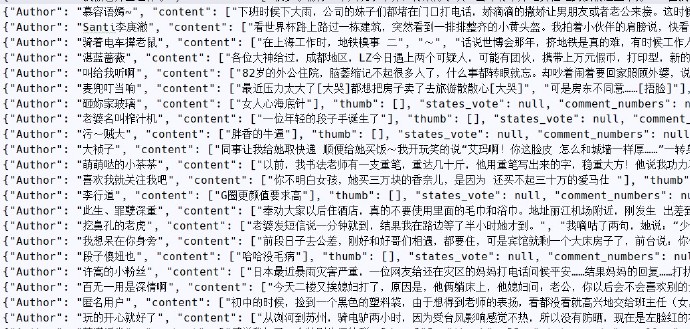
*# coding=utf-8  
# 爬虫最简模板***import**requests  
**from**lxml **import**etree  
**import**json  
  
**class QiubaiSpyder:  
def \_\_init\_\_**(self)**:**self.url\_temp **= "https://www.qiushibaike.com/8hr/page/{}/"**self.headers **=**{**"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/67.0.3396.99 Safari/537.36"**}  
  
**def get\_url\_list**(self)**:***# 1*url\_list **=**[self.url\_temp.format(i) **for**i **in**range(1, 13)]  
**return**url\_list  
  
**def parse\_url**(self, *url*)**:***# 2.*print(**"url = "**, *url*)  
response **=**requests.get(*url*, headers**=**self.headers)  
**return**response.content.decode()  
  
**def get\_content\_list**(self, *html\_str*)**:***# 3.*html **=**etree.HTML(*html\_str*)  
*# 1.分组*div\_list **=**htl.xpath(**"//div[@id='content-left']/div"**)  
content\_list **=**[]  
**for**div **in**div\_list**:**item **=**{}  
*# .strip()使“\n”去掉*item[**"Author"**] **=**div.xpath(**".//h2/text()"**)[0].strip() **if**len(div.xpath(**".//h2/text()"**)) **>**0 **else None**item[**"content"**] **=**div.xpath(**".//div[@class= 'content']/span/text()"**)  
item[**"content"**] **=**[i.strip() **for**i **in**item[**"content"**]]  
item[**"thumb"**] **=**div.xpath(**".//div[@class= 'thumb']/img/@src"**)  
  
item[**"states\_vote"**] **=**div.xpath(**".//span[@class= 'states-vote']/i/text()"**)  
item[**"states\_vote"**] **=**item[**"states\_vote"**][0] **if**len(item[**"states\_vote"**]) **>**0 **else None**item[**"comment\_numbers"**] **=**div.xpath(**".//span[@class= 'states-comments']//i/text()"**)  
item[**"comment\_numbers"**] **=**item[**"comment\_numbers"**] **if**len(item[**"comment\_numbers"**]) **>**0 **else None**content\_list.append(item)  
**return**content\_list  
  
**def save\_content\_list**(self, *content\_list*)**:  
with**open(**"qiubai.txt"**, **"a"**, encoding**= 'utf-8'**) **as**f**:  
for**content **in***content\_list***:**f.write(json.dumps(content, ensure\_ascii**=False**)) *# 转换成json字符串才能写入字典当中*f.write(**"**\n**"**)  
print(**"保存成功"**)  
  
**def run**(self)**:***# 实现主要逻辑  
# 1.根据url地址规律，构造url list*url\_list **=**self.get\_url\_list()  
  
**for**url **in**url\_list**:***# 2.发送请求，获取响应*html\_str **=**self.parse\_url(url)  
*# 3.提取数据*content\_list **=**self.get\_content\_list(html\_str)  
*# 4.保存数据*self.save\_content\_list(content\_list)  
  
**if**\_\_name\_\_ **== '\_\_main\_\_':**qiubai **=**QiubaiSpyder()  
qiubai.run()

这是写的一个简单的爬取糗事百科的程序，非常简单，复制粘贴即可使用，属于一个爬虫的基本框架，还可以进一步修改。

  
这个就是爬取的数据，以字典的形式展示出来。   
一个爬虫的基本流程就是：   
1.url   
--知道你所要爬取的网站url地址，或url地址的规律和页数，构造url地址列表   
--开始创建start\_url列表   
2.发送请求，获取响应   
--requests发送请求，get响应   
3.提取数据   
--返回json，使用json模块   
--返回html，lxml模块，并使用xpath提取数据   
--返回其他，如图片，可以直接保存为jpg或png文件   
4.保存数据   
--可以按列表的形式保存为.csv文件   
--或者直接保存为.txt文件   
  后面的3，4步可以在响应中进行，也可单独进行。   
  主程序为run()函数，在该函数下进行细分为四个子函数，每个子函数完成一个功能。   
**def get\_content\_list**(self, *html\_str*)函数，使用了之前讲解过的lxml模块，这个模块用于从html中提取数据。

窗体顶端