山东大学 软件 学院

Web数据管理 课程实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学号：202000300125 | 姓名：贾星宇 | | 班级：2020级5班 |
| 实验题目：可视化爬取的数据 | | | |
| 实验学时：20 | | 实验日期：2023年4月16日 | |
| 实验目的：  对实验1爬取的数据进行分析。  选择结构化数据或者文本  若实验1，2失败，可以选择现成的数据集  设计简单的web 网站  能够采用各种可视化方法展示数据分析的结果 | | | |
| 硬件环境：  Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz | | | |
| 软件环境：  Intellj IDEA 2022.2 | | | |
| 实验步骤与内容：   1. 网站数据来源：   本次实验设计的网站数据主要来自于大疆无人机论坛：  <https://bbs.dji.com/forum.php?mod=forumdisplay&fid=84&page=1>  网页截图如下：     1. 网站数据类型:   本网站数据使用了实验二中爬取到的大疆论坛中的相关数据，总计12万条，使用了5万条。与实验二不同的是，实验二中的数据被存储在了txt文件中，而在本次实验中，我们使用python中对mysql数据库的操作接口，将其中的5万条数据存储到了MySQL数据库中，具体的代码和存储结果如下：  from bs4 import BeautifulSoup import requests import pymysql import time  *# 打开数据库连接* db = pymysql.connect(host='localhost',  port=3306,  user='root',  password='1111',  database='webshujuguanli') *# 使用cursor()方法获取操作游标* cursor = db.cursor() bbs\_id = 0 *#128431条数据（论坛，航拍无人机板块），得到了53894条* for page in range(2,2143):*#2143  # 爬取基本网页* url = 'https://bbs.dji.com/forum.php?mod=forumdisplay&fid=84&page='+str(page)  headers = {  "user-agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/112.0.0.0 Safari/537.36 Edg/112.0.1722.64"  }  response = requests.get(url, headers=headers)  soup = BeautifulSoup(response.text, "html.parser")   *# 寻找有用价值* soup = soup.find(id='threadlisttableid').find\_all('tbody')    drone\_name = ''  link\_drone\_name = ''  bbs\_content = ''  link\_content = ''  bbs\_user\_name = ''  link\_user = ''  bbs\_comment = ''  link\_comment = ''  bbs\_publish\_time = ''    for each\_body in soup:  get\_a = each\_body.find\_all('a')  if not get\_a:  continue  index = 0  for each\_a in get\_a:  if each\_a.get\_text().strip() != '' and each\_a.get\_text().strip() != 'New':  if index==0:  drone\_name = each\_a.get\_text()  link\_drone\_name = "https://bbs.dji.com/"+str(each\_a.attrs['href'])  elif index==1:  bbs\_content = each\_a.get\_text()  link\_content = "https://bbs.dji.com/" + str(each\_a.attrs['href'])  elif index==2:  bbs\_user\_name = each\_a.get\_text()  link\_user = "https://bbs.dji.com/" + str(each\_a.attrs['href'])  elif index==3:  bbs\_comment = each\_a.get\_text()  link\_comment = "https://bbs.dji.com/" + str(each\_a.attrs['href'])  elif index==4:  bbs\_publish\_time = each\_a.get\_text()  else:  print('+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++')  index+=1  print(each\_a.get\_text() + ' ' + 'https://bbs.dji.com/' + each\_a.attrs['href'])   sql = "INSERT INTO bbs\_camera\_drone(`drone\_name`,`link\_drone\_name`,`bbs\_content`,`link\_content`,`bbs\_user\_name`," \  "`link\_user`,`bbs\_comment`,`link\_comment`,`bbs\_publish\_time`) VALUES (" + "\"" + drone\_name + "\""+\  ",\"" + link\_drone\_name + "\""+ \  ",\"" + bbs\_content + "\"" + \  ",\"" + link\_content + "\"" + \  ",\"" + bbs\_user\_name + "\"" + \  ",\"" + link\_user + "\"" + \  ",\"" + bbs\_comment + "\"" + \  ",\"" + link\_comment + "\"" + \  ",\"" + bbs\_publish\_time + "\"" + \  ");"  print(sql)  try:  *# 执行sql语句* cursor.execute(sql)  *# 提交到数据库执行* db.commit()  except:  *# 如果发生错误则回滚* db.rollback()  time.sleep(0.2)  bbs\_id+=1  print()  db.close()  存储结果：    存储结果计数：     1. 网页可视化结果： 2. 搜索结果展示   打开网页，首先向数据库请求21页内容，展示到搜索栏，上方显示了热门机型：    对于每一条数据，从上到下、从左到右分别为论坛名称、发布用户、发布时间、相关无人机机型、回复数量。每一项皆可以点击。  对于搜索栏下方的按钮，点击对应的按钮可以直接搜搜与相关机型有关的信息：    在搜索栏中输入数据支持对文章的模糊搜索，如输入“官方”：    分页逻辑如下：  当用户提供搜索需求后，前端会向后端请求数据量；前端获取数据量后计算页码数并进行页号的生成；之后请求第一页的数据并显示到主页中。当用户点击相应的页面号码后，前端进行修饰后向后端请求此页对应的数据，并显示。   1. 数据显示     下滑页面进入数据显示区域，根据5万条数据进行了相关的统计，包括：  无人机相关文章数量的统计。后端总共统计了41种无人机的文章数量，我们展示出了热门的前十名无人机，他们的相关文章数量从10173条到1414条。  用户文章数量。统计了一共13967个用户编写的文章的数量，并将前十名的用户显示到页面。   1. 开发工具   对于网站，使用了BS架构，前端使用Vue框架进行可视化展示，使用echarts以及vuetify进行组件优化；后端使用springboot框架+mybatisplus进行数据操作及传输。 | | | |
| 结论分析与体会：  通过此次实验，我对web数据爬取、反爬以及数据统计、大量数据的处理、存储及整合进行了深入地理论的学习，有了一个较为系统的认识。 | | | |