# 物联网公司爬虫脚本使用说明

简介

物联网公司爬虫脚本是一个使用python脚本自动从百度百聘、华展云、脉脉、百度搜索与公司官网抓取公司数据的自动爬虫。名称叫做物联网爬虫，但是只要修改关键词也可以爬取物联网以外的公司。本项目旨在为销售提供便捷的公司查询系统，包括公司更新、公司过滤、查看公司联系方式等。本项目分为两大部分，后台的爬虫脚本，与前端的用户界面，下面的文档将详细介绍这两部分。本项目在python3环境下开发。

1. 爬虫脚本说明

爬虫脚本在开发时按功能分为了几个脚本，后来整合为一个主脚本，这里将对脚本的作用和实现做出说明。

* 1. 主爬虫脚本 crawler\_main.py

整合了百度百聘、华展云、脉脉、公司官网信息提取的主要脚本。

用户在网页上更新数据库就会自动运行此脚本，用户一般只需要运行此脚本，但是前端仍然保留了单独运行其他某一脚本的选项。

该脚本的主要运行流程为：

1. 首先从百度百聘按照关键词获得招聘相关职位的公司名称、招聘职位、公司简介、公司所在地区、公司官网。
2. 在华展云搜索该公司，搜到就可以获得公司内部人员联系方式、公司所属行业等信息，搜不到则该信息为空。将公司信息插入company表。
3. 用公司官网链接抓取公司官网文本，从文本中用jieba库提取关键词，插入数据库关键词表。
4. 登陆脉脉网站；在脉脉搜索公司名称，如果搜索到该公司员工，就把员工姓名、职位、领域添加到数据库脉脉表；没有搜索到则跳过。
5. 再次在百度百聘搜索该公司名称，获得该公司全部的招聘信息，插入到数据库百度百聘表。
6. 从公司招聘信息与公司简介中提取关键词，插入到关键词表。

百度百聘、华展云、脉脉搜索的功能被打包成类放在crawler\_main\_head.py中，作为对象引入。

对于低质量（没有联系方式、官网）的公司信息的过滤则主要放在前端中，对于老旧公司信息（可能已经失效信息）的过滤用前端从新至旧展示结果的方法可以起到一定效果。

注：百度百聘对于持续大量爬取会有一段时间封禁ip的操作，脉脉对于持续大量爬取可能会导致账号无法登陆，华展云目前没有发现任何异常。

因为前端中爬虫目录的路径固定为/crawler，需要把爬虫脚本放到/crawler文件夹内才能使前端正常工作

python连接数据库密码全部使用用户名root，密码misakaxindex，请使用全局搜索修改

需要的python库：urllib requests pymysql bs4 jieba synonyms

* 1. 华展云单独脚本 crawler\_huazhanyun.py

单独爬取华展云信息，因为以百度百聘为基础的信息质量层次不齐，不如直接在华展云搜索关键词，抓取到的公司全部都会有联系方式和主页链接

工作流程：

1. 登录华展云
2. 按顺序搜索每一个关键词
3. 获得搜索结果中公司的详细信息
4. 插入数据库

注1：华展云的排序方式有时间排序(time)，综合排序(auto)两种，可以在配置文件中修改

注2：华展云的登陆方式比较特殊，只需要用用户名密码和一个phpsid调用华展云接口就可以完成登陆，之后使用该sid调用其它华展云接口，不涉及cookie

* 1. 招聘信息脚本 crawler\_huazhanyun\_hireinfo.py

给从华展云单独爬取到的公司搜索招聘信息。

因为正常流程中会先搜索百度百聘，直接获得招聘信息；但是如果是使用华展云脚本，会没有百度百聘和脉脉信息，所以需要再运行此脚本获得招聘信息。

工作流程：

1. 从数据库倒序获取公司名称和所在城市
2. 用公司名称和城市搜索百度百聘
3. 将结果中的招聘数据插入数据库百度百聘表

注：如果是从华展云脚本添加的没有招聘数据的公司，应该正常搜索到招聘数据然后插入；但是如果是从主脚本添加的公司，已经搜索过了一遍招聘信息，会再次搜索一遍，但是招聘信息不会重复插入，因为百度招聘给出的sid在数据库中设置是唯一的。

* 1. 公司简介脚本 crawler\_huazhanyun\_description.py

给从华展云单独爬取到的公司搜索公司简介。

需要爬取公司简介的原因基本同上，公司简介来自于百度招聘。

工作流程：

1. 从数据库倒叙获得公司名称和所在城市
2. 用公司名称和城市搜索百度百聘
3. 用搜索结果中的公司简介更新数据库中的公司简介

注：百度百聘搜索公司全名只会出现此公司的相关招聘信息或者就是没有结果，不会出现其他公司信息。

* 1. 公司关键词脚本 crawler\_company\_keyword.py

给没有添加关键词的公司手动添加关键词。

工作流程：

1. 从数据库获得招聘信息
2. 提取招聘信息工作名称的关键词
3. 插入到关键词表

注：关键词表中，关键词唯一，公司id与公司名称对应公司表(company)

* 1. 公司主页信息脚本 crawler\_company\_homepage\_information.py

给没有爬取过主页的公司爬取一遍主页中的信息。

工作流程：

1. 从数据库获取公司名称和链接
2. 如果已经有官网链接则直接使用；没有则尝试用百度搜索官网
3. 用官网链接爬取官网文本，跟随超链接爬取两层
4. 从文本中提取<p><h1~h6><label>标签中的文字
5. 对提取出的文字用jieba库提取关键词
6. 将关键词和比率、公司名称、id插入到关键词表
   1. 脉脉脚本 crawler\_maimai.py

根据公司名称爬取脉脉信息

工作流程：

1. 从数据库获得公司名称
2. 登陆脉脉，在脉脉搜索改公司
3. 将搜索到的结果添加回数据库，搜索结果为空或错误则跳过

注1：脉脉数据库表中mmid唯一，避免重复插入

注2：脉脉登陆使用cookies，每次使用账户密码登陆，长期爬取数据可能会导致账户异常无法登陆

* 1. 爬虫配置文件 crawler\_config.json

上面每一个脚本都会按需读取这个配置文件，配置文件包括代理设置，输出设置，等待时间，用户名密码，关键词等

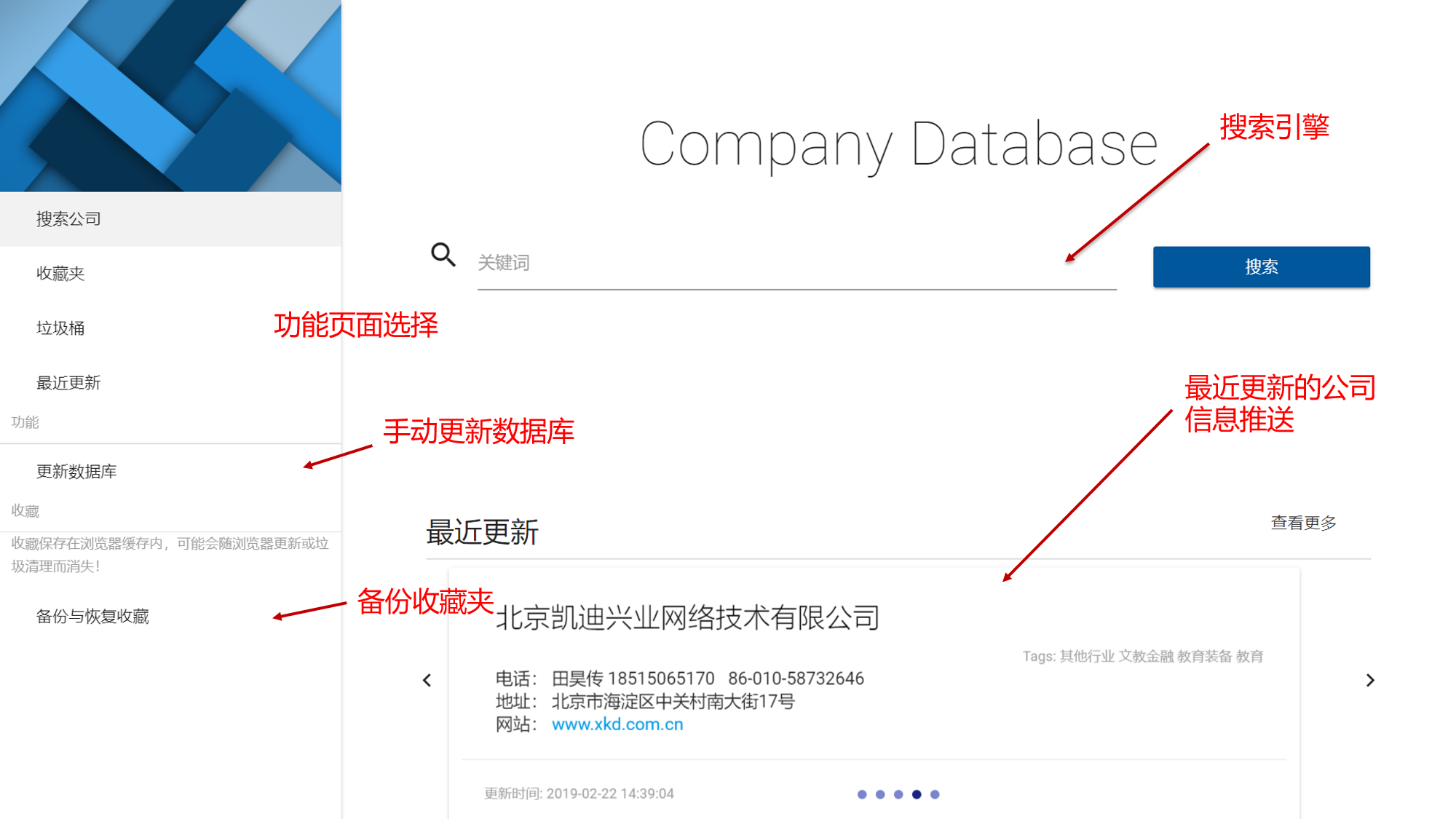
文件说明：

* + - proxy: intel/none/socks5 *设置不同的代理<intel> child-prc.inte.com:913 <socks5>127.0.0.1:1080 <none>无代理*
    - print\_method: terminal/file *选择爬虫输出到终端或文件(php在线执行脚本时需要从文件读取输出)*
    - time\_out: 20 *(second)设置爬虫请求超时时间，服务器部署在海外可以调长一点*
    - BAIDU *百度百聘爬虫设置*
      * time\_sleep: 60 *(second)设置百度百聘爬虫延时，建议60秒以上*
      * fail\_count\_limit: 10 *最大重试次数（已废弃）*
      * fail\_sleep: 100 *超出最大重试次数后的睡眠时间（已废弃）*
      * keywords: [“计算机视觉”, “物联网”] *百度百聘搜索的关键词 （只用于百度百聘脚本）*
      * keycities: [“上海”, “深圳”, “广州”, “北京”] *百度百聘搜索的城市（只用于百度百聘脚本）*
      * row\_wanted: 100 *对百度百聘的搜索结果插入前多少条（只用于百度百聘脚本）*
      * headers: {} *request设置的头*
    - HUAZHAN *华展云爬虫设置*
      * time\_sleep: 20 *华展云爬虫延时*
      * fail\_count\_limit: 10 *最大失败重试次数（已废弃）*
      * fail\_sleep: 100 *超出最大失败重试次数后的睡眠时间（已废弃）*
      * keywords: [“物联网”,“计算机视觉”] *华展云搜索关键词*
      * page\_start: 0 *华展云搜索结果起始页*
      * page\_end: 500 *华展云搜索结果结束页*
      * sort\_type: time/auto *华展云结果排序方式<auto>综合排序，<time>时间倒序*
      * headers: {} *华展云的request头*
      * username: 13818209932 *脉脉用户名*
      * password: intel@123 *脉脉密码*
    - MAIMAI *脉脉配置文件*
      * time\_sleep: 20 *脉脉爬虫延时*
      * username: 13818209932 *脉脉用户名*
      * password: intel@123 *脉脉密码*
    - MYSQL *数据库设置*
      * db\_username*:* root *数据库用户名*
      * db\_password*:* misakaxindex *数据库密码*
      * db\_dbname: findcompany *数据库名称*

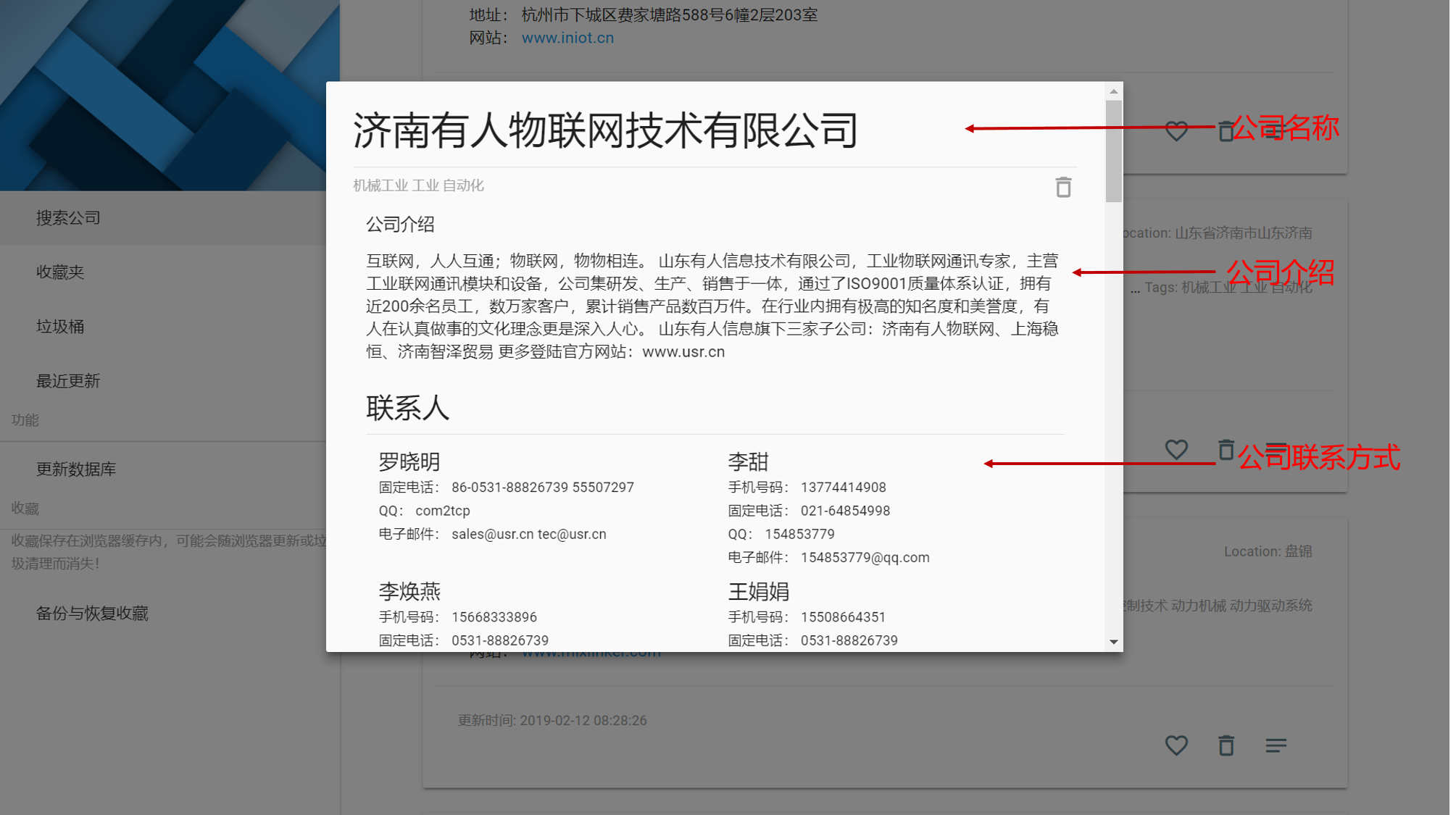
1. 前端使用与结构说明
   1. 前端使用

前端页面分为六个主要页面：

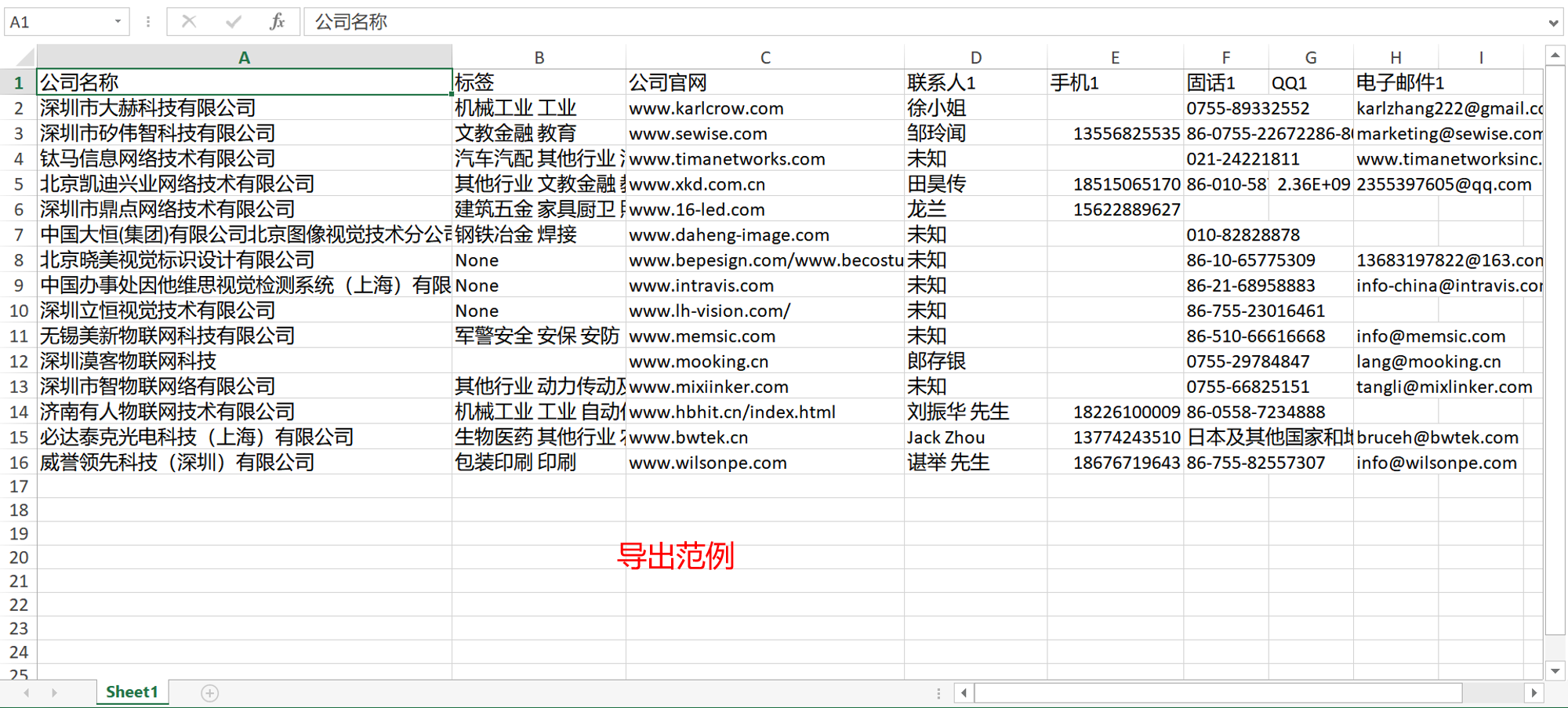
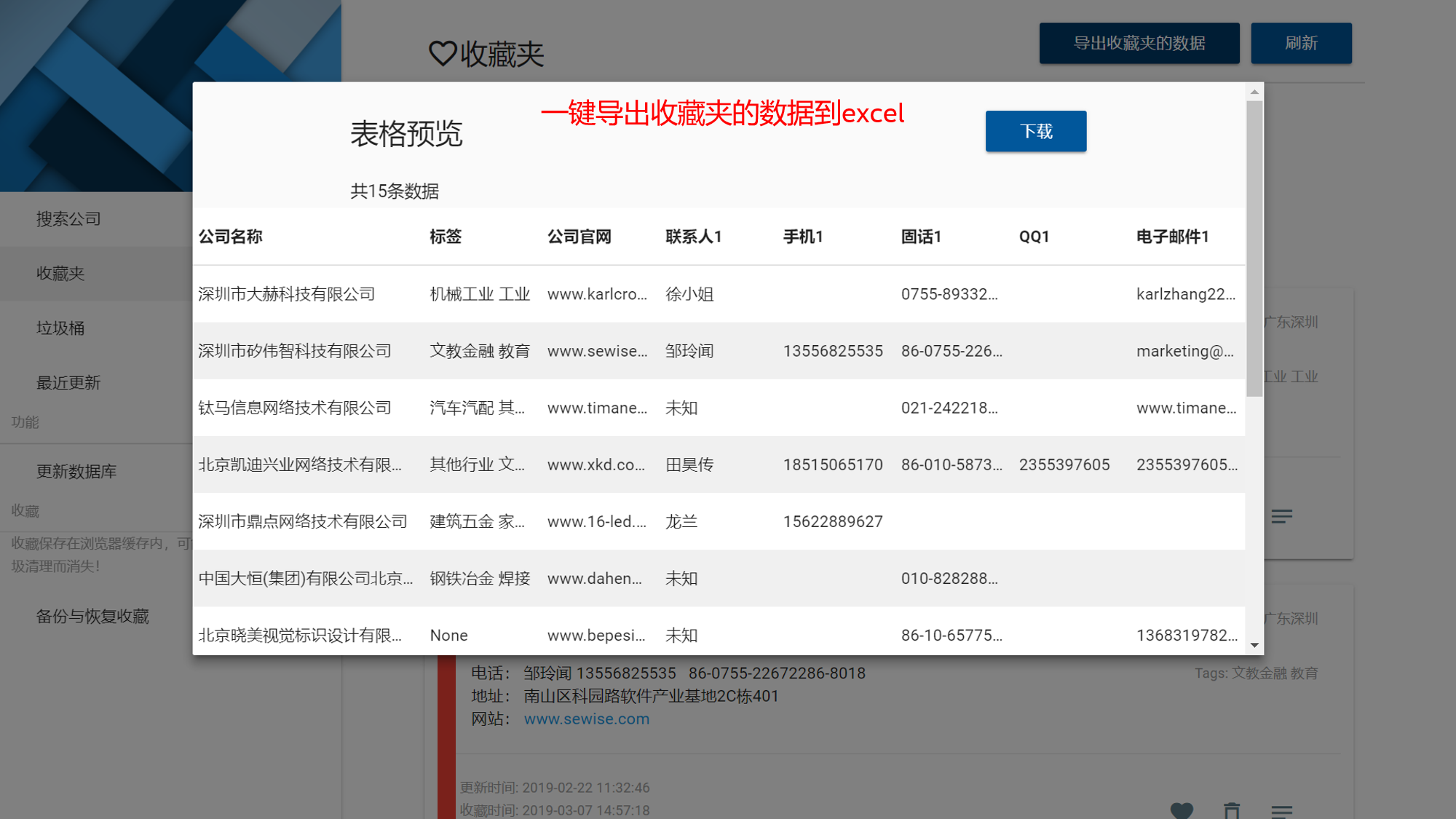
* + 1. 搜索引擎







* + 1. 收藏夹



* + 1. 垃圾桶



* + 1. 最近更新



* + 1. 更新数据库



更新设置：



* + 1. 备份恢复与收藏



* 1. 前端代码结构说明
     1. 文件结构说明

前端的编写直接由html+js完成，主要使用jquery库和vue.js库，没有使用node。后端使用php实现大部分功能，重写的近似搜索方法则使用python的flask。

* + 1. js代码说明

每个页面独自的功能存放在与页面的index.html同文件夹下的js/main.js中，为了在渲染搜索结果使用v-for方法，搜索结果的列表被作为了一个vue组件。同样，在其他出现的公司列表中使用的都是vue，但是因为他们之间的细微差别，用的不是同一套代码而是相似的代码放在main.js中。

搜藏夹和垃圾公司过滤都使用浏览器的localstrorage存放数据，可以不需要用户登陆，缺点是只能存放在本地。所以设计了备份与恢复页面，可以将localstorage中的数据导出/导入。转换csv下载的功能实现函数在/js/function.js中的JSONToCSVConvertor，函数本身已经十分清楚了，注意在windows下生成的文件需要加bom头。

* + 1. php代码说明

所有php文件存放在/php目录下，其中的功能有搜索、执行数据更新、停止数据更新、清空数据库、获得更新历史、查询更新进程、查询最近更新数据等功能，具体脚本对应的功能在代码中注释。

php数据库用户名密码配置在mysql.php中

php.ini中需要激活extension mysqli和curl

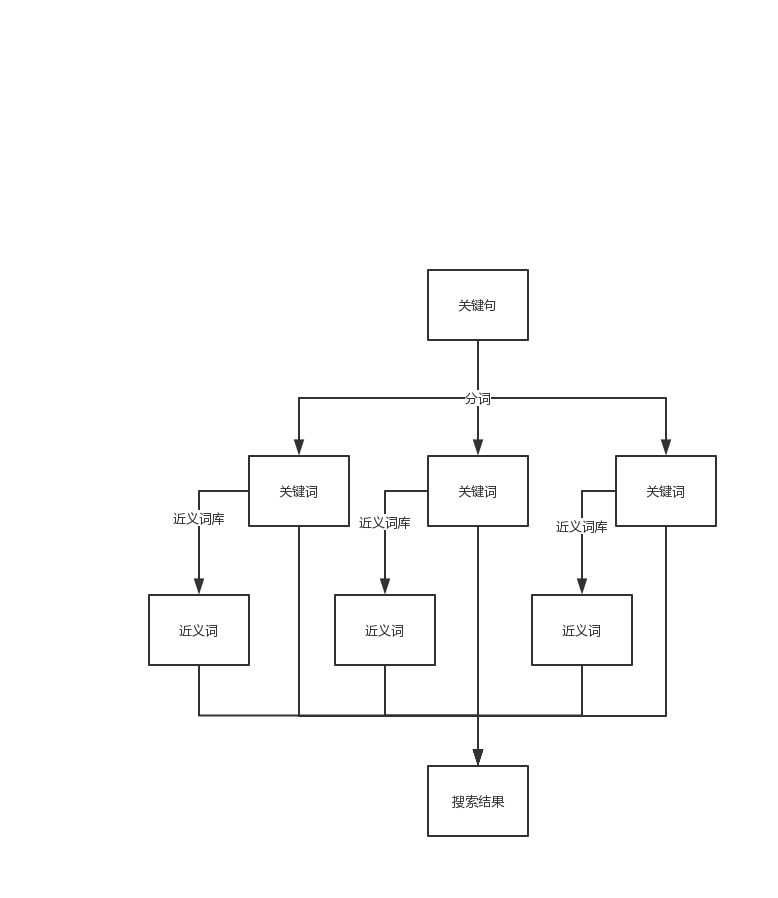
* + 1. 搜索方法说明

php搜索有两种方法，简单的search.php是数据库company表内直接全局搜索，返回速度快，出错概率小；第二种方法是search\_new.php中，数据库联合company表与keyword表搜索，使用了一个数据库过程，数据库设置出错（特别是字符集）可能导致无法运行的可能性很大。（前端会自动先尝试search\_new，超时/错误返回后尝试search）

* + 1. 近似搜索说明

还有一种搜索方法是新设计的php转发到python的flask服务器，这里使用python是为了调用jieba分词与synonym中文近义词，对搜索关键句先分词，再查找每一个关键词的近义词，去重，再去数据库中全部搜索在汇总返回。这样可以实现类似近似搜索的效果。flask服务的python脚本存放在/flask文件夹中。

近似搜索逻辑：



* + 1. 使用的css样式库

使用Materialize CSS(MIT License)，<https://materializecss.com/>

1. 数据库
   1. 数据库初始化

数据库可以使用爬虫脚本代码中附带的database.sql文件初始化，数据库名为findcompany，表的定义在sql文件中，另外包括了高级搜索使用的数据库过程的定义。

初始化的数据库没有数据，如果需要样例数据，请把/old/example database.sql中的数据导入数据库。

1. 其他

脚本调试信息的输出：脚本设置输出为terminal时，脚本print输出到命令行；设置为file时，脚本print输出到文件crawler\_log.txt。前端网页通过读取此文件来得到脚本的输出。

脚本启动时会把自己的pid与启动时间输出到mainPID.txt，同时向数据库添加一次更新记录。前端通过判断kill pid -0的返回值确认脚本是否正在运行，用kill pid -9杀死脚本。

1. 总结

脚本中使用了jieba分词和synonym中文近义词库来提取关键词、实现近义词搜索，使用了request和bs4来抓取解析http request。具体的抓取接口反向工程为自己分析实现，抓取后的数据处理流程也是自己实现。前端中主要使用materialize作为css库，以及vue jquery辅助，页面设计、前端功能代码、后端php与flask代码为自己实现。