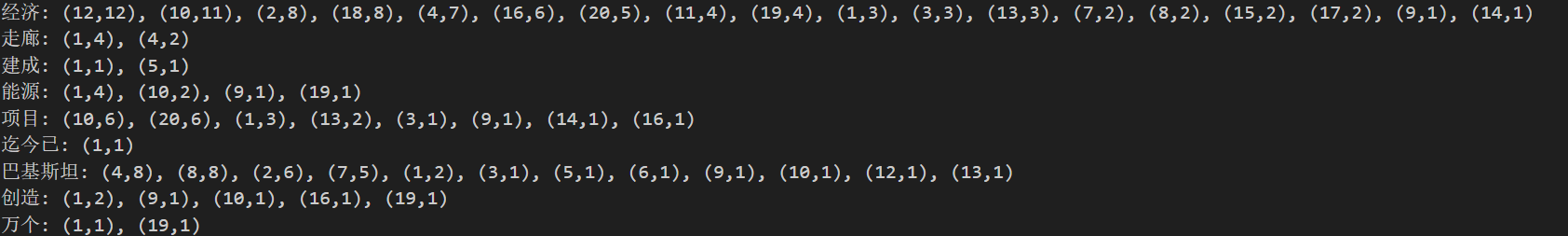
1.1单个词语查询



查询成功示例

查询失败案例

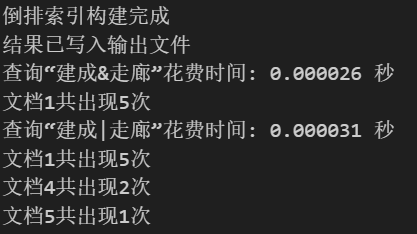
1.2 输出倒排索引文件（output.txt）



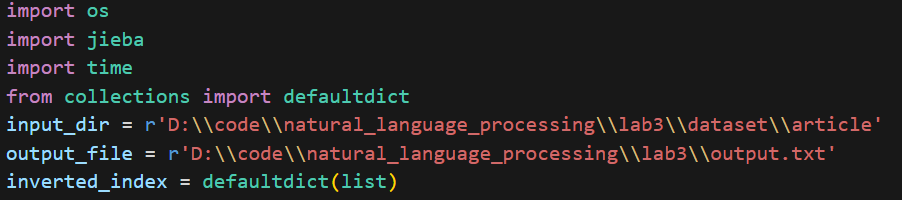
(output.txt部分截图，其中括号记录文档名和词语出现次数，如经济（10，11）代表了“经济”一词在文档10出现了11次)

1.3多个单词进行查找

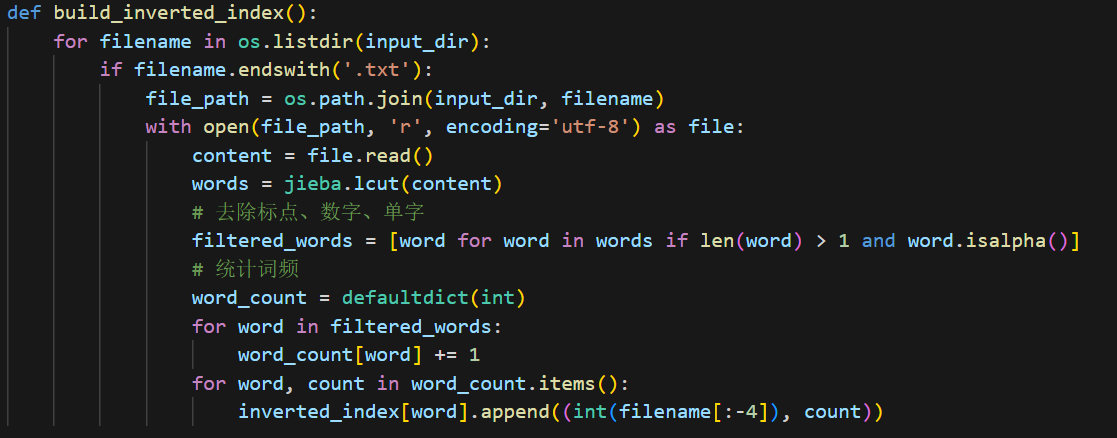
以“走廊”和“建成”为例，为查询走廊和建成共同出现的文档，输入“走廊&建成”；为查询走廊和建成共同出现的文档，输入“走廊|建成”



2.实现细节



必要的库的导入和文件路径，并创建一个 defaultdict，键为词，值为包含 (doc\_id, count) 元组的列表，用于存储倒排索引。



Build\_inverted\_index函数用于构建倒排索引，首先通过os库遍历目录下的所有txt文档，然后调用jieba库 分词，去掉单个字（len（word）>1）和去除符号（word.isalpha）,再统计

每个单词再文档中出现的次数，然后将单词和文档id和词频存入全局变量inverted\_index