《数据科学与数据分析》实验课作业

Week 10_Exercise 4

2020 年 11 月 20 号 23:55 之前提交

一. 问题

（此处有 5 道题，请注意）

（1）实验一： 淘宝网商品信息采集

- 搜寻“华为手机”，对相关信息进行采集，由于数据大小受限，**只需要采集前 100 条记录**。
- 采集数据的属性，尽可能多。
- 查看/检索获得商品数据信息页面，记录 URL；
- 输入 URL，根据信息采集流程采集相关数据；
- 注意循环翻页、循环列表的时滞与循环终止条件、数据的屏蔽与截断。

1、数据采集

（1）工具：Python，bs4、requests、webdriver 等 py 库

（2）数据源：https://s.taobao.com

（3）步骤：

①在 search_product 函数中利用 webdriver 完成打开浏览器、获取淘宝界面、进行商品检索、获取商品页数等准备任务。

②在 get_product 函数中利用 xpath 方法获取每一页展示商品的框目节点，并进行遍历，访问其中存储商品信息的元素节点，并提取相应文本，添加至列表中。

注： 据观察，淘宝商品页面的 url 命名规律为：

“https://s.taobao.com/search?q=华为手机&imgfile=&

commend=all&ssid=s5-e&search_type=item&sourceId=tb.index&s={0}&ie=utf8”**.format(page*44)**

③根据 search_product 函数返回的商品总页数遍历执行 get_product，最后生成总的 DataFrame，其中存储了检索所得前 100 页的所有“华为手机”商品信息，并导出为本地 csv 文件。

2、数据分析

最后获得有关商品的：商品名（name）、商品价格（price），商品成交量（deal），商家店铺（shop）、店铺地址（loc）等 **5 大基本属性**，共计 **4408 行数据**。

>>> taobao.head(16)

	name	price	deal	shop	loc
0	【花呗 12 期免息】Huawei/华为畅享 20 Plus5g 手机华为官方旗舰店正品全网通 nova7 直降 mate30 荣耀 x10 新款 p40pro	2599.00 元	182 人付款	京联通达数码旗舰店	北京
1	【大电池 长续航】Huawei/华为 畅享 20 5G 手机华为官方旗舰店正品 nova7se 荣耀 30 荣耀 X10 学生 P30	1499.00 元	1229 人付款	京联通达数码旗舰店	北京
2	【百亿补贴】Huawei/华为 华为 畅享 Z 5G 全网通手机	1899.00 元	958 人付款	深港通信专卖	广东 深圳
3	【24 期 0 首付 咨询客服享优惠】Huawei/华为畅享 20 5G 手机 10Plus 畅想官方旗舰店 P40Pro 正新品官网 mate40Pro	1499.00 元	764 人付款	华为莫问专卖店	浙江 杭州
4	【限时 12 期免息】华为旗下荣耀 30 5G 手机 50 倍超稳远摄麒麟 985 芯片同款智能手机官方旗舰	2999.00 元	1.0 万+人付款	荣耀官方旗舰店	广东 深圳
5	【当天发 24 期分期】Huawei/华为 Mate 40 Pro 5G 手机 mate40pro+官方旗舰店 30e 官网正品新直降折叠屏保时捷 RS	7399.00 元	1564 人付款	华为莫问专卖店	浙江 杭州

（2）实验二： 地图信息采集

- 打开百度地图，分别在北京、上海、深圳、南京、杭州五座城市地图中输入“书店”；
- 分别采集书店名称、地址、经纬度；
- 分别统计五座城市的书店数量，并计算分布密度，进行简单统计分析（带统计图）。

1、数据采集

（1）工具：Python，bs4、requests、webdriver 等 py 库

（2）数据源：http://api.map.baidu.com/lbsapi/getpoint/index.html

（3）爬取步骤

①在 get_page 函数中利用 webdriver 完成打开浏览器、输入检索关键字(eg.上海市，书店)、检索等模拟操作，并得到检索页面。

②在 get_bookstore 函数中利用 find_elements_by_xpath 搜索存储书店信息的 div 元素块，并提取其中的书店名称、地址、经纬度等数据添加进列表。

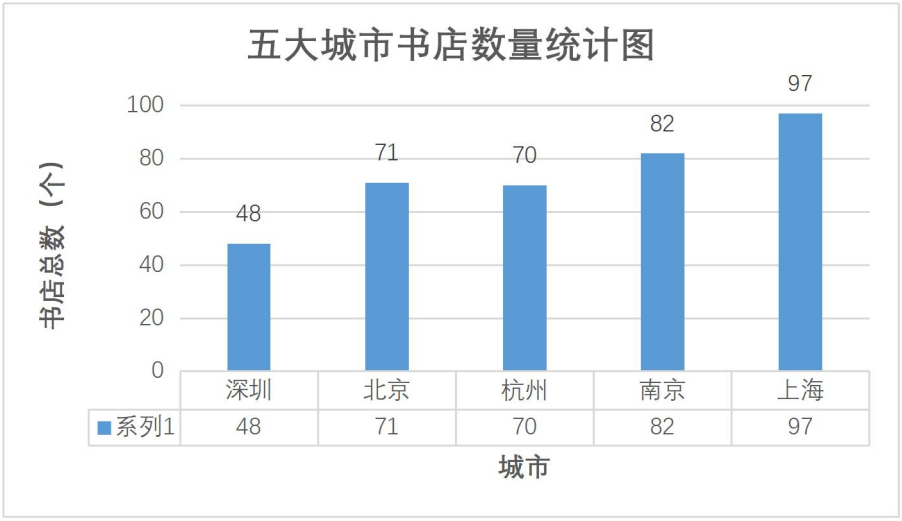
③对每一个页面进行 div 块的遍历与书店数据信息提取，最终得到总的 DataFrame，并生成本地 csv 文件。

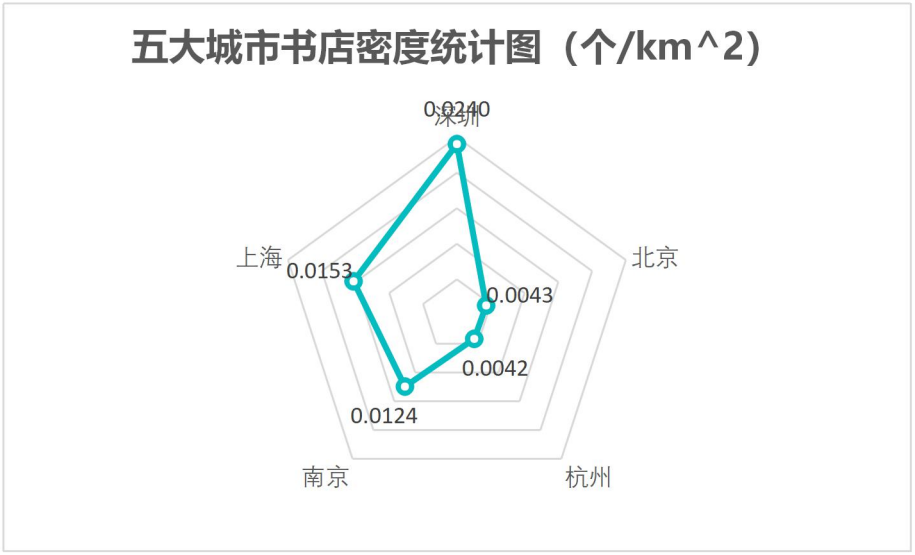
2、数据分析

得到上海、深圳、杭州、南京、北京五大城市检索所得的共 365 个书店的书店名（name）、地址（position）、经、纬度（lon、lat）4 大属性的数据。

	name	position	lon	lat
0	北京市新华书店(王府井书店)	北京市东城区王府井大街 218 号	116.418373	39.9166
1	新华书店总店(城	北京市西城区北礼士路 135 号 7 号楼一层	116.35880	39.93529

	市书房店）		6	6
2	万圣书园	成府路 59-1 号	116.332565	39.998637
3	北京市新华书店（良乡店）	北京市房山区良乡镇良乡中路 37 号	116.150671	39.737555
4	北京市雨丝书店	北京市丰台区东大街 2 号	116.30086	39.87049
5	北京市新华书店（沙河店）	北京市昌平区沙河镇工商南街 12 号	116.27671	40.132002





有统计数据与统计图表可知：

①在上海、深圳、杭州、南京、北京五大城市里，**百度地图可以检索到**书店总量上海排名第一，其次是南京，深圳最少。

②关于书店分布密度（城市书店总数除以城市面积），深圳书店

密度最大，其次是上海和南京，杭州和北京分布最为稀疏。

（3）实验三： 政府政策数据采集

- 选定“南京市”，根据城市政府门户网站的政府信息公开栏目，获取政策文件的分布网页，考虑数据大小，只采集前 1000 条；
- 通过上述网页地址，和政策记录的元数据格式，获取政策的相关记录；
- 注意政策的翻页与标题列表中的“……”。

1、数据采集

（1）工具：Python、requests 等 py 库

（2）数据源：http://www.nanjing.gov.cn/xxgkn/zcfgk/

（3）爬取步骤

①在 get_page 函数中获取政策总 page 页数，用于 url 的循环遍历；

②总结发现政策数据库文件 url 命名规律，更改 url 中的 p 值作为遍历依据

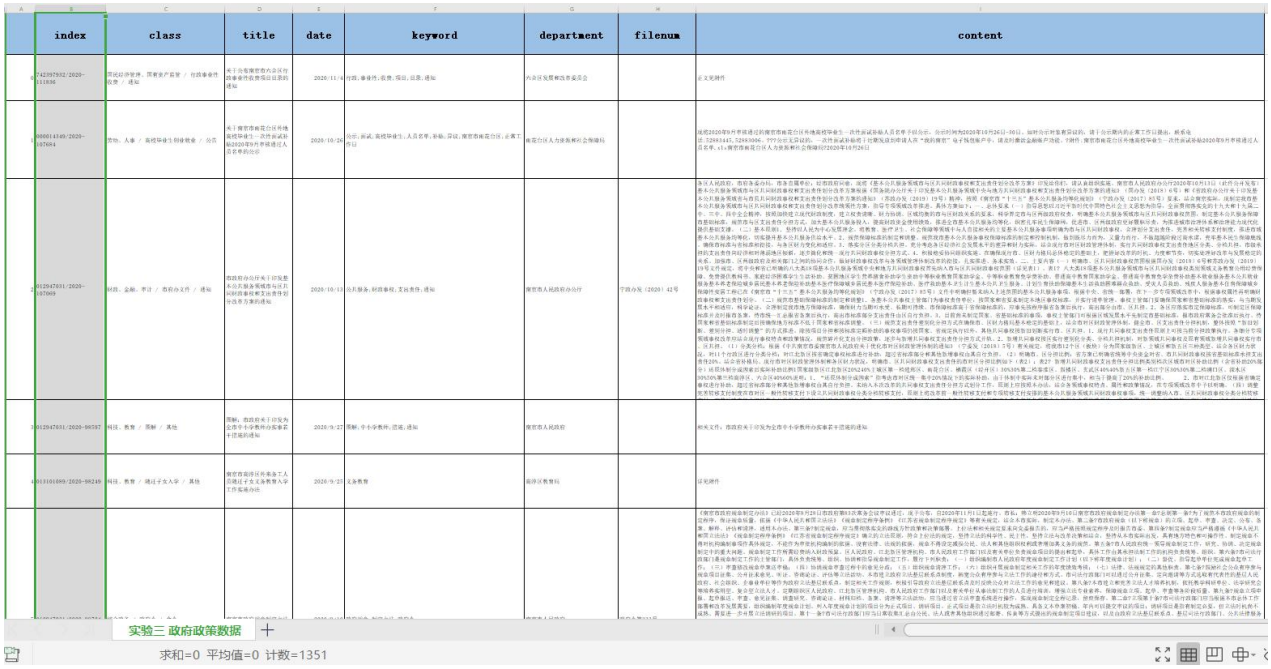
③在 get_policy 函数中，根据传入的 url，利用 requests 库模拟 get 操作向对应服务器发送请求，返回 response 也就是 html 的源代码，解码后利用正则表达式/美丽汤获取政策元数据信息。

④循环调用 get_policy 函数，得到总的 DataFrame，存储为本地 CSV 文件

2、数据分析

最终得到南京市人民政府共计发布的 **1350 余份**政策文件信息，

每行记录包括政策索引号、发文机构、发布部门、发布时间、关键词、政策分类、政策概述、政策内容共计 **8 个属性值**。



	index	class	title	date	keyword	department	filenum	content

实验三 政府政策数据

求和=0 平均值=0 计数=1351

（4）实验四： 新浪微博数据采集

- 对“外交部发言人办公室”官方微博进行数据采集；
- 搜寻和获取官方微博地址，注意观察 URL 命名规则；
- 采用一定的策略完成官方微博中的标题、发布时间、好友列表等数据的获取。

1、数据采集

(1) 工具：python，webdriver、requests、bs4 等 py 库

(2) 数据源：

https://weibo.com/u/7099422177?is_search=0&visible=0&is_tag=0&profile_ftype=1&page=10&is_all=1

(3) 爬取步骤

①在 get_page 函数中利用 webdriver 完成打开浏览器、输入检索

关键字(外交部发言人办公室)、检索等模拟操作，并得到检索页面。

②在 get_bookstore 函数中利用 find_elements_by_xpath 搜索存储外交部发言人办公室微博博文信息的 div 元素块，并提取其中的博文相关数据添加进列表。

③对每一个页面进行 div 块的遍历与博文信息提取，最终得到总的 DataFrame，并生成本地 csv 文件。

2、数据分析

最终得到外交部发言人办公室近半年发布的 **216 篇**微博博文的数据信息，每条记录包含发文用户（user）、发文时间（time）、来源（src）、博文内容（content）、转发量（forward）、点赞数（like）、评论数（comment）共 **7 个属性值**。

	user	time	src	content	forward	like	comment
0	外交部发言人办公室	2020/11/14 10:49	微博 weibo.com	【双语】例行记者会/Regular Press Conference (2020-11-13) #跟我看外交# #与发言人相约蓝厅# ° 【双语】例行记者会/Regular Press Conferenc...	16	87	15
1	外交部发言人办公室	2020/11/13 21:07	微博视频	【汪文斌：#中非合作论坛成为引领中非合作的旗帜，践行多边主义的楷模，坚持互利共赢的典范#】“外交部发言人办公室”消息，在 11 月 13 日外交部例行记者会上，有记者问：昨天，中方举办了中非合作论坛成立 20 周年纪念招待会。中方如何评价 20 年来中非合作论坛对促进中非友好与合作发挥的作用？当前形势下， ...展开全文 c	22	248	13
2	外交部发言人办公室	2020/11/13 18:58	微博视频	【汪文斌：#任何损害中国核心利益、干涉中国内政的行径都会遭到中方坚决回击#】“外交部发言人办公室”消息，在 11 月 13 日外交部例行记者会上，有记者问：据报道，美国务卿蓬佩奥 11 月 12 日接受采访时称，台湾不是中国的一部分。中方对此有何评论？ 汪文斌：世界上只有一个中国，台湾是中国领土不可分 ...展开全文 c	57	402	47

3	外交部发言人办公室	2020/11/13 18:47	微博 weibo.com	#中国—东盟（10+1）领导人会议取得丰硕成果# #跟我看外交# #与发言人相约蓝厅# ° 中国—东盟（10+1）领导人会议取得丰硕成果	12	184	9
4	外交部发言人办公室	2020/11/13 17:46	微博视频	【汪文斌：#难道英国、澳大利亚指望港人不认同“一国”，不效忠国家和特区吗？#】“外交部发言人办公室”消息，在 11 月 13 日外交部例行记者会上，有记者问：关于中国全国人大常委会通过关于香港特别行政区立法会议员资格问题的决定，英国昨天称这一决定违反了《中英联合声明》。中方有何回应？ 汪文 ...展开全文 c	54	467	63
5	外交部发言人办公室	2020/11/13 10:31	微博 weibo.com	【双语】例行记者会/Regular Press Conference（2020-11-12） #跟我看外交# #与发言人相约蓝厅# °【双语】例行记者会/Regular Press Conferenc...	26	134	20
6	外交部发言人办公室	2020/11/12 19:00	微博视频	【汪文斌：#巴西监管机构允许科兴公司恢复新冠疫苗临床试验#】#跟我看外交##与发言人相约蓝厅# @人民视频 L 外交部发言人办公室的微博视频	19	238	22

（5）实验五： 实验总结

- **爬虫首次运行过程中的主要问题，以及你的解决方法？**

Q1：服务器有反爬机制，操作频繁会封杀 IP 拒绝访问

A1：

①更改 USER-AGENT，伪装成真人爬取；

②设置页面爬取休眠时间，避免 get 请求过于频繁；

③启用代理 IP 爬取信息。

Q2：部分网址（例如百度地图 API）的数据获取会因网络不畅而出现异常，导致数据缺失

A2：

①利用 Python “try-except”语句发现并处理异常网页；

②使用 webdriver 库机器模拟操作的同时加入真人监督，进行及时的异常处理；

③利用 python 的伪多线程机制爬取。

Q3：有些网页会出现解码错误问题

A3：

查询网页的源代码，搜索”charset”属性值，即网页对应的编码，并相应地进行解码

…

- **当自己无法解决时，如何获取相关的帮助信息？**

①在 GOOGLE 中搜索相关文档

②在 CSDN、B 站等平台搜索相关爬虫教程

③多尝试爬取不同的网站，发掘其中规律

- **其它工具的使用。**

①pyregex 在线正则表达式工具 http://www.pyregex.com/

②Webdriver、bs4 等 python 库

…

【附：部分 Python 代码展示（以淘宝为例）】

```
if __name__ == "__main__":
    # 进入浏览器设置
    keyword = input("输入你要搜索的关键词：")
    options = webdriver.ChromeOptions()
    # 更换头部
    options.add_argument('user-agent=Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/84.0.4147.105 Safari/537.36')
    driver = webdriver.Chrome(executable_path='C:\Program Files (x86)\Google\Chrome\Application\chromedriver.exe',options=options)
    # driver = webdriver.Chrome(executable_path='F:\Anaconda3\chromedriver.exe') #???
    driver.get("https://www.taobao.com/")
    main()
```

```
    mydata=pd.DataFrame({'name':namelst,'price':pricelst,'deal':deallst,'shop':shoplst,'loc':loclst})
    mydata.to_csv('E:\\Desktop\\data2.csv')

urllib3.contrib.pyopenssl.inject_into_urllib3()
headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/70.0.3538.25 Safari/537.36 Core/1.70.3766.400 QQBrowser/10.6.4163.400',
    'Connection': 'close'
}   #注意格式为字典，不能有空格

def search_product(key):

    #向搜索框传入关键字
    driver.find_element_by_id('q').send_keys(key)
    #点击搜索框的搜索按钮
    driver.find_element_by_class_name('btn-search').click()
    #扫码登录
    driver.maximize_window()
    time.sleep(25)

    page = driver.find_element_by_xpath('//*[@id="mainsrp-pager"]/div/div/div/div[1]').text   #提取页数
    page = re.findall('(\d+)',page)[0]

    return int(page)

urllst=[]
def get_product():
    """解析网页，得到想要收集的数据"""
    #第一层数据解析 → 找到 item 块标签，并返回可迭代对象
    driver.find_element_by_xpath('//*[@id="tabFilterMall"]').click()
    divs = driver.find_elements_by_xpath('//div[@class="items"]/div[@class="item J_MouserOnverReq
"]')


    for div in divs:
        #基本信息
        try:
            name = div.find_element_by_xpath('.//div[@class="row row-2 title"]/a').text #ITEM NAME
            price = div.find_element_by_xpath('.//div[@class="price g_price
g_price-highlight"]/strong').text+"元" #ITEM PRICE
            deal = div.find_element_by_xpath('.//div[@class="deal-cnt"]').text #ITEM DEAL
            shop = div.find_element_by_xpath('.//div[@class="shop"]/a').text    #ITEM SHOP
            loc    = div.find_element_by_xpath('.//div[@class="location"]').text       #ITEM LOCATION
```

```
            url = div.find_element_by_xpath('.//a[@class="J_ClickStat"]').get_attribute('href')
            print('*****hello!*****')
            print(url)

            urllst.append(url)
            namelst.append(name)
            pricelst.append(price)
            deallst.append(deal)
            shoplst.append(shop)
            loclst.append(loc)
            print(name,price,deal,shop,loc,sep=' | ',end='\n')

            response = requests.get(url, headers=headers).content.decode('utf-8')
            soup = bs4.BeautifulSoup(response)
            index = soup.find_all(name='td', align="left", class_='c2')[0].text.strip()

        except Exception as e:
            #异常通报
            print(e)
            pass
```

二. 作业要求

（1）题目类型

总共 5 道大题，前 4 道是操作题，最后 1 道是实践总结。做题之前先学会使用爬虫软件，爬虫工具不限于“八爪鱼”。

（2）提交要点

- 所有爬取结果分别存放在 Excel 中，每一个 Excel 文件命名为“**实验 X 数据**”；
- **建立 Word 文件，按顺序回答 5 个题，第 1、3、4 主要是操作，需要在文档中阐明操作简介，并对爬取的数据进行介绍（采用列表或者截图形式，并结合文字叙述的方式呈现）；第 2 题有部分分析，需要先阐明操作和数据介绍，再进行分析；第 5 题，是纯论述，字数不限，内容不限。**
- 完成后另存为PDF 文件，上传到教学立方中（注意文件命名与截止时间）。