山东大学 软件 学院

信息检索 课程实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学号：202000300125 | 姓名：贾星宇 | | 班级：2020级5班 |
| 实验题目：使用开源软件构建垂直搜索引擎 | | | |
| 实验学时：16 | | 实验日期：2022年11月7日 | |
| 实验目的：  收集数据  文本预处理  建立索引  查询交互界面 | | | |
| 硬件环境：  Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz | | | |
| 软件环境：  Windows10 Intellij IDEA | | | |
| 实验步骤与内容：  概要：  本实验为童话故事书籍的检索，其中收集数据利用了python爬虫的BeautifulSoup对网页进行爬取，搭建网页利用BS架构，前端使用Vue框架并利用Vuetify插件绘制界面，后端使用SpringBoot框架并利用MyBatisPlus与数据库连接，数据库使用mySQL数据库。在建立索引时采用了lucence7.4.0版本，采用SmartChineseAnalyzer中文分词器。   1. 收集数据：   本次信息检索主题定为童话故事检索，因此首先利用python的BeautifylSoup爬虫爬取相应网页数据。由于每本小说有多个章节分多个页面，因此首先分析页面网址规则，如小王子页面：    可以发现其从第一张开始url地址从6992递减。  随后进入每一章浏览其页面文本规则：    发现其内容被限制在class=“articleContent”下，因此根据此设计脚本：    最终爬取到26章的内容并存储到相应txt文件中：     1. 文本预处理：由于接收到的文本首先为HTML格式，因此首先利用get\_text方法分割文本，随后将文本编码方式转换为utf-8，方便索引。最后，将段落处添加分段标签，以方便在前端显示：      1. 建立索引。在建立索引时利用了lucence7.4.4版本，并通过SmartChineseAnalyzer分词器进行分词。首先程序通过BufferedReader读取每一本书每一章的内容，并存储到相应数组中；随后将每一章内容添加到doc中，每一章有三个属性：书本名称、章节号以及内容。最后建立索引：      1. 网页搭建   建立索引后，使用BS架构进行网页搭建。其中前端使用Vue框架并利用Vuetify插件绘制界面，后端使用SpringBoot框架并利用MyBatisPlus与数据库连接，数据库使用mySQL数据库。网页快照：    首先上方为按钮序列，表示支持的书本，点击可以进入页面浏览相应书籍  内容：    在输入框输入想要检索的内容，前端会将请求发送给后端，后端处理好后将结果以JSON方式发送给前端。前端经过处理（添加div头、添加div标识）后显示到页面中。其中div标识主要标识所显示内容属于的书本和章节，这样可以实现点击相应章节进入相应书本的相应章节，如点击上述第二条结果，会跳转到小王子第14章：    同时，每次搜索后端返回的结果亦包含搜索市场及搜索内容数量，标记在搜索框右下侧： | | | |
| 结论分析与体会：  通过本次实验，我不仅学到了luence如何通过建立倒排索引实现快速查询，也学到了如何将luence与BS架构，springboot框架结合起来，实现检索网页，增长了许多知识。  但是，我的网站依然存在问题。由于仅有75万字，因此数据量较小。当数据量过大时，需要考虑分页的方法检索。  在日后的学习中，我必将继续改进，继续向前！ | | | |