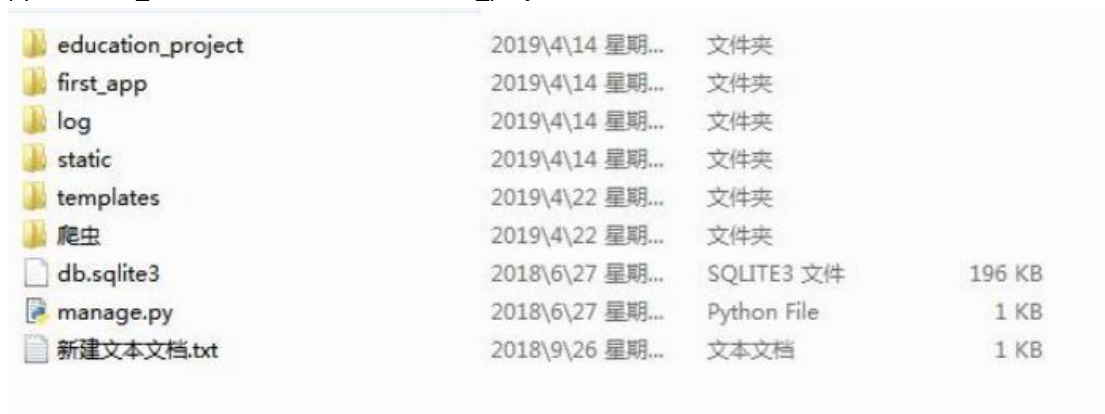


爬虫期末作业



pythonWeb_venv 为虚拟目录，education_project 是项目目录



Templates 存放 html，static 存放资源文件，爬虫存放爬虫文件，

名称	修改日期	类型	大小
ad.html	2018\9\9 星期日 ...	360 se HTML Do...	2 KB
article.html	2018\9\26 星期...	360 se HTML Do...	4 KB
articleAdd.html	2018\9\26 星期...	360 se HTML Do...	9 KB
author.html	2018\9\23 星期...	360 se HTML Do...	1 KB
base.html	2018\9\26 星期...	360 se HTML Do...	6 KB
course.html	2018\9\10 星期...	360 se HTML Do...	9 KB
courseVideo.html	2018\9\25 星期...	360 se HTML Do...	10 KB
douban.html	2019\6\16 星期...	360 se HTML Do...	11 KB
failure.html	2015\6\18 星期...	360 se HTML Do...	1 KB
index.html	2019\6\16 星期...	360 se HTML Do...	23 KB
jingdong.html	2019\4\22 星期...	360 se HTML Do...	10 KB
login.html	2018\9\26 星期...	360 se HTML Do...	5 KB
login12306.html	2019\6\16 星期...	360 se HTML Do...	9 KB
my.html	2018\9\17 星期...	360 se HTML Do...	1 KB
ouwang.html	2019\4\22 星期...	360 se HTML Do...	10 KB
pagination.html	2019\4\14 星期...	360 se HTML Do...	2 KB
pagination2.html	2019\4\14 星期...	360 se HTML Do...	3 KB
remen.html	2018\9\24 星期...	360 se HTML Do...	1 KB
userLead.html	2018\9\24 星期...	360 se HTML Do...	1 KB

名称	修改日期	类型	大小
assets	2019\4\14 星期...	文件夹	
css	2019\4\14 星期...	文件夹	
images	2019\4\14 星期...	文件夹	
js	2019\4\14 星期...	文件夹	
lib	2019\4\14 星期...	文件夹	
mp4	2019\4\14 星期...	文件夹	

Get_ip 从网上爬取免费 ip

scrapy 静态网站：瓯网使用 scrapy 爬取，

Scrapy 爬取豆瓣

登陆 12306

名称	修改日期	类型	大小
jidong	2019\4\22 星期...	文件夹	
login12306	2019\6\16 星期...	文件夹	
scrapy	2019\4\22 星期...	文件夹	
ScrapyDouban	2019\6\16 星期...	文件夹	
get_ip	2019\6\18 星期...	文件夹	

修改 scrapy 的_requests_to_follow 方法，把 url 写入到 redis.

```
def _requests_to_follow(self, response):
    if not isinstance(response, HtmlResponse):
        return
    reg = r'article\d*show.html'
    seen = set()
    for n, rule in enumerate(self._rules):
        links = [lnk for lnk in rule.link_extractor.extract_links(response)
                  if lnk not in seen]
        if links and rule.process_links:
            links = rule.process_links(links)
        for link in links:
            match = re.findall(reg, str(link.url))
            if match == []:
                if self.redis.is_existArticle(link.url):
                    continue
                else:
                    self.redis.putArticle(link.url)
                    seen.add(link)
                    r = self._build_request(n, link)
                    yield rule.process_request(r)
            else:
                self.redis.put(link.url)
```

解析网页时使用多线程,使用新闻通用 filter。

```

if __name__ == "__main__":
    print('Parent process %s.' % os.getpid())
    p = Pool(10)

    for i in range(10):
        p.apply_async(main, args=())
    print('Waiting for all subprocesses done...')
    p.close()
    p.join()
    print('All subprocesses done.')

```

__pycache__	2019\4\22 星期...	文件夹	
myfilter.py	2019\4\20 星期...	Python File	6 KB
myMongodb.py	2018\7\8 星期日 ...	Python File	2 KB
mysqlDb.py	2019\4\20 星期...	Python File	3 KB
ouwang.py	2019\4\22 星期...	Python File	3 KB
ouwang2.py	2018\7\1 星期日 ...	Python File	2 KB
redisDB.py	2018\7\1 星期日 ...	Python File	2 KB
setting.py	2019\4\20 星期...	Python File	1 KB
textExcel.py	2018\7\1 星期日 ...	Python File	1 KB

京东使用 selenium 爬取

名称	修改日期	类型	大小
__pycache__	2019\4\22 星期...	文件夹	
1.py	2019\4\14 星期...	Python File	3 KB
2.py	2019\4\17 星期...	Python File	0 KB
chromedriver.exe	2018\3\20 星期...	应用程序	6,207 KB
db.py	2018\7\23 星期...	Python File	2 KB
jingdong.py	2019\4\19 星期...	Python File	3 KB
mysqlDb.py	2019\4\18 星期...	Python File	3 KB
setting.py	2019\4\10 星期...	Python File	1 KB

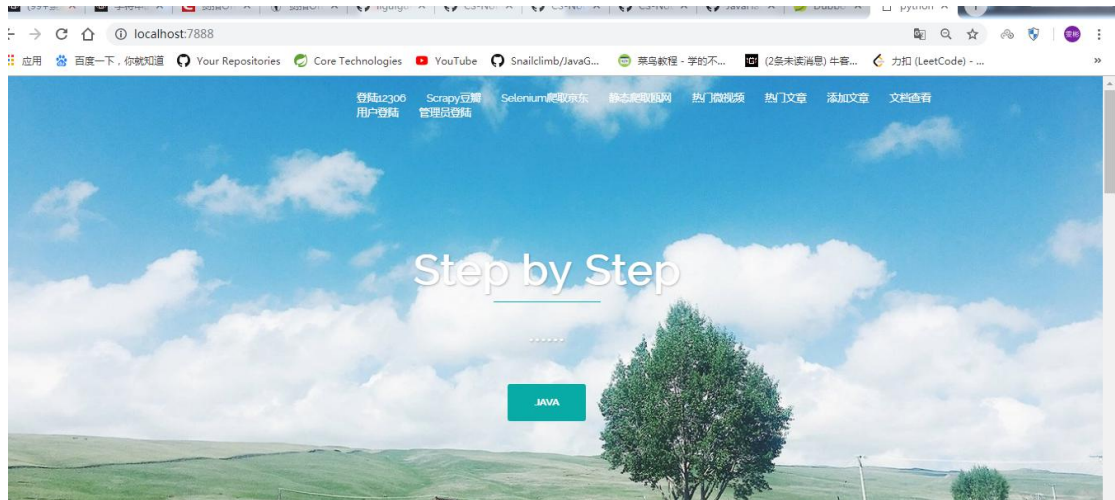
```

temp += 1

except Exception as e:
    print(e)
finally:
    driver.close()

if __name__ == '__main__':
    keyword = "电脑"
    #设置爬取页面
    page = 30
    page = page * 2
    print("开始爬取京东")
    base_url = 'https://search.jd.com/Search?keyword={key}&enc=utf-8&qrst=1&rt=1&stop=1&vt=2&wq={key}'
    for i in range(1, page, 2):
        url = base_url.format(key=keyword,page=i);
        get_jingdong(url, i)
        time.sleep(5)

```



◆ Data Capture

首页 后台登录 更新 扫码 帮助

Search..

公众号 日期 X 搜索 Go!

搜索结果 search result 177

商品名	店铺名	价格	评论数
【Pencil套装版】Apple iPad 平板电脑 9.7英寸 (128G WLAN版) 银色及 Pencil套装 MR7K2CH/A	Apple产品京东自营旗舰店	3699.00	113万+条评价
【亮铂金属键盘版】微软 (Microsoft) Surface Pro (第五代) 二合一平板电脑笔记本 (Core M3 4G 128G)	微软京东自营官方旗舰店	4649.00	1.4万+条评价
【新品上市】APPLE苹果2019新款iPad mini5/mini4平板电脑7.9英寸 金色【新品】mini5-64G-WLAN版	科隆数码专营店	2749.00	2.9万+条评价
【领券立减】Apple iPad Pro 11英寸平板电脑 2018年款(256G WLAN版/全面屏/A12X/FaceID MTXQ2CH/A)深空灰色	Apple产品京东自营旗舰店	7669.00	3.6万+条评价
【领券立减】Apple iPad Pro 11英寸平板电脑 2018年款(64G WLAN版/全面屏/A12X/Face ID MTXN2CH/A) 深空灰色	Apple产品京东自营旗舰店	6469.00	3.6万+条评价
【领券立减】Apple iPad 平板电脑 2018年新款9.7英寸 (128G WLAN版/A10 芯片/Retina显示屏 MR7K2CH/A) 银色	Apple产品京东自营旗舰店	3268.00	113万+条评价
【领券立减】Apple iPad 平板电脑 2018年新款9.7英寸 (128G WLAN版/A10 芯片/Retina显示屏 MR7K2CH/A) 深空灰色	Apple产品京东自营旗舰店	3268.00	113万+条评价

Selenium爬取京东 静态爬取网站 热门微视频 热门文章 添加文章 文档查看 用户登录 管理员登录

热门文章

虚拟化信息传播与服务融合

python
推送一些python小道理，个性化操作，项目介绍，实战推荐

爬虫
介绍一些各式网站不同的爬取，全站爬取，搜索引擎等

Java
Java SE基础知识，SSM框架，springBoot,springCloud微服务等热门技术分享



个人博客

专注Python开发，欢迎和大家交流

[首页](#)[关于我](#)[文章](#)[添加文章](#)[用户排行榜](#)

Python读写Excel表格 就是这么简单粗暴又好用

关注我



CSDN



简书



公众号



邮箱

站长推荐

1 如何玩转微服务

2 Python3 File(文件) 方法

3 (转) JavaEE基础知识回顾

4 Java SE基础知识30问

5 (转)Spring AOP中定义切点 (PointCut)

用户文章

如何玩转微服务

如何玩转微服务

Java 2018-09-26

评论 (5) 浏览 (14)

2018-09-26 10:00:00

[首页](#)[关于我](#)[文章](#)[添加文章](#)[用户排行榜](#)

如何玩转微服务

作者: 语老师 javaSE 2018-09-26

关注我



CSDN

务，软件应用开发的新纪元

14年 Martin Fowler 在《MicroServices》论文中首次提出了微服务的概念。近些年，伴随互联网的日益发展，微服务在国内、甚至国际上的发展已达到一个新高潮。

微服务流行之前，SOA (Service Oriented Architecture) 被广泛熟知与采用。微服务基于IA 发展而来，但与之相比，微服务更易于理解，也更利于设计者、开发者的实践落地，它“面向服务”的设计思想实现得更加彻底。

标签云

Python

显示功能

您的评论或留言（必填）

评论

#2

万老师(Sept. 26, 2018, 12:29 p.m.)

文章很好！

#2

廖学长(Sept. 26, 2018, 12:30 p.m.)

搭配着项目一起进行就更好了

微视频



python基础

第一个python程序与数据存储01
print&input与变量和运算符01
字符串与循环中的while01
布尔&list与条件循环语句与tuple01
元组&字符串&字典01



爬虫

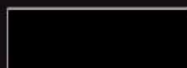
爬虫基本原理讲解 Urllib库基本使用
Requests库基本使用 正则表达式基础
BeautifulSoup库详解



Django

Django基本流程走通01
Django中的模型01
Django中的模板01
Django中的视图01
Django的高级使用01

文档简介



后台管理界面

Site administration

AUTHENTICATION AND AUTHORIZATION		
Groups	+ Add	Change
FIRST_APP		
分类	+ Add	Change
教程	+ Add	Change
教程目录	+ Add	Change
教程视频	+ Add	Change
教程视频目录内容	+ Add	Change
文章	+ Add	Change
普通用户	+ Add	Change
标签	+ Add	Change
用户	+ Add	Change

Recent actions

My actions

- ✖ python基础文章
- ✖ python基础文章
- + 第7讲图书查询功能实现教程视频目录内容
- + 第6讲图书添加功能实现教程视频目录内容
- + 第5讲图书类别修改功能实现教程视频目录内容
- + 第4讲图书类别查询功能实现教程视频目录内容
- + 第3讲Debug 详解教程视频目录内容
- + 第2讲图书类别添加功能实现