**爬虫工作汇报**

|  |  |
| --- | --- |
| **序号** | **工作内容** |
| 1 | 搭建爬虫开发环境。 |
| 2 | 测试开发环境安装是否成功 |
| 3 | 研究“裁判文书网”的反爬策略 |
| 4 | 编写页面信息的抓取脚本 |
| 5 | 抓取信息的保存结构设计 |
| 6 | 验证码处理 |
| 7 | JS混淆和解混淆 |
| 8 | 网站商业加密的研究 |
| 9 | Chrome和Fiddler抓包分析 |
| 10 | 海马模拟器装入文书APP |
| 11 | 尝试抓app包 |
| 12 | 尝试对网页截屏，通过图片识别出需要的信息 |
| 13 | 国家企业信用信息公示系统(江苏)的反爬虫策略 |
| 14 | 编写脚本爬取免费IP代理 |
| 15 | 测试爬取的IP稳定性 |
| 16 | 拖拽型验证码研究 |
| 17 | 编写具体企业的信息爬取脚本 |
| 18 | 数据存储结构设计 |
| 19 | JS模拟点击事件 |
| 20 | 网站分页技术的研究 |
| 21 | 国家企业信用信息公示系统(全国)的爬取策略 |
| 22 | 编写全国版具体企业的信息爬取脚本 |
| 23 | 设计数据的抓取结构 |
| 24 | 清洗数据 |

遇到的困难：

1.全国版的公示系统，具体企业页面的URL会随着每次搜索动态变化。

访问频率高会返回操作错误，需要重新搜索进入。

极简验证码破解难度高

2.网上抓取的代理IP可用性和稳定性都比较差

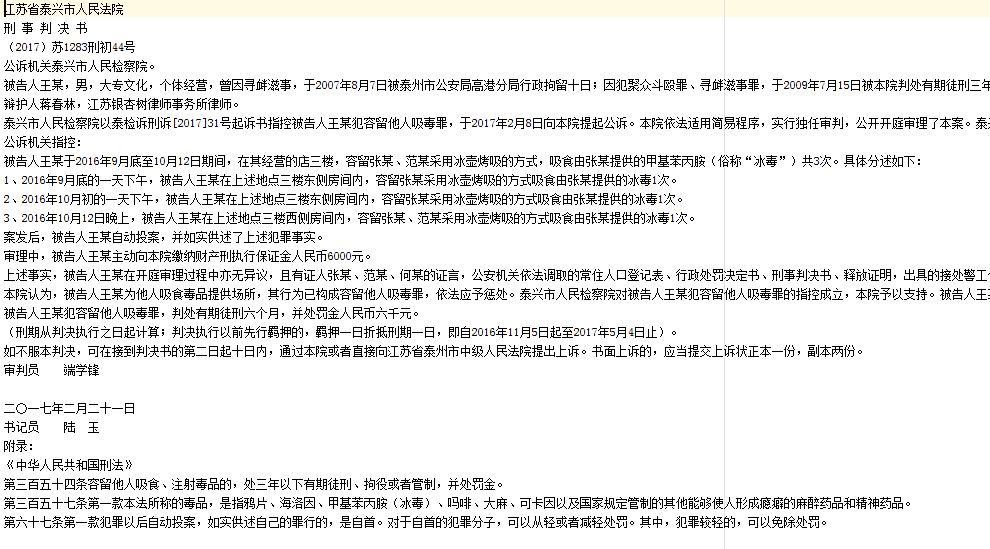
1. 文书裁判网的随机验证码破解难度也高

接下来的工作：1.整合工商全国的抓取脚本，脚本的调优工作

2.针对验证码问题，做一个专门的研究

3.bug的修复工作，维护具体网页爬虫的稳定

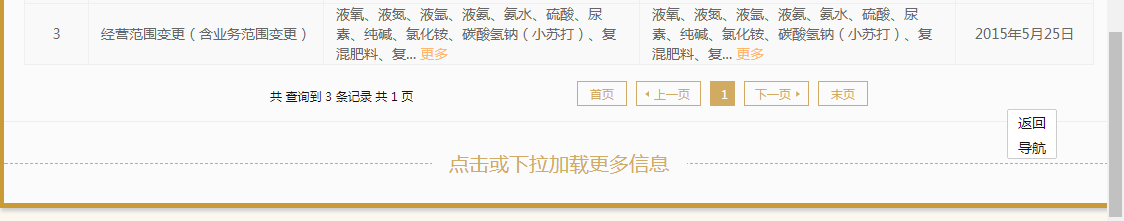
文书裁判页面信息写入文件：



工商全国的具体页面抓取

营业执照信息

下拉加载，用JS和selenium实现自动下拉，翻页处理



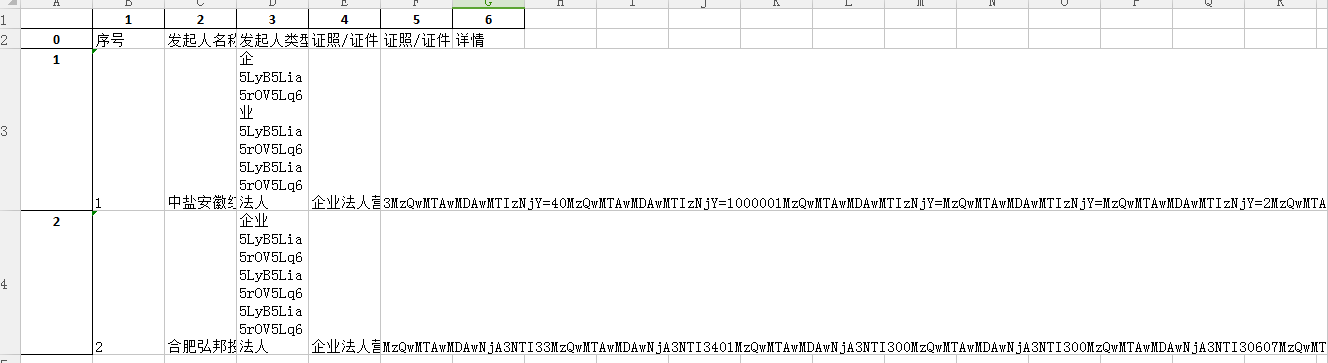


抓取展示：

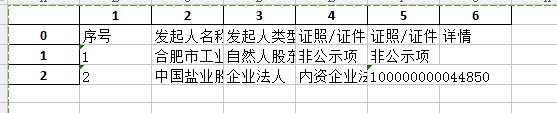


股东及出资信息





清洗过的数据：



主要人员信息



