书单 ELK Spark CSDN Zookeeper Python

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)
    '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'}
   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
       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

```
" 包扮,我现在举扬我自己打逼找电话,请付这直察罪立动吗?" 你还别说,真智用;本来要拘留安半个月的,最后就拘留了我一个用;我真是太机智了; 295 1
fnte//22/220622/ thash/70169320146525.jpg/image/jok/2/1/w/90/b/90
                                                                            高兴,爸爸就套路伦 262 0
tem/avtnew/1071/10713660/thumb/20151004091244.jpg?imageView2/1/w/90/h/90
                                                                            7. 款不能走进矩的生活里。只能在这里发一下 73 19
tem/avtnew/3553/35537457/thumb/20171205102120.JPEG?imageView2/1/w/90/h/90
                                                                            -张100元钱,非常高兴的多去买粮地。宣到挨了店主一巴拿之后,我才知道了有一种钱叫做宴币 865 9
tem/avtnew/3125/31258013/thumb/20160222093838.jpg?linageView2/1/w/90/h/90
                                                                        管交待,刘绍驼好好带一下这个老皮。当时看刘珉驳脸色有些不好看!年纪大款秋老皮嘛!!! 结果入职卖一看,人家姓皮。 263 2
stan/avtnew/3298/32984261/thumb/2016111723234998.3PEG7inageView2/1/w/99/h/90
                       - Globialus (Uniformation of Control of Con
                            念。... 背景不割......老板说光话后突然放了个层,为了延翰尴尬。$$身对表说: " 2000. 读你了! " 安一时懵了,一本正眨地说: 老板,表....表............................
pshbaike.com/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/system/sy
                         a 対漢得时候不知道地と了,突然一拍头板大声得叫了的奶奶,我们就主任正蓝得回头应了。( 关键我们就主任正是个五十多半老头[突哭])接着全班供堂大奖! 346 0
Jushibalke.com/system/avtnew/756/7561589/thumb/2013120115639-jpg/lbaspk/be/2/1//950/h/99
                              8
1分,我听到女同客在大喊,出来了,读了好多水出来了,旁夜当时就走起一震,赶紧跑过去,推淌,不就是小非掉气管积了点水,温琛油门轰出来端,说的这么强气冲天。 290 1
shubalke.com/system/a/thoe//371/3711158/thumb/20171699213131.3FG/hangeNae/1/b/99/h/99
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。
```

Copyright ©2018 小破孩92