Python程序设计实验报告

数据获取实验

姓　名：王仕和

学　号：2019213681

日　期：2021.12.10

Python程序设计实验报告
数据获取实验

本实验完整代码附在压缩包中，并在下列链接可以查看：https://github.com/wshprimy/scrapy-spiders

1 作业1：北京链家

1.1 项目创建

首先创建项目

```
scrapy startproject Lianjia
```

然后进入项目目录Lianjia下，生成一个spider模板

```
cd Lianjia
scrapy genspider lianjia bj.lianjia.com
```

1.2 Item模块

编辑items.py，定义一个LianjiaItem类，用于描述爬取到数据的结构。

```
class LianjiaItem(scrapy.Item):
    name = scrapy.Field()
    total_price = scrapy.Field()
    area = scrapy.Field()
    unit_price = scrapy.Field()
    district = scrapy.Field()
    pass
```

1.3 spider模块

编辑spiders/lianjia.py，书写爬虫主程序。

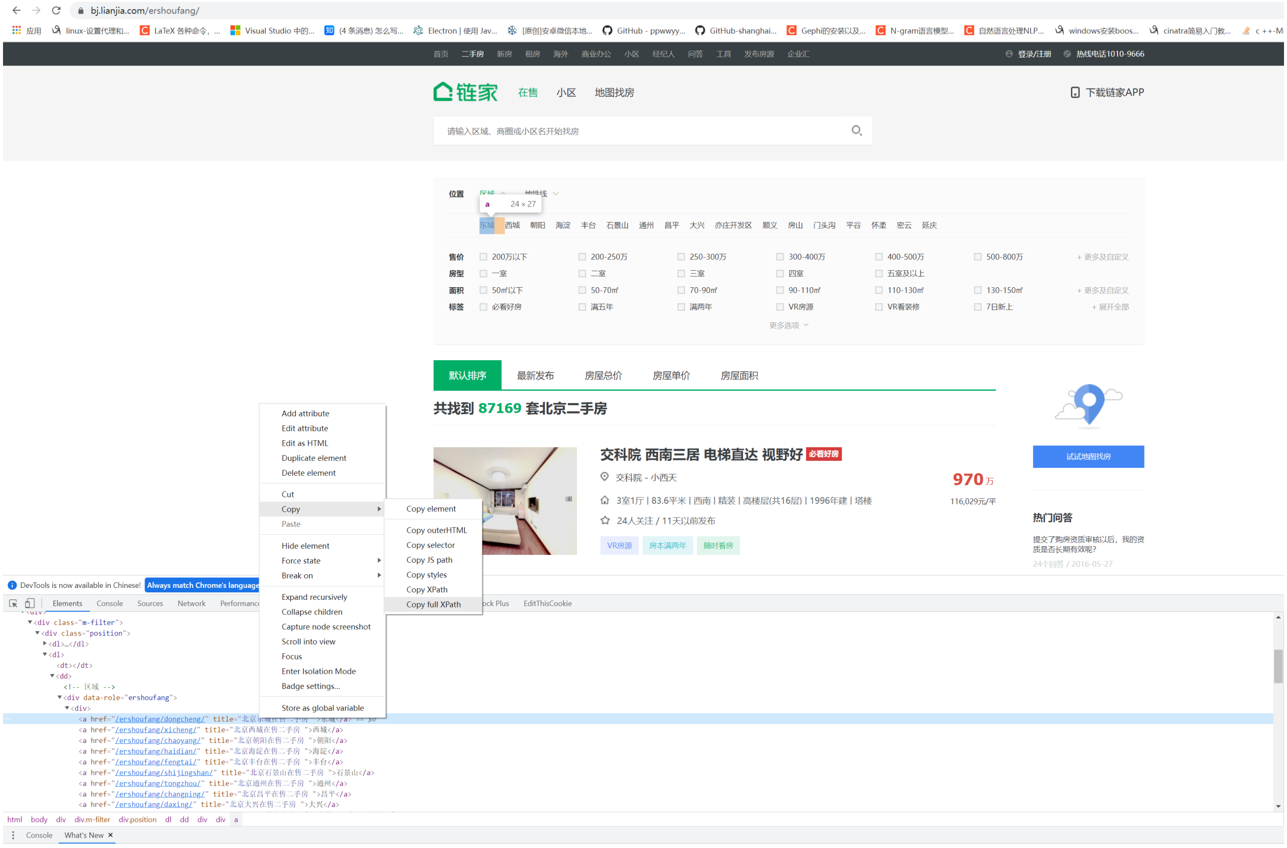
首先根据题目要求，输入起始url，并定义两个参数：

1. district_limit用于描述爬取的区数，我发现前四个区恰好为东城、西城、朝阳、海淀，故采取循环的方式。
2. page_limit用于描述每个区爬取的页面数。

```
class LianjiaSpider(scrapy.Spider):
    name = 'lianjia'
    allowed_domains = ['bj.lianjia.com']
    start_urls = ['https://bj.lianjia.com/ershoufang/']
    district_limit = 4
    page_limit = 5
```

在编写主程序时，我首先获取了题目要求爬取的四个区的xpath。

现在讲解通过浏览器获取xpath的过程，由于获取过程类似，以后将不再讲解。



```
<div class="m-filter">
  <div class="position">
    <dl>…</dl>
    <dl>
      <dt></dt>
      <dd>
        <!-- 区域 -->
        <div data-role="ershoufang">
          <div>
            <a href="/ershoufang/dongcheng/" title="北京东城在售二手房 ">东城</a> == $0
            <a href="/ershoufang/xicheng/" title="北京西城在售二手房 ">西城</a>
            <a href="/ershoufang/chaoyang/" title="北京朝阳在售二手房 ">朝阳</a>
            <a href="/ershoufang/haidian/" title="北京海淀在售二手房 ">海淀</a>
            <a href="/ershoufang/fengtai/" title="北京丰台在售二手房 ">丰台</a>
            <a href="/ershoufang/shijingshan/" title="北京石景山在售二手房 ">石景山</a>
            <a href="/ershoufang/tongzhou/" title="北京通州在售二手房 ">通州</a>
            <a href="/ershoufang/changping/" title="北京昌平在售二手房 ">昌平</a>
            <a href="/ershoufang/daxing/" title="北京大兴在售二手房 ">大兴</a>
```

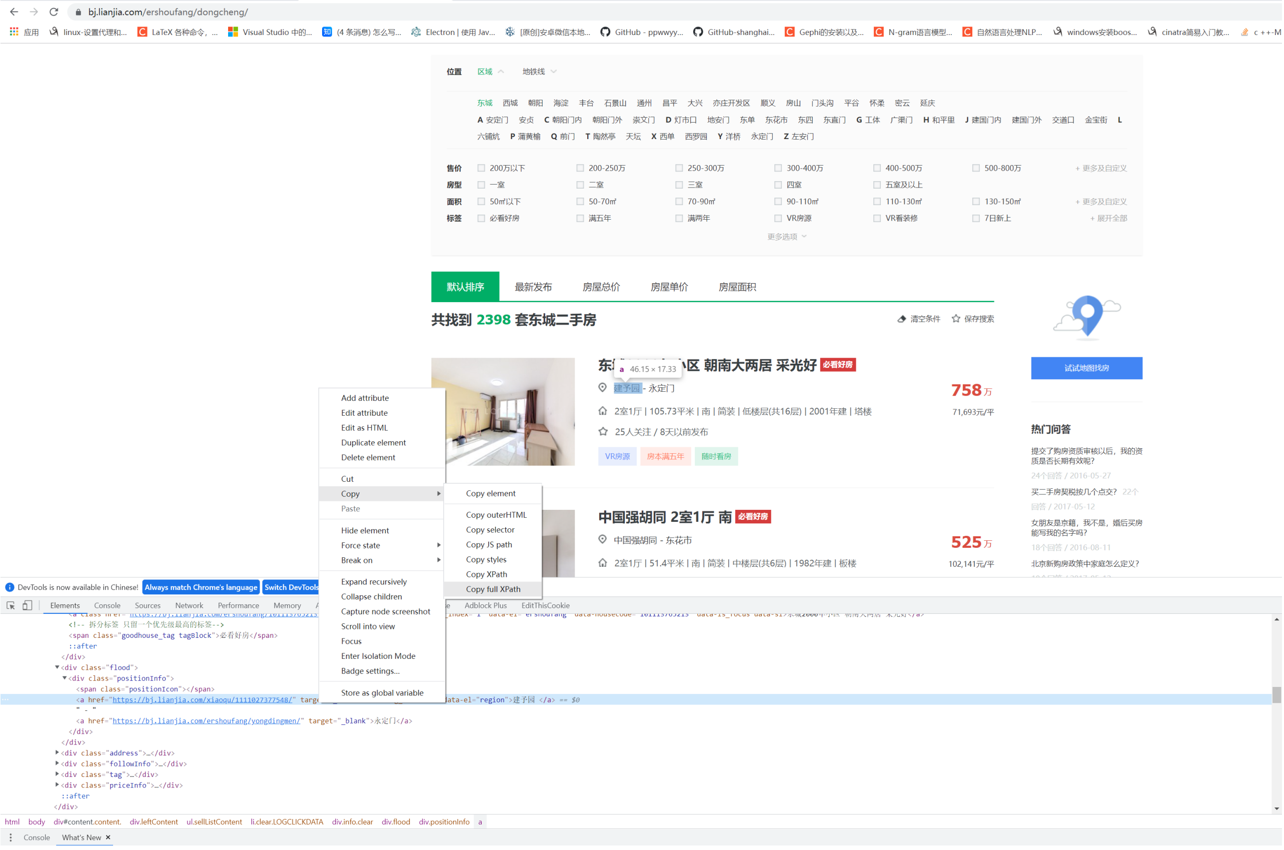
打开Console(F12)，使用元素选择器选中希望获得xpath的元素，右键对应的html语句，选择复制full xpath即可获得原始路径。

通过分析，发现a[]表示第几个区，子路径中第一个href标签表示指向了对应区的html链接，于是书写为下面代码中的xpath路径。

然后将next_page存储为需要的不同区的html链接，调用一个新的函数parse_district，并传递区名和page_id（第一次进入即第一页）。

```python
def parse(self, response):
    self.logger.info(f'Recieved url: {response.url}')
    for district in range(1, self.district_limit + 1):
        district_name = response.xpath(f'/html/body/div[3]/div/div[1]/dl[2]/dd/div[1]/div/a[{district}]/@href').get()
        district_name = district_name.split('/')[-2]
        next_page = response.url + district_name + '/'
        if next_page is not None:
            self.logger.info(f'Next url: {next_page}')
            yield response.follow(next_page, callback=self.parse_district, cb_kwargs={'district': district_name, 'page_id': 1})
```

进入对应区的网页后，首先确定各个房源的xpath。


bj.lianjia.com/ershoufang/dongcheng/
位置 区域 地铁线
东城 西城 朝阳 海淀 丰台 石景山 通州 昌平 大兴 亦庄开发区 顺义 房山 门头沟 平谷 怀柔 密云 延庆
默认排序 最新发布 房屋总价 房屋单价 房屋面积
共找到 2398 套东城二手房
朝南大两居 采光好 必看好房
建予园 - 永定门
758万
71,693元/平
2室1厅 | 105.73平米 | 南 | 简装 | 低楼层(共16层) | 2001年建 | 塔楼
25人关注 / 8天以前发布
中国强胡同 2室1厅 南 必看好房
中国强胡同 - 东花市
525万
102,141元/平
2室1厅 | 51.4平米 | 南 | 简装 | 中楼层(共6层) | 1982年建 | 板楼
试试地图找房
热门问答
Add attribute
Edit attribute
Edit as HTML
Duplicate element
Delete element
Cut
Copy
Paste
Hide element
Force state
Break on
Expand recursively
Collapse children
Capture node screenshot
Scroll into view
Focus
Enter Isolation Mode
Badge settings...
Store as global variable
Copy element
Copy outerHTML
Copy selector
Copy JS path
Copy styles
Copy XPath
Copy full XPath
DevTools is now available in Chinese! Always match Chrome's language Switch DevTools
Elements Console Sources Network Performance Memory Adblock Plus EditThisCookie
html body div#content.content. div.leftContent ul.sellListContent li.clear.LOGCLICKDATA div.info.clear div.flood div.positionInfo a
Console What's New

```
      <!-- 拆分标签 只留一个优先级最高的标签-->
      <span class="goodhouse_tag tagBlock">必看好房</span>
      ::after
    </div>
  ▼<div class="flood">
    ▼<div class="positionInfo">
        <span class="positionIcon"></span>
        <a href="https://bj.lianjia.com/xiaoqu/1111027377548/" targe          data-el="region">建予园 </a> == $0
        " - "
        <a href="https://bj.lianjia.com/ershoufang/yongdingmen/" target="_blank">永定门</a>
      </div>
    </div>
  ▶<div class="address">…</div>
  ▶<div class="followInfo">…</div>
  ▶<div class="tag">…</div>
  ▶<div class="priceInfo">…</div>
    ::after
  </div>
```

通过两个房源的xpath差异分析，获得公共url前缀，随后通过for循环爬取各个房源。

类似的，获得我们需要获取的各个数据的xpath，然后使用get方法获取，对于不同的数据，需要进行拼接或分离。

判断当前数据不全为空，yield返回结果，程序会自动进入piplines，这部分将在下一节中讲解。

然后获得进入网页时传入的page_id，若当前已大于等于最大页数限制，则不进行任何操作，随后自动退出。

若未达到最大页数，获得下一页的xpath，获取下一页html链接后yield返回下一页（会由scrapy自动执行）。

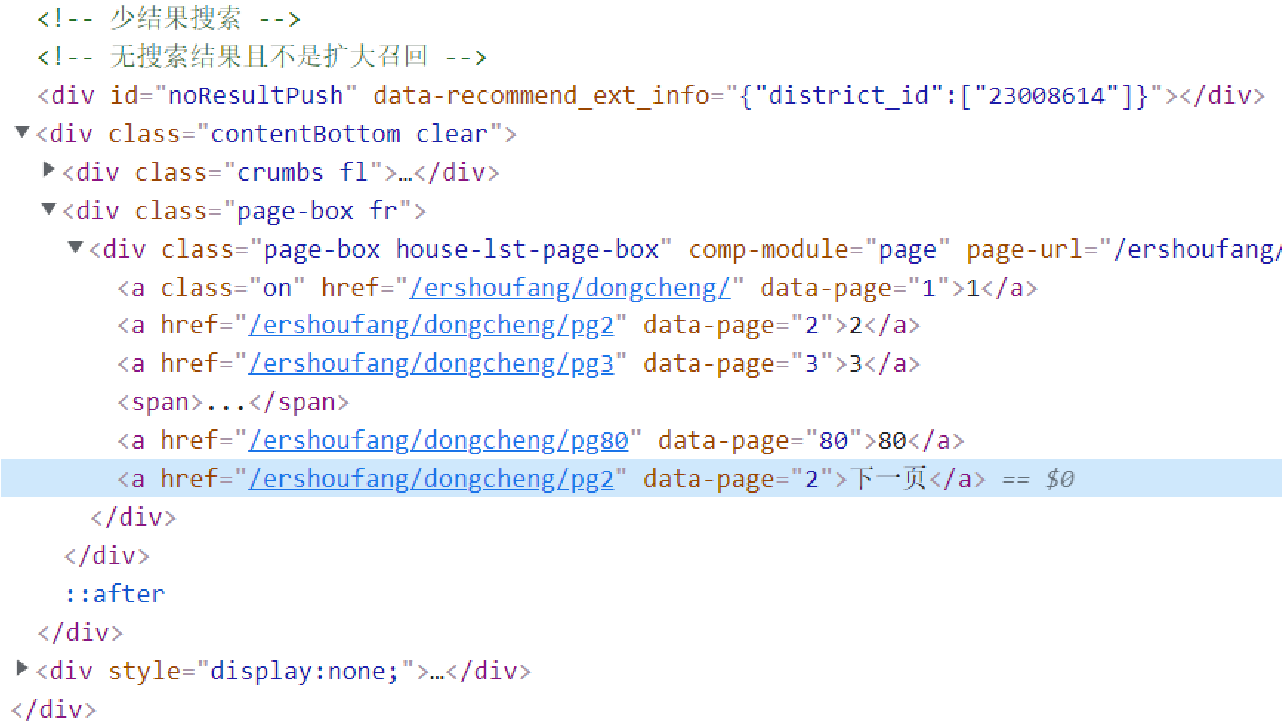
但发现执行失败，强制关闭浏览器html后发现下一页由js渲染得到一个html链接，但同时发现无js时{page}可以替换为当前页码 + 1，即下一页。

于是我们采取replace操作，详见下列代码。

在yield时提供下一页url，仍执行当前函数，参数：

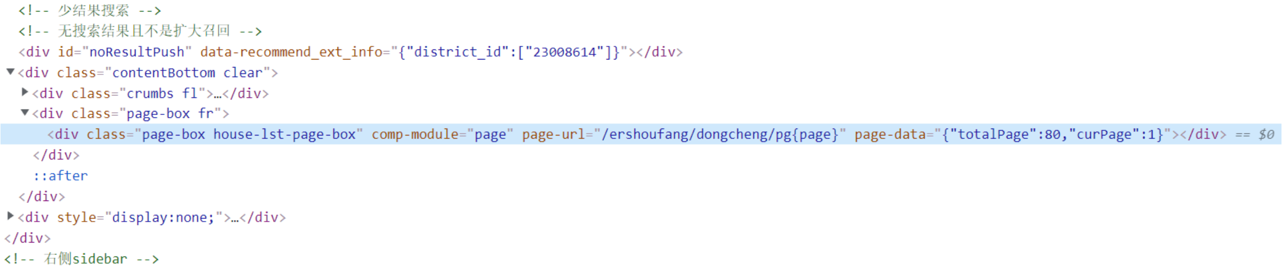
1. 区名
2. 页码（即当前页码+1）

关闭js前：



```
<!-- 少结果搜索 -->
<!-- 无搜索结果且不是扩大召回 -->
<div id="noResultPush" data-recommend_ext_info="{"district_id":["23008614"]}"></div>
▼<div class="contentBottom clear">
  ▶<div class="crumbs fl">…</div>
  ▼<div class="page-box fr">
    ▼<div class="page-box house-lst-page-box" comp-module="page" page-url="/ershoufang/
        <a class="on" href="/ershoufang/dongcheng/" data-page="1">1</a>
        <a href="/ershoufang/dongcheng/pg2" data-page="2">2</a>
        <a href="/ershoufang/dongcheng/pg3" data-page="3">3</a>
        <span>...</span>
        <a href="/ershoufang/dongcheng/pg80" data-page="80">80</a>
        <a href="/ershoufang/dongcheng/pg2" data-page="2">下一页</a> == $0
      </div>
    </div>
    ::after
  </div>
 ▶<div style="display:none;">…</div>
</div>
```

关闭js后：



```
<!-- 少结果搜索 -->
<!-- 无搜索结果且不是扩大召回 -->
<div id="noResultPush" data-recommend_ext_info="{"district_id":["23008614"]}"></div>
▼<div class="contentBottom clear">
  ▶<div class="crumbs fl">…</div>
  ▼<div class="page-box fr">
      <div class="page-box house-lst-page-box" comp-module="page" page-url="/ershoufang/dongcheng/pg{page}" page-data="{"totalPage":80,"curPage":1}"></div> == $0
    </div>
    ::after
  </div>
 ▶<div style="display:none;">…</div>
</div>
<!-- 右侧sidebar -->
```

```python
def parse_district(self, response, **kwargs):
    self.logger.info(f'Recieved url: {response.url}')
    district_name = kwargs['district']
    houses = response.xpath('/html/body/div[4]/div[1]/ul/li')
    for house in houses:
        item = LianjiaItem()
        item['name'] = house.xpath('./div[1]/div[2]/div/a[1]/text()').get().strip()
        item['total_price'] = house.xpath('./div[1]/div[6]/div[1]/span/text()').get() + \
            house.xpath('./div[1]/div[6]/div[1]/i[last()]/text()').get()
        item['area'] = house.xpath('./div[1]/div[3]/div/text()').get().split('|')[1].strip()
        item['unit_price'] = house.xpath('./div[1]/div[6]/div[2]/span/text()').get()
        item['district'] = district_name
        if item['name'] and item['total_price'] and item['area'] and item['unit_price']:
            yield item

    page_now = kwargs['page_id']
    self.logger.info(f'page_now = {page_now}')
    if page_now >= self.page_limit:
        pass
    else:
        next_page = response.url
        if page_now != 1:
            next_page = next_page[:-4]
        next_page = response.xpath('/html/body/div[4]/div[1]/div[7]/div[2]/div/@page-url').get()
        next_page = next_page.replace('{page}', str(page_now + 1))
        next_page = response.urljoin(next_page) + '/'
        # next_page = response.xpath(f'/html/body/div[10]/a[{page_now + 1}]/@href').get()
        # next_page = response.urljoin(next_page) + '/'
        self.logger.info(f'Now url: {response.url}')
        self.logger.info(f'Next url: {next_page}')
        yield response.follow(next_page, callback=self.parse_district, cb_kwargs={'district': district_name, 'page_id': page_now + 1})
```

1.4 pipelines模块

编辑pipelines.py，定义一个LianjiaPipeline类，用于处理yield item返回的已爬取数据。

首先打开settings.py，将pipelines启用，优先级为默认。

```
ITEM_PIPELINES = {
   'Lianjia.pipelines.LianjiaPipeline': 300,
}
```

LianjiaPipeline共有两个函数：

1. process_item用于处理每次yield item返回的已爬取数据
2. close_spider用于最后将数据写入文件

process_item首先将item_temp转换为dict，然后通过判断items字典中是否存在值为district的key，若没有则创建一个，values为一个list。
然后将不需要获取的district去除后append进入对应数组。

close_spider首先创建一个csv表头，然后将表头和数据分别写入CSV中。
这里我为了方便在Windows下使用office-excel查看，使用了GBK编码。

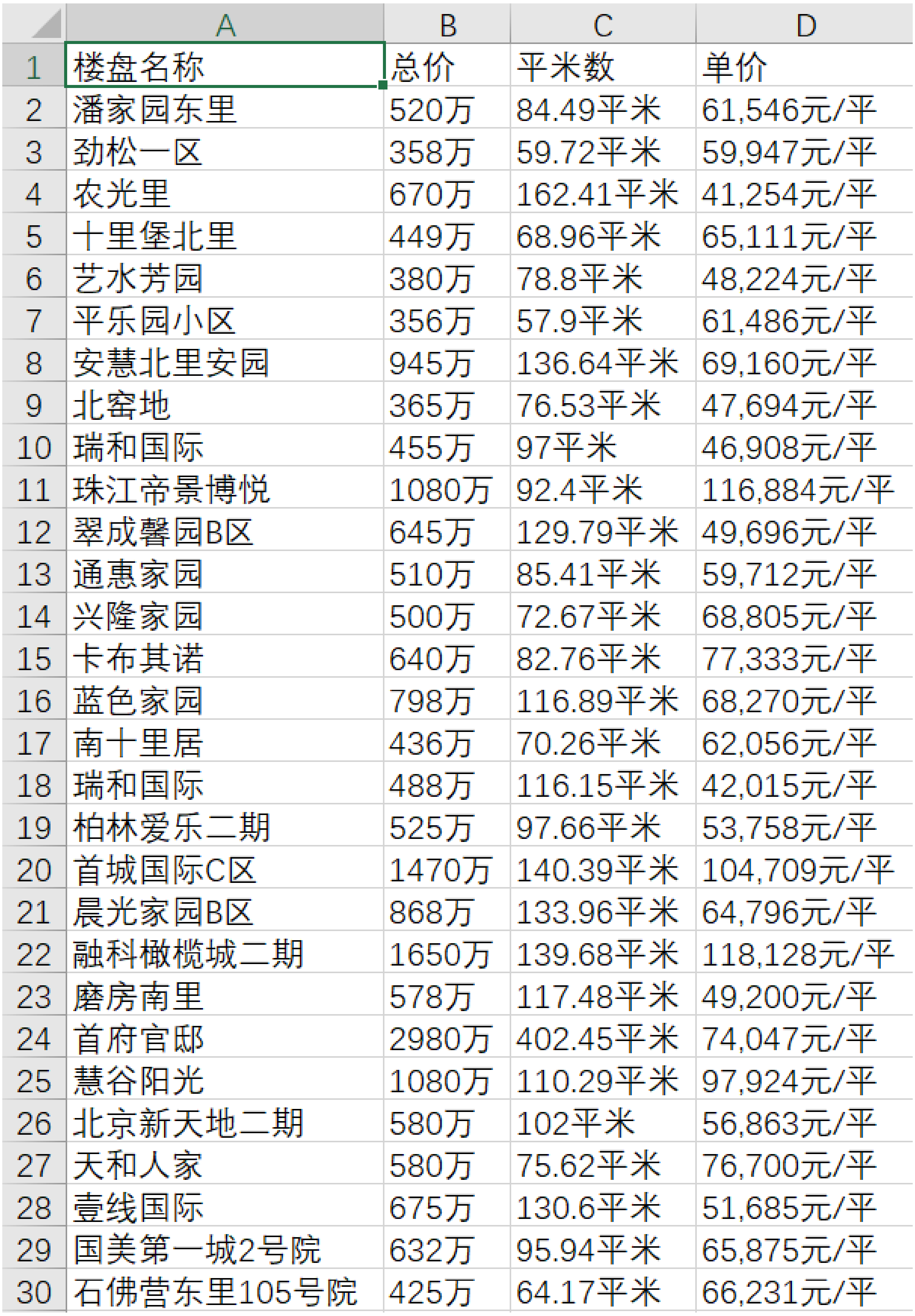
```
class LianjiaPipeline:
    items = {}

    def process_item(self, item, spider):
        item_temp = ItemAdapter(item).asdict()
        district = item_temp['district']
        if district not in self.items:
            self.items[district] = []
        item_list = list(item_temp.values())[:-1]
        self.items[district].append(item_list)
        return item

    def close_spider(self, spider):
        headers = ['楼盘名称', '总价', '平米数', '单价']
        for district_name, items in self.items.items():
            with open(f'{district_name}.csv', 'w', newline='', encoding='GBK') as file:
                file_csv = csv.writer(file)
                file_csv.writerow(headers)
                file_csv.writerows(items)
                file.close()
```

至此，北京链家的爬取到此结束。

1.5 数据



	A	B	C	D
1	楼盘名称	总价	平米数	单价
2	潘家园东里	520万	84.49平米	61,546元/平
3	劲松一区	358万	59.72平米	59,947元/平
4	农光里	670万	162.41平米	41,254元/平
5	十里堡北里	449万	68.96平米	65,111元/平
6	艺水芳园	380万	78.8平米	48,224元/平
7	平乐园小区	356万	57.9平米	61,486元/平
8	安慧北里安园	945万	136.64平米	69,160元/平
9	北窑地	365万	76.53平米	47,694元/平
10	瑞和国际	455万	97平米	46,908元/平
11	珠江帝景博悦	1080万	92.4平米	116,884元/平
12	翠成馨园B区	645万	129.79平米	49,696元/平
13	通惠家园	510万	85.41平米	59,712元/平
14	兴隆家园	500万	72.67平米	68,805元/平
15	卡布其诺	640万	82.76平米	77,333元/平
16	蓝色家园	798万	116.89平米	68,270元/平
17	南十里居	436万	70.26平米	62,056元/平
18	瑞和国际	488万	116.15平米	42,015元/平
19	柏林爱乐二期	525万	97.66平米	53,758元/平
20	首城国际C区	1470万	140.39平米	104,709元/平
21	晨光家园B区	868万	133.96平米	64,796元/平
22	融科橄榄城二期	1650万	139.68平米	118,128元/平
23	磨房南里	578万	117.48平米	49,200元/平
24	首府官邸	2980万	402.45平米	74,047元/平
25	慧谷阳光	1080万	110.29平米	97,924元/平
26	北京新天地二期	580万	102平米	56,863元/平
27	天和人家	580万	75.62平米	76,700元/平
28	壹线国际	675万	130.6平米	51,685元/平
29	国美第一城2号院	632万	95.94平米	65,875元/平
30	石佛营东里105号院	425万	64.17平米	66,231元/平

各区前50条数据附在随后的报告中，完整数据参见完整代码。

1.5.1 朝阳区

```
楼盘名称,总价,平米数,单价
潘家园东里,520万,84.49平米,"61,546元/平"
劲松一区,358万,59.72平米,"59,947元/平"
农光里,670万,162.41平米,"41,254元/平"
十里堡北里,449万,68.96平米,"65,111元/平"
艺水芳园,380万,78.8平米,"48,224元/平"
平乐园小区,356万,57.9平米,"61,486元/平"
安慧北里安园,945万,136.64平米,"69,160元/平"
北窑地,365万,76.53平米,"47,694元/平"
瑞和国际,455万,97平米,"46,908元/平"
珠江帝景博悦,1080万,92.4平米,"116,884元/平"
翠成馨园B区,645万,129.79平米,"49,696元/平"
通惠家园,510万,85.41平米,"59,712元/平"
兴隆家园,500万,72.67平米,"68,805元/平"
卡布其诺,640万,82.76平米,"77,333元/平"
蓝色家园,798万,116.89平米,"68,270元/平"
南十里居,436万,70.26平米,"62,056元/平"
瑞和国际,488万,116.15平米,"42,015元/平"
柏林爱乐二期,525万,97.66平米,"53,758元/平"
首城国际C区,1470万,140.39平米,"104,709元/平"
晨光家园B区,868万,133.96平米,"64,796元/平"
融科橄榄城二期,1650万,139.68平米,"118,128元/平"
磨房南里,578万,117.48平米,"49,200元/平"
首府官邸,2980万,402.45平米,"74,047元/平"
慧谷阳光,1080万,110.29平米,"97,924元/平"
北京新天地二期,580万,102平米,"56,863元/平"
天和人家,580万,75.62平米,"76,700元/平"
壹线国际,675万,130.6平米,"51,685元/平"
国美第一城2号院,632万,95.94平米,"65,875元/平"
石佛营东里105号院,425万,64.17平米,"66,231元/平"
胜古南里,415万,56平米,"74,108元/平"
小关北里,539万,64.36平米,"83,748元/平"
枣营南里,420万,54.4平米,"77,206元/平"
劲松七区,295万,50.8平米,"58,071元/平"
石佛营西里小区,398万,61.07平米,"65,172元/平"
柏林爱乐二期,565万,100.67平米,"56,124元/平"
安慧里一区,560万,68.47平米,"81,788元/平"
华威西里,285万,46.76平米,"60,950元/平"
晨光家园A区,545万,82.89平米,"65,750元/平"
定福庄北里2号院,265万,57.84平米,"45,817元/平"
电建南院,380万,69.13平米,"54,969元/平"
和平街十二区,325万,38.41平米,"84,614元/平"
美景东方,860万,143.77平米,"59,818元/平"
后现代城C区,778万,99.19平米,"78,436元/平"
南太平庄北巷,218万,64.36平米,"33,872元/平"
华严北里中科院,610万,61.17平米,"99,723元/平"
```

```
国典华园,1370万,139.71平米,"98,061元/平"
翠成馨园E区,415万,97.08平米,"42,749元/平"
农光东里,355万,48.41平米,"73,332元/平"
望京新城,795万,115.45平米,"68,861元/平"
望京西园三区,610万,82.6平米,"73,850元/平"
```

1.5.2 东城区

```
楼盘名称,总价,平米数,单价
建予园,758万,105.73平米,"71,693元/平"
中国强胡同,525万,51.4平米,"102,141元/平"
东交民巷32号院,596万,49.98平米,"119,248元/平"
本家润园一期,1850万,154.09平米,"120,060元/平"
东直门内北小街8号院,1040万,90.76平米,"114,588元/平"
新景家园东区,880万,79.93平米,"110,097元/平"
金鱼池西区,670万,61.48平米,"108,979元/平"
工体北里,655万,57.02平米,"114,872元/平"
安馨园,2750万,311.87平米,"88,178元/平"
东花市北里中区,1100万,139.04平米,"79,114元/平"
华龙美晟,582万,75.51平米,"77,076元/平"
三元街甲19号院,435万,52.93平米,"82,185元/平"
忠实里,720万,82.2平米,"87,592元/平"
龙潭北里,625万,65.37平米,"95,610元/平"
中海紫御公馆,1200万,91.44平米,"131,234元/平"
阳光都市,2200万,240.55平米,"91,458元/平"
太华公寓,1249万,231.13平米,"54,039元/平"
国瑞城中区,1699万,157.42平米,"107,928元/平"
西革新里110号院,370万,42.32平米,"87,430元/平"
民旺园,730万,56.64平米,"128,885元/平"
安外花园,679万,60.76平米,"111,752元/平"
富贵园一区,2030万,144.71平米,"140,281元/平"
小黄庄一区,1130万,90.28平米,"125,167元/平"
景泰西里西区,445万,56.82平米,"78,318元/平"
望陶园小区,680万,82.42平米,"82,505元/平"
景泰西里西区,705万,88.23平米,"79,905元/平"
城市亮点,763万,84.46平米,"90,339元/平"
左安浦园,768万,89.45平米,"85,859元/平"
复康南里,450万,47.55平米,"94,638元/平"
安德路乙61号院,696万,59.32平米,"117,330元/平"
培新街乙5号院,625万,55.15平米,"113,328元/平"
光明楼,518万,43.27平米,"119,714元/平"
金鱼池西区,660万,61.33平米,"107,615元/平"
望陶园小区,675万,82.42平米,"81,898元/平"
富贵园三区,728万,76.46平米,"95,214元/平"
东直门北大街,542万,48.26平米,"112,309元/平"
东直门内大街,815万,71.45平米,"114,066元/平"
沙滩后街55号院,780万,56.91平米,"137,059元/平"
北河沿大街,640万,53.03平米,"120,687元/平"
```

```
左安浦园,880万,97.63平米,"90,137元/平"
菊儿胡同,1280万,92.35平米,"138,604元/平"
新景家园东区,1160万,88.45平米,"131,148元/平"
西水井胡同,1150万,96.74平米,"118,876元/平"
海运仓小区,1120万,95.18平米,"117,672元/平"
花市枣苑二期,786万,73.01平米,"107,657元/平"
丽水湾畔家园,1100万,130平米,"84,616元/平"
华龙美晟,698万,67.53平米,"103,362元/平"
兴隆都市馨园,405万,29.36平米,"137,943元/平"
建国门北大街5号,3100万,508.74平米,"60,935元/平"
京城仁合,1050万,114.51平米,"91,696元/平"
```

1.5.3 海淀区

```
楼盘名称,总价,平米数,单价
交科院,970万,83.6平米,"116,029元/平"
会城门小区,415万,41.5平米,"100,000元/平"
缘溪堂,4200万,418.05平米,"100,467元/平"
普惠北里,595万,60.9平米,"97,702元/平"
铭科苑,645万,114.24平米,"56,461元/平"
信悦华庭,1288万,147.87平米,"87,104元/平"
学院南路60号院,628万,58.9平米,"106,622元/平"
中纺宿舍南院,670万,65.1平米,"102,919元/平"
御墅临枫锦园一区,2050万,212.53平米,"96,457元/平"
橙色年代,669万,94.22平米,"71,005元/平"
碧森里,406万,41.59平米,"97,620元/平"
阜北小区,500万,56.73平米,"88,137元/平"
翠微东里,730万,53.6平米,"136,195元/平"
建设部大院,800万,69.02平米,"115,909元/平"
门头馨村北二区,720万,126.12平米,"57,089元/平"
清华园,679万,50.9平米,"133,399元/平"
永泰东里,650万,80.56平米,"80,686元/平"
皂君东里,838万,76.3平米,"109,830元/平"
富海中心,1460万,146.56平米,"99,618元/平"
北洼西里,1130万,116.2平米,"97,247元/平"
玉泉路16号院,750万,81.9平米,"91,576元/平"
幸福时光,445万,58.77平米,"75,719元/平"
燕北园,588万,58.3平米,"100,858元/平"
永泰庄甲6号院,540万,63.6平米,"84,906元/平"
西木小区,535万,53.5平米,"100,000元/平"
华盛家园,1320万,135.71平米,"97,267元/平"
中关村南大街乙8号,1314万,95.54平米,"137,535元/平"
苏州街77号院,635万,57.8平米,"109,862元/平"
恩济里小区,585万,63.4平米,"92,272元/平"
百朗园,930万,155.93平米,"59,643元/平"
10号名邸,1697万,207.02平米,"81,973元/平"
万寿路12号院,690万,64.1平米,"107,645元/平"
智学苑,745万,114.5平米,"65,066元/平"
```

```
铁东小区,510万,51.1平米,"99,805元/平"
紫金庄园,950万,92.8平米,"102,371元/平"
人济山庄,1568万,161.54平米,"97,066元/平"
昌运宫,658万,72.6平米,"90,634元/平"
复兴路79号院,540万,66.4平米,"81,326元/平"
北辰香麓,1360万,248.02平米,"54,835元/平"
正源尚峰尚水,965万,124.11平米,"77,754元/平"
国悦府,3250万,396.45平米,"81,978元/平"
领秀新硅谷2号院,1688万,215.97平米,"78,160元/平"
知春路82号院,799万,54.3平米,"147,146元/平"
甘家口小区,680万,62平米,"109,678元/平"
宝盛里,699万,147.74平米,"47,313元/平"
羊坊店路3号院,600万,57.27平米,"104,767元/平"
复兴路61号院,660万,57.6平米,"114,584元/平"
文慧园,670万,61.2平米,"109,478元/平"
德胜门西大街甲5号,930万,90.5平米,"102,763元/平"
豪景大厦,1180万,167.54平米,"70,431元/平"
```

1.5.4 西城区

```
楼盘名称,总价,平米数,单价
羊肉胡同120号院,1120万,81.7平米,"137,087元/平"
贵都国际,650万,181.53平米,"35,807元/平"
小红庙,580万,56.96平米,"101,826元/平"
新德街20号,996万,67.5平米,"147,556元/平"
新文化街,850万,67.3平米,"126,301元/平"
华龙大厦,1600万,210.44平米,"76,032元/平"
新街口西里一区,1180万,84.59平米,"139,497元/平"
北礼士路,790万,63.5平米,"124,410元/平"
富国里小区,1155万,80平米,"144,375元/平"
月坛西街西里,1259万,81.8平米,"153,912元/平"
新街口西里三区,1005万,84.86平米,"118,431元/平"
北营房西里,750万,57.5平米,"130,435元/平"
北礼士路66号院,750万,52.1平米,"143,954元/平"
百万庄中里,605万,41.7平米,"145,084元/平"
三义里,638万,57.21平米,"111,519元/平"
德胜里一区,1200万,79.7平米,"150,565元/平"
裕中西里,900万,63.3平米,"142,181元/平"
铁二区,1050万,69.9平米,"150,215元/平"
陶然居,1150万,115.49平米,"99,576元/平"
新街口西里三区,750万,55.92平米,"134,121元/平"
六铺炕二区,1060万,81.9平米,"129,427元/平"
展览路,910万,70.4平米,"129,262元/平"
信和嘉园,2980万,308.5平米,"96,597元/平"
立恒名苑,1100万,147.19平米,"74,734元/平"
阳光丽景,2390万,146.71平米,"162,907元/平"
马连道中里,649万,59.35平米,"109,352元/平"
新安北里,690万,61.32平米,"112,525元/平"
```

```
新外大街10号院,765万,60.4平米,"126,656元/平"
新外大街6号院,830万,51.3平米,"161,794元/平"
六铺炕一区6号院,1260万,150.1平米,"83,945元/平"
阜成门南大街,615万,40.1平米,"153,367元/平"
南纬路南巷9号院,860万,75.23平米,"114,317元/平"
政泰家园,555万,67.44平米,"82,296元/平"
西便门东里,660万,56.66平米,"116,485元/平"
三里河二区A区,1750万,117.6平米,"148,810元/平"
三里河南二巷,789万,54.9平米,"143,716元/平"
格调小区,1590万,138.79平米,"114,562元/平"
车站西街17号院,490万,46.8平米,"104,701元/平"
车站西街13号院,390万,48.97平米,"79,641元/平"
玉桃园三区,720万,57.7平米,"124,784元/平"
新街口西里二区,1200万,101.6平米,"118,111元/平"
车站西街15号院,653万,57.41平米,"113,744元/平"
小红庙,720万,66.96平米,"107,527元/平"
木樨地,985万,69.4平米,"141,931元/平"
车站西街,516万,49.28平米,"104,708元/平"
红居南街,888万,93.31平米,"95,167元/平"
天宁寺西里,580万,58.05平米,"99,914元/平"
新文化街,660万,53.63平米,"123,066元/平"
新德街35号院,740万,45.3平米,"163,356元/平"
温馨家园,1416万,80.5平米,"175,901元/平"
```

2 作业2：学堂在线

2.1 项目创建

首先创建项目

```
scrapy startproject Xtzx
```

然后进入项目目录Xtzx下，生成一个spider模板

```
cd Xtzx
scrapy genspider xtzx www.xuetangx.com
```

2.2 Item模块

编辑items.py，定义一个XtzxItem类，用于描述爬取到数据的结构。

```
class XtzxItem(scrapy.Item):
    school_name = scrapy.Field()
    total_courses = scrapy.Field()
    pass
```

2.3 spider模块

编辑spiders/xtzx.py，书写爬虫主程序。

首先根据题目要求，输入起始url，并定义以一个参数：

1. page_limit用于描述爬取的页面总数。

```python
class XtzxSpider(scrapy.Spider):
    name = 'xtzx'
    allowed_domains = ['www.xuetangx.com']
    start_urls = ['https://www.xuetangx.com/university/all/']
    page_limit = 36
```

进入页面获取xpath后发现爬取失败，强制关闭浏览器js后发现该网页为纯动态网页，故采用selenium进行爬取。

引入下列库：

```python
import time
import scrapy
from Xtzx.items import XtzxItem
from scrapy.http import HtmlResponse
from selenium import webdriver
```

首先定义启动参数，然后根据启动参数启动webdriver
首先使用浏览器get起始网页，隐式等待最长五秒后开始执行，
第一个循环表示点击下一页按钮的次数，通过真实网页观察发现其共有36的页面，已定义为变量page_limit，n个页面共需n - 1次下一页，故使用range(0, self.page_limit)进行循环。
然后通过HtmlResponse获得当前webdriver的current_url作为网页url，driver.page_source作为网页html，编码格式采用utf-8。
随后的爬取各个学校过程与“作业1：北京链家”中类似，详见上文。

爬取结束后，发现下一页按钮为一个button，没有指向任何html链接，故使用webdriver提供的根据xpath定位位置，执行点击，模拟人类下一页操作。
因实验中implicitly_wait为出现部分问题，故采取强制睡眠2s等待js渲染后继续爬取。

```python
def parse(self, response):
    driver_options = webdriver.ChromeOptions()
    driver_options.add_argument('--headless')
    driver_options.add_argument('--disable-gpu')
    driver = webdriver.Chrome(options = driver_options)

    driver.get(response.url)
    driver.implicitly_wait(5)
    for i in range(0, self.page_limit):
        response = HtmlResponse(url=driver.current_url, body=driver.page_source, encoding='utf-8')
        schools = response.xpath('/html/body/div[1]/div/div[2]/div[1]/div[2]/div')
        for school in schools:
            item = XtzxItem()
            item['school_name'] = school.xpath('./div[1]/p[1]/text()').get()
            item['total_courses'] = school.xpath('./p/text()').get()
            if item['school_name'] and item['total_courses']:
                yield item
        button = driver.find_element_by_xpath('/html/body/div[1]/div/div[2]/div[1]/div[3]/button[2]')
        button.click()
        time.sleep(2)
```

2.4 pipelines模块

编辑pipelines.py，定义一个XtzxPipeline类，用于处理yield item返回的已爬取数据。

首先打开settings.py，将pipelines启用，优先级为默认。

```
ITEM_PIPELINES = {
   'Xtzx.pipelines.XtzxPipeline': 300,
}
```

XtzxPipeline共有三个函数：

1. open_spider用于初始化，将self.result定义为list，用于保存爬取的数据，并在最后转换为json
2. process_item用于处理每次yield item返回的已爬取数据
3. close_spider用于最后将数据写入文件

process_item首先将item_temp转换为dict，然后将其append进入数组。
然后直接使用json.dumps将包含若干dict的数据转换为json类型字符串，
随后使用file.write将字符串写入文件中。

```
class XtzxPipeline:
    def open_spider(self, spider):
        self.result = []

    def process_item(self, item, spider):
        item_temp = ItemAdapter(item).asdict()
        print(type(item_temp))
        self.result.append(item_temp)
        return item

    def close_spider(self, spider):
        with open('xuetangzaixian.json', 'w', encoding='utf-8') as file:
            json_str = json.dumps(self.result, ensure_ascii=False, indent=4)
            file.write(json_str)
            file.close()
```

至此，学堂在线的爬取到此结束。

2.5 数据

前50条数据附在随后的报告中，完整数据参见完整代码。

```
[
    {
        "school_name": "清华大学",
        "total_courses": "426门课程"
    },
    {
        "school_name": "北京大学",
        "total_courses": "24门课程"
    },
    {
        "school_name": "北京师范大学",
        "total_courses": "63门课程"
    },
    {
        "school_name": "国防科技大学",
        "total_courses": "15门课程"
    },
    {
        "school_name": "西安交通大学",
        "total_courses": "102门课程"
    },
    {
        "school_name": "哈尔滨工业大学",
        "total_courses": "38门课程"
    },
    {
        "school_name": "华南理工大学",
        "total_courses": "49门课程"
    },
    {
        "school_name": "南开大学",
        "total_courses": "41门课程"
    },
    {
        "school_name": "复旦大学",
        "total_courses": "9门课程"
    },
    {
        "school_name": "南京大学",
        "total_courses": "9门课程"
    },
    {
        "school_name": "中国科学技术大学",
        "total_courses": "7门课程"
    },
    {
        "school_name": "圣彼得堡国立大学",
        "total_courses": "15门课程"
    },
    {
        "school_name": "重庆大学",
        "total_courses": "47门课程"
    },
    {
        "school_name": "暨南大学",
        "total_courses": "75门课程"
    },
    {
        "school_name": "东北大学",
        "total_courses": "34门课程"
    },
    {
        "school_name": "中南大学",
        "total_courses": "41门课程"
    },
    {
        "school_name": "中国农业大学",
        "total_courses": "37门课程"
    },
    {
        "school_name": "云南大学",
        "total_courses": "39门课程"
    },
    {
        "school_name": "山东大学",
        "total_courses": "69门课程"
    },
    {
        "school_name": "西北工业大学",
        "total_courses": "28门课程"
    },
    {
        "school_name": "四川大学",
        "total_courses": "34门课程"
    },
    {
        "school_name": "大连理工大学",
        "total_courses": "14门课程"
    },
    {
        "school_name": "SDGAcademy",
        "total_courses": "13门课程"
    },
    {
        "school_name": "湖南大学",
        "total_courses": "17门课程"
    },
    {
        "school_name": "天津大学",
        "total_courses": "16门课程"
    },
    {
        "school_name": "武汉大学",
        "total_courses": "9门课程"
    },
    {
        "school_name": "上海交通大学",
        "total_courses": "4门课程"
    },
    {
        "school_name": "浙江大学",
        "total_courses": "5门课程"
    },
    {
        "school_name": "北京体育大学",
        "total_courses": "29门课程"
    },
    {
        "school_name": "河北工业大学",
        "total_courses": "22门课程"
    },
    {
        "school_name": "中国石油大学（北京）",
        "total_courses": "23门课程"
    },
    {
        "school_name": "南昌大学",
        "total_courses": "24门课程"
    },
    {
        "school_name": "北京理工大学",
        "total_courses": "51门课程"
    },
    {
        "school_name": "中国传媒大学",
        "total_courses": "23门课程"
    },
    {
        "school_name": "北京交通大学",
        "total_courses": "37门课程"
    },
    {
        "school_name": "宁夏大学",
        "total_courses": "14门课程"
    },
    {
        "school_name": "郑州大学",
        "total_courses": "21门课程"
    },
    {
        "school_name": "北京林业大学",
        "total_courses": "11门课程"
    },
    {
        "school_name": "大连海事大学",
        "total_courses": "18门课程"
    },
    {
        "school_name": "中央民族大学",
        "total_courses": "18门课程"
    },
    {
        "school_name": "华北电力大学",
        "total_courses": "33门课程"
    },
    {
        "school_name": "天津医科大学",
        "total_courses": "6门课程"
    },
    {
        "school_name": "武汉理工大学",
        "total_courses": "14门课程"
    },
    {
        "school_name": "中南财经政法大学",
        "total_courses": "8门课程"
    },
    {
        "school_name": "苏州大学",
        "total_courses": "11门课程"
    },
    {
        "school_name": "台湾交通大学",
        "total_courses": "1门课程"
    },
    {
        "school_name": "国际关系学院",
        "total_courses": "6门课程"
    },
    {
        "school_name": "东南大学",
        "total_courses": "6门课程"
    },
    {
        "school_name": "青海大学",
        "total_courses": "7门课程"
    },
    {
        "school_name": "辽宁对外经贸学院",
        "total_courses": "10门课程"
    },
    {
        "school_name": "深圳大学",
        "total_courses": "4门课程"
    }
]
```