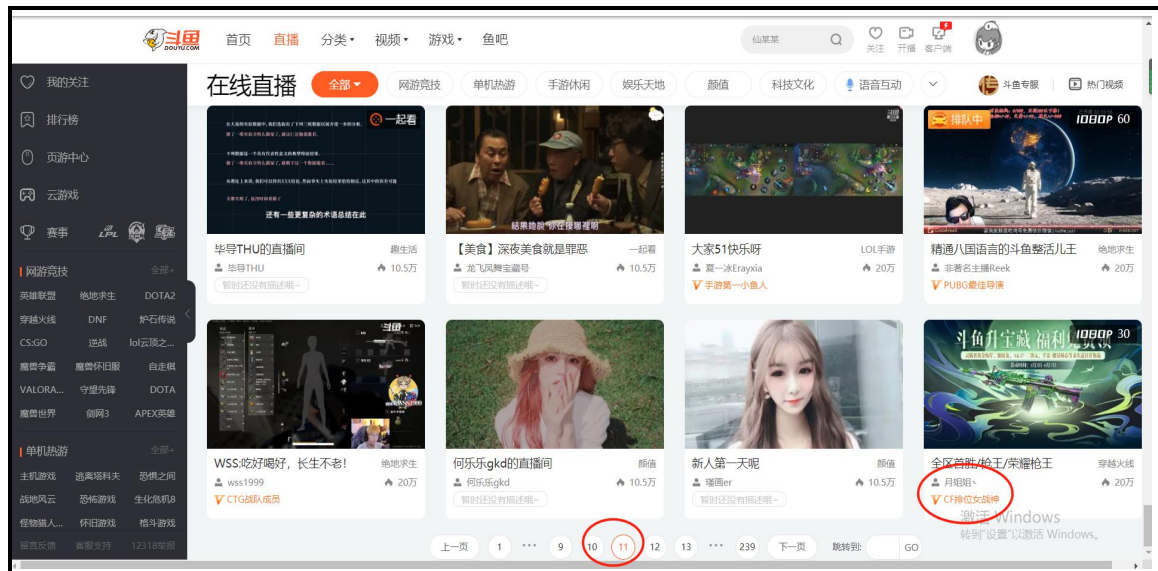


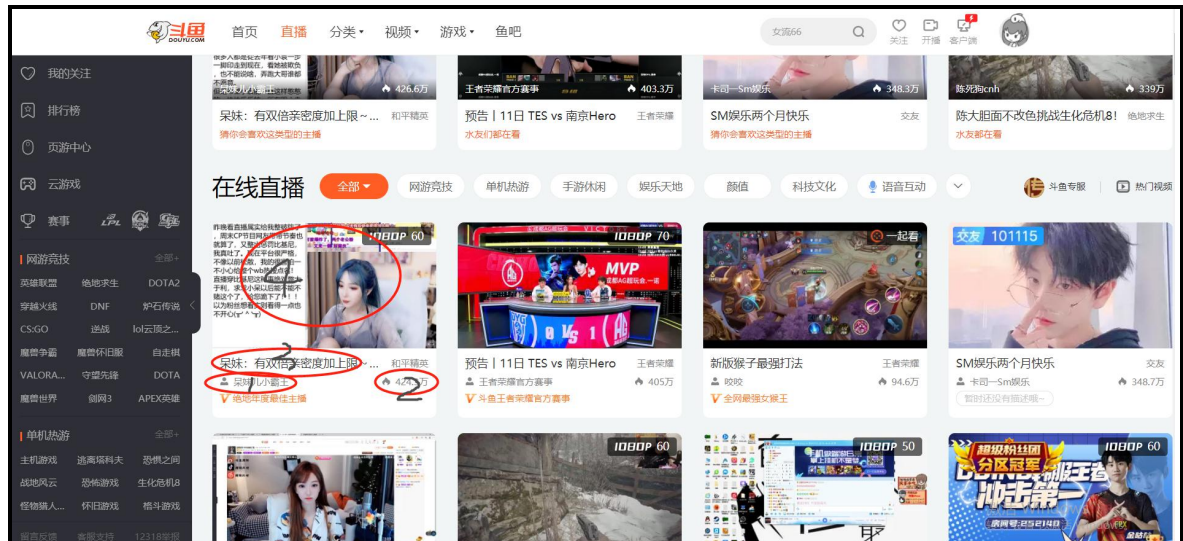
斗鱼房间信息采集

伙计们,request是不是用的多了想用新技术,今天它来了,我们使用 selenium 动态点击页面采集斗鱼直播间房间信息.结果如下,由于页面非常多,这里我们的程序就只采集了 11 个页面的数据,如下图所示.



一:需求分析

1. 目标 url: "<https://www.douyu.com/directory/all>"
2. 分析页面,打开网站.
3. 抓取信息,主播(name),人气(popularity),主题(title),图片链接(pic_link)
4. 使用技术 selenium



二.发起请求

首先我们需要定义一个类,

```
class Douyu():

    def __init__(self,url,single):

        self.url = url

        self.driver = webdriver.Chrome() # 谷歌的 webdriver

        self.driver.maximize_window() # 窗口最大化

        self.driver.get(self.url) # 请求

        self.single = single # 定义一个属性,后面用到
```

当我们实例化一共对象时,就会自动打开网页.

三.解析页面

```
def getData(self):

    time.sleep(3)
```

```

lis = self.driver.find_elements_by_xpath("//div[@class='layout-Module-container layout-Cover ListContent']/ul/li")
info_list = []
for item in lis:
    dic = {}
    name = item.find_element_by_xpath("./div/a/div[2]/div[2]/h2").text # 主播名字
    popularity = item.find_element_by_xpath("./div/a/div[2]/div[2]/span").text # 人
    title = item.find_element_by_xpath("./div/a/div[2]/div[1]/h3").text # 主题
    pic = item.find_element_by_xpath("./div/a/div[1]/div[1]/img").get_attribute('src') # 图片链接
    dic['name'] = name
    dic['popu'] = popularity
    dic['title'] = title
    dic['pic'] = pic
    print(dic)
    info_list.append(dic)
return info_list

```



解析页面是用的 xpath 语法,这里就不详细说明.

四.保存数据

```
def saveData(self, infolist):

    f = open('douyu.txt', 'a', encoding='utf-8')

    for content in infolist:

        json.dump(content, f, ensure_ascii=False, indent=2)

        f.write('\n')

    f.close()
```

因为是重复抓取,如果是 csv 文件,头部(head)不太好写入(除非只爬取一页数据),这里我们就使用 txt 文件保存数据.

五.重复抓取

因为查看每页的 url,发现一样,所以需要动态点击切换页面.

```
def run(self):

    contents = self.getData()

    self.saveData(contents)

    count = 0

    while self.single:

        self.driver.find_element_by_xpath("//li[@title='下一页']/span").click() # 点击下一页的按钮

        time.sleep(5) # 休息 5s,切换页面需要时间

        cont = self.getData()

        self.saveData(cont)

        count += 1 # 定义的 count 变量控制循环停止

        if count == 10: # 我们规定爬取 11 页数据就停止.

            self.single = False
```

六.运行程序

```
问题 输出 终端 调试控制台
Python
{'name': '大秦洛阳', 'popu': '20.3万', 'title': '【直播】狼人杀大师场 预女猎白', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '花子婆家', 'popu': '20.3万', 'title': '峡谷第一吊桶女王', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '阿若Arro', 'popu': '20.3万', 'title': '不想长大的人，却总在问向长大', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '是迷人路', 'popu': '10.7万', 'title': '让耳朵休息一下叭', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '今天想宁了吗', 'popu': '10.7万', 'title': '24k纯新主播', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': 'DNF彩粉', 'popu': '20.3万', 'title': '5.1 强化增幅龙袍龙盒 秒上号', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': 'BM\DJ调', 'popu': '20.3万', 'title': 'DJ调：街头飞机上号~~~~', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '球仔baby', 'popu': '20.3万', 'title': '星球：恐怖之夜（茶桥头）', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '桑桑子zz', 'popu': '10.6万', 'title': '是plan A 是必答题 是盛开的花', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '允恩Ye', 'popu': '10.6万', 'title': '现在开跳着跳舞8241253', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '悠然ccc', 'popu': '20.2万', 'title': '新的一周，VX区冲冲冲，有五排车位', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '津津有味大西瓜', 'popu': '20.2万', 'title': '睡觉一定不玩手机', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '泥鸽鸽J', 'popu': '20.2万', 'title': '斗鱼最细致打野教学！！', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '李思德Daisy', 'popu': '10.6万', 'title': '双倍经验搬', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '夏顾顾', 'popu': '10.6万', 'title': '明眸善睐，眉眼承欢', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': 'E3C 俊王爷', 'popu': '20.2万', 'title': '王爷：飞机上号特位一小段+精细理行阵容！', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '安雅', 'popu': '20.2万', 'title': '【安雅】最高难度魅影通关！', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '大魔王小楠', 'popu': '20.2万', 'title': '国服瑶瑶教学', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '表妹要吃肉', 'popu': '10.6万', 'title': '喜欢吃就多吃点', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '年雨琳', 'popu': '10.6万', 'title': '新出炉的主播，请多关照~', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '芥大狗理性消费', 'popu': '20.2万', 'title': '天才少年来啦\\(ㄙ)/~', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '仙童·Mr', 'popu': '20.1万', 'title': '国服瑶瑶带粉，飞机上十直', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': '鹿叫cz', 'popu': '20.1万', 'title': '【cz】帮水友定级/se上分实战教学', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': 'LC一百事可乐', 'popu': '10.5万', 'title': '本厅由咆哮哥独家冠名/招主理人', 'pic': 'https://shark2.douyucdn.cn/front-publish/sdk-file-master/room_list_default@2x_a960c5b.png'},
{'name': 'Lulu没烦恼', 'popu': '10.5万', 'title': '5月15号晚上8点周年庆，p关注链接 (Ctrl + 单击) douyucdn.cn/live-cover/appcovers/2021/04/08/8725440_20210408222254_small.jpg/webpdy1'}
{'name': '很吧的薇薇', 'popu': '20.1万', 'title': '很雨薇♥豹子潘神', 'pic': 'https://rpic.douyucdn.cn/asrpic/210510/2856169_2017.png/webpdy1'}
{'name': '、吊哥哥', 'popu': '20.1万', 'title': '吊哥：圆梦的神！', 'pic': 'https://rpic.douyucdn.cn/asrpic/210510/280730_2022.png/webpdy1'}
{'name': '主理猪六', 'popu': '20万', 'title': '村庄8 最高难度 在线直播', 'pic': 'https://rpic.douyucdn.cn/asrpic/210510/546621_2022.png/webpdy1'}
{'name': '毕哥mU', 'popu': '10.5万', 'title': '毕哥mU的直播间', 'pic': 'https://rpic.douyucdn.cn/asrpic/210510/9356436_2018.png/webpdy1'}
{'name': '龙飞凤舞宝藏号', 'popu': '10.5万', 'title': '【美食】深夜美食就是罪恶', 'pic': 'https://rpic.douyucdn.cn/asrpic/210510/9650860_2018.png/webpdy1'}
{'name': '夏一冰Eraxia', 'popu': '20万', 'title': '大家51快乐呀', 'pic': 'https://rpic.douyucdn.cn/asrpic/210510/16322_2022.png/webpdy1'}
{'name': 'wss1999', 'popu': '20万', 'title': 'wss:吃得好，长生不老！', 'pic': 'https://rpic.douyucdn.cn/asrpic/210510/6167549_2017.png/webpdy1'}
{'name': '何乐乐gkd', 'popu': '10.5万', 'title': '何乐乐gkd的直播间', 'pic': 'https://rpic.douyucdn.cn/live-cover/roomcover/2021/04/16/b2024e6cc26a240701f52a707f7971b2_big.png/webpdy1'}
{'name': '理画师', 'popu': '10.5万', 'title': '新人第一大招', 'pic': 'https://rpic.douyucdn.cn/live-cover/roomcover/2021/05/03/f4024a0162263f60d4c4e1ef5c62c0f_big.png/webpdy1'}
{'name': '月姐姐', 'popu': '20万', 'title': '全区首胜/枪王/荣耀枪王', 'pic': 'https://rpic.douyucdn.cn/live-cover/roomcover/coverupdate_2021-04-30_d15f6acb81986cb96e448013547d8761_500/webpdy1'}
(venv) E:\PythonProject2>
```

感兴趣的小伙伴可以去试一试！