爬虫案例5(讲数据存入数据库中 就存mysql)

https://blog.csdn.net/qq_25046261/article/details/79746561

存入mysql

就爬这个吧:http://download.java1234.com

顺便温习哈操作

1.新建项目

```
MichaelYun:PycharmProjects Yun$ scrapy startproject java1234
New Scrapy project 'java1234', using template directory '/usr/local/lib/python3.
7/site-packages/scrapy/templates/project', created in:
    /Users/Yun/PycharmProjects/java1234

You can start your first spider with:
    cd java1234
    scrapy genspider example example.com
MichaelYun:PycharmProjects Yun$ cd java1234
MichaelYun:java1234 Yun$ scrapy genspider tutu http://download.java1234.com/
Created spider 'tutu' using template 'basic' in module:
  java1234.spiders.tutu
MichaelYun:java1234 Yun$
```

2.编写逻辑代码

```python
# -*- coding: utf-8 -*-
import scrapy
from java1234.items import Java1234Item

class TutuSpider(scrapy.Spider):
    name = 'tutu'
    allowed_domains = ['download.java1234.com']
    start_urls = ['http://download.java1234.com']

    def parse(self, response):
        #单页资源列表名字
        rnames = response.css('body > div:nth-child(3) > div.pLeft > div:nth-child(2) > div.layui-form > table > tbody > tr > td:nth-child(1) > a::text').extract()
        #单页的查看次数
        rcounts = response.css('body > div:nth-child(3) > div.pLeft > div:nth-child(2) > div.layui-form > table > tbody > tr > td:nth-child(2)::text').extract()
        #单页的上传者
        rauthors = response.css('body > div:nth-child(3) > div.pLeft > div:nth-child(2) > div.layui-form > table > tbody > tr > td:nth-child(3)::text').extract()
        for rname,rcount,author in zip(rnames,rcounts,rauthors):
            msg = Java1234Item()
            msg['rname'] = rname
```

```
            msg['rcount'] = rcount
            msg['author'] = author
            yield msg
        #获取下一页的class属性，判断是否完全相等 如果可以就提取下一页的链接，然后拼接，并且再次请求
        urlClassName = response.css('#layui-laypage-1 > a.layui-laypage-next::attr(class)').extract_first()
        if urlClassName == 'layui-laypage-next':
            urlhref = response.css('#layui-laypage-1 > a.layui-laypage-next::attr(href)').extract_first()
            url = response.urljoin(urlhref)
            print('url的地址'+url)
            yield  scrapy.Request(url,callback=self.parse)


```

```
Stored csv feed (1366 items) in: tutu.csv
```

总共是1366条数据

3.接下来就是新建数据库表,存入到mysql数据库中：

在pipelines.py中，新建一个处理mysql的管道

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

# Define your item pipelines here
#
# Don't forget to add your pipeline to the ITEM_PIPELINES setting
# See: https://doc.scrapy.org/en/latest/topics/item-pipeline.html
import  pymysql

class Java1234Pipeline(object):
    def process_item(self, item, spider):
        return item

class MySqlPipeline(object):
    #链接数据库基本信息,并且获取游标对象
    def connect_db(self):
        self.conn = pymysql.connect(host='localhost',
                                    port=3306, user='root',
                                    password='123456',
                                    database='wwj',
                                    charset='utf8')
```

```
        self.cursor = self.conn.cursor()

    #链接数据库
    def open_spider(self,spider):
        self.connect_db()
    #关闭数据库连接
    def close_spider(self,spider):
        self.cursor.close()
        self.conn.close()

    #写入数据库
    def process_item(self, item, spider):
        sql = 'insert into msgs (author, downloadcount,downloadcontent) values ("%s","%s","%s")' % (item['author'], item['rcount'],item['rname'])
        self.cursor.execute(sql)
        self.conn.commit()
        return item


```

然后在setting.py文件中:

```
ITEM_PIPELINES = {
    'java1234.pipelines.Java1234Pipeline': 300,
    'java1234.pipelines.MySqlPipeline':200
}
```

开启对象处理管道。

----------最后运行

截取部分:

id	author	downloadcount	downloadcontent
1	开心	140 次	WEB前端教程 教程 下载
2	开心	158 次	Socket网络编程进阶与实战完整无秘 教程 下载
3	小九	244 次	传智播客大数据就业班完整版 教程 下载
4	小九	184 次	移动web携程旅行网实战开发 教程 下载
5	Voitor	201 次	Hadoop权威指南(第4版)(修订版&升级版) 中文完整pd
6	星期七	226 次	大数据开发之Hadoop工程师 教程 下载
7	星期七	274 次	36套Java技术提升及架构精华课 教程 下载
8	幸运儿	224 次	Mysql实战45讲 教程 下载
9	幸运儿	216 次	互联网架构师完整不加密版 教程 下载
10	lieying701	250 次	黑马python人工智能基础加就业
11	心情乌云	216 次	622079 Ionic学习手册 教程 下载
12	心情乌云	214 次	React16.4 开发简书项目 从零基础入门到实战 教程

scrapy crawl tutu -o tutu.csv

不仅可以到处csv 然后也同时保存到数据库中

当然还有点要考虑，就是插入数据前，先要数据去重