**数博科技**

数据组-爬虫技术文档

*中国裁判文书网*

|  |  |  |
| --- | --- | --- |
| 2019-05-23 | 1.0 | 陈子昂 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

中国裁判文书网爬取大致思路：

1. 发起附带”User-Agent”的请求，去访问所需页面，此时会返回得到一段加密的JavaScript代码。此段代码主要作用就是重定向URL，通过解析JS代码，得到重定向所需要的动态URL与参数，手动构造URL，发起附带”User-Agent” 和”Cookie”的请求，去访问新的URL
2. 重定向之后，返回的Cookie中会附带一个”vjkl5”的参数，可通过正则等方法取出具体的值，此数据是加密了的，需要继续解析JS，通过此数据得到POST请求的参数”vl5x”, ”number”, ”guid”这三个参数，进行后序的请求，请求连接为

”http://wenshu.court.gov.cn/List/ListContent”

1. 通过POST请求得到一段JSON数据，里面有两个所需参数

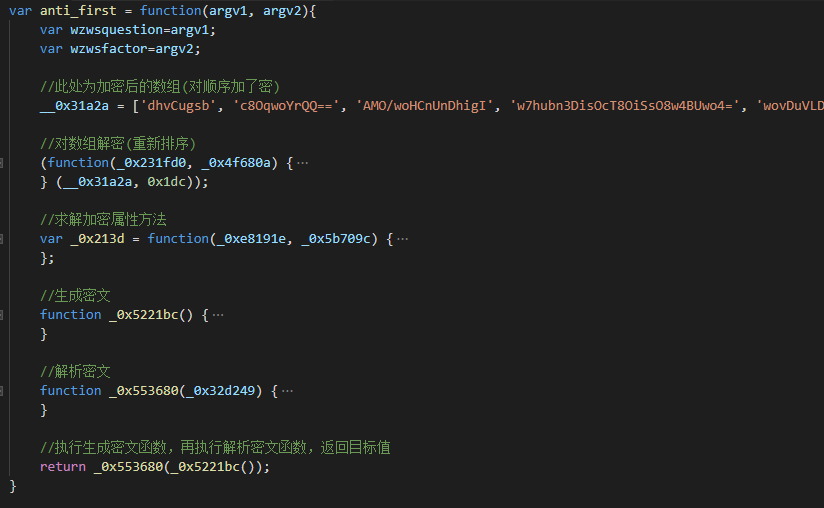
”RunEval”, ”文书ID”

从第一个值是唯一的，文书ID则对应着每篇文章，继续解析JS，其中第一个参数，是为第二个参数准备环境的，而从第二个参数中，我们可以拿到一个”DocID”，这个值就是详情页面所需要的了，详情页面连接为”http://wenshu.court.gov.cn/content/content”

1. 具体的文章内容是异步加载的，如果需要，可以直接访问相应的连接，上面所有访问都需要带上第一步所得Cookie

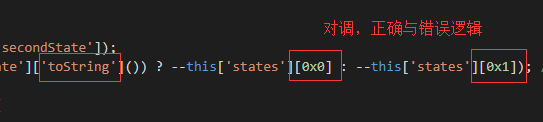
”http://wenshu.court.gov.cn/CreateContentJS/CreateContentJS.aspx”

1、anti\_first



注意事项：

1.1 求解加密属性方法，这里”toString”在本地环境会报错，需要对调源码中正确与错误的处理逻辑



1.2 返回的JS代码中，包含动态URL和一些参数，这些参数需要用于解密，有”wzwsquestion” 和 “wzwsfactor”，这两参数可用正则取出

以下代码都是需要从网页中慢慢扣的，执行过程中解决相应问题即可，如某些函数未定义，则找出其定义函数，copy过来即可



调用方法：

>>>import execjs

>>>with open(“zgcpwsw.js”, “r”) as f\_r:

>>> js\_dode = f\_r.read()

>>> anti\_first = execjs.compile(js\_code)

>>>wzwschallange = anti\_first.call(“anti\_first”, argv1, argv2)

其中argv1是”wzwsquestion”而argv2是“wzwsfactor”