

---

# Python 课程大作业

## 高校就业信息的爬取和处理

刘昱辉

2025 年 1 月 2 日



---

Contents:

---

<b>1</b>	<b>任务描述及分析</b>	<b>1</b>
<b>2</b>	<b>系统设计</b>	<b>3</b>
2.1	爬虫 . . . . .	3
2.2	信息处理 . . . . .	3
2.3	JobInfo.spiders.BUPT 模块 . . . . .	3
2.4	JobInfo.spiders.XiDian 模块 . . . . .	4
2.5	JobInfo.items 模块 . . . . .	5
2.6	JobInfo.pipelines 模块 . . . . .	5
2.7	main.py . . . . .	6
<b>3</b>	<b>测试</b>	<b>7</b>
3.1	测试环境 . . . . .	7
3.2	测试执行 . . . . .	7
<b>4</b>	<b>总结</b>	<b>11</b>



## 任务描述及分析

本次大作业的主要任务及分析如下：

### 1. 信息爬取

- 抓取某两所性质相近的高校自 2024 年 9 月 1 日以来发布的就业相关信息的网页，各存储为 1 个 csv 文件。获取字段包括：招聘主题、发布日期、浏览次数。
- 我采用 scrapy 框架，针对不同网页，分析其网页元素，编写爬虫爬取。

我选择了北邮和西电两所高校，它们的招聘信息网页都是动态加载的，所以不能直接请求得到数据。

对于北邮信息，我用开发者工具找到了数据源 json，直接请求 json 数据，解析 json 数据。

对于西电信息，我未能找到数据源，所以使用 playwright 模拟浏览器行为，获取 html 数据，用 css 选择器解析数据。

我发现北邮的信息是按照时间顺序倒序排列的，所以只要找到第一个早于 2024-09-01 的信息，就可以停止爬取。而西电的信息中间有一部分是乱序的，所以必须爬取所有页面并筛选。

- 使用 csv 库将爬取到的数据存储到 csv 文件中。

### 2. 信息处理：

- 首先对以上爬取到的数据预处理，然后将信息整合到 Excel 文件中，两所高校各占一个工作表；再在新的工作表中汇总信息。

#### • 预处理

- 去掉招聘主题两端的空格等无关字符。
- 将招聘信息的日期格式修改为：YYYY-MM-DD。
- 将访问次数为空的置为 0，并将字符串统一转为整数格式。

#### • 整合汇总

- 将上面预处理完成的结果，存入一个名为“就业信息汇总.xlsx”的 Excel 文件中，分别放在 < 高校一>、< 高校二> 两个工作表上。

- 上面两个标签页表头字段一致，包括：序号（数字序号）、招聘主题、发布日期、浏览次数。四列对应的数据类型分别设定为：文本、文本、日期（YYYY-MM-DD）、数字（整数）。
- 在“就业信息汇总.xlsx”文件前面两个工作表之后，再添加三个工作表，分别为：
  - \* < 高校一 > 招聘 TOP10：填写最受 < 高校一 > 学生关注的招聘信息 TOP10（根据浏览次数的多少进行排序）。
  - \* < 高校二 > 招聘 TOP10：填写最受 < 高校二 > 学生关注的招聘信息 TOP10。
  - \* < 两校 > 招聘 TOP10：在选定的两所高校都出现，并且按浏览次数从高到低排序，取前十个（如果有的话）；只考虑招聘信息的标题同名，不用解析企业名称。
- 信息的预处理涉及到字符串处理，较为简单。
- 我使用 openpyxl 库处理 excel 文件。我首先将 csv 文件读取为 list，方便处理。然后将处理后的数据写入 excel 文件。

有 2 个主要模块：**爬虫**和**信息处理**。

## 2.1 爬虫

爬虫位于 `JobInfo/spiders`，有 BUPT、XiDian 两个爬虫。

两个文件中都含有一个继承自 `scrapy.Spider` 的类，类中有两个起主要作用的函数，`parse` 和 `parse_views`。前者是对招聘信息列表的请求的回调函数，从响应中得到主题、日期两个信息；后者是对具体页面的请求的回调函数，是为了得到浏览次数并返回 `item`。

`JobInfo/items.py` 中定义了类 `JobinfoItem`，作为爬取的目标。

`JobInfo/pipelines.py` 中定义了类 `JobinfoPipeline`，对 `parse_views` 返回的 `JobinfoItem` 对象进行预处理，去除首尾空白，存入 csv 文件。

## 2.2 信息处理

信息处理位于 `main.py`。首先读取 csv，将 csv 转换成 list，通过排序得到 top10 的 list，将 list 写入表格。以上每一操作都由相应函数来实现。具体的处理函数见下文。

以下是用 `sphinx` 生成的接口说明。

## 2.3 JobInfo.spiders.BUPT 模块

```
class JobInfo.spiders.BUPT.BuptSpider (*args: Any, **kwargs: Any)
```

基类: `Spider`

爬取北邮招聘信息的爬虫。

变量

- `name (str)` -- 爬虫名称。

- `allowed_domains(list[str])` -- 允许的域。只会爬取该域下的 url。
- `timestamp(int)` -- 当前时间戳。
- `base_url(str)` -- 爬取的基本 url。
- `current_page(int)` -- 当前正在爬的页数。
- `start_urls(list[str])` -- 加上参数的 url。
- `done(bool)` -- 是否已经爬到了 2024-09-01。

`parse(response)`

该生成器函数为 `scrapy.Request` 默认回调函数，用来处理获取到的 json。由于 json 中只有主题和日期，还要进一步获取浏览次数。

**参数**

`response` -- 请求得到的响应。

**Yield**

对浏览次数的请求，或对下一页的请求。

`parse_views(response, topic, date)`

该生成器函数为 `parse` 中调用的请求浏览次数的回调函数，用于从响应中得到浏览次数。

**参数**

- `response` -- 请求得到的响应。
- `topic` -- 招聘主题。由 `cb_kwargs` 传递。
- `date` -- 发布日期。由 `cb_kwargs` 传递。

**Yield**

一个 `JobinfoItem` 对象。

## 2.4 JobInfo.spiders.XiDian 模块

`class JobInfo.spiders.XiDian.XidianSpider(*args: Any, **kwargs: Any)`

基类: `Spider`

爬取西电招聘信息的爬虫。

**变量**

- `name(str)` -- 爬虫名称。
- `allowed_domains(list[str])` -- 允许的域。只会爬取该域下的 url。

`parse(response)`

该生成器函数为 `scrapy.Request` 默认回调函数，用来处理获取到的 html。由于 html 中只有主题和日期，还要进一步获取浏览次数。

**参数**

`response` -- 请求得到的响应。

**Yield**

对浏览次数的请求，或对下一页的请求。

`parse_views(response, topic, date)`

该生成器函数为 `parse` 中调用的请求浏览次数的回调函数，用于从响应中得到浏览次数。

**参数**



- **response** -- 请求得到的响应。
- **topic** -- 招聘主题。由 *cb\_kwargs* 传递。
- **date** -- 发布日期。由 *cb\_kwargs* 传递。

#### Yield

一个 *JobinfoItem* 对象。

#### **start\_requests()**

开始爬取第一页，生成一个请求，使用 playwright 作为 Downloader。

#### Yield

对第一页的请求。

## 2.5 JobInfo.items 模块

```
class JobInfo.items.JobinfoItem(*args: Any, **kwargs: Any)
```

基类: *Item*

招聘信息的 item。

#### 变量

- **topic** -- 招聘主题。
- **date** -- 发布日期。
- **views** -- 浏览次数。

## 2.6 JobInfo.pipelines 模块

```
class JobInfo.pipelines.JobinfoPipeline
```

基类: *object*

将获取到的招聘信息预处理后存入 csv 文件。

#### 变量

- **file** -- csv 文件。
- **writer** -- csv.writer。

```
close_spider(spider)
```

关闭爬虫时关闭文件。

```
open_spider(spider)
```

打开爬虫时打开 {爬虫名.csv} 文件，并用 csv writer 打开。

```
process_item(item, spider)
```

处理爬虫返回的 item。

#### 参数

- **item** -- 爬虫返回的 item。
- **spider** -- 爬虫。

#### 返回

None

## 2.7 main.py

`adjust_col_width(sheet: Worksheet, col: str, wid: int) → None`

调整列宽。

### 参数

- **sheet** -- 待调整的工作表。
- **col** -- 待调整的列。(单个字母, 如果不是, 则不改变)
- **wid** -- 调整后的宽度。

### 返回

None

`csv_to_list(csv_file_path: str) → list`

将 csv 文件转换成列表, 方便进一步处理。

### 参数

**csv\_file\_path** -- csv 文件的路径。

### 返回

转换后的列表。

`get_both_top_10(info_list_1: list, info_list_2: list) → list`

获取两个列表中都出现的、浏览次数前十的招聘信息。

### 参数

- **info\_list\_1** -- 招聘信息列表。
- **info\_list\_2** -- 另一个招聘信息列表。

### 返回

前十浏览次数的招聘信息的列表。

`get_top_10(info_list: list) → list`

从列表中获得前十浏览次数的招聘信息。

### 参数

**info\_list** -- 招聘信息列表。

### 返回

前十浏览次数的招聘信息的列表。

`list_to_xlsx(info: list, sheet: Worksheet, sheet_header: tuple | None = None) → None`

将列表写入工作表。

### 参数

- **info** -- 待写入的列表。
- **sheet** -- 待被写入的工作表。
- **sheet\_header** -- 可选的表头。

### 返回

None

`str_to_date(date_str: str) → date`

将形如'2024-12-17' 的字符串转换成 datetime.date 对象: param date\_str: 日期字符串。:return: date 对象。

### 3.1 测试环境

- 操作系统：Windows 11；
- Python 版本：Python 3.11.1；
- IDE：Visual Studio Code；
- packages：见 requirements.txt。

### 3.2 测试执行

首先激活虚拟环境，命令如 `.venv\Scripts\activate`（使用 IDE 会自动激活，则该步骤省略），如图1。

```
D:\projects\python\pythonProject>.venv\Scripts\activate  
(.venv) D:\projects\python\pythonProject>
```

图 1: 激活虚拟环境

进入 JobInfo 文件夹，运行 `python main.py`，即可爬取两所高校的招聘信息，并将信息整合到 Excel 文件中。

开始输出“正在爬取北邮信息”，如图2。

```
D:\projects\python\pythonProject\.venv\Scripts\python.exe D:\projects\python\pythonProject\JobInfo\main.py  
正在爬取北邮信息...  
2025-01-02 08:40:02 [scrapy.utils.log] INFO: Scrapy 2.12.0 started (bot: JobInfo)
```

图 2: 正在爬取北邮信息

一段时间后，输出“正在爬取西电信息”，如图3。

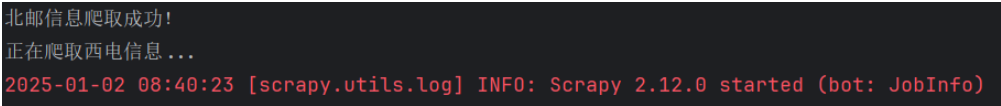


图 3: 正在爬取西电信息

西电信息较多，爬取时间较长，以下是爬取过程中的截图，如图4。

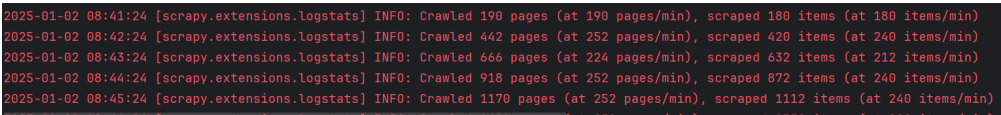


图 4: 正在爬取西电信息

最后输出 “All Done!”, 如图5。

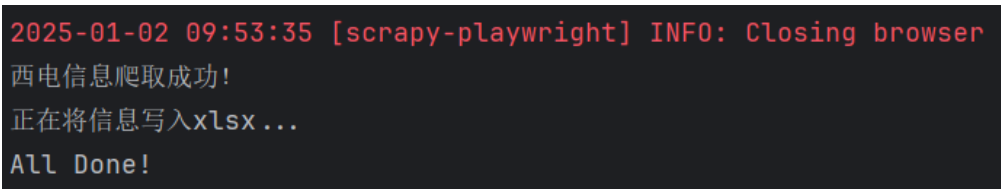


图 5: 爬取完成

结果如下：在文件夹下生成了 BUPT.csv、XiDian.csv、就业信息汇总.xlsx 三个文件，如图6。





 BUPT.csv	2025/1/2 8:40	Microsoft Excel ...	41 KB
 XiDian.csv	2025/1/2 9:53	Microsoft Excel ...	121 KB
 就业信息汇总.xlsx	2025/1/2 9:53	Microsoft Excel ...	112 KB
 main.py	2025/1/2 14:28	Python 源文件	5 KB

图 6: 生成的文件

将 csv 文件用记事本打开，如图7 8。注意，由于 csv 文件中含有中文，用 Excel 打开可能会乱码。

将 Excel 文件用 Excel 打开，如图9 10 11 12 13。

可以看到，成功爬取了两所高校的招聘信息，并且将招聘信息浏览次数前十以及两校共同的浏览次数前十整合到 Excel 文件中，达成了本次大作业的目标。

BUPT.csv				XiDian.csv					
文件	编辑	查看		文件	编辑	查看			
上海电机学院教师招聘公告, 2024-12-25, 97				广西医科大学诚聘海内外优秀博士人才, 2025-01-02, 1					
大连商品交易所2024年校园招聘启事, 2024-12-26, 146				厦门大学嘉庚学院2024-2025学年人才招聘启事, 2025-01-02, 16					
北京亦庄实验中学《北京市十一学校经济技术开发区分校》2025年校园招聘, 2024-12-26, 153				宿州航空职业学院招聘简章, 2024-12-31, 87					
电子科技大学长三角研究院(衢州)招聘公告, 2024-12-26, 157				无锡高新人才发展集团有限公司, 2024-12-31, 125					
东北电力大学2025年招聘公告, 2024-12-26, 85				贵州医科大学诚聘海内外优秀人才, 2025-01-02, 1					
四川英杰电气股份有限公司2025校园招聘启事, 2024-12-26, 150				广州软件学院中青年博士招聘公告, 2025-01-02, 10					
长沙新奥燃气2024校园招聘, 2024-12-26, 156				2024年宁波幼儿师范高等专科学校公开招聘高层次人才公告, 2024-12-31, 147					
一汽解放汽车有限公司2025年度校园招聘启事, 2024-12-26, 100				重庆电力高等专科学校2025年高层次人才博士招聘公告, 2025-01-02, 28					
无人机运行管理工程师, 2024-12-26, 190				2024年闽南师范大学引进人才招聘启事, 2024-12-31, 101					
浙大宁波理工学院2025年高层次人才招聘公告, 2024-12-26, 133				中国电科第十二所2025年校园招聘, 2025-01-02, 67					
中国科学院长春应用化学研究所2025年招聘公告, 2024-12-26, 142				东北电力大学2025年人才招聘公告, 2025-01-02, 29					
广西机电职业技术学院人才招聘公告, 2024-12-26, 167				公安部道路交通安全研究中心, 2025-01-02, 48					
城银清算服务有限公司2025年度校园招聘, 2024-12-26, 173				派驻华群能源集团2025校园招聘项目正式启动, 2025-01-02, 37					
中国人民保险集团股份有限公司博士后科研工作站2025年博士后研究人员招收简章, 2024-12-26, 182				江门市宝士制冷电器有限公司, 2025-01-02, 22					
作业帮2025校招“AI专项”补录中, 2024-12-26, 195				广安理工学院筹建处2024年12月直接考核招聘博士公告, 2025-01-02, 14					
九江学院2025年高层次人才招聘公告, 2024-12-26, 151				绥化学院常年招聘博士公告, 2025-01-02, 11					
厦门松霖科技股份有限公司2025校园招聘, 2024-12-26, 204				空中能力装备研究院, 2025-01-02, 159					
中信招标2025秋季校园招聘, 2024-12-03, 298				河北银行2025年度校园招聘启事, 2025-01-02, 25					
海外销售校园招聘(2024/2025届), 2024-12-04, 270				浙江工业职业技术学院公开招聘公告(专任教师岗位), 2025-01-02, 13					
深圳大学组立讲席教授团队招聘博士后和专职研究人员, 2024-12-04, 268				上海启源电力科技有限公司2025寒假实习计划, 2025-01-02, 16					
金蝶2025校园招聘简章, 2024-12-04, 303				中建六局, 2024-12-30, 148					
汉口学院2025年招聘公告, 2024-12-26, 265				合肥他们方田教育科技有限公司, 2024-12-30, 76					
深圳联通2025校园招聘, 2024-12-03, 333				桂林理工大学2024招聘公告, 2024-12-30, 88					
闽江学院2024年度高层次人才招聘公告, 2024-12-04, 268				内蒙古大学人才招聘公告, 2024-12-30, 116					
行 7, 列 31	21,709 个字符	100%	Windows (CRLF)	UTF-8	行 1, 列 1	62,323 个字符	100%	Windows (CRLF)	UTF-8

图 7: 北邮.csv

图 8: 西电.csv

A	B	C	D	A	B	C	D
序号	招聘主题	发布日期	浏览次	序号	招聘主题	发布日期	浏览次数
1	上海电机	2024-12-25		1	广西医科	2025-01-02	1
2	大连商品	2024-12-26	1	2	厦门大学	2025-01-02	16
3	北京亦庄	2024-12-26	1	3	宿州航空	2024-12-31	87
4	电子科技	2024-12-26	1	4	无锡高新	2024-12-31	125
5	东北电力	2024-12-26		5	贵州医科	2025-01-02	1
6	四川英杰	2024-12-26	1	6	广州软件	2025-01-02	10
7	长沙新奥	2024-12-26	1	7	2024年宁	2024-12-31	147
8	一汽解放	2024-12-26	1	8	重庆电力	2025-01-02	28
9	无人机运	2024-12-26	1	9	2024年闽	2024-12-31	101
10	浙大宁波	2024-12-26	1	10	中国电科	2025-01-02	67
11	中国科学	2024-12-26	1	11	东北电力	2025-01-02	29
12	广西机电	2024-12-26	1	12	公安部道	2025-01-02	48
13	城银清算	2024-12-26	1	13	派驻华群	2025-01-02	37
14	中国人民	2024-12-26	1	14	江门市宝	2025-01-02	22
15	作业帮20	2024-12-26	1	15	广安理工	2025-01-02	14
16	九江学院	2024-12-26	1	16	绥化学院	2025-01-02	11
17	厦门松霖	2024-12-26	2	17	空中能力	2025-01-02	159
18	中信招标	2024-12-03	2	18	河北银行	2025-01-02	25
19	海外销售	2024-12-04	2	19	浙江工业	2025-01-02	13
20	深圳大学	2024-12-04	2	20	上海启源	2025-01-02	16
21	金蝶2025	2024-12-04	3	21	中建六局	2024-12-30	148
22	汉口学院	2024-12-26	2	22	合肥他们	2024-12-30	76
23	深圳联通	2024-12-03	3	23	桂林理工	2024-12-30	88
24	闽江学院	2024-12-04	2	24	内蒙古大	2024-12-30	116
25	芯联集成	2024-12-06	2	25	前海国际	2024-12-30	89
26	杭州高新	2024-12-05	3	26	中国石化	2024-12-30	130
27	菁英计划	2024-12-04	3	27	中信证券	2024-12-30	106
28	航天时代	2024-12-04	3	28	烟台东方	2024-12-30	51
29	华佗集团	2024-12-05	2	29	云南电网	2024-12-30	61

图 9: 就业信息汇总.xlsx-北邮

图 10: 就业信息汇总.xlsx-西电

A	B	C	D	A	B	C	D
序号	招聘主题	发布日期	浏览次数	序号	招聘主题	发布日期	浏览次数
1	首开集团	2024-12-13	1318	1	中国邮政	2024-09-26	33854
2	小学数学	2024-09-07	823	2	京东方科	2024-09-09	12235
3	泰山学院	2024-11-20	790	3	西门子（	2024-09-25	10479
4	比亚迪20	2024-12-13	759	4	中国华电	2024-10-25	8974
5	中智集团	2024-12-11	753	5	爱立信（	2024-09-02	7608
6	招聘岗位	2024-12-19	707	6	乾元国家	2024-09-18	7562
7	国网江西	2024-10-27	706	7	东方财富	2024-09-09	6979
8	中国铁路	2024-11-18	688	8	陕西省高	2024-09-14	5727
9	中国兵器	2024-12-12	687	9	中国电建	2024-09-03	5675
10	中学老师	2024-09-01	687	10	拼多多集	2024-09-06	5147

图 11: 就业信息汇总.xlsx-北邮招聘 TOP10    图 12: 就业信息汇总.xlsx-西电招聘 TOP10

A	B	C	D	E	F	G	H
序号	招聘信息	北邮发布日期	北邮浏览	西电发布日期	西电浏览	浏览次数之和	
1	西安电子	2024-09-02	283	2024-09-11	3266	3549	
2	中国信息	2024-09-26	546	2024-09-25	2450	2996	
3	国航机务	2024-09-03	295	2024-09-03	2668	2963	
4	中国电子	2024-09-11	424	2024-09-09	2103	2527	
5	央企大厂	2024-09-06	261	2024-09-04	1846	2107	
6	中国南山	2024-09-18	262	2024-09-14	1708	1970	
7	联通（广	2024-10-31	330	2024-10-24	1613	1943	
8	国家电投	2024-10-04	441	2024-10-11	1403	1844	
9	中国移动	2024-09-18	276	2024-09-14	1555	1831	
10	物产中大	2024-09-23	228	2024-09-21	1587	1815	

图 13: 就业信息汇总.xlsx-两校 TOP10

---

### 总结

---

本次大作业，我首先通过 `scrapy` 框架爬取了两所高校的招聘信息，然后对爬取到的数据进行了预处理，最后将信息整合到 `Excel` 文件中。在这个过程中，我学到了很多关于爬虫和数据处理的知识，也提高了自己的编程能力。

起初，我尝试直接爬取北邮的招聘网站，但发现网站是动态加载的，无法直接获取到数据。后来，我查找“如何爬取动态加载的网页”的方法，发现可以通过查看开发者工具，找到 `json` 数据的 `url`，然后直接爬取 `json` 数据。分析 `api` 调用的格式耗费了我一定时间，但最终，我成功爬取到了北邮的招聘信息。

爬取西电的招聘信息时，我发现西电的招聘信息也是动态加载的，但是在开发者工具-网络-*XHR* 中没有找到 `json` 数据的 `url`。我尝试了很多方法，还尝试直接分析 `js`，但都没有成功。于是，我尝试使用可以模拟浏览器的库，我使用的是 `scrapy-playwright`，利用 `css` 选择器获取信息，最终成功爬取到了西电的招聘信息。

预处理过程比较简单，主要是字符串处理。

我使用了 `openpyxl` 库处理 `excel` 文件。处理过程比较顺利，因为之前做过了类似的作业。

我主要参考了以下资料：

- [scrapy 官方文档](#)
- [python csv](#)

总的来说，这次大作业让我学到了很多，例如：

- 爬虫可以分析网页元素，或分析 `api` 调用；
- 起初我使用的是 `scrapy-splash` 这个库，但是由于其长时间没有维护，落后于 `scrapy` 的版本，导致无法正常使用，于是我转而使用 `scrapy-playwright` 这个库，这也让我明白有时候需要灵活变通。

同时提高了我解决问题的能力，希望在以后的学习中，能够更好地运用所学知识。