#### 不涉及验证码的破解工作

**【业务需求】**

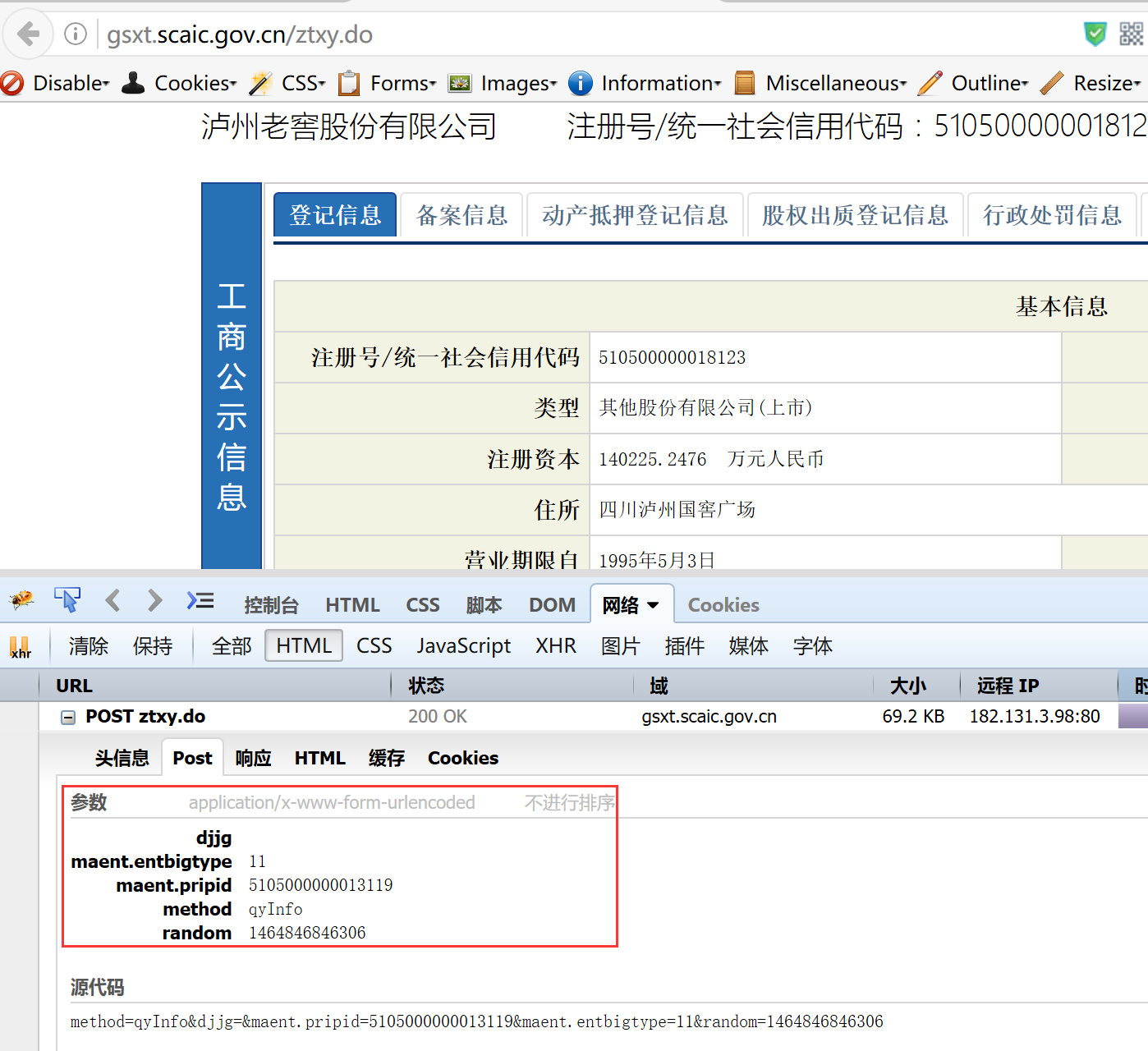
**以四川作为例子。**

目标网站：http://gsxt.scaic.gov.cn/，以“泸州老窖股份有限公司”作为采集对象示例。要求以唯一或一组参数作为该企业采集的参数入口，程序入口提供任意一组参数发起采集，参数示例：

“泸州老窖股份有限公司”在网站中的参数为：

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | djjg |  | | maent.entbigtype | 11 | | maent.pripid | 5105000000013119 | | method | qyInfo | | random | 1464846846306 | |  |

截图如下：



Main函数入口参数（5个参数）则为该程序外部唯一输入参数，伪代码组织格式请看附件java工程中Demo文件。

按照下文采集内容顺序完成该企业各种信息采集，除main函数输入参数外要求不能从外部输入任何参数，在采集过程中涉及到的各个参数由程序自动解析获取到并应用于后续各种处理。

采集具体内容如下：

（一）工商公示信息

1、登记信息，包括三块小栏目：基本信息、股东信息（若有股东详情则也需采集对应股东详情）、变更信息。**--涉及到该类信息为多页的请完成处理，确保信息完整性**

2、备案信息，包括两块小栏目：主要人员信息、分支机构信息。**--涉及到该类信息为多页的请完成处理，确保信息完整性**

3、经营异常信息，要求剔除“作出决定机关(列入)”列内容中包含的超链接部分再行入库。**--涉及到该类信息为多页的请完成处理，确保信息完整性**

4、严重违法信息。**--涉及到该类信息为多页的请完成处理，确保信息完整性**

5、抽查检查信息。**--涉及到该类信息为多页的请完成处理，确保信息完整性**

（二）企业公示信息

1、企业年报，提取年报列表，为后面对每个年报详情采集做准备。

2、\*\*\*\*年度报告，包括五块小栏目：企业基本信息（包括“企业资产状况信息”小栏目）、股东及出资信息、对外提供保证担保信息、修改信息。**--涉及到该类信息为多页的请完成处理，确保信息完整性**

**特别说明：**

**1、凡以上栏目内容中涉及到分页情况的，所有分页点击进去对应栏目内容都需要进行完整采集入库（每个栏目内容都已经建立对应mongodb表，请看mongodb表标题头文字说明和字段说明，看附件压缩包）。**

**2、用**SAICCompanySpider**表同步记录被采集企业url特征，根据**SAICCompanySpider**表字段记录格式要求同步保存。**

**3、使用附件提供mongodb工具类进行mongodb库交互处理，使用附件Http工具类完成网络访问交互。--（若有更好，请沟通确认）**

**4、使用开源工具JSOUP 开发包（附件中已自带jar）完成解析并提取目标网站特定企业指定栏目界面中字段对应内容，并根据附件中提供已定义好mongodb java表对象填写入对应值进行入库保存。--指定开发包为模块整合需要**

**5、要求在配置文件中可设置相邻两次对目标网站任意HTTP请求的时间间隔。**

**6、以附件java工程和开发包为基础进行开发，若有新增jar，请及时沟通说明。**

**【人员技能要求】**

有java语言采集网页解析方面开发经验，了解jsoup开发包和mongodb java基本操作。

**【交付要求】**

1、提交源代码前需提供不同类型企业（个体、有下属分公司企业、分公司）采集验证自测通过截图确认，后续我方验证通过。

2、需要提交源代码，要求代码编写规范化、结构化良好，易于阅读，并在源代码中进行相关注释说明便于后续理解跟踪，源代码一次性交付。

3、配合我方验证通过，交付后15个工作日内程序相关问题咨询支持。