2017年8月4号周报

本周的任务是研究“裁判文书网”的反爬策略，以及编写爬虫脚本。目前对“裁判文书网”的反爬策略了解如下：

1访问频率略高会提示输入验证码。

2需要伪装成浏览器进行访问

3请求得到的页面数据是经过js加密混淆的，无法获得直接数据

4页面的详情内容是<input type=”hidden”.........>,这些内容都是执行时js生成的，有对应的文档获取借口，但是接口被加密，无法获得。同样列表页也是接口被加密。

存在困难：由于网站的加密算法十分少见，暂时没有获得可行的破解办法。

下一步的策略是：想通过对页面截屏，然后用图片解析器，解析出页面内容，再进行处理。通过chromedriver对页面截屏，由于只能截取屏幕上可以看到的内容，超出部分无法抓取。

下周准备：1研究下能否将页面缩小使它完全显示出来再截屏。2分析js等前段技术，反混淆技术，努力破解加密页面。3。完成分配的其他网站的爬虫脚本编写。