第15课--建立一个组织有序的爬虫

明确目标

先爬取每个榜单中的10家公司，四个榜单一共就是40家公司。再跳转到这40家公司的招聘信息页面，爬取到公司名称、职位、工作地点和招聘要求等细节信息。

分析过程

1、企业排行榜的公司信息

1.首先，要看企业排行榜里的公司信息藏在了哪里。

1、现在让我们来使用“检查”工具，点击Network，刷新页面。点开第0个请求company/，看Response，找一下有没有榜单的公司信息在里面
2、经过查找可以发现：四个榜单的所有公司信息都在里面。说明企业排行榜的公司信息就藏在html里
3、现在请你点击Elements，点亮光标，再把鼠标移到【阿里巴巴集团】，这时就会定位到含有这家公司信息的<a>标签上。

4、点击href="/company/281097/"，会跳转到阿里巴巴这家公司的详情页面。详情页面的网址是：

https://www.jobui.com/company/281097/，所以可以得出详情页url的拼接方式

```
https://www.jobui.com/company/+数字/
```

5、仔细观察网页html的结构，你会发现，每个公司信息都藏在一个<li>标签里，而每5个<li>标签都从属与一个<ul>标签。这是一个层层嵌套的关系
6、在这里我们首先抓取最外层的<ul>标签，再抓取<ul>标签里的<a>标签，最后提取到<a>标签href属性的值
7、

公司详情页面的招聘信息

代码实现

创建项目
定义item
创建和编写爬虫文件
存储文件

修改设置

代码实操