央行罚单信息爬虫使用说明

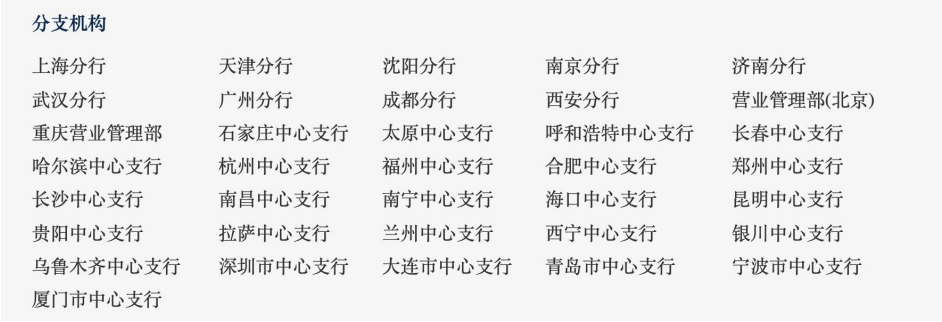
零、背景介绍

目标

收集中国人民银行（后文简称央行）所有**罚单数据**并整理，根据需求提取相关信息。

结构

央行的官网不像银保监会的那样收录了全国的罚单信息。全国 36 家央行附属分支机构分别在其不同网址上公开罚单信息：



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				

(http://www.pbc.gov.cn/rmyh/105226/105442/index.html)

各央行分支机构的网站源代码结构也略有不同，所以开发爬虫软件时不可以像教科书上写的那样根据网站结构的规律一致性编写程序。我使用笨方法，直接写了 36 个爬虫。或许你是一位有经验的开发者，你可以尝试优化该爬虫程序，使其更简洁、更易用。但央行网站的结构是我们无法改变的。因此，更好地理解央行组织结构或许会帮到你：

~ 地区分行 ~

上海、天津、沈阳、南京、广州、济南、武汉、成都、西安

共 9 家

~ 营业管理部 ~

北京、重庆

共 2 家

~ 省会支行 ~

石家庄、太原、呼和浩特、长春、哈尔滨、杭州、福州、合肥、郑州、长沙、南昌…

共 20 家

~ 非省会支行 ~

深圳、大连、青岛、宁波、厦门

共 5 家

以上是我根据自己的理解分类的央行机构，并非官方名称。但也不是瞎分的，其依据是：

- 地区分行：XX 分行
- 营业管理部（官方名称）
- 省会支行：XX 支行 且 XX 处城市皆为省会
- 非省会支行：XX 市 支行 且 XX 处城市均为非省会大城市

证毕。Quod erat demonstrandum。

网址

所有央行罚单网址都存于文档 pbcs.txt（文本格式）或 pbcs.xlsx（Excel 格式）。

特殊

需要特别注意的是以下网址的 HTML 源码：

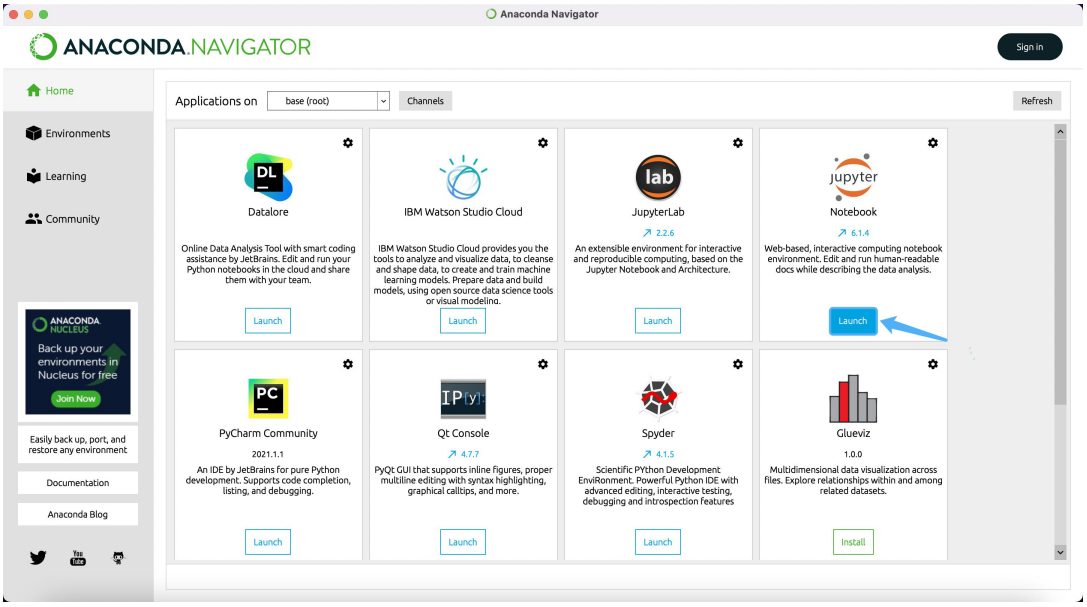
重庆	暂时没有找到有效方法爬取罚单（网站结构不同）
成都	注意 width="66"
北京	注意 width="140"

备注：其余网站 width 均为“100”。

一、使用方法

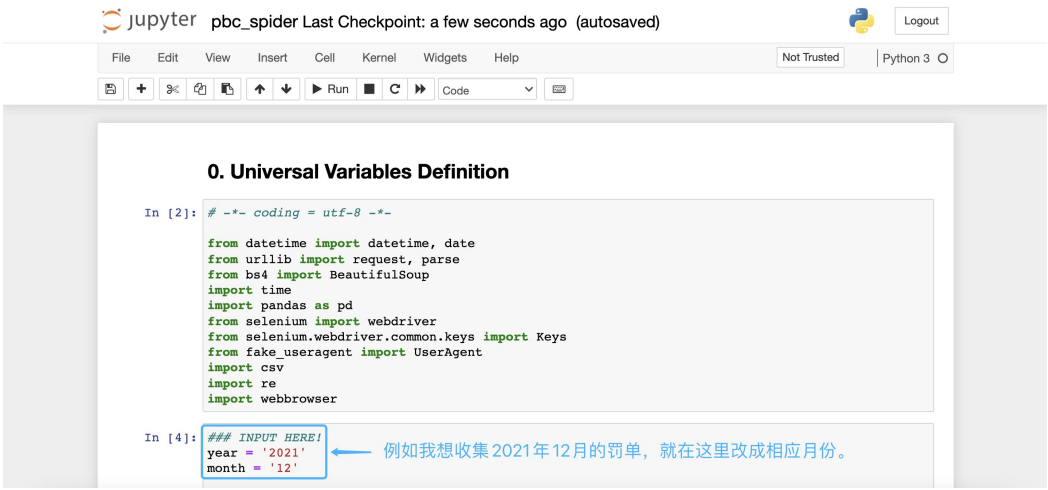
· 安装 Anaconda。详见 https://www.anaconda.com/。

· 打开 Anaconda，进入 Jupyter Notebook：

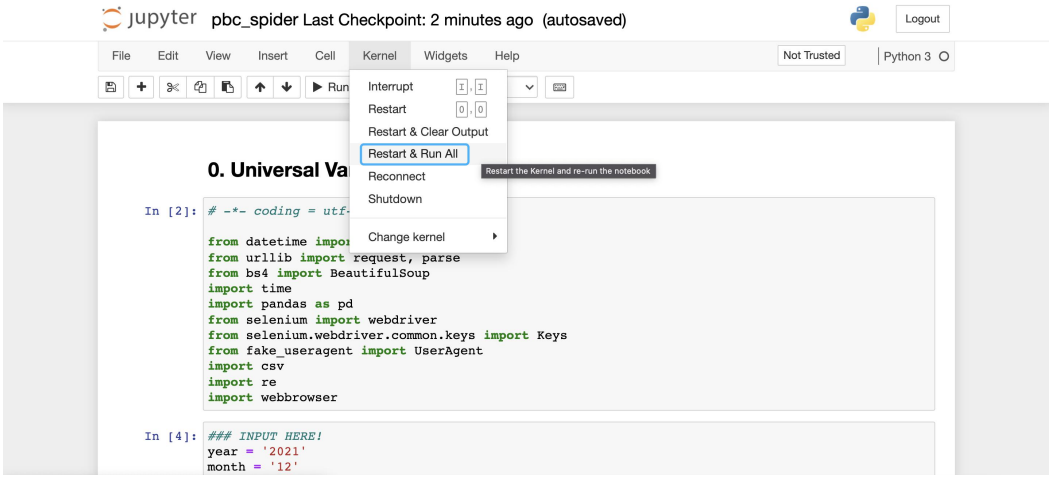


· 进入 Jupyter Notebook 后，找到 pbc_spider.ipynb 文件所在地。
（找文件的方法就和平时在电脑硬盘里找文件一样，打开一看你就懂了）

· 打开 pbc_spider.ipynb 文件，修改变量。比如想要 2021 年 12 月的罚单：



· 修改好日期后，点击 pbc_spider.ipynb 文件上方的“Cell”，“Restart & Run All”：



· 即可在文件夹中发现一堆（36 个）输出文件【年-月-地区.csv】：