第四周报告

1. 实验概览

本实验旨在理解建立索引的基本步骤（分析、索引、统计词频），理解索引查询的基本原理和步骤（语义理解并语言处理，查找索引，相关性排序）；基于Luence库和jieba中文分词库，实现中文网页索引的创建和搜索。

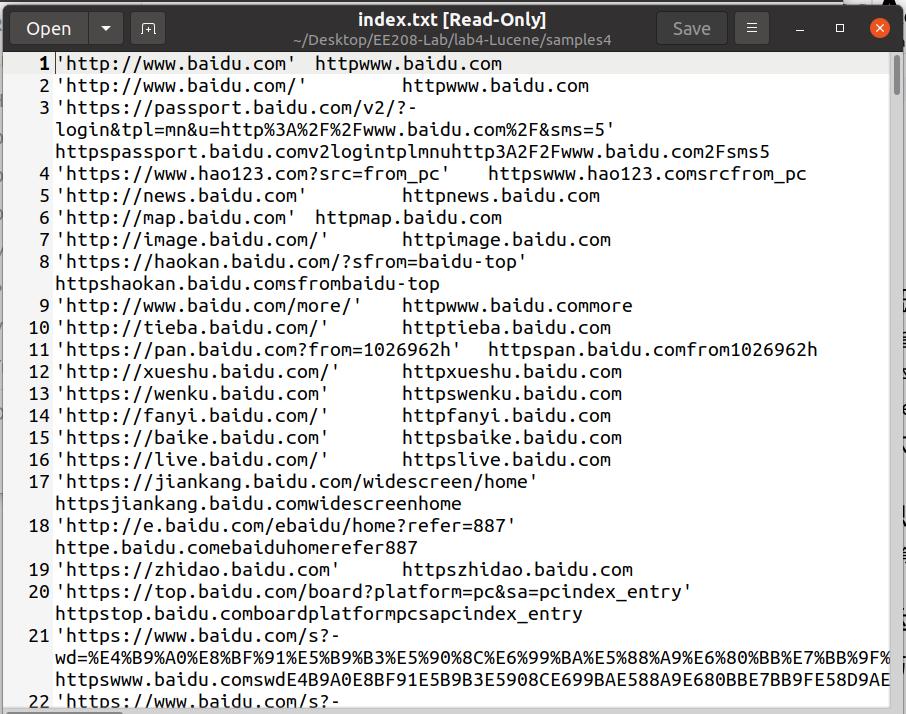
1. 实验环境

Docker: sjtumic/ee208

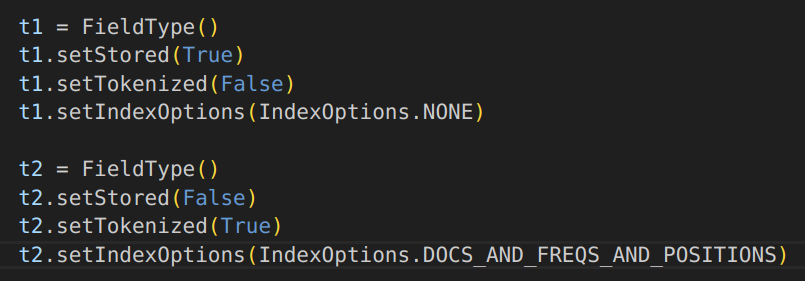
1. 解决思路

创建索引：

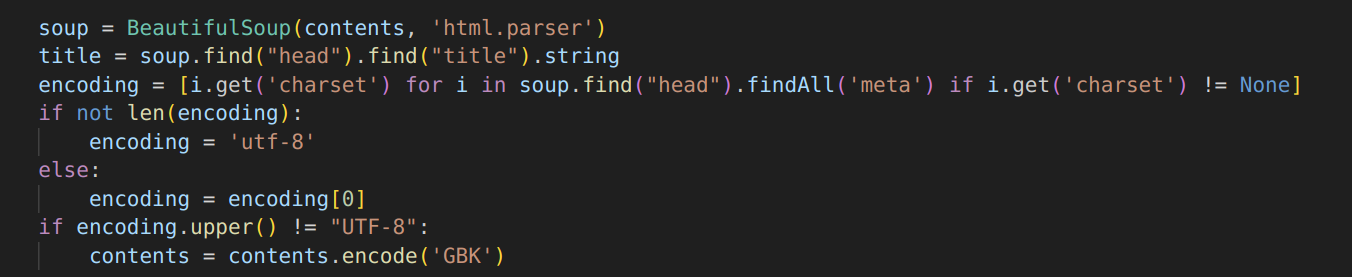
利用第三次实验中的并发爬虫爬取10k个网页（图为index.txt截图）



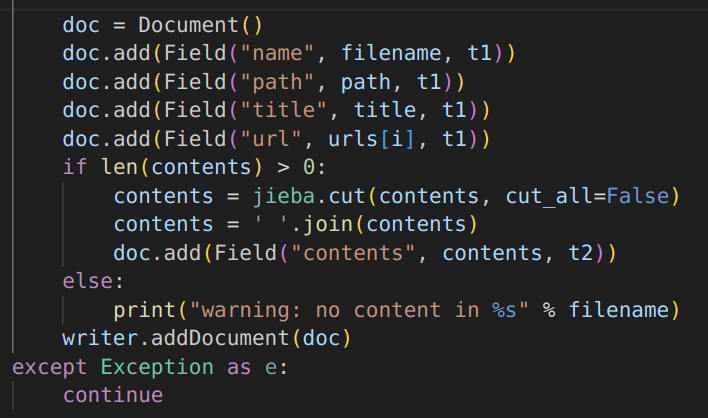
创建Field，分别用于无词频统计内容（文件名等）和含词频统计和位置记录的内容（网页正文内容）



读取index文件，查找每个文件名；利用Beautifulsoup解析网页内容，获得网页的标题和编码方法，利用编码方法重编码contents

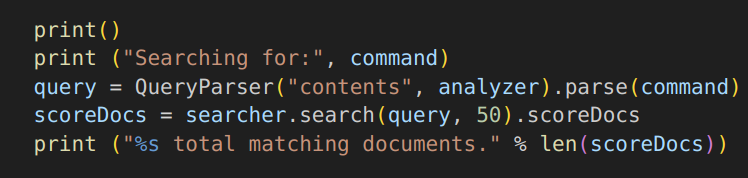


利用正则表达式去除contents中所有的汉字以外的字符lab4

创建Document实例，添加相关内容；其中，利用jieba库实现中文分词并中文隔开，利用WhitespaceAnalyzer作为分析器根据空格分词，实现索引的建立

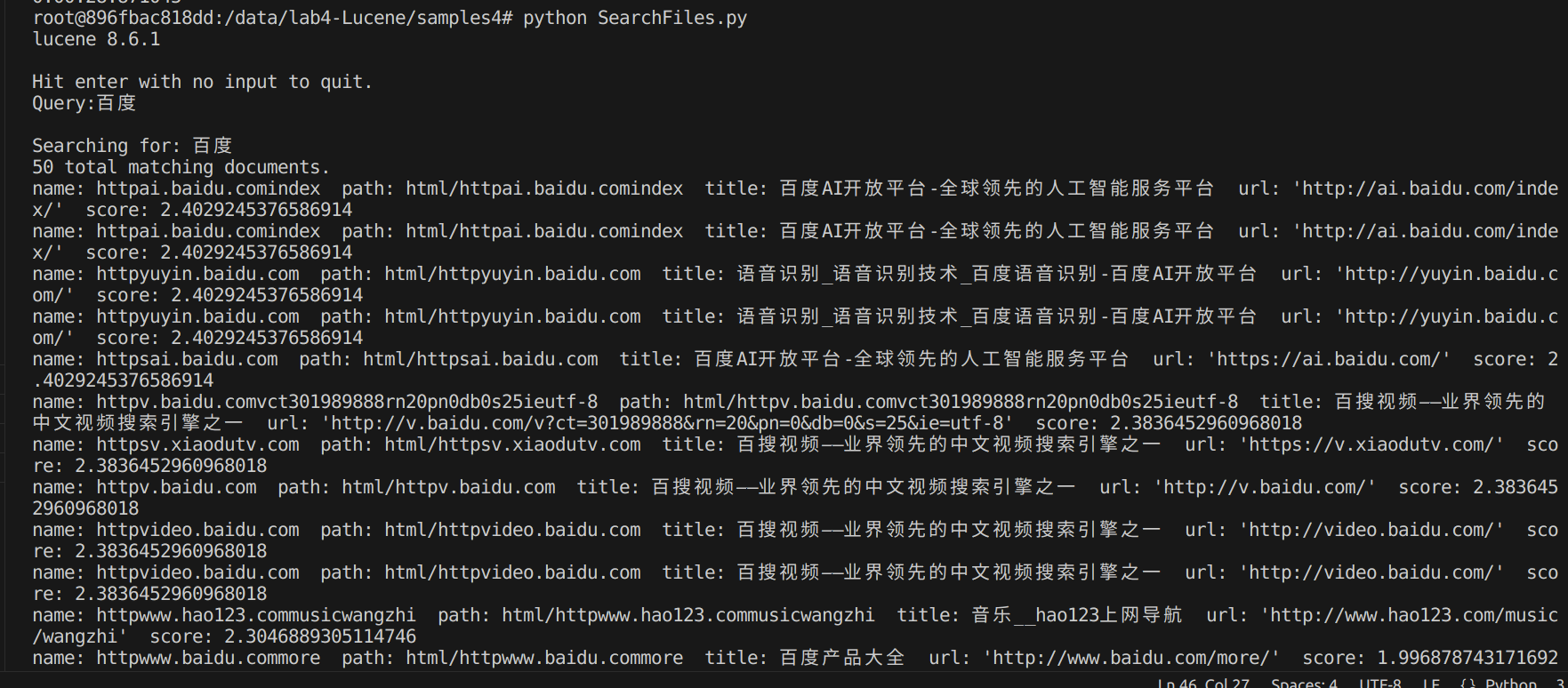
查找索引：

直接基于lucene库中的QueryParser对查询命令进行解析并调取索引查询



1. 代码运行结果

以百度为根节点爬取10000个网页，索引结果如下



1. 分析与思考

Lucene的实现：

存储：基于FST字典格式

查询：在 lucene 中查询是基于 segment。每个 segment 可以看做是一个独立的 subindex，在建立索引的过程中，lucene会不断的 flush 内存中的数据持久化形成新的 segment。每个segment是不可变的，只有在多个 segment也会不断的被合成时才被真正删除。