

如何利用Python网络爬虫抓取微信朋友圈的动态

Python进阶者

每日一个Linux、Python干货 ▲ 关注的人都加薪了

来源：程序人生

ID: coder_life



图片源自网络

作者

Python进阶者

如需转载，请联系原作者授权。

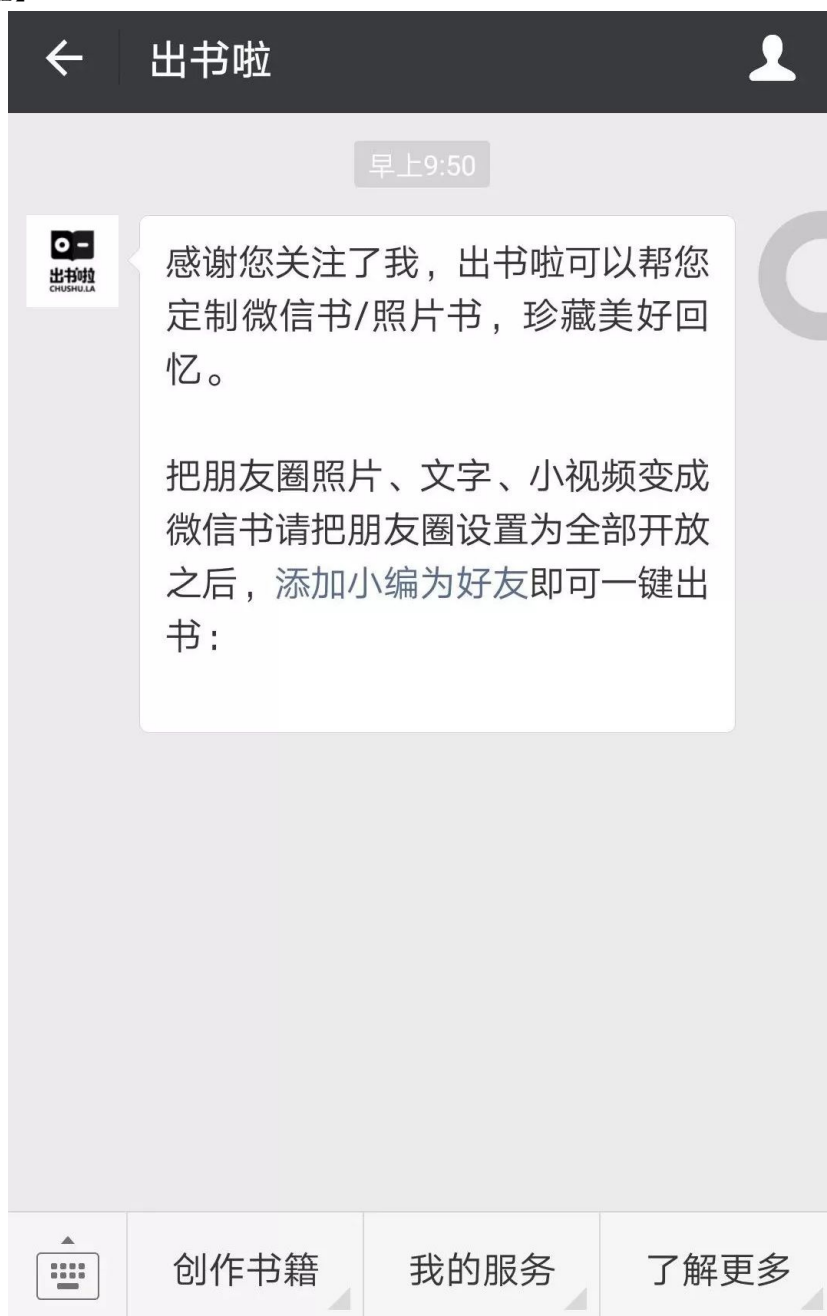
今天小编给大家分享一下如何利用Python网络爬虫抓取微信朋友圈的动态信息，实际上如果单独的去爬取朋友圈的话，难度会非常大，因为微信没有提供向网易云音乐这样的API接口，所以很容易找不到门。不过不要慌，小编在网上找到了第三方工具，它可以将朋友圈进行导出，之后便可以像我们正常爬虫网页一样进行抓取信息了。

【出书啦】就提供了这样一种服务，支持朋友圈导出，并排版生成微信书。本文的主要参考资料来源于这篇博文：

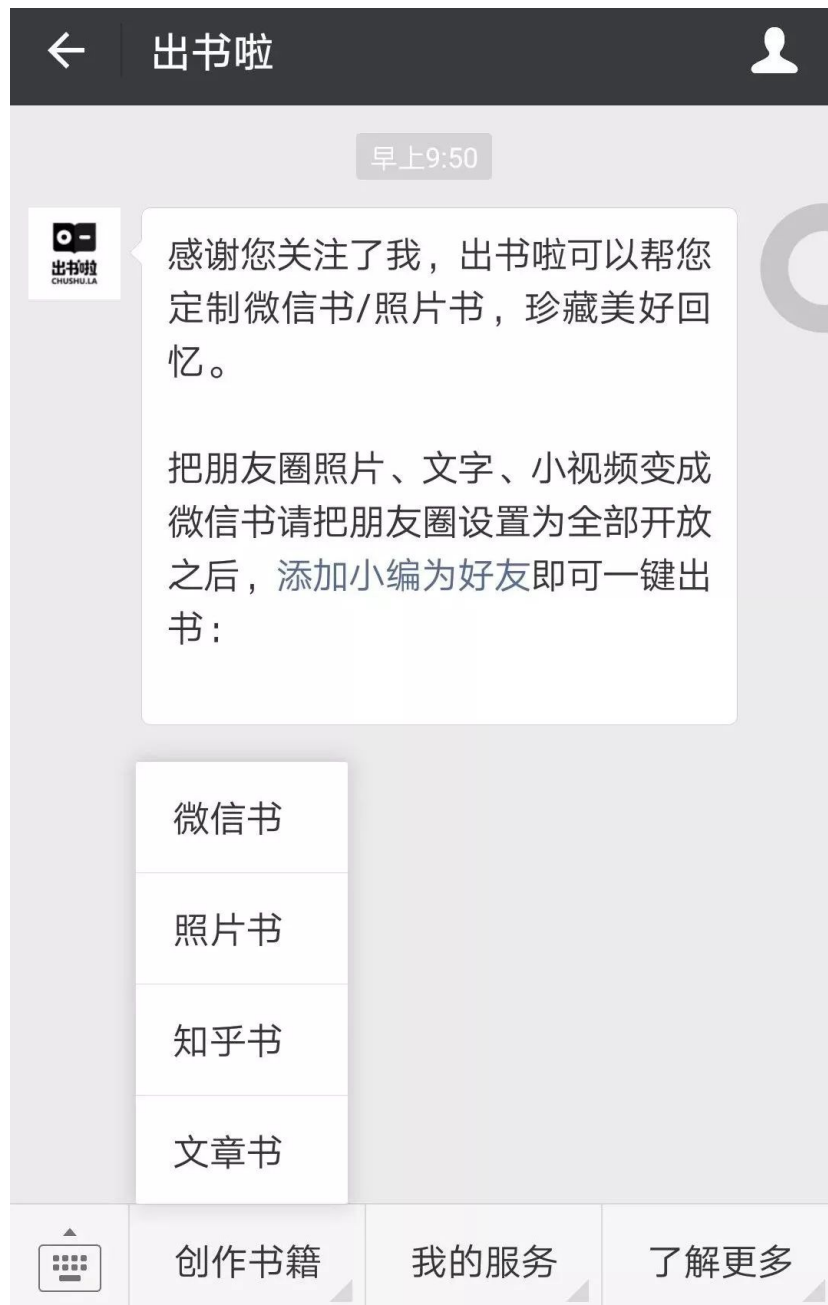
<https://www.cnblogs.com/sheng-jie/p/7776495.html>，感谢大佬提供的接口和思路。具体的教程如下。

一、获取朋友圈数据入口

1、关注公众号【出书啦】



2、之后在主页中点击【创作书籍】-->【微信书】。



3、点击【开始制作】-->【添加随机分配的出书啦小编为好友即可】，长按二维码之后便可以添加好友了。

4、之后耐心等待微信书制作，待完成之后，会收到小编发送的消息提醒，如下图所示。

至此，我们已经将微信朋友圈的数据入口搞定了，并且获取了外链。

确保朋友圈设置为【全部开放】，默认就是全部开放，如果不知道怎么设置的话，请自行百度吧。

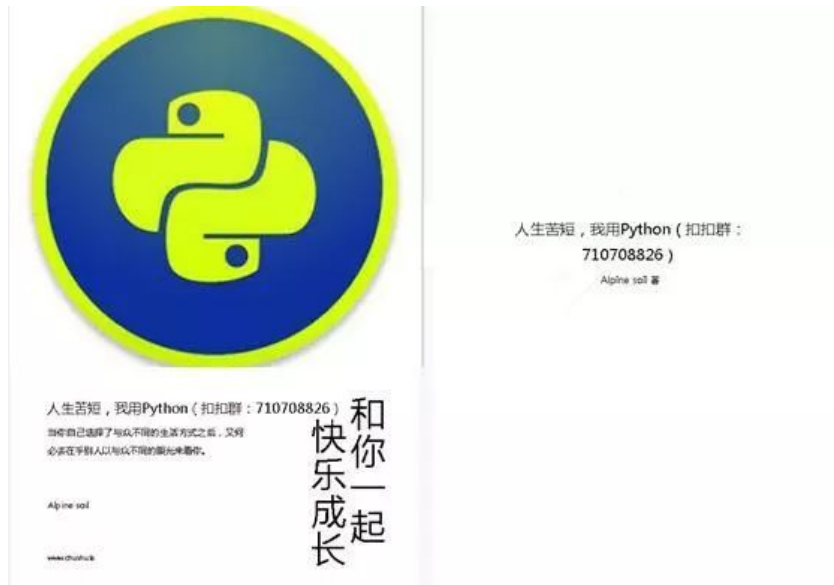


5、点击该外链，之后进入网页，需要使用微信扫码授权登录。

6、扫码授权之后，就可以进入到微信书网页版了，如下图所示。



7、接下来我们就可以正常的写爬虫程序进行抓取信息了。在这里，小编采用的是Scrapy爬虫框架，Python用的是3版本，集成开发环境用的是Pycharm。下图是微信书的首页，图片是小编自己自定义的。



二、创建爬虫项目

1、确保您的电脑上已经安装好了Scrapy。之后选定一个文件夹，在该文件夹下进入命令行，输入执行命令：

```
scrapy startproject weixin_moment
```

，等待生成Scrapy爬虫项目。

2、在命令行中输入`cd weixin_moment`，进入创建的weixin_moment目录。之后输入命令：

```
scrapy genspider 'moment' 'chushu.la'
```

，创建朋友圈爬虫，如下图所示。

```
C:\Windows\System32\cmd.exe
Microsoft Windows [版本 6.1.7601]
版权所有 (c) 2009 Microsoft Corporation. 保留所有权利。

D:\pythonDemo\2018\May\5.9>scrapy startproject weixin_moment
New Scrapy project 'weixin_moment', using template directory 'c:\users\lenovo\anaconda3\lib\site-packages\scrapy\templates\project'
D:\pythonDemo\2018\May\5.9\weixin_moment

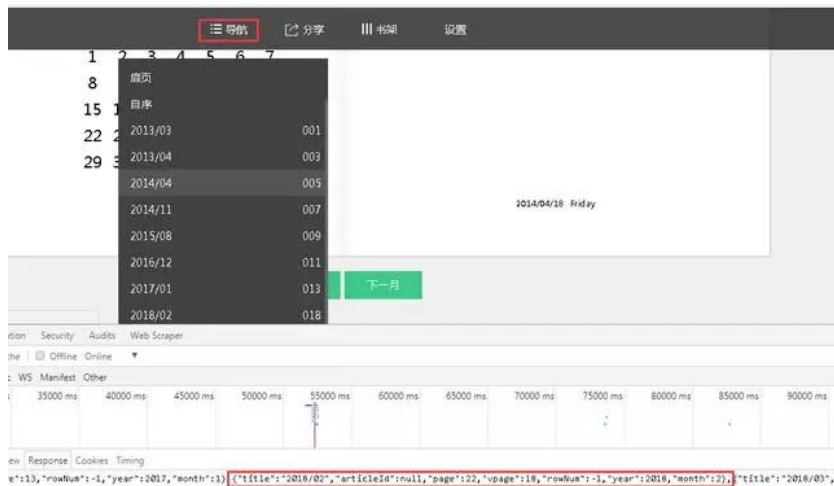
You can start your first spider with:
cd weixin_moment
scrapy genspider example example.com

D:\pythonDemo\2018\May\5.9>cd weixin_moment

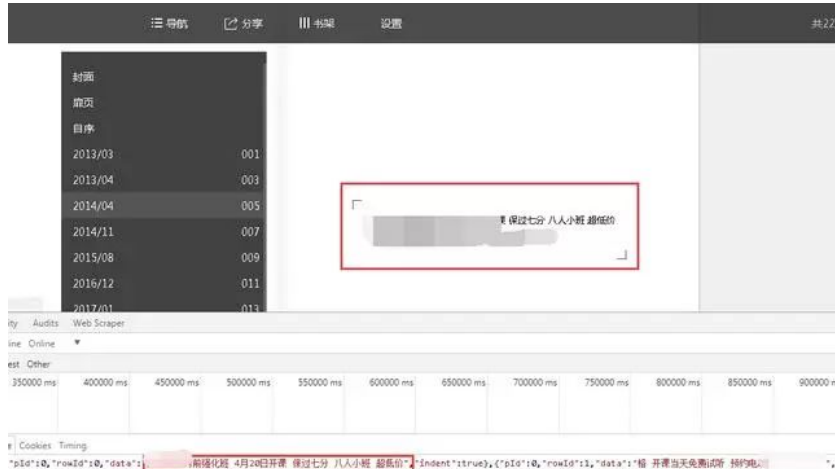
D:\pythonDemo\2018\May\5.9\weixin_moment> scrapy genspider 'moment' 'chushu.la'
Created spider "moment" using template 'basic' in module:
weixin_moment.spiders.a'moment'

D:\pythonDemo\2018\May\5.9\weixin_moment>
```

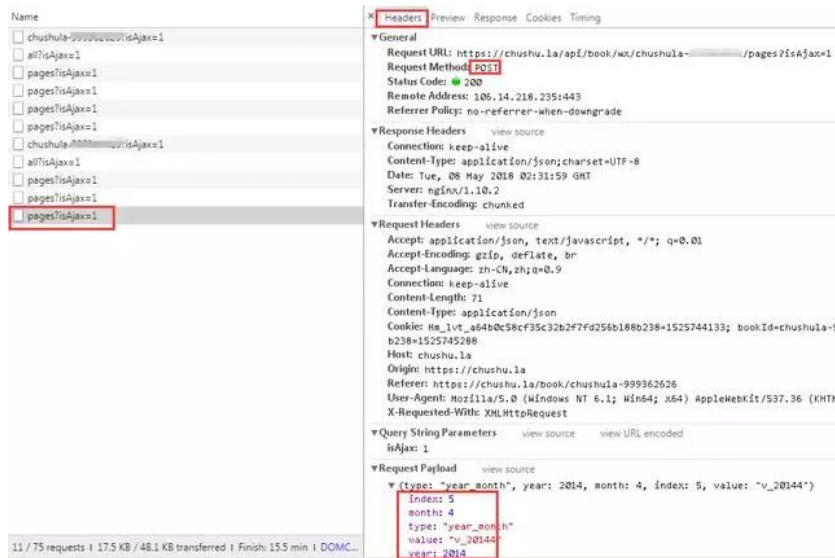
3、执行以上两步后的文件夹结构如下：



4、当点击【2014/04】月份，之后查看服务器响应数据，可以看到页面上显示的数据和服务器的响应是相对应的。



5、查看请求方式，可以看到此时的请求方式变成了POST。细心的伙伴可以看到在点击“下个月”或者其他导航月份的时候，主页的URL是始终没有变化的，说明该网页是动态加载的。之后对比多个网页请求，我们可以看到在“Request Payload”下边的数据包参数不断的发生变化，如下图所示。



6、展开服务器响应的数据，将数据放到JSON在线解析器里，如下图所示：


```

36     'imgs': [],
37     'paras': [{
38         'rows': [{
39             'pId': 0,
40             'rowId': 0,
41             'data': '强化班 4月20日开课 保过七分 八人小班 超低价',
42             'indent': true
43         }, {
44             'pId': 0,
45             'rowId': 1,
46             'data': '格 开课当天免费试听 预约电话...',
47             'indent': false
48         }
49     ]
50     }, {
51         'imgDivStyle': {
52             'width': -2147483648,
53             'height': -2147483648

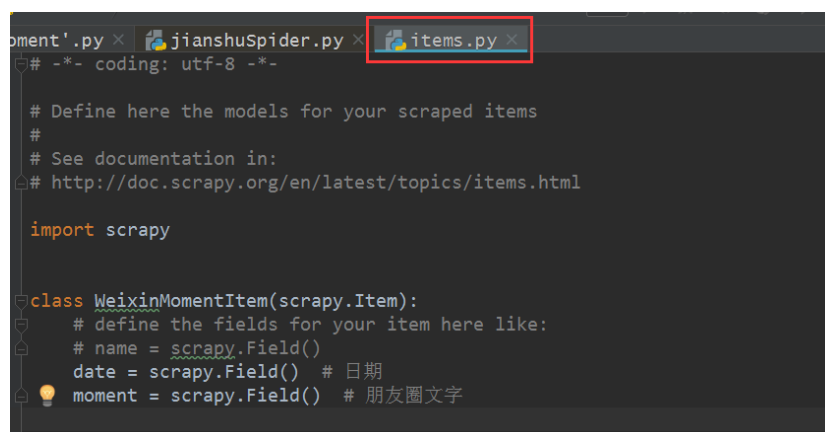
```

可以看到朋友圈的数据存储在paras /data节点下。

接下来将写程序，进行数据抓取。接着往下继续深入。

四、代码实现

1、修改Scrapy项目中的items.py文件。我们需要获取的数据是朋友圈和发布日期，因此在这里定义好日期和动态两个属性，如下图所示。



```

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

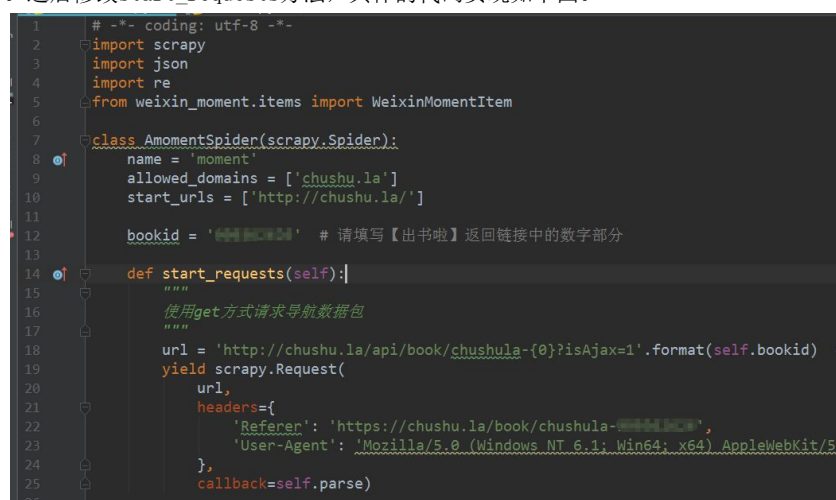
# Define here the models for your scraped items
#
# See documentation in:
# http://doc.scrapy.org/en/latest/topics/items.html

import scrapy

class WeixinMomentItem(scrapy.Item):
    # define the fields for your item here like:
    # name = scrapy.Field()
    date = scrapy.Field() # 日期
    moment = scrapy.Field() # 朋友圈文字

```

2、修改实现爬虫逻辑的主文件moment.py，首先要导入模块，尤其是要主要将items.py中的WeixinMomentItem类导入进来，这点要特别小心别被遗漏了。之后修改start_requests方法，具体的代码实现如下图。



```

1 # -*- coding: utf-8 -*-
2 import scrapy
3 import json
4 import re
5 from weixin_moment.items import WeixinMomentItem
6
7 class AmomentSpider(scrapy.Spider):
8     name = 'moment'
9     allowed_domains = ['chushu.la']
10    start_urls = ['http://chushu.la/']
11
12    bookid = '444444' # 请填写【出书啦】返回链接中的数字部分
13
14    def start_requests(self):
15        """
16        使用get方式请求导航数据包
17        """
18        url = 'http://chushu.la/api/book/chushula-{}?isAjax=1'.format(self.bookid)
19        yield scrapy.Request(
20            url,
21            headers={
22                'Referer': 'https://chushu.la/book/chushula-{}'.format(self.bookid),
23                'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/54.0.2840.99 Safari/537.36'
24            },
25            callback=self.parse)

```

3、修改parse方法，对导航数据包进行解析，代码实现稍微复杂一些，如下图所示。


```

28 def parse(self, response):
29     # 处理获取到的导航数据包
30     json_body = json.loads(str(response.text)) # 加载json数据包
31     catalogs = json_body['book']['catalogs'] # 获取json中的目录数据包
32     url = 'https://chushu.la/api/book/wx/chushula-{0}/pages?isAjax=1'.format(self.bookid) # 分页数据url
33     start_page = int(catalogs[0]['month']) # 获取起始月份作为index传值
34     for catalog in catalogs:
35         year = catalog['year']
36         month = catalog['month']
37         formdata = {
38             "type": 'year_month',
39             "year": str(year),
40             "month": str(month),
41             "index": str(start_page),
42             "value": 'v_{0}{1}'.format(year, month)
43         }
44         start_page += 1
45     yield scrapy.FormRequest(
46         url,
47         method='POST',
48         body=json.dumps(formdata),
49         headers={
50             'Host': 'chushu.la',
51             'Connection': 'keep-alive',
52             'Accept': 'application/json, text/javascript, */*; q=0.01',
53             'Origin': 'https://chushu.la',
54             'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; Win64; x64) AppleWebKit/537.36 (KHTML, like
55             'Content-Type': 'application/json',
56             'Referer': 'https://chushu.la/book/chushula-{}'.format(self.bookid),
57             'Accept-Encoding': 'gzip, deflate, br',
58             'Accept-Language': 'zh-CN,zh;q=0.9',
59             'Cookie': 'Hm_lvt_a64b8c58cf35c32b2f7fd256b188b238=1525744133; tokensac9ba3520ffd752044
60         },
61         callback=self.parse_moment)

```

- 需要注意的是从网页中获取的response是bytes类型，需要显示的转为str类型才可以进行解析，否则会报错。
- 在POST请求的限定下，需要构造参数，需要特别注意的是参数中的年、月和索引都需要是字符串类型的，否则服务器会返回400状态码，表示请求参数错误，导致程序运行的时候报错。
- 在请求参数还需要加入请求头，尤其是Referer（反盗链）务必要加上，否则在重定向的时候找不到网页入口，导致报错。
- 上述的代码构造方式并不是唯一的写法，也可以是其他的。

4、定义parse_moment函数，来抽取朋友圈数据，返回的数据以JSON加载的，用JSON去提取数据，具体的代码实现如下图所示。

```

63 def parse_moment(self, response):
64     """
65     朋友圈数据处理
66     """
67     # json_body = json.loads(response.body)
68     json_body = json.loads(response.text)
69     pages = json_body['pages']
70     # print(pages)
71     pattern = re.compile(u"[\u4e00-\u9fa5]+") # 匹配中文
72     items = WeixinMomentItem()
73     for page in pages:
74         if (page['type'] == "weixin_moment_page"): # 仅抽取朋友圈分页数据
75             paras = page['data']['paras']
76             if paras:
77                 moment = ''
78                 for content in paras[0]['rows']:
79                     result = re.findall(pattern, content['data']) # 使用正则匹配所有中文朋友圈
80                     moment += ''.join(result)
81                 items['moment'] = moment
82                 items['date'] = page['data']['dateText'] # 获取时间
83             yield items
84

```

5、在setting.py文件中将ITEM_PIPELINES取消注释，表示数据通过该管道进行处理。

```

56 # 'weixin_moment.middlewares.MyCustomDownloaderMiddleware': 543,
57 #}
58
59 # Enable or disable extensions
60 # See http://scrapy.readthedocs.org/en/latest/topics/extensions.html
61 #EXTENSIONS = {
62 # 'scrapy.extensions.telnet.TelnetConsole': None,
63 #}
64
65 # Configure item pipelines
66 # See http://scrapy.readthedocs.org/en/latest/topics/item-pipeline.html
67 ITEM_PIPELINES = {
68     'weixin_moment.pipelines.WeixinMomentPipeline': 300,
69 }
70
71 # Enable and configure the AutoThrottle extension (disabled by default)

```

6、之后就可以在命令行中进行程序运行了，在命令行中输入

```
scrapy crawl moment -o moment.json
```

，之后可以得到朋友圈的数据，在控制台上输出的信息如下图所示。

```
<'date': '2013/03/22', 'moment': '花艺课上我和...同学插花合照'>
2018-05-09 10:01:04 [scrapy.core.scraper] DEBUG: Scraped from <200 https://chushu.la/api/book/ux/chushula.../pages?isAjax=1>
<'date': '2014/04/18', 'moment': '...日开课保过七八人小班超低价格开课当天免费试听预约电话'>
2018-05-09 10:01:04 [scrapy.core.scraper] DEBUG: Scraped from <200 https://chushu.la/api/book/ux/chushula.../pages?isAjax=1>
<'date': '2014/11/14', 'moment': '...>
'moment': '期间宣传下我也做过一个礼拜赚了多有意私聊招聘单位超分期工作地点走寝宣传工作时间晚上一个小时学生要求性格开朗工作报酬张件'
2018-05-09 10:01:04 [scrapy.core.scraper] DEBUG: Scraped from <200 https://chushu.la/api/book/ux/chushula.../pages?isAjax=1>
<'date': '2018/02/16', 'moment': '...>
'moment': '童心未泯哈哈和小朋友们一起偷偷的玩鞭炮跨年已来到祝福大家新的一年大吉大利身体健康工作顺利感谢生命中遇到的每一位人因为有你'
2018-05-09 10:01:04 [scrapy.core.scraper] DEBUG: Scraped from <200 https://chushu.la/api/book/ux/chushula.../pages?isAjax=1>
<'date': '2017/01/28', 'moment': '广州过年花城看花行花街舞龙耍龙求福'>
2018-05-09 10:01:04 [scrapy.core.scraper] DEBUG: Scraped from <200 https://chushu.la/api/book/ux/chushula.../pages?isAjax=1>
<'date': '2013/04/28', 'moment': '宿舍的蝎子今天好恐怖受不了阿'>
2018-05-09 10:01:04 [scrapy.core.engine] INFO: Closing spider (finished)
2018-05-09 10:01:04 [scrapy.extensions.feedexport] INFO: Stored json feed (7 items) in: moment.json
2018-05-09 10:01:04 [scrapy.statscollectors] INFO: Dumping Scrapy stats:
{'downloader/request_bytes': 5906,
 'downloader/request_count': 11,
 'downloader/request_method_count/GET': 2,
 'downloader/request_method_count/POST': 9,
 'downloader/response_bytes': 35269,
 'downloader/response_count': 11,
 'downloader/response_status_count/200': 10,
 'downloader/response_status_count/301': 1,
 'finish_reason': 'finished',
 'finish_time': datetime.datetime(2018, 5, 9, 2, 1, 4, 264403),
 'item_scraped_count': 7,
 'log_count/DEBUG': 19,
 'log_count/INFO': 9,
 'request_depth_max': 1,
 'response_received_count': 10,
 'scheduler/dequeued': 11,
 'scheduler/dequeued/memory': 11,
 'scheduler/enqueued': 11,
 'scheduler/enqueued/memory': 11,
 'start_time': datetime.datetime(2018, 5, 9, 2, 1, 2, 916326)}
2018-05-09 10:01:04 [scrapy.core.engine] INFO: Spider closed (finished)

D:\pythonDemo\2018\May\5.7\weixin_moment>
北:
```

7、尔后我们得到一个moment.json文件，里面存储的是我们朋友圈数据，如下图所示。

8、嗯，你确实没有看错，里边得到的数据确实让人看不懂，但是这个并不是乱码，而是编码的问题。解决这个问题的方式是将原来的moment.json文件删除，之后重新在命令行中输入下面的命令：

```
scrapy crawl moment -o moment.json -s FEED_EXPORT_ENCODING=utf-8,
```

此时可以看到编码问题已经解决了，如下图所示。

[《Linux云计算及运维架构师高薪实战班》2018年08月27日即将开课中，120天冲击Linux运维年薪30万，改变速约~~~~](#)

*声明：推送内容及图片来源于网络，部分内容会有所改动，版权归原作者所有，如来源信息有误或侵犯权益，请联系我们删除或授权事宜。

免 费 好 礼



糖豆

推荐一个福利包

Python福利包

主讲人：上市公司十年开发经理

福利1：15册Python入门书籍

福利2：30集Python入门视频

福利3：50个Python商业项目源代码



长按识别二维码，即刻获取



每天精选技术干货，十万Linux人订阅

◀ **Linux人充电第一站**

长按识别二维码 关注马哥Linux运维

更多Linux好文请点击【[阅读原文](#)】哦

↓↓↓

[阅读原文](#)