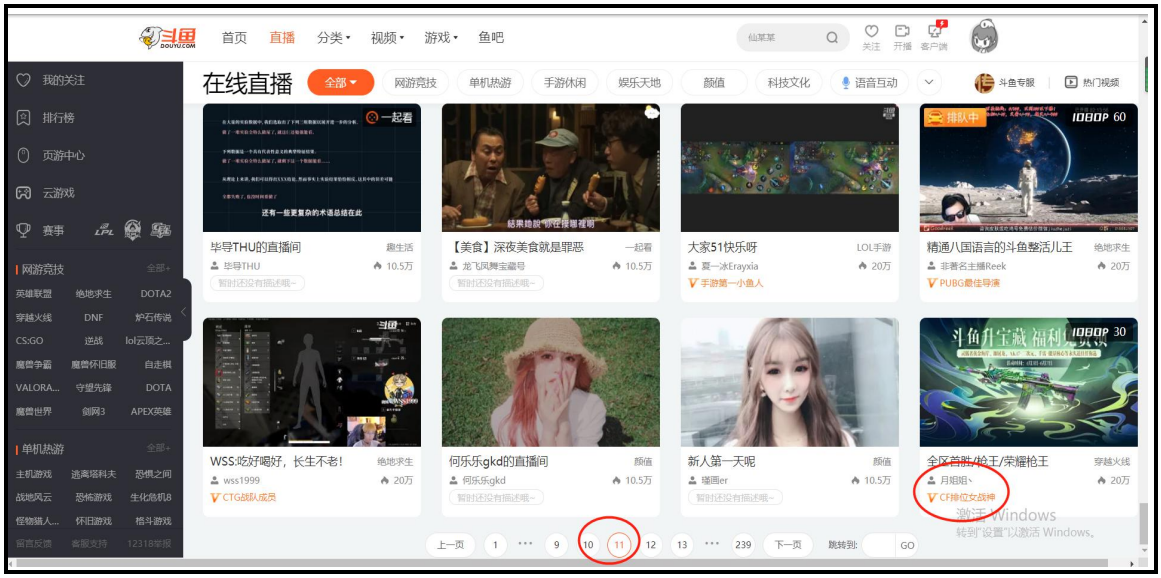
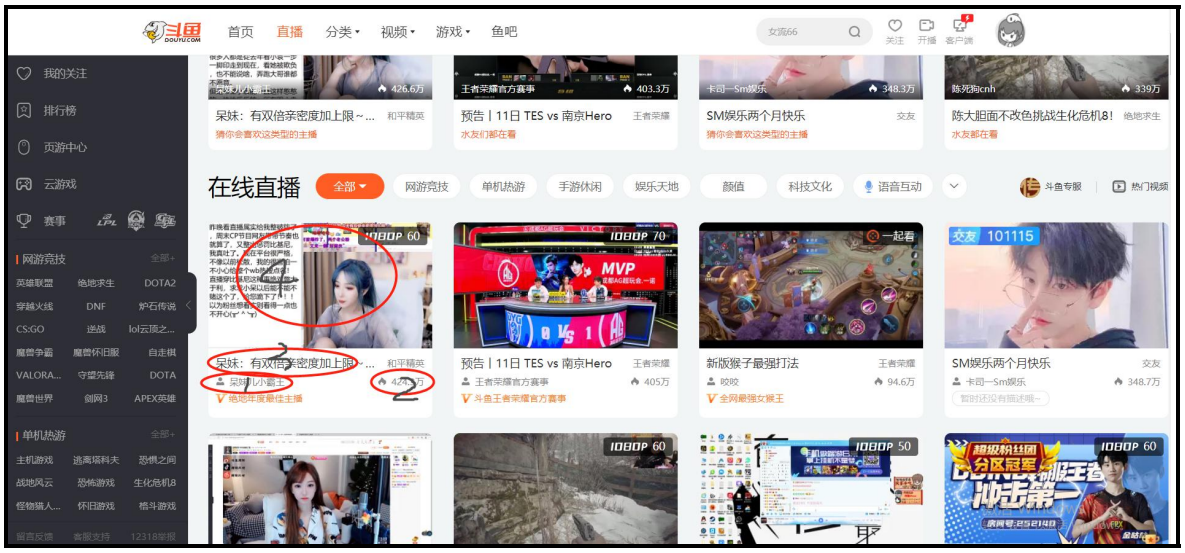
斗鱼房间信息采集

伙计们,request 是不是用的多了想用用新技术,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

感兴趣的小伙伴可以去试一试!