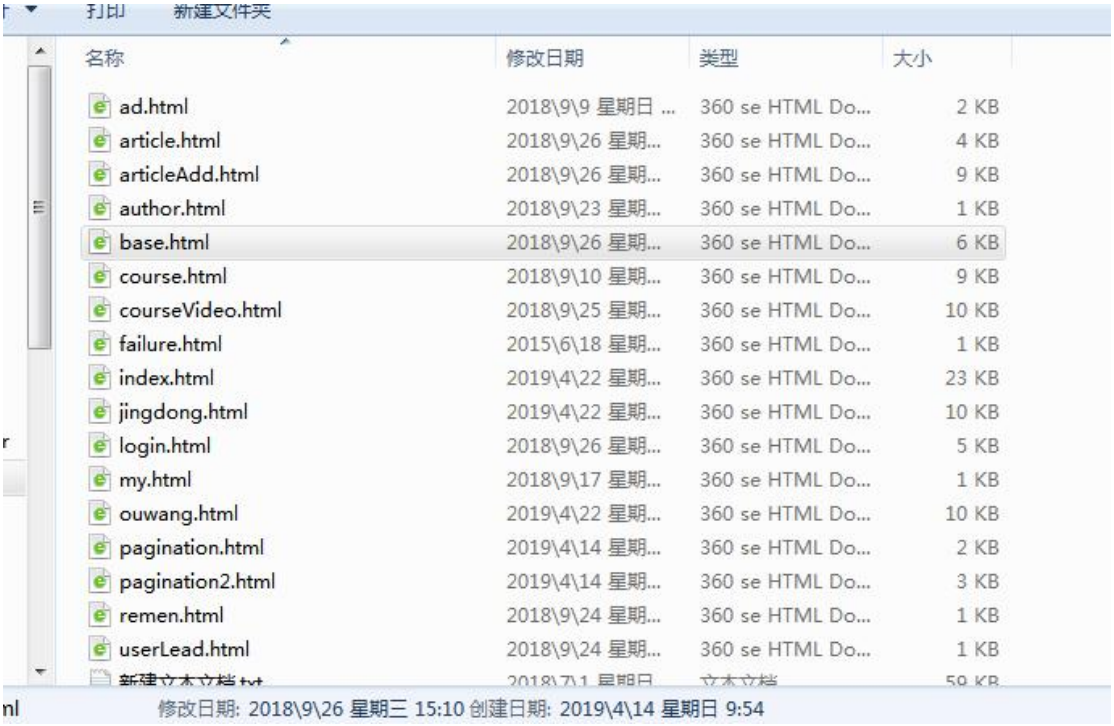
爬虫期中作业

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，



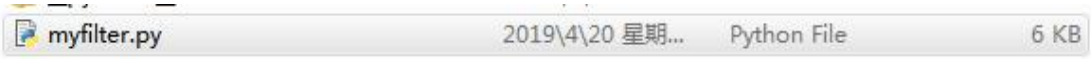
Get_ip 从网上爬取免费 ip

静态网站：瓯网使用 scrapy 爬取，

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。



```python
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.')
```

名称	修改日期	类型	大小
__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

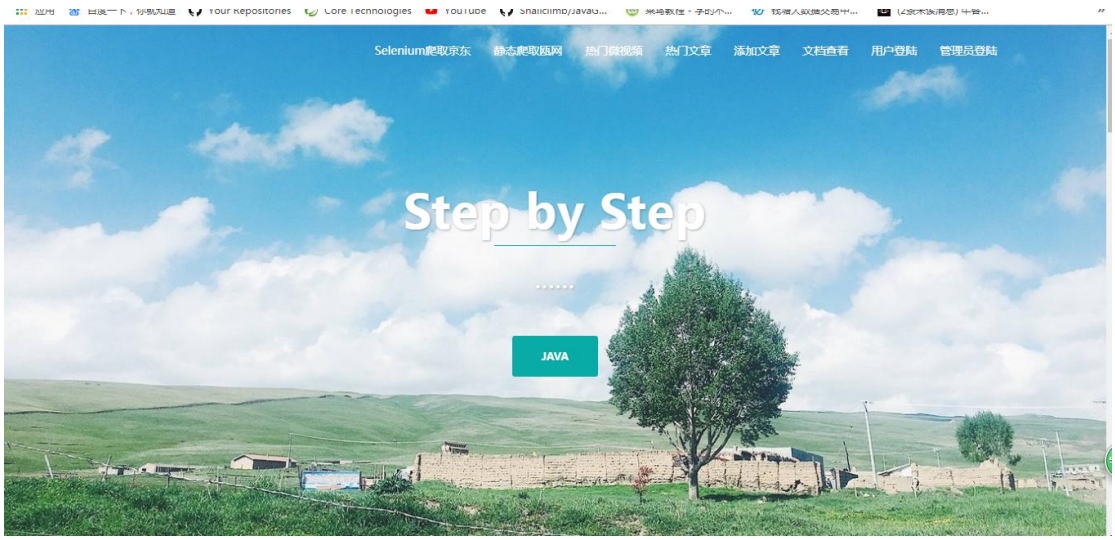
名称	修改日期	类型	大小
__pycache__	2019\4\22 星期...	文件夹	
1.py	2019\4\14 星期...	Python File	3 KB
2.py	2019\4\17 星期...	Python File	0 KB
chromedriver.exe	2018\3\20 星期...	应用程序	6,207 KB
db.py	2018\7\23 星期...	Python File	2 KB
jingdong.py	2019\4\19 星期...	Python File	3 KB
mysqlDb.py	2019\4\18 星期...	Python File	3 KB
setting.py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

```python
            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)
```

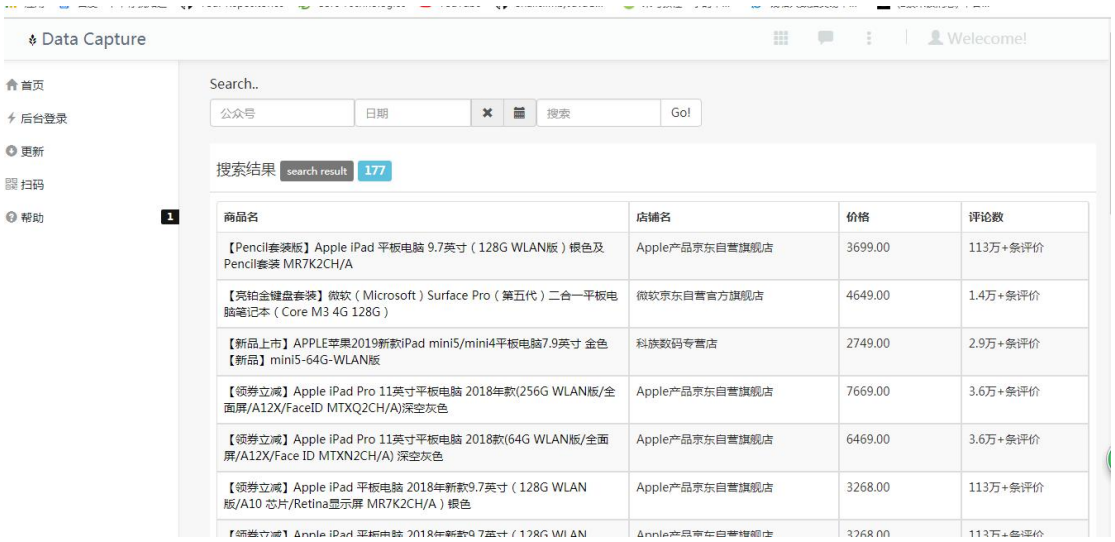
首页



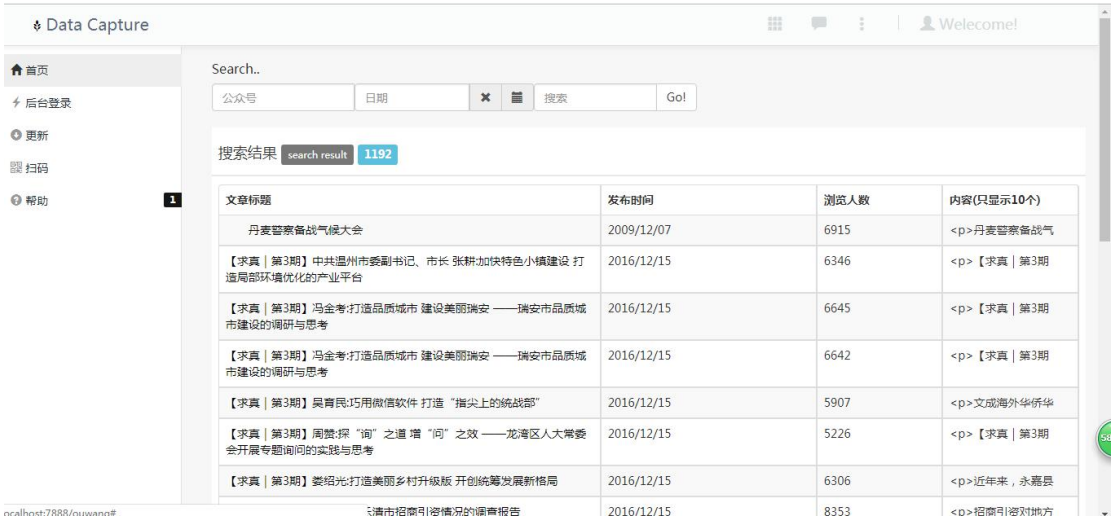
爬取京东

```
def jingdong(request):
    jingdong_list = Jingdong.objects.all().order_by('productname')
    totalNum = 0
    for i in jingdong_list:
        totalNum += 1
    #jingdong_list,thisPage1,thisPage2,thisPage3,thisPage4,thisPage5 = getPage(request, jingdong_list)
    paginator = Paginator(jingdong_list, 25) # 每页显示25条

    page = request.GET.get('page')
    try:
        jingdong_list = paginator.page(page)
    except PageNotAnInteger:
        # 如果请求的页数不是整数，返回第一页。
        jingdong_list = paginator.page(1)
    except EmptyPage:
        # 如果请求的页数不在合法的页数范围内，返回结果的最后一页。
        jingdong_list = paginator.page(paginator.num_pages)
    return render(request, 'jingdong.html', locals())
```



静态瓯网


Data Capture
Welcome!
首页
后台登录
更新
扫码
帮助
1
Search..
公众号
日期
搜索
Go!
搜索结果
search result
1192
文章标题
发布时间
浏览人数
内容(只显示10个)
丹麦警察备战气候大会
2009/12/07
6915
<p>丹麦警察备战气
【求真｜第3期】中共温州市委副书记、市长 张耕:加快特色小镇建设 打造局部环境优化的产业平台
2016/12/15
6346
<p>【求真｜第3期
【求真｜第3期】冯金考:打造品质城市 建设美丽瑞安 ——瑞安市品质城市建设的调研与思考
2016/12/15
6645
<p>【求真｜第3期
【求真｜第3期】冯金考:打造品质城市 建设美丽瑞安 ——瑞安市品质城市建设的调研与思考
2016/12/15
6642
<p>【求真｜第3期
【求真｜第3期】吴育民:巧用微信软件 打造“指尖上的统战部”
2016/12/15
5907
<p>文成海外华侨华
【求真｜第3期】周赞:探“询”之道 增“问”之效 ——龙湾区人大常委会开展专题询问的实践与思考
2016/12/15
5226
<p>【求真｜第3期
【求真｜第3期】娄绍光:打造美丽乡村升级版 开创统筹发展新格局
2016/12/15
6306
<p>近年来，永嘉县
2016/12/15
8353
<p>招商引资对地方


Selenium爬取京东
静态爬取瓯网
热门微视频
热门文章
添加文章
文档查看
用户登陆
管理员登陆
热门文章
集成化信息传播与服务博客
python
推送一些python小资源，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,springCould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
1
2
3
4
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


首页
关于我
文章
添加文章
用户排行榜
如何玩转微服务
老师
javaSE
2018-09-26
微服务，软件应用开发的新纪元
2014年 Martin Fowler 在《MicroServices》论文中首次提出了微服务的概念。近些年，伴随着互联网的日益发展，微服务在国内、甚至国际上的发展已达到一个新高潮。
在微服务流行之前，SOA（Service Oriented Architecture）被广泛熟知与采用。微服务基于SOA 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，它把“面向服务”的设计思想实现得更加彻底。
关注我
CSDN
标签云
Python

评论功能

评论 1 comment

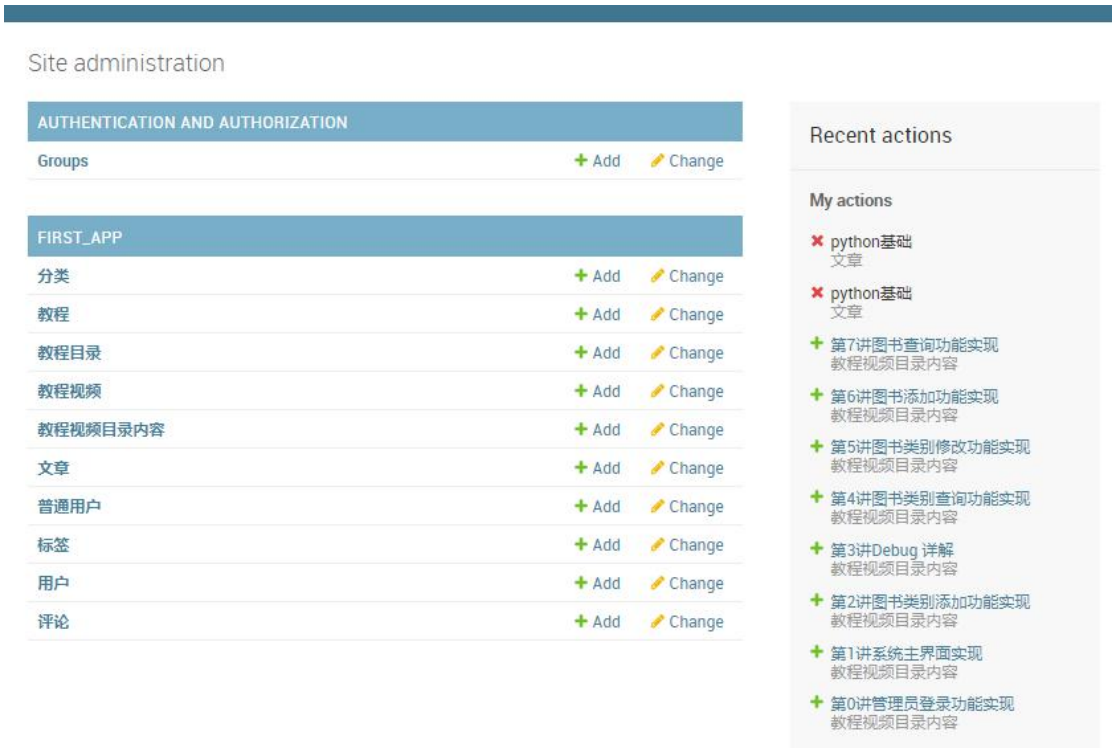
您的评论或留言（必填）

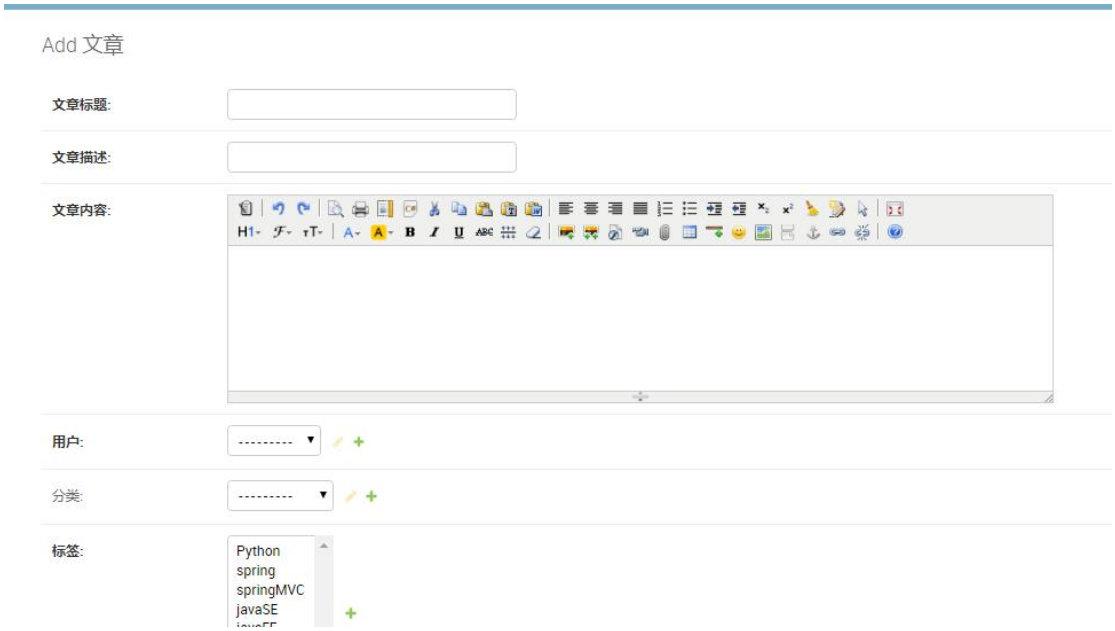
评论

#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

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
评论 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 教程视频目录内容
第1讲系统主界面实现 教程视频目录内容
第0讲管理员登录功能实现 教程视频目录内容


Add 文章
文章标题:
文章描述:
文章内容:
用户:
分类:
标签:
Python
spring
springMVC
javaSE