# 辽宁数据爬取说明文档

## 一 数据来源

|  |  |  |  |
| --- | --- | --- | --- |
| 名称 | 网址 | 内容 | 说明 |
| 沈阳市科技条件平台 | <http://www.sykjtjpt.cn/h/equip/largeEquipment> | 《大型仪器》栏目下数据 | 网页做了重定向，可以抓取。  网站已经挂掉 |
| 辽宁省大型科学仪器共享服务平台 | <http://www.liaoninglab.com/index/dxyqfront/instrument/list> | 《大型仪器》栏目下数据 | post请求，带参数请求，目前看可以抓取。 |
| 沈阳市科学技术局 | <http://kjj.shenyang.gov.cn/kxjsj/zwgk/zcfg/glist.html> | 政策文件栏目 | 索引页正常，详情页为图片。 |
| 大连市科学技术局大连市外国专家局 | <http://www.kjj.dl.gov.cn> | 政策法规 | 索引页做了参数配置，详情页正常，目前看可以抓取。 |
| 鞍山市科学技术局 | <http://kjj.anshan.gov.cn> | 政务公开，双创信息栏目。 | 可以抓取，附件可以抓取。 |
| 抚顺市科学技术局 | <http://fskjj.fushun.gov.cn/fskjj> | 政策法规 | 可以抓取(详情页有问题） |
| 本溪市人民政府 | <http://www.benxi.gov.cn/publicity_bxskxjsj/jgzn> | 无任何说明 | 政府网站，无爬取说明。 |
| 丹东市科学技术局 | <http://ddkjj.dandong.gov.cn/> | 政策法规 | 索引页有索引，详情页为pdf文件或者doc文件 |
| 锦州市科学技术局 | <http://kjj.jz.gov.cn/> | 政策与解读栏目 | 可以抓取 |
| 营口市科学技术局 | <http://kjj.yingkou.gov.cn/dynamic/zw/openlist.html?categorynum=003004> | 政策法规 | 索引页网页源代码无数据，真实数据在重定向url中，该重定向url为post请求，携带数据中出现siteGuid,无法判断是什么数据疑似md5或者uuid生成的字符串。暂时无法解决问题，详情页正常。链接文件可以抓取。 |
| 阜新市科学技术局 | <http://kjj.fuxin.gov.cn/newslist.jsp?id=12352> | 科技政策栏目 | 可以抓取 网站响应速度极慢 |
| 辽阳市科学技术局 | <http://kjj.liaoyang.gov.cn/lyskjxxw/xxgk/zcfg/glist.html> | 政策法规 | 可以抓取 |
| 盘锦市科学技术局 | <http://kjj.panjin.gov.cn/8288/> | 政策文件 | 可以抓取 |
| 铁岭市科学技术局 | <http://www.tieling.gov.cn/tieling/zwgk62/xzqlgk94/skxjsj/index.html> | 页面下数据 | 有图片 |
| 朝阳市 | <http://kjj.zgcy.gov.cn/cyskjj/czfg/> | 政策法规模块 | 可以抓取 |
| 葫芦岛市科学技术局 | <http://kjj.hld.gov.cn/zwgkzdgz/jcgk/gzwj/> | 政策文件 | 可以抓取 |
| 沈阳市科技条件平台 | <http://www.sykjtjpt.cn/h/talent/talentService> | 人才服务 | 网站打不开 |
| 东北科技专家服务云平台 | [http://www.dbzjk.com/#/Mall/ExpertList?type=1](http://www.dbzjk.com/" \l "/Mall/ExpertList?type=1) | 专家 | 可以抓取 |
| 科易网 | <https://www.1633.com/tec/> | 科技成果下技术成果 | 可以抓取 |
| 专利检索及分析系统 | <http://218.60.146.4/pubsearch/portal/uiIndex.shtml> | 按**常规检索**，通过大市场提供的企业名称信息获取专利数据 账号：DBKJDSC2020;密码：DBKJDSC2020 |  |

## 数据库字段说明

表2.1 仪器设备表

|  |  |
| --- | --- |
| **仪器设备** | **字段** |
| 仪器名称 | equipment\_name |
| 型号规格 | specification\_model |
| 仪器类别 | instrument\_classification |
| 仪器原值 | original\_instrument\_value |
| **所属单位** | **workCompany** |
| 购置日期 | purchase\_code |
| 资产编码 | asset\_code |
| 分类编码 | classification\_code |
| 生产厂商 | manufacturer |
| 产地国别 | production\_country |
| 供货商 | supplier |
| 仪器所在地 | located\_area\_instrument |
| 仪器安放地 | instrument\_placement\_address |
| 联系人 | contact\_person |
| 联系电话 | contact\_number |
| 收费标准 | charges |
| 每小时占用费 | hourly\_occupancy\_fee |
| 电子邮箱 | email |
| 设备编号 | device\_id |
| 年有效机时 | effective\_time\_per\_year |
| 对外共享机时 | external\_sharing\_machine\_time |
| 使用状态 | status\_of\_use |
| 运行状态 | operating\_status |
| 技术指标 | technical\_index |
| 主要功能 | main\_function |
| 仪器简介 | instrument\_introduction |
| 主要附件 | main\_accessories |
| 技术特色 | technical\_characteristics |
| 仪器已服务数量 | number\_of\_services |
| 好评率 | praise\_rate |
| 成交量 | volume |

2.2 专家表

|  |  |
| --- | --- |
| **姓名** | **name** |
| 性别 | gender |
| 出生年月 | data\_of\_birth |
| 学历学位 | degrees |
| 专业领域 | professional\_field |
| **工作单位** | **workCompany** |
| 所属部门 | department |
| 专家职称 | expert\_title |
| 擅长领域 | areas\_of\_expertise |
| 人才简介 | talent\_profile |
| 工作或学习经历 | work\_study\_experience |
| 研究方向 | research\_direction |
| 研究成果 | research\_results |
| 获奖情况 | awards |
| 照片 | photo |

2.3 科技成果表

|  |  |
| --- | --- |
| **成果** | **字段** |
| 成果名称 | achievement\_name |
| **专利申请号** | **patent\_application\_no** |
| 专利类型 | patent\_type |
| 应用领域 | application\_filed |
| 技术成熟度 | technology\_maturity |
| 最近更新日期 | last\_update\_date |
| 来源 | source |
| 所在地 | location |
| 行业 | industry |
| 所有人 | owner |
| 项目简介 | project\_description |
| 图片 | image |

2.4专利表

|  |  |
| --- | --- |
| **专利** | **字段** |
| 专利名称 | patent\_name |
| **申请号** | **application\_number** |
| 申请日 | application\_date |
| 公开（公告）号 | public\_number |
| 公开（公告）日 | public\_date |
| IPC分类号 | IPC\_classification\_number |
| **申请（专利权）人** | **applicant** |
| **发明人** | **inventor** |
| 代理机构 | agency |
| 优先权号 | priority\_number |
| 优先权日 | priority\_date |
| 申请人地址 | applicant\_address |
| 申请人邮编 | applicant\_zip\_code |
| 申请人所在国（省） | applicant\_country |
| 摘要 | abstract |
| 权利要求书 | claims |
| 说明书 | manual |
| 全文图像 | full\_text\_image |
| 法律状态 | legal\_status |
| 同族文献 | family\_documents |
| 引证文献 | cited\_literature |

2.5 政策表

|  |  |
| --- | --- |
| **政策** | **字段** |
| **标题** | **Title** |
| 索引号 | index\_number |
| 发文机关 | issuing\_authority |
| 发文字号 | text\_number |
| 主题词 | subject\_heading |
| 主题分类 | subject\_classification |
| 成文日期 | data\_of\_writing |
| 发布日期 | release\_data |
| 来源 | source |
| 作者 | author |
| 内容详情 | details |
| 附件 | annex |
| 附件 |  |
| 原文链接 | original\_link |

## 项目部署说明

### 3.1 项目文件夹说明



图3.1 项目文件说明

### 3.2 环境说明

软件环境：爬虫模块基于python3.6，所需其余环境依赖在requirements.txt 中。谷歌浏览器，谷歌浏览器插件chromedriver。

硬件环境：linux服务器一台。

### 3.3 项目问题阐述

1. 所有数据均保存在MYSQL数据库中；
2. 代码会进行数据去重判断，将新数据保存在数据库中；
3. 部分网站会存在无法响应问题；
4. 专利检索依据公司，详情见下图。总共11502



图3.2 所需检索专利公司

1. 专利部分数据爬取存在反爬虫机制，主要为浏览器cookies缓存，其值失效时间为半天左右，必须重新更换cookies值，确保不被反爬虫机制发现。现在使用登录帐号信息进行cookies值维护。因为登录帐号需要输入验证码，而验证码使用代码请求，则会和登录验证码不一致。采用截图登录页面，并进行验证码裁剪，识别。最后登录。
2. 同一账号存在登录次数过多帐号异常被管理员封掉的问题，同一账号登录失败次数超过六次存在30分钟无法登录问题。



### 3.4 项目部署

1. 确保所需环境为python3（>=3.6即可）；
2. 使用虚拟环境；

2.1 创建虚拟环境 virtualenv -p python3 虚拟环境名称

2.2 进入虚拟环境 source ./bin/activate

3.在虚拟环境中安装所需依赖；pip install -r requirements.txt

4.编写sh脚本；当前python绝对路径 执行的python文件

5.给sh脚本添加可执行命令 chmod +x 名字.sh

6.执行vi ./etc/crontab 进行定时执行命令设置定时执行任务