WEB 爬虫基础

Zhenghui Wang *

International School, Beijing University of Posts & Telecommunications

2020 年 10 月 25 日

1 动因与概述

近段时间在学习一些 web 前后端的基础知识，其中一个核心项目是制作一个中国疫情实时数据页面。这个页面要求获取最新的疫情数据，因此只能通过爬虫技术来实时获取。同时，我的与深度学习相关的毕业设计也需要一些数据，这类数据虽然没有对实时性的要求，但所需的数据量非常大，如果不利用爬虫将浪费大量时间成本。

综合了以上的两个需求，我考虑稍系统一些地学习爬虫。我理解的爬虫，即一种在整个互联网络范围内搜集信息的技术。利用爬虫技术可以快速、实时地获取目标数据。这些数据可以被广泛运用在数据分析、数据挖掘、深度学习等领域当中。

整个学习过程中，我尝试爬取了：安徽省卫建委的每日疫情数据，百度疫情实时地图，英雄联盟官方网站所有英雄的全部皮肤、所有英雄的属性、技能数据，新浪微博热搜数据、Nike 官网、Adidas 官网、Lining 官网鞋类图片以及 YOHO 有货官网的鞋类图片。除了 Nike 官网的图片没能获取外（Nike 官网的图片链接应该是经过 base64 的加密了，解密失败后放弃），其它的数据都成功获取并保存到本地。

环境配置：python：3.8；requests：2.24.0；beautifulsoup4：4.9.3

其他需求：Google Chrome 浏览器以及内置的开发者工具

你可以从这里下载项目源码。

2 实验过程

2.1 主要流程

2.1.1 考虑需求、定位 html 位置

在正式开始项目之前，先考虑对数据的具体需求。例如：我希望获得疫情期间（持续更新）安徽省每一天的总确诊病例、新增确诊病例、新增疑似病例、死亡病例的数量。

紧接着考虑数据的可能来源。例如：我可以通过访问安徽省卫建委官网的相关内容得到我想要的信息，我也可以通过访问百度实时疫情地图来找到安徽省的相关数据。以下就是我调取的安徽省卫建委的疫情数据页面：

*Corresponding author: wangzhenghui@bupt.edu.cn

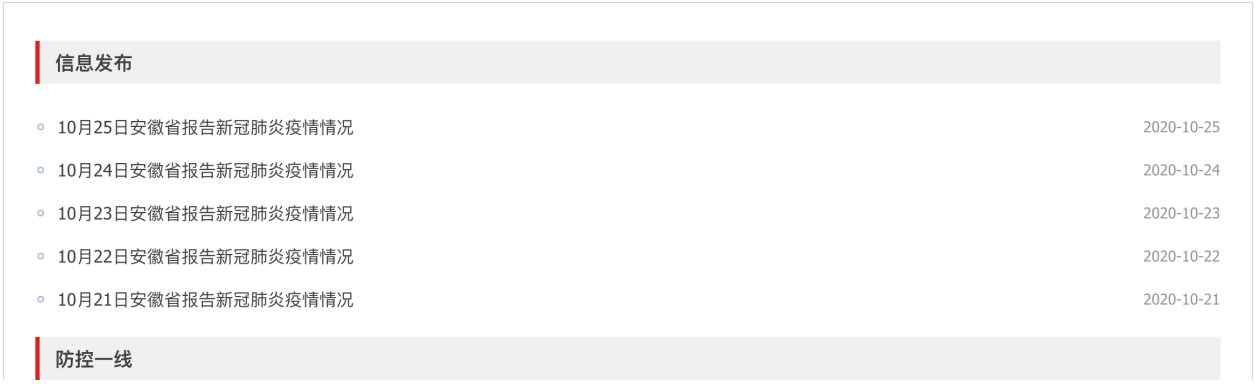

图 1: 安徽省卫建委官方网站

2.1.2 利用 python 相关库文件获取并解析 html 文本

确定了数据所在位置后，利用 python 扩展库解析 html 以及搜索目的数据位置。其中，requests 库是用于发送 HTTP 请求（模拟人访问网页的行为），服务器端收到请求后将 html 源代码传回，转为字符串后准备进一步处理。接着用 beautifulsoup 库对提取到的字符串类型的 html 文件执行查找操作得到所需数据（如果对正则表达式熟悉，也可以用 python 内置的正则化表达式库 re 搜索数据，效果可能更好~）。给出一段简单的实现代码：

```
import requests
from bs4 import BeautifulSoup
url = 'https://www.yohobuy.com/product/51592386.html'
res = requests.get(url)  # 通过requests获取html数据
soup = BeautifulSoup(res.text, features='html.parser')  # 解析获取的html数据
for item in soup.find_all('a'):  # 在soup中找到html中所有的<a>标签
    print(item)
```

2.1.3 通过浏览器开发者工具抓包获取数据

实际上，不是每一个网页都能这么轻松地获取其中的内容。很多时候，我们查看网页源代码的时候会发现其中“省略”了很多内容，这一般是由于网页中的内容是通过向外部发送请求动态获取到的数据，因此这部分的数据无法直接从静态的 html 文件中获得。典型案例：LOL 官方网站的英雄数据网页的源代码

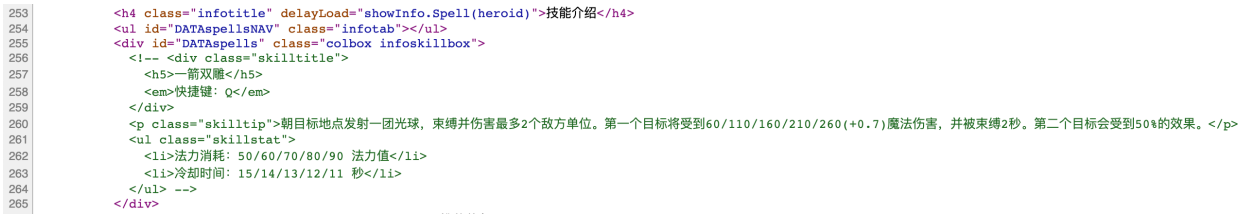

图 2: 英雄联盟官方网站英雄资料页面源码

上面的代码中，在应该显示英雄技能的位置上，空空如也啥都没有（只有拉克丝的 q 技能而且居然还被注释掉了）（实际上这里是因为加载了 showInfo.Spell() 这个函数动态地获取了数据）。这种情况下，我们可以用另外的办法来获取数据。如果 html 页面中没有所要的数据，那 html 页面必然与外部建立了链接，从而获取了外部数据，因此只要我们追踪到这个链接，就能获得数据了（抓包）。这个过程需要借助 GoogleChrome 内置的开发者工具完成。通过开发者工具我追踪到了在加载英雄列表界面时的一个外部链接：hero_list.js

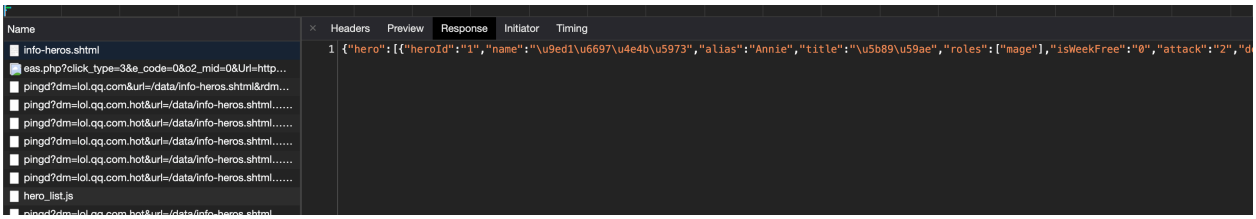

图 3: 找到的带有英雄信息的 js 文件

2.1.4 灵活使用 HTTP 头部作请求

所谓道高一尺，魔高一丈。现在的很多成熟网站都会有特别的“反爬虫”手段以阻止任何非法请求。出现得比较频繁的是重复访问某一页面后，会有要求输入验证码的操作（恍然大悟）。以 Adidas 官网为例，如果你要遍历访问每一款 Adidas 产品，那么当你访问次数超过某一阈值，网页将会禁止你继续访问（把你视为恶意爬虫）。这一问题可以通过修改头部信息（user-agent）得到暂时解决。除此以外还会有更多其他手段识别爬虫行为并进行封锁。

如果你需要以某一特定用户的身份访问网站以获取信息，比如：我需要获取我的 leetcode 账号中所有 ac 的题目。那么也需要在头部放入相关的内容（cookie）。

2.2 遇到的问题

在实际操作过程中遇到过非常多问题，多数在上面已被提及，这里列出一个尚未解决的问题在尝试获取 Nike 官网鞋类图片过程中，遇到过一种非常奇怪的现象：我在观察网页源代码时，图片的链接是正常的以“http”开头的格式，可当我用 requests 请求后，图片链接居然变了：

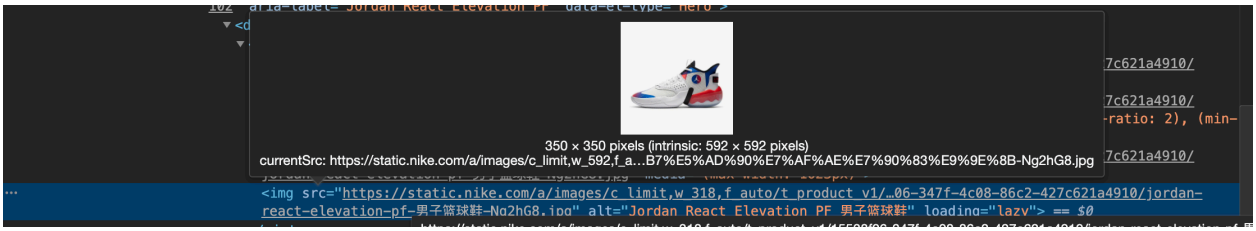

图 4: 正常访问得到的 html 文件中的路径

图 5: 经 requests 请求后得到的路径

3 项目内容展示

Beijing

	date	confirm	recover	death	newConfirm
0	2020-01-26	72	2	0	4
1	2020-01-27	91	2	1	19
2	2020-01-28	102	4	1	11
3	2020-01-29	111	4	1	12
4	2020-01-30	121	5	1	18
5	2020-01-31	139	5	1	24
6	2020-02-01	183	9	1	27
7	2020-02-02	212	12	1	32
8	2020-02-03	228	23	1	16
9	2020-02-04	253	24	1	25
10	2020-02-05	274	31	1	21
11	2020-02-06	297	33	1	23
12	2020-02-07	315	34	2	18
13	2020-02-08	326	37	2	11
14	2020-02-09	337	44	2	11
15	2020-02-10	342	48	3	5
16	2020-02-11	352	56	3	10
17	2020-02-12	366	68	3	14
18	2020-02-13	372	79	3	6

sina_hot

	hot_topic	hot_point	hot_label
0	镜头中的脱贫故事		热
1	白冰	3182835	热
2	取消申请驾驶证70周岁年龄上限	3037620	
3	王一博丝缎衬衫大片	3036894	荐
4	佟丽娅被陈奕迅cue到开心一整天	3036168	新
5	杨迪期望妈妈早日单飞	1419684	新
6	羡慕罗永浩周鸿祎的友情	1409435	荐
7	印度一ICU病房爬满老鼠	1346475	
8	正确的喝奶茶技术	906450	新
9	首批iPhone12发货	518595	
10	奥巴马炮轰特朗普	465533	
11	江苏大学杨凯	453262	
12	付定金与付尾款时的区别	441981	沸
13	巴西总统不同意买中国疫苗	378998	
14	老人组团龟蛇爬行走红	350140	
15	爱是跟世界沟通的方式	349941	
16	曹骏加上了尔冬升微信	340186	热
17	四川北川4.7级地震	324073	
18	蒙古向中国捐赠羊交接仪式举行	301759	

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
2														
3		1	2	狂战士	Olaf	['fighter', 'ta	狂战之怒: 奥	逆流投掷: 奥拉夫将战斧投至目标区域，对所有被战斧穿过的敌人造成伤害并减速。如果奥拉夫捡起斧头，那么该技能冷却时间就会减少4.5秒。奥拉夫朝目标区域扔出一柄战斧，对战斧穿过的敌人造成80/125/170/215/260(+1*额外AD)物理伤害，并减少他们29/33/37/41/45%的速度，最多持续2.5/2.5/2.5/2.5/2.5秒。斧子飞行的距离越远，减速效果的持续时间越长，但持续时间不会短于1.5/1.5/1.5/1.5/1.5秒。 如果奥拉夫将斧头捡起，那么该技能的冷却将减少4.5秒。 距离：1000	残暴打击: 奥拉夫的攻击速度得到提升，并获得生命偷取，且他损失的生命值越多，所受的治疗效果也越多。在6秒的持续时间内，奥拉夫获得14/16/18/20/22%（4月26日6.8版本更新，此前为9/12/15/18/21%）生命偷取，并且攻击速度会提升55/65/75/85/95%。 在技能持续期间，奥拉夫每损失2/2/2/2/2%生命值，就会获得1%的治疗效果加成。	鲁莽挥击: 奥拉夫以破釜沉舟之势发动进攻，对目标造成真实伤害（不受护甲与魔抗减免），同时，自身也会根据敌方所受的伤害而受到真实伤害的反噬。奥拉夫狂野地挥舞他的双斧，对他的目标造成70/115/160/205/250(+0.5*总AD)真实伤害。这个技能的消耗相当于此技能所造成伤害的30/30/30/30/30%，但如果此技能将目标击杀，则会返还所有的施法消耗。 普通攻击会减少鲁莽挥击1秒冷却时间。 距离：325	诸神黄昏: 奥拉夫暂时免疫控制技能。被动：奥拉夫获得20/30/40护甲和20/30/40魔法抗性。 主动：奥拉夫移除身上的所有控制效果，并在6/6/6秒的持续时间里免疫任何限制技能。奥拉夫还会在跑向敌方英雄时获得20/45/70%移动速度加成，持续1秒。技能持续期间里，奥拉夫会损失此技能被动部分的抗性加成，并获得15/20/25攻击力。 【新增】：在持续时间内，奥拉夫现在还会获得30%总攻击力的额外攻击力。				
		2	3	正义巨像	Galio	['tank', 'mag	巨像重击: 每	战争罡风: 加里奥发射两道罡风，罡风在汇聚后会形成一团大型龙卷风，造成持续伤害。加里奥发射两道罡风，造成70/105/140/175/210(+0.75 法术强度)	杜朗护盾: 加里奥在防御姿态下蓄力，同时移动速度减慢。在蓄力得以释放时，加里奥将嘲讽并伤害附近的敌人。被动：如果加里奥在12秒内没有收到伤害，那么他就会获得一层护盾，可吸收（+0.8*最大生命值）魔法伤害。 首次施放：加里奥开始蓄力，并获得20/25/30/35/40%(每100点额外魔抗+8%)(+每100点法术强度+5%)伤	正义冲拳: 加里奥将暂时后跳然后冲锋，击飞他遇到的第一个敌方英雄。加里奥挟着一股强劲向前猛冲，对敌人造成 80/115/150/185/220（+0.9*AP）魔法伤害并将他们击飞0.75秒。加里奥将在命中一名敌方英雄或地形时停下。	英雄登场: 加里奥将一名友军的当前位置作为他的着陆点，并为区域内的所有友军提供一个魔法护盾。在短暂的延迟后，加里奥会落到该位置上，然后击退附近的敌人。加里奥选定一个友方英雄的当前位置作为他的着陆位置。				

图 6: 英雄联盟英雄数据统计

图 7: 英雄联盟 ezreal 皮肤

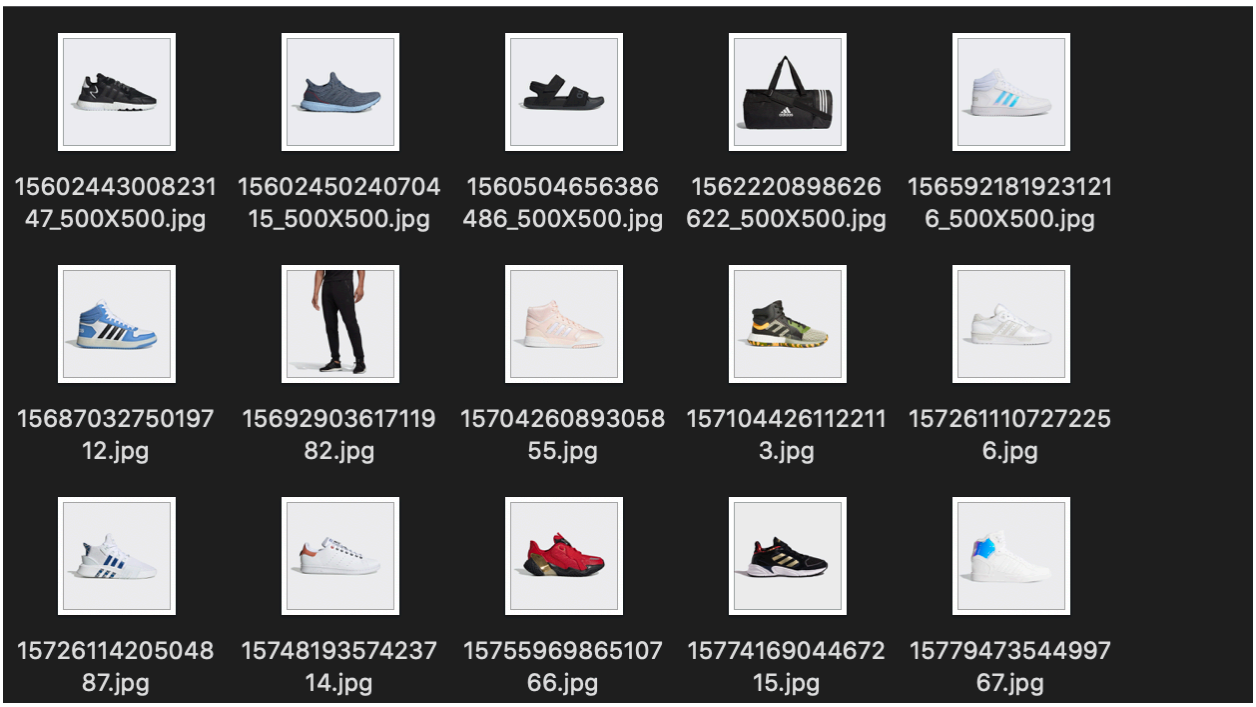

图 8: adidas 鞋类图片