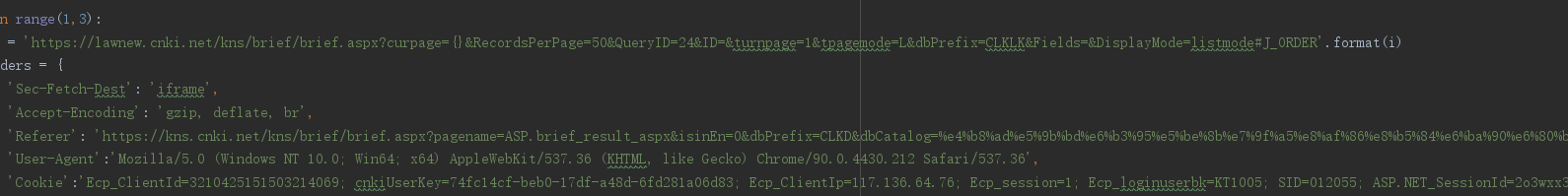
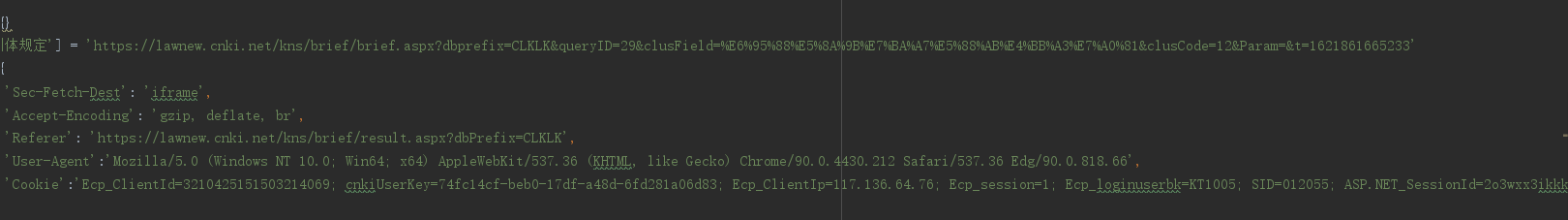
1. 对知网进行前期调研发现，网站会将部分数据隐藏不显示在前端页面，需要对网站前端进行一定的分析。要爬的是‘地方性法规及文件’，‘团体规定’，‘行业规定’，‘地方司法文件’，‘地方政府规章及文件’等不同属性的产业文件。可按照如下步骤进行数据爬取操作。
2. 由于网站进行了特殊设置，需要对所爬取的页面进行抓包分析，找到客户端请求网站服务器端的真实URL链接。并为了更好的模拟浏览器端真实操作，设置Cookies等请求头。具体的抓包链接和请求头如下所示：

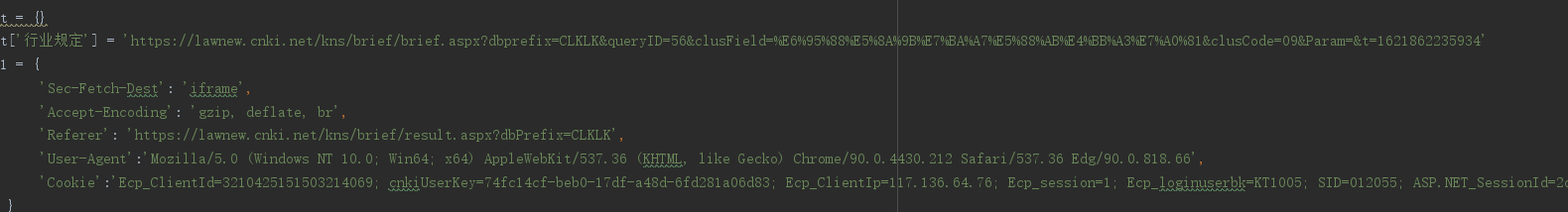
‘地方性法规及文件’页面链接



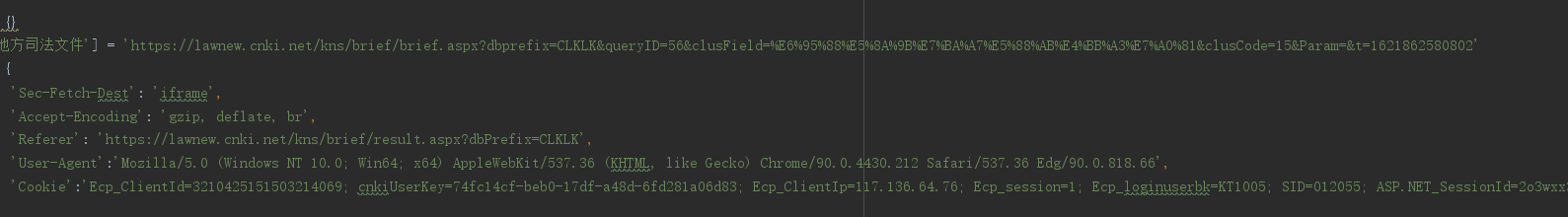
‘团体规定’



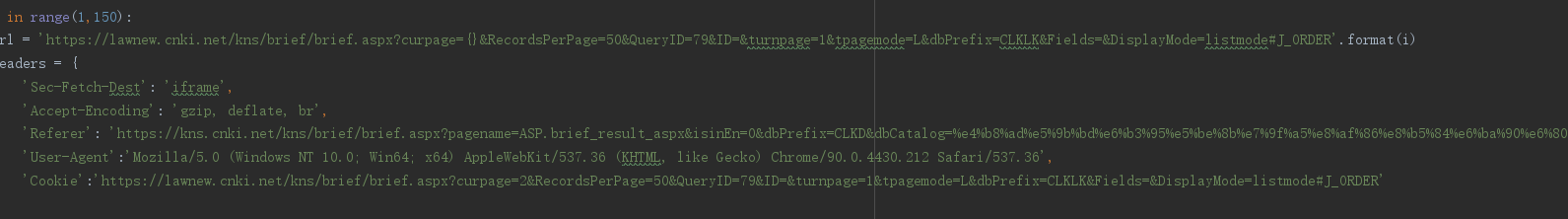
‘行业规定’



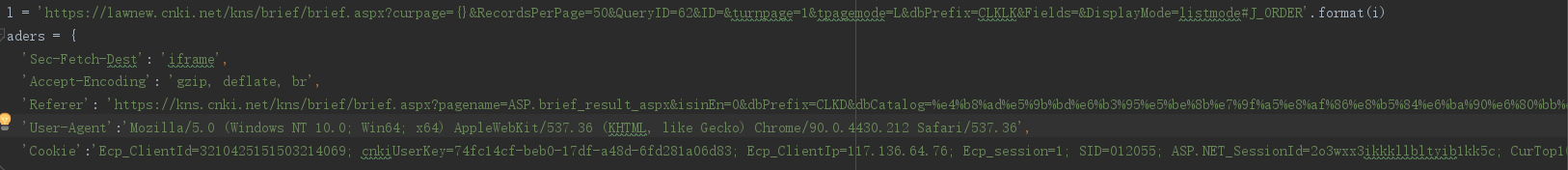
‘地方司法文件’



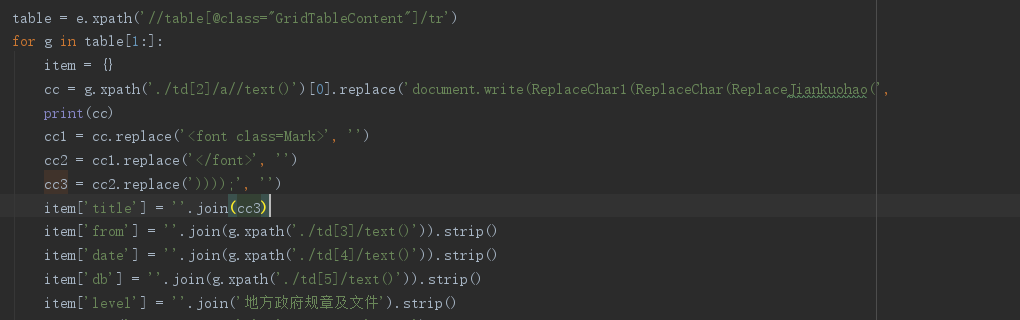
‘地方政府规章及文件’（2012-3-15-2020-12-31）



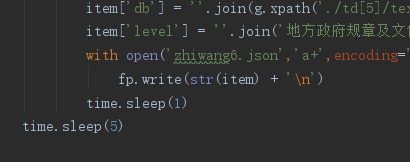
‘地方政府规章及文件’（2012-3-14-2020-12-31）



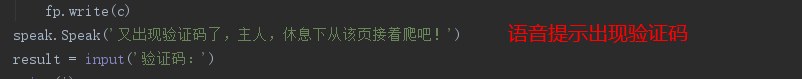
1. 通过浏览器开发者工具进行Xpath操作，对html页面进行解析，找到数据存放的路径位置，如下所示：



1. 在请求较为频繁的条件下，服务器端会进行检测请求端的IP地址，频繁操作后可能进行ip地址的封禁。为了避免该情况出现，尽可能让电脑间歇性的请求，因而设置请求休眠时间，每爬取完一个页面，暂停5秒。

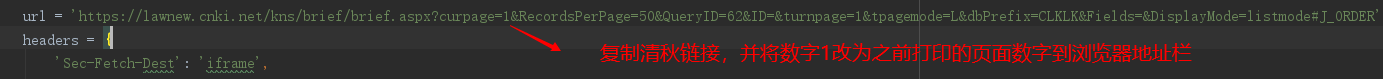


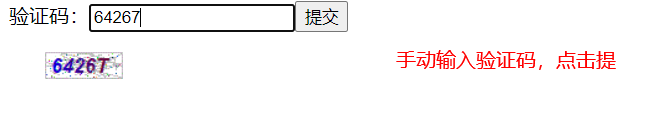
1. 在数据进行爬取的过程中，会出现验证码的情况。通常的验证码破解处理方式是OCR光学字符识别，或者直接对接远程打码平台API识别。但在本次数据爬取中，经过前期简单的测试，验证码出现的并不平凡，可直接在浏览器端进行手动输入操作，通过验证码。具体的操作方式为：当爬取过程中出现验证码时，程序设置了电脑语音控制提示信息，电脑会发出语音信号来提示你验证码出现了，并且此时程序会中段运行，并打印该页面页数，然后通过在浏览器端更改链接数字，手动输入验证码进行提交，即可再次启动程序进行运行，直到数据爬取完成。

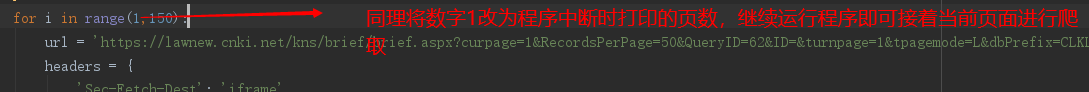








，点击完提交之后，即可重新启动程序从当前页接着爬取数据。



1. 通过爬取数据并保存，得到了两个json数据文件，分别为zhiwang2.json, zhiwang6.json 文件，需要对文件进行数据清洗，清洗要求为按省份进行划分。通过对爬取的数据进行调研分析，有很多市，县是不带有省份标志的，为了判别所属省份，需要下载一份全国各省市县的文件列表，为1622205159175.json文件。数据i清洗程序为：data\_process.py文件。清洗流程如下：

