生活到处充满激情

我在普通的一天

上着普通的网

不普通的兄弟发来了看似普通的网站

<https://www.mzitu.com/>

打开后



嗯？我这个马赛克打的



突然有了不普通的心思

这么多妹子

Emmm

不爬出来怪可惜了

开整！

熟练的打开pycharm

开始分析网站



可以发现页面的所有妹子都放进了li标签

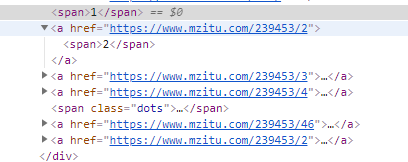
我们点击span标签里的网站

跳转进这个妹子的图库



有46张图

查看源码

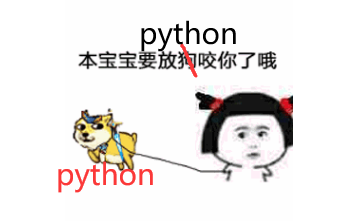


如我所愿

<https://www.mzitu.com/239453/2>

最后面的数字就是第几张图

OK关门 放Python



先弄出主页登录和分析的代码

*def* denglu():  
 url="https://www.mzitu.com/"  
 date={  
 'Host': 'max-l.mediav.com',  
 'cookie': 'Hm\_lvt\_cb7f29be3c304cd3bb0c65a4faa96c30=1593671101,1593672512; Hm\_lpvt\_cb7f29be3c304cd3bb0c65a4faa96c30=1593673214',  
 'user - agent': 'Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.1.6) Gecko/20091201 Firefox/3.5.6',  
 }  
 html=requests.post(url,headers=date)  
 print(html.text)  
 *return* html.text  
*def* fenxi(a):  
 html\_lxlm=BeautifulSoup(html,'lxml')  
 html\_ul=html\_lxlm.find('ul',id='pins')  
 html\_li=html\_ul.find\_all('li')  
 *for* tag *in* html\_li:  
 tag\_name=tag.find('span').a  
 print(tag\_name.get\_text())  
 print(tag\_name['href'])

打印出来看看



得到了想要的数据

接下来我们得爬这些链接里的照片了

上面已经分析过链接后面加个/数字

就是第几章照片

我们先获取一下每个链接里有多少张

*def* tupian(*tag\_href*,*path*):  
 html2=denglu(*tag\_href*)  
 html\_lxml2 = BeautifulSoup(html2, 'lxml')  
 html\_page=html\_lxml2.find('div',class\_='pagenavi')  
 html\_total=html\_page.find\_all('a')[-2].span.get\_text()  
 print(html\_total)  
 *for* i *in* range (1,int(html\_total)):  
 page=str(*tag\_href*+'/'+str(i))  
 page2=denglu(page)  
 page\_lxml=BeautifulSoup(page2,'lxml')  
 page\_jpg=page\_lxml.find('div',class\_='main-image').p.a.img['src']  
 print(page\_jpg)

打印出来

获取到每个图集里面的图片链接和图片数量



这个竟然有71张图片



好获取了链接我们就下载图片

*def* download(*page\_jpg*,*path*,*i*):  
 date = {

#抓包发现需要有这个Referer  
 'Referer': *page\_jpg*,  
 'User-Agent': "Mozilla/5.0 (X11; Linux i686) AppleWebKit/535.7 (KHTML, like Gecko) Ubuntu/11.04 Chromium/16.0.912.77 Chrome/16.0.912.77 Safari/535.7",  
 }  
 r=requests.get(*page\_jpg*,headers=date)  
 path=*path*+'/'+str(*i*)+'.jpg'  
 *with* open(*path*,'wb') *as* img:  
 img.write(r.content)

激动跑一下



？？我的妹子呢

肯定是服务器反爬虫发现不是浏览器

这个时候又到我放技能了



我们先加一个延时函数

time.sleep(random.randint(1,3))

打印一下还是403

那就多换几个浏览器头伪装一下

meizi\_headers=[  
 "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/39.0.2171.95 Safari/537.36",  
 "Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_9\_2) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36",  
 "Mozilla/5.0 (Windows NT 6.1; WOW64; rv:30.0) Gecko/20100101 Firefox/30.0",  
 "Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_9\_2) AppleWebKit/537.75.14 (KHTML, like Gecko) Version/7.0.3 Safari/537.75.14",  
 "Mozilla/5.0 (compatible; MSIE 10.0; Windows NT 6.2; Win64; x64; Trident/6.0)",  
 'Mozilla/5.0 (Windows; U; Windows NT 5.1; it; rv:1.8.1.11) Gecko/20071127 Firefox/2.0.0.11',  
 'Opera/9.25 (Windows NT 5.1; U; en)',  
 'Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322; .NET CLR 2.0.50727)',  
 'Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.5 (like Gecko) (Kubuntu)',  
 'Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.8.0.12) Gecko/20070731 Ubuntu/dapper-security Firefox/1.5.0.12',  
 'Lynx/2.8.5rel.1 libwww-FM/2.14 SSL-MM/1.4.1 GNUTLS/1.2.9',  
 "Mozilla/5.0 (X11; Linux i686) AppleWebKit/535.7 (KHTML, like Gecko) Ubuntu/11.04 Chromium/16.0.912.77 Chrome/16.0.912.77 Safari/535.7",  
 "Mozilla/5.0 (X11; Ubuntu; Linux i686; rv:10.0) Gecko/20100101 Firefox/10.0",  
 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/54.0.2840.99 Safari/537.36'  
]

够多了吧

我们随机调用一下

'user - agent': random.choice(meizi\_headers),

然后我们跑一下



带劲



好兄弟，还有好的资源可以分享给我

知道好兄弟想要源码

我已经放到GitHub上

https://github.com/robertshuai/test2/blob/master/%E7%88%AC%E5%9B%BE%E7%89%87.py