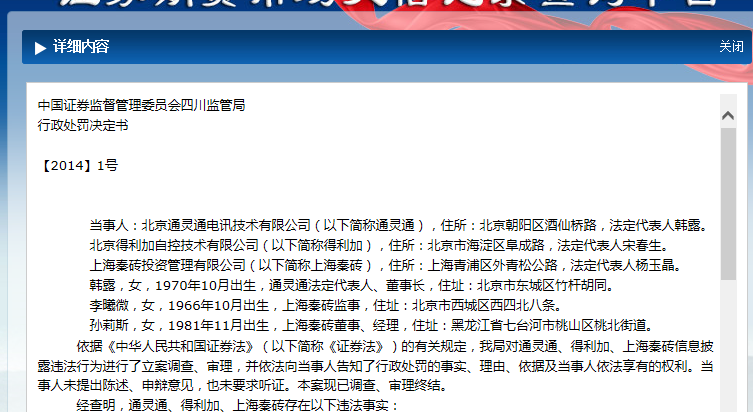
| 证监会证券期货市场失信记录 | | |
| --- | --- | --- |
| **URL** | http://shixin.csrc.gov.cn/honestypub/honestyObj/query.do | |
| **数据规模** | **存量数据** | 分别输入‘中国’、‘国际’以及32个省级行政单位名称进行搜素，对搜索到的列表和详细信息页面进行抓取 |
| **增量数据** |  |
| **更新频率** | **存量数据** |  |
| **增量数据** |  |
| **数据过期判定标准** |  | |
| **增量数据识别方式** | 全量更新判断 | |
| **数据结构化程度（结构化、半结构化、非结构化）** | 非结构化 | |
| **身份验证方式** | 无 | |
| 数据采集项 | | |
| **信息项** | | **样例** |
| 企业名称 | | 北京得利加自控技术有限公司 |
| 组织机构代码 | | 70012017-7 |
| 处罚处理时间 | | 2014/06/05 |
| 处罚处理机构 | | 中国证监会 |
| 行政 处罚种类 | | 行政处罚(责令改正),行政处罚(警告),行政处罚(罚款) |
| 处罚内容详情 | | 中国证券监督管理委员会四川监管局  行政处罚决定书  　　　　韩露，女，1970年10月出生，通灵通法定代表人、董事长，住址：北京市东城区竹杆胡同。  　　　　李曦微，女，1966年10月出生，上海秦砖监事，住址：北京市西城区西四北八条。  　　　和本条第三款的规定，通灵通与上海秦砖构成一致行动人，  ................  　　　 中国证券监督管理委员会四川监管局  　　　 2014年6月5日 |







验收条件：

1 完整的代码（需要必要的注释）和sql语句

爬取的数据保存到到数据库，数据库样式设计自定，但要注意在表名前t\_

2 注意不要用Python自带的mysqldb，而是使用附件中的Python工具箱

使用工具包导入SQLAlchemy

3 要确保所编写程序的稳定性，调节合适的访问频率（注意网站的反爬取机制），以确保能够实现连续的爬取，程序不中断（重点）。

4 数据获取考核任务考核期为30天

语言选择：python（编译器推荐pycharm）

推荐入门学习材料：

极客学院Python初级课程：会员课程-知识体系图-Python课程

廖雪峰的博客：推荐看完函数部分即可，后面有选择的看。

百度：你要相信百度上的东西还是很全面的

推荐程序包：

Python推荐包：Selenium，Requests