爬虫(十七)_糗事百科案例

糗事百科实例

爬取糗事百科段子，假设页面的URL是: http://www.qiushibaike.com/8hr/page/1

要求:

1. 使用requests获取页面信息，用XPath/re做数据提取
2. 获取每个帖子里的用户头像连接、用户姓名、段子内容、点赞次数和评论次数
3. 保存到json文件内

参考代码

```
#-*- coding:utf-8 -*-

import requests
from lxml import etree

page = 1
url = 'http://www.qiushibaike.com/8hr/page/' + str(page)
headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/52.0.2743.116
Safari/537.36',
    'Accept-Language': 'zh-CN,zh;q=0.8'}

try:
    response = requests.get(url, headers=headers)
    resHtml = response.text

    html = etree.HTML(resHtml)
    result = html.xpath('//div[contains(@id,"qiushi_tag")]')

    for site in result:
        item = {}

        imgUrl = site.xpath('./div//img/@src')[0].encode('utf-8')

        # print(imgUrl)
        username = site.xpath('./div//h2')[0].text
        # print(username)
        content = site.xpath('.//div[@class="content"]/span')[0].text.strip().encode('utf-8')
        # print(content)
        # 投票次数
        vote = site.xpath('.//i')[0].text
        # print(vote)
        #print site.xpath('.//*[@class="number"]')[0].text
        # 评论信息
        comments = site.xpath('.//i')[1].text
        # print(comments)
        print imgUrl, username, content, vote, comments

except Exception, e:
    print e
```

演示效果

分类: python

+加关注

0

« 上一篇：Python爬虫(十六)_JSON模块与JsonPath

» 下一篇：Python爬虫(十八)_多线程糗事百科案例

posted @ 2017-12-21 18:26 小破孩92 阅读(194) 评论(0) 编辑 收藏

刷新评论 刷新页面 返回顶部

注册用户登录后才能发表评论，请 登录 或 注册，访问网站首页。