**项目背景**

此爬虫项目主要实现依据搜索法律相关关键字爬取法律数据的功能（当然也可以获取所爬服务器上所有法律数据）。

**项目实现**

该项目使用Python语言编写，通过Python requests模块发送请求获取响应数据，以脚本的方式简单实现，未使用爬虫框架。

脚本定义了一个LawSpider类

类中\_\_init\_\_方法，初始化构造数据，包含：

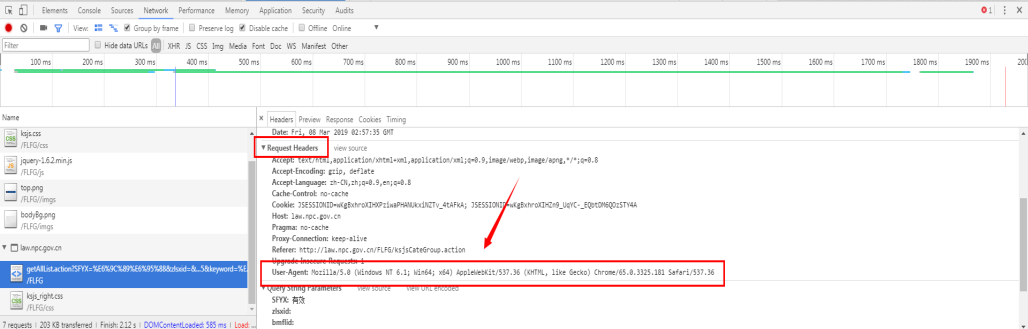
1.获取关键字搜索链接（若要爬取服务器上所有数据，链接：http://law.npc.gov.cn/FLFG/getAllList.action）

获取关键字搜索链接通过浏览器开发者工具获取，如下：

检索“婚姻法”



2.请求头信息，通过运行程序的浏览器获取，目的是不让爬取服务器认为当前程序是一个爬虫。



3.构造翻页获取数据，此处需要通过构造post请求的data数据

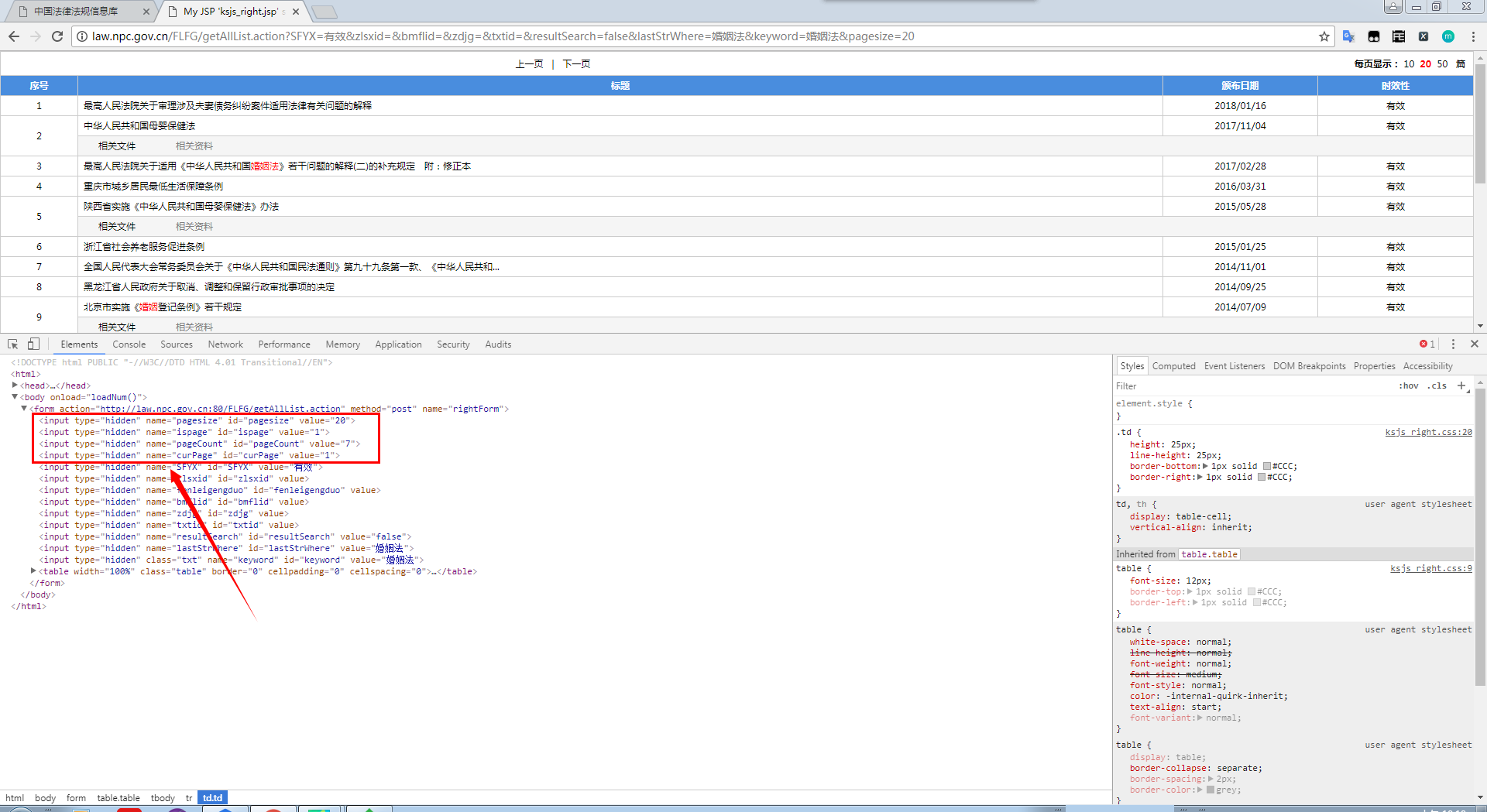
首先使用获取到的关键字链接发送请求，使用开发者工具，在Element中可以发现：

Ⅰ、pagesize 的value值= 20，表示每页显示20条数据

Ⅱ、pageCount 的value值= 7，表示获取的数据共计7页

Ⅲ、curPage的value值，每一次翻页操作，curPage的value值会+1

由此可以构造循环获取到关键字搜索的所有数据，即curPage的value=1时，获取第一页的20条数据，之后使curPage的 value=2，获取第二页数据，以此类推，循环次数为pageCount，获取到所有数据。





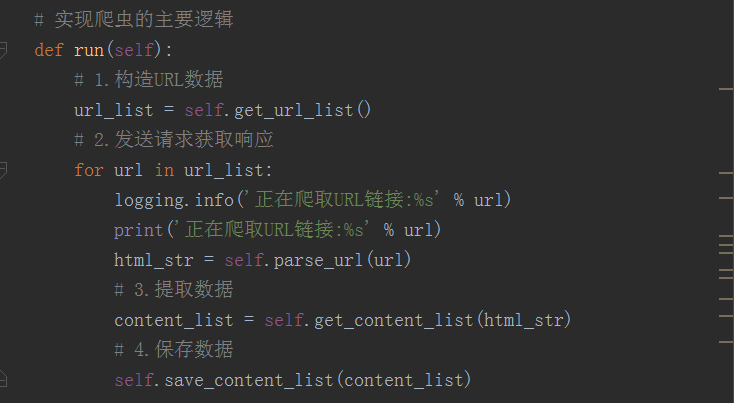
类中run方法实现爬虫的主要逻辑，包含：

1.构造爬取URL链接

2.发送URL链接请求获取响应数据

3.提取响应数据

4.保存数据



另外定义了一个日志配置，用于记录程序运行过程中的log

