第五周实验报告

1. 实验概览

本实验旨在进一步理解索引中“域”的概念，实现浏览器中的组合搜索（“site:”）；理解Luence索引的更新和删除；理解图片搜索的过程，基于Luence库和jieba中文分词库，实现组合搜索和图片索引。

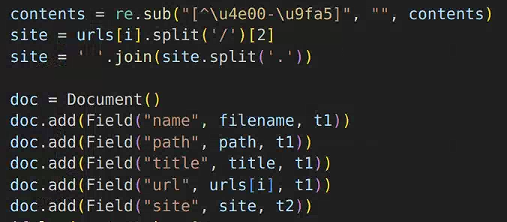
1. 实验环境

Docker: sjtumic/ee208

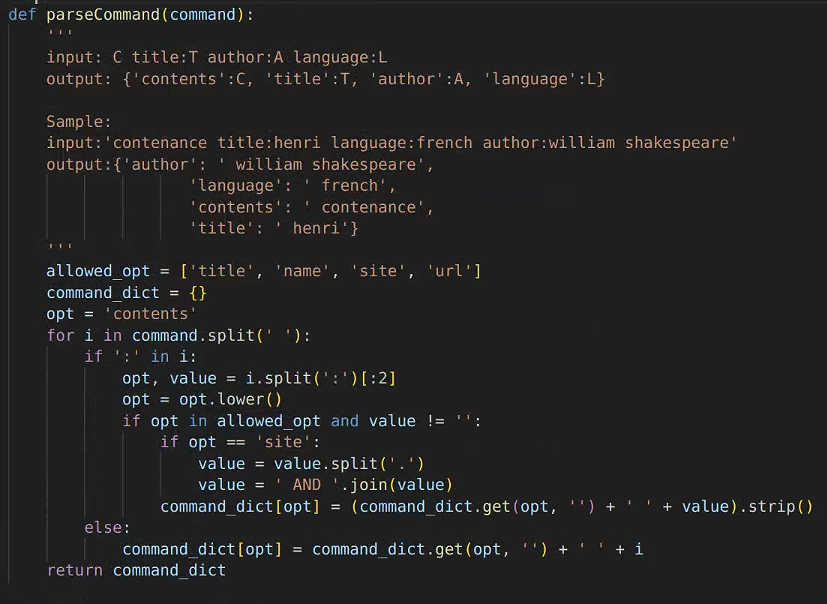
1. 解决思路

组合搜索：

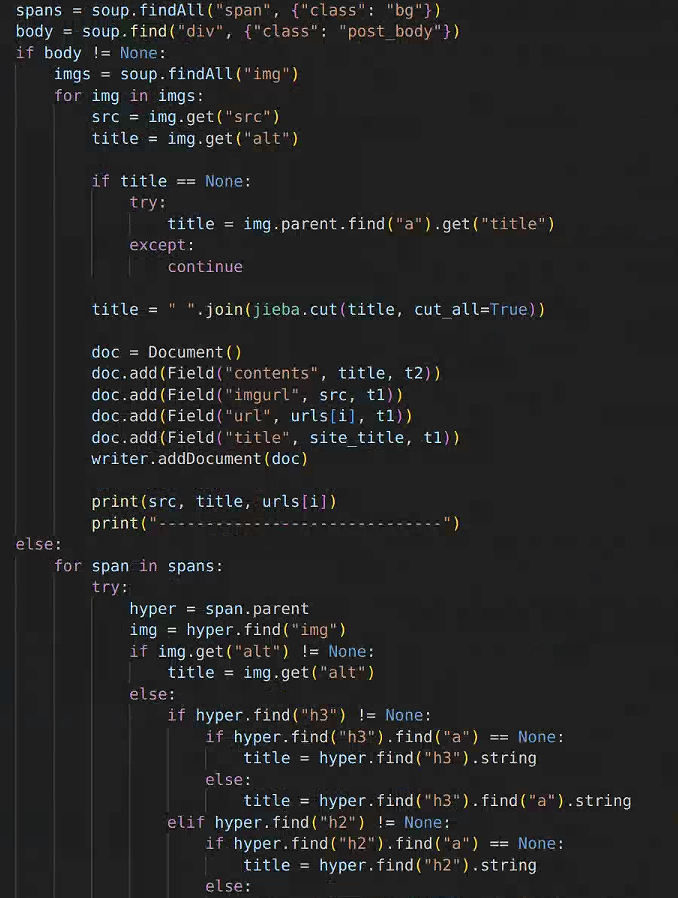
索引建立过程整体类似于lab4，添加site域提取网站的域名

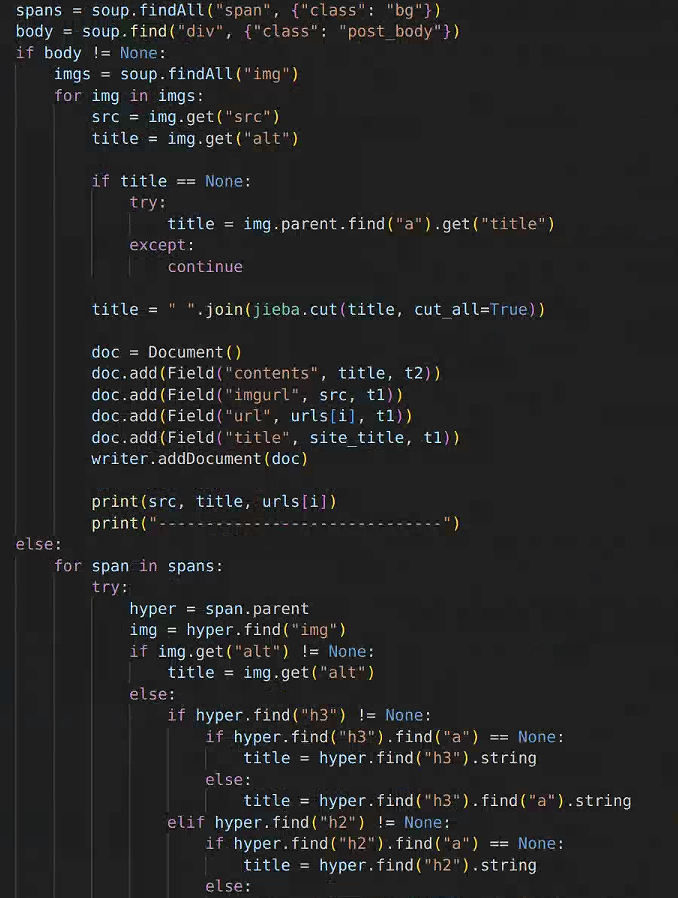


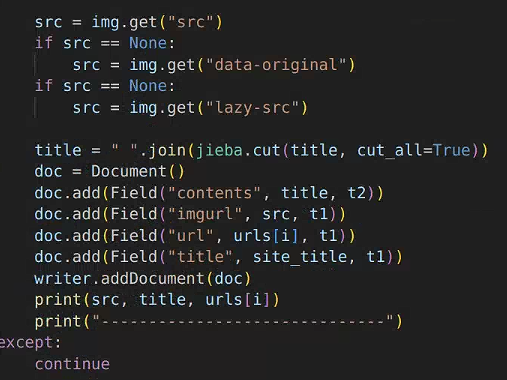
查找的过程中加入代码按照格式分析输入命令



图片索引：

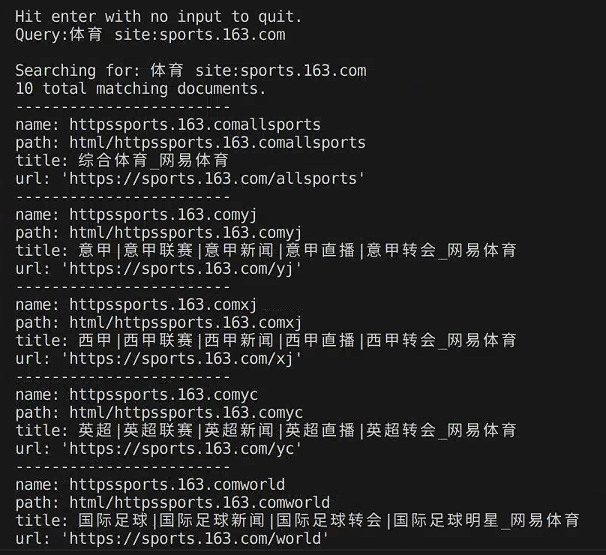
实验以网易新闻作为根节点，根据对于网页的分析，得到图片和对应标题的结构；基于bs4中的parent、next\_sibling等针对树结构的分析，得到图片和对应的文字描述。基于这些信息创建索引，实现搜索





1. 代码运行结果

组合搜索：



图片索引：

