大作业报告

张驰 2022010754 zhang-ch22@mails.tsinghua.edu.cn

一、爬虫

本次大作业选择便携新浪新闻科技板块滚动网页 *https://news.sina.com.cn/* 的爬虫。分为两步，首先使用 `Selenium` 爬取 *https://news.sina.com.cn/roll/#pageid=153&lid=2515&etime={`etime`}&stime={`stime`}&ctime={`stime`}&date={`datestr`}&k=&num=50&page={`page`}*，获得新闻网页链接，再根据链接爬取具体网页信息。

直接使用 `request.get()` 函数只能获得网页的静态信息。为了获得新闻的评论数和参与用户数（评论用户和给评论点赞的用户数量），需要进一步爬取 *"https://comment5.news.sina.com.cn/page/info?version=1&format=json&channel=cj&newsid=comos-{`newsid`}&group=undefined&compress=0&ie=utf-8&oe=utf-8&page=1&page_size=3&t_size=3&h_size=3&thread=1"* 和 *"https://comment5.news.sina.com.cn/page/info?version=1&format=json&channel=kj&newsid=comos-{`newsid`}&group=undefined&compress=0&ie=utf-8&oe=utf-8&page=1&page_size=3&t_size=3&h_size=3&thread=1"* 这两个链接。

本次大作业爬取了新浪新闻科技板块滚动网页上能获得的的**所有新闻，时间范围为从2014-10-16到2023-8-28，共142485条。**

爬取的所有新闻分存在以下两个文件夹中。

位置:	C:\ZhangChi-THU\1-Summer\Programming Practice
大小:	371 MB (389,134,620 字节)
占用空间:	549 MB (576,389,120 字节)
包含:	86,314 个文件，0 个文件夹

位置:	C:\ZhangChi-THU\1-Summer\Programming Practice
大小:	254 MB (266,458,969 字节)
占用空间:	368 MB (385,937,408 字节)
包含:	56,137 个文件，0 个文件夹

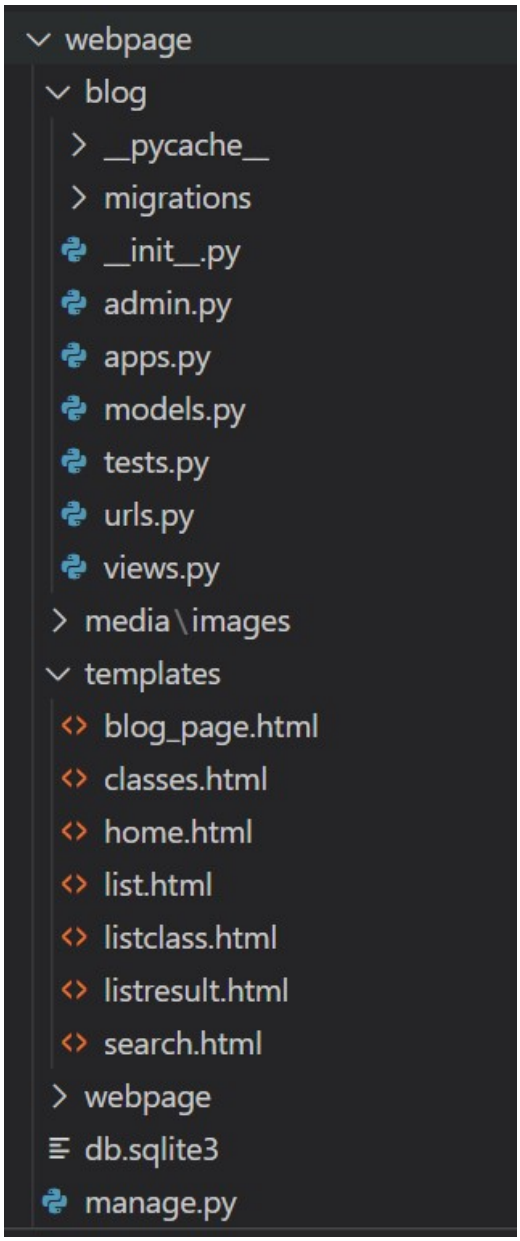
> 说明：部分爬取结果因信息不全被删除，故总数略少。

二、网页设计

1. 数据结构和算法

网页设计采用Django设计。虽然爬虫共爬取了14万余条新闻，但由于将新闻添加到sqlite数据库过于缓慢，故只往网站里添加了5000条新闻。

网页大致的文件结构如图：



为了存放网页，我建立了一个`Blog`类，里面成员变量有标题、发表日期、作者、评论数、热度（根据原始网页的评论数和参与用户数加入一定随机性计算得到）、链接等；另外建立`Comment`类存放评论、`Para`类存放正文段落，通过多对一的外键连到`Blog`类；建立`Img_para`和`Text_para`两个类，通过一对一外键连到`Para`类。这样完成数据库结构的建立。

网页用到的所有图片全都存储在本地，通过本地路径索引在网站上显示。存储的文件夹如图：

位置:	C:\ZhangChi-THU\1-Summer\Programming Practice
大小:	6.61 GB (7,102,268,080 字节)
占用空间:	6.69 GB (7,184,510,976 字节)
包含:	40,999 个文件，0 个文件夹

网站的搜索功能有三个子功能，分别为按时间、热度、匹配度排序。时间和热度排序使用精确搜索（即搜索结果必须包含完整的查询内容），这是因为如果对查询内容再做分词的话，容易查找到过多内容，按照时间/热度排序容易让不相关内容出现在较高位置。

按匹配度排序搜索使用倒排表和tf-idf算法完成。先将搜索内容进行分词，然后查找这些分词对每篇文章的tf-idf值，最后求和，按照和从大到小排列。含有tf-idf的倒排表存储在本地。

网站使用`django.core.cache`对搜索结果进行缓存，这样搜索出结果之后，翻页不需要再重新搜索。提高了搜索速度。

2. 页面设计

前端全部为手写html，未使用任何网上组件，通过`<style>`中的css样式进行美化。效果大致如图：

> 首页，当前鼠标处于顶部栏“分组”按钮处（鼠标经过按钮产生变色效果）：

> 列表页：

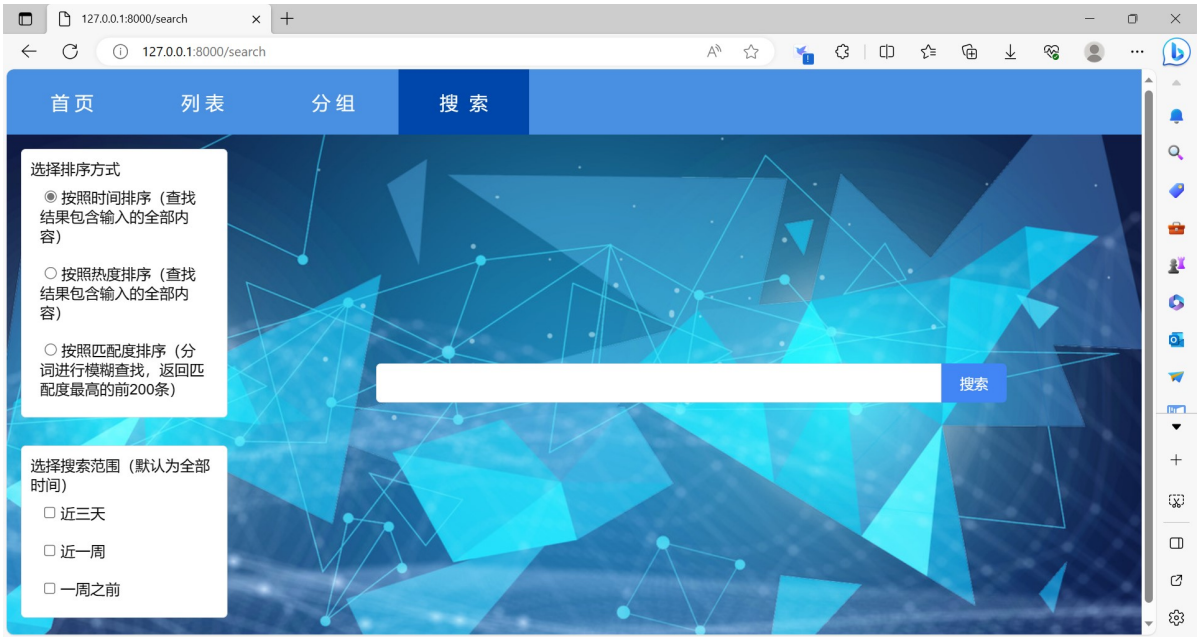
> 下方按钮也有鼠标经过变色功能。输入页面跳转自带页数限制，不合法输入无法跳转：


127.0.0.1:8000/list?page=1
127.0.0.1:8000/list?page=1
评论数：0， 热度：5
摘要：IT之家 8 月 28 日消息，腾讯 QQ 团队官方去年底发布了 QQ Linux 版 3.0.0 ……
香港小姐10号意难平"登热搜：这么好看都没进三甲"
发布时间：2023-08-28 16:19:13， 作者：快科技
评论数：0， 热度：63
摘要：2023香港小姐"前三甲诞生。15号选手庄子璇夺得冠军桂冠，来自江苏的4号选手王怡然获亚军和"最上镜……
装冷凝壁挂炉是交智商税么?
发布时间：2023-08-28 16:09:07， 作者：中国家电网
评论数：0， 热度：6
摘要： 秋天已至，冬天还会远么?若不想坐在南方阴冷的屋子里瑟瑟发抖，那么是时候考虑装一台壁挂炉了。但是且……
多人地牢大逃杀游戏《GREED IS GOOD》9 月 4 日开启招募测试，支持简中
发布时间：2023-08-28 16:07:53， 作者：IT之家
评论数：0， 热度：1
摘要：IT之家 8 月 28 日消息，由 The Fool's Studio 开发发行的地牢寻宝大逃杀游戏……
首页
1
2
3
后一页
尾页
值必须小于或等于 250。
666
跳转

分组页：


127.0.0.1:8000/classes
127.0.0.1:8000/classes
首 页
列 表
分 组
搜 索
近三天
近三天（8.28往前到8.26，共113条新闻）
近一周
近一周（不含近三天的新闻，8.25往前到8.22，共678条新闻）
一周之前
一周之前（8.21及更早，共4209条）

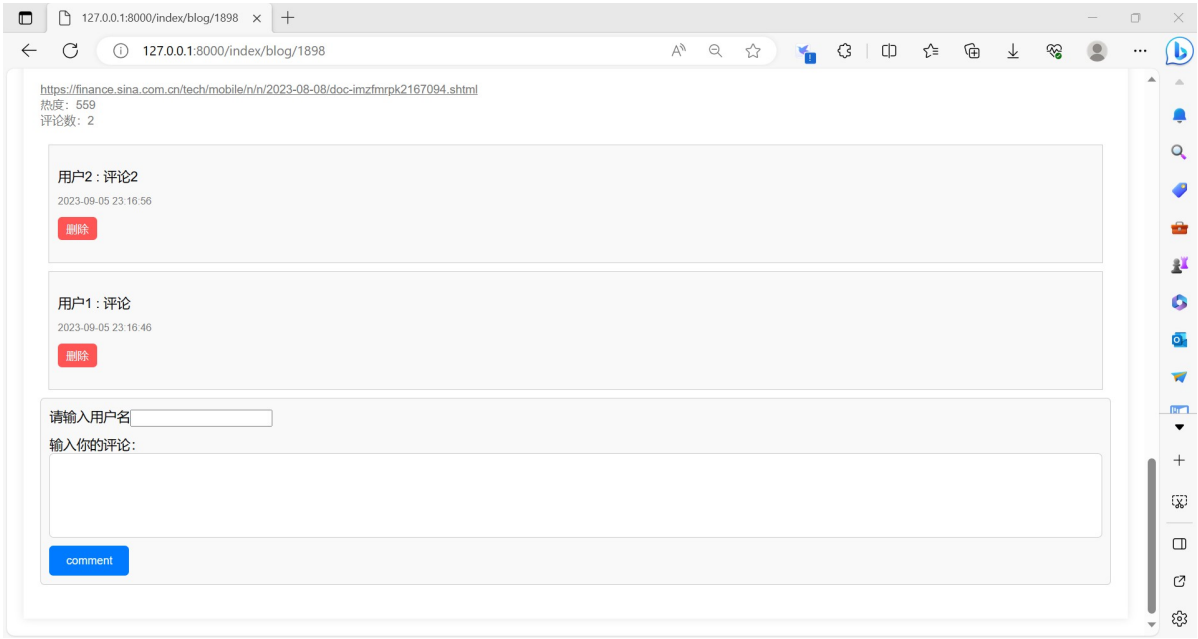
搜索页：


127.0.0.1:8000/search
127.0.0.1:8000/search
首 页
列 表
分 组
搜 索
选择排序方式
按照时间排序（查找结果包含输入的全部内容）
按照热度排序（查找结果包含输入的全部内容）
按照匹配度排序（分词进行模糊查找，返回匹配度最高的前200条）
搜索
选择搜索范围（默认为全部时间）
近三天
近一周
一周之前

搜索结果页：

新闻正文页：

新闻页评论区：



三、代码介绍

作业代码比较多，下面介绍一下几个主要文件的功能：

1. `sina_crawler.ipynb`：使用`Selenium`爬取 *https://news.sina.com.cn/roll/#pageid=153&lid=2515&etime={`etime`}&stime={`stime`}&ctime={`stime`}&date={`datestr`}&k=&num=50&page={`page`}* 中的链接的代码；
2. `crawl.py`：爬取各个链接得到新闻内容的代码；
3. `get_img.py`：通过爬到的新闻中的图片链接爬取图片的代码；
4. `order.py`：将爬到的新闻按照时间排序的代码；
5. `webpage/blog/views.py`：后端核心，负责各个网页的后端计算和显示功能；
6. `webpage/blog/models.py`：定义数据库里的各个模型；
7. `store_blog.py`：将新闻储存到`sqlite`数据库中的代码；
8. `tfidf.ipynb`：计算tfidf信息的代码。

四、感想

感觉作业很有难度，一开始毫无头绪，中间页走了很多弯路（比如学会`Selenium`后试图用它来爬正文，后来因为太慢放弃；还有所有的文件一开始都村成了`.csv`格式，这样复杂的词典（包含`list`元素）的读取就需要把输出形式的字符串转化回`list`，读取速度极慢，后来都改成了`.json`文件后快了很多）。大多数问题都在同学和GPT的帮助下解决了，也有少量没有解决的（比如如何把`tf_idf.json`文件存在内存中让每次搜索不用反复读取，如何使用Elementui库等）。虽然做的过程很累，但这个工程下来对爬虫、网页前后端都有了一个初步的了解，很有收获。