中国人民银行罚单信息爬虫程序可行性报告

李昌浩

2021.12.27

本报告只旨在指出项目**困难之处**及**解决方案**。作为学生，我希望能发挥价值，**不说“做不到”**。

一、难题

网页格式较不统一

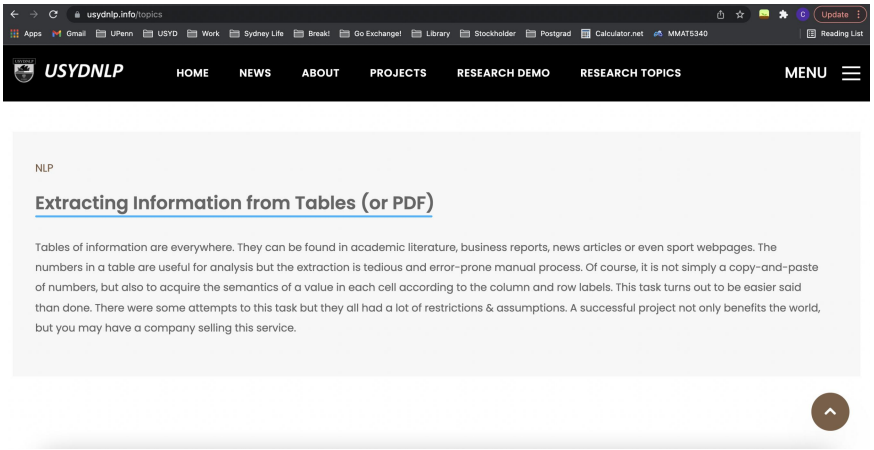
相比起银保监会，人民银行罚单信息的网页格式较为复杂。首先，银保监所有罚单网页格式一致，只需要用两次 Python 语言 Beautiful Soup 库 find_all 函数即可找到相关信息，而央行罚单的页面上只有一条条链接，其指向需要**另外点击**才能下载的文件（这里应该有自动下载的方法，正在研究）；其次，有些地区的罚单文件是 Excel 格式，有些是 Word 格式，也有许多文件是 PDF 格式，还有的干脆直接在网页上写罚单信息，非常混乱。此外，人民银行罚单信息网页标题不显示处罚什么公司，而只是写明了罚单号和日期，无法通过标题过滤有效信息。

各地罚单分别公布

同样对比银保监会官网：银保监会只有一个网站，而全国各省都有当地人民银行支行网站；另外，罚单信息也由**各省分别公布**，而非国家统一公布。这也就意味着我们需要针对每一个人民银行支行网站开发不一样的爬虫程序，这极大地加重了工作量（全国央行网站列表见文末）。

附件文件无法解析

文件内容读取，尤其是中文 PDF 文件的解析，到目前为止仍是难题。比如，悉尼大学自然语言处理小组（USYD NLP Group）在官网公开了 2021 年度招聘博士生候选人信息，其中一项目需要候选人在 4-8 年时间里写出 PDF 表解析的解决方法提案、程序及论文，其难度不言而喻。



二、流程


打开各地央行罚单信息网址
（网址链接可见附表）
抓取所需日期范围内罚单页链接
通过各罚单链接自动下载罚单文件
（PDF、Word、Excel）
从已下载文件中提取相关文字信息
保存相关罚单信息，删除下载文件

三、应对

上策：部分程序自动化

当我们无法完全通过一键抓取网络信息时，我考虑的是将不可程序自动完成的操作的改为延续传统人工操作。比如，第四步“从已下载文件中提取相关文字信息”较为不可行：

提取 PDF 内信息的 Python 库及其可行性		
pdf2htmlEX	极不可行	转换格式不符合
pdfminer	较不可行	无法读取中文字符
PyPDF2	极不可行	无法读取中文字符
tabula	极不可行	无法读取中文字符
camelot	较不可行	可以处理汉字，但读取后文字顺序被完全打乱

在此情况下，该步骤依旧维持人工执行，其他步骤继续尝试自动化即可。现在已经可以将步骤一自动化，接下来继续尝试第二、三、五步的 Python 程序方法。

中策：尝试完全程序化

该方法比较消耗时间：这是由于全国一共有 8+2+20+5 = **35** 个人民银行分支机构，每个网站的结构又不完全相同，所以需要编写 35 个不同的程序（尽管有部分网站结构相似，可以减轻一些工作量）。编程的时间远大于人工采集数据的时间，沉没成本较高。

下策：废弃该爬虫程序

完全由人手动搜索需要耗费更大的事件。现已完成的程序已拥有许多功能。比如，若使用者输入想要查询的日期范围（例：2021 年 7 月至今）：

使用者即可得到相关日期范围内罚单的链接。本例中人民银行广州分行自 2021 年 7 月以来只发布了 2 张罚单，这与广州分行官方网站信息相吻合：

四、参考

全国人民银行支行罚单信息网页链接

分支机构

上海分行	天津分行	沈阳分行	南京分行	济南分行
武汉分行	广州分行	成都分行	西安分行	营业管理部(北京)
重庆营业管理部	石家庄中心支行	太原中心支行	呼和浩特中心支行	长春中心支行
哈尔滨中心支行	杭州中心支行	福州中心支行	合肥中心支行	郑州中心支行
长沙中心支行	南昌中心支行	南宁中心支行	海口中心支行	昆明中心支行
贵阳中心支行	拉萨中心支行	兰州中心支行	西宁中心支行	银川中心支行
乌鲁木齐中心支行	深圳市中心支行	大连市中心支行	青岛市中心支行	宁波市中心支行
厦门市中心支行				

~ 地区分行 ~

上海分行：http://shanghai.pbc.gov.cn/fzhshanghai/113577/114832/114918/index.html
天津分行：http://tianjin.pbc.gov.cn/fzhtianjin/113682/113700/113707/index.html
沈阳分行：http://shenyang.pbc.gov.cn/shenyfh/108074/108127/108208/index.html
南京分行：http://nanjing.pbc.gov.cn/nanjing/117542/117560/117567/index.html
广州分行：http://guangzhou.pbc.gov.cn/guangzhou/129142/129159/129166/index.html
济南分行：http://jinan.pbc.gov.cn/jinan/120967/120985/120994/index.html
武汉分行：http://wuhan.pbc.gov.cn/wuhan/123472/123493/2164231/index.html
成都分行：http://chengdu.pbc.gov.cn/chengdu/129320/129341/129350/index.html
西安分行：http://xian.pbc.gov.cn/xian/129428/129449/129458/index.html

~ 营业管理部 ~

北京营业管理部：http://beijing.pbc.gov.cn/beijing/132030/132052/132059/index.html
重庆营业管理部：http://chongqing.pbc.gov.cn/chongqing/107680/107897/107909/index.html

~ 省会支行 ~

杭州中心支行：http://hangzhou.pbc.gov.cn/hangzhou/125268/125286/125293/index.html
福州中心支行：http://fuzhou.pbc.gov.cn/fuzhou/126805/126823/126830/index.html
合肥中心支行：http://hefei.pbc.gov.cn/hefei/122364/122382/122389/index.html
郑州中心支行：http://zhengzhou.pbc.gov.cn/zhengzhou/124182/124200/124207/index.html
……（略）……
呼和浩特中心支行：
http://huhehaote.pbc.gov.cn/huhehaote/129797/129815/129822/index.html

~ 非省会支行 ~

深圳市中心支行：http://shenzhen.pbc.gov.cn/shenzhen/122811/122833/122840/index.html
大连市中心支行：http://dalian.pbc.gov.cn/dalian/123812/123830/123837/index.html
青岛市中心支行：http://qingdao.pbc.gov.cn/qingdao/126166/126184/126191/index.html
宁波市中心支行：http://ningbo.pbc.gov.cn/ningbo/127076/127098/127105/index.html
厦门市中心支行：http://xiamen.pbc.gov.cn/xiamen/127703/127721/127728/index.html