**拉勾网订制网站开发说明**

###### 基本信息

* 开发人：杨羽飞
* 开发日期：2018-12-4~2018-12-6
* 网站名称：拉勾网
* 网站url：https://www.lagou.com
* 网址类型：招聘

###### **网站抓取接口说明**

**2.1 获取行业类型接口**

* 抓取需求：获取拉勾网的行业类型数据、融资类型数据和公司规模数据
* 请求接口：https://www.lagou.com/gongsi/
* 请求方式：GET
* 参数说明：无
* 接口返回数据类型：html
* Headers设置：无
* 其它说明：无

**2.2 获取招聘城市数据**

* 抓取需求：获取拉勾网的所有招聘城市的id和名字
* 请求接口：<https://www.lagou.com/gongsi/allCity.html?option=0-0-0>
* 请求方式：GET
* 参数说明：
  + option：必须，固定写0-0-0即可
* 返回数据类型：html
* Headers设置：无
* 其它说明：无

**2.3 获取指定城市的公司列表**

* 抓取需求：获取拉勾网的指定城市的公司导航页信息（主要是公司id字段）
* 请求接口：https://www.lagou.com/gongsi/${cityCode}-${rongziCode}-${industryCode}-0.json
* 请求方式：POST
* 参数说明：

Url中的参数：

* + cityCode：城市id；0表示全国；
  + rongziCode：融资类型id；0表示不限；
  + industryCode：行业类型id；0表示不限；

POST参数（application/x-www-form-urlencoded）：

* first：boolean类型；false即可；
* pn：int类型；page页码；
* sortField：int类型；1表示按公司发布岗位数量从大到小排序；
* havemark：int类型；固定为0；
* 返回数据类型：json
* Headers设置：需要UA和Referer；
* 其它说明：发起请求的页面为https://www.lagou.com/gongsi/

**2.4 获取公司信息**

* 抓取需求：根据公司id获取其详细信息
* 请求地址：http://www.lagou.com/gongsi/${corpId}.html
* 请求方式：GET
* 参数说明：
  + corpId：公司id
* 返回数据类型：html
* Headers设置：需要UA和Referer；
* 其它说明：数据位于css标签“#companyInfoData”的节点内；内容为json格式；

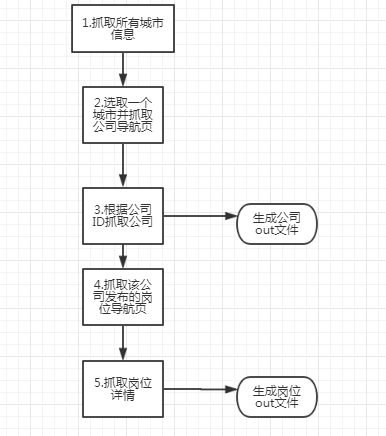
**2.5 获取职位列表**

* 抓取需求：获取指定公司发布的岗位列表（主要是岗位id）
* 请求地址：https://www.lagou.com/gongsi/searchPosition.json
* 请求方式：POST
* 参数说明（application/x-www-form-urlencoded）：
  + companyId：公司id
  + positionFirstType：string类型；固定为“全部”；
  + schoolJob：boolean类型；是否为校招类型；
  + pageNo：int类型；页码；
  + pageSize：每页多少条数据（每页最多可返回100条数据）；
* 返回数据类型：json
* Headers设置：需要UA和Referer；
* 其它说明：发起请求的页面为http://www.lagou.com/gongsi/${corpId}.html

**2.6 获取职位信息**

* 抓取需求：获取指定职位的详细信息
* 请求地址：http://www.lagou.com/jobs/${jobId}.html
* 请求方式：GET
* 参数说明：
  + jobId：职位id
* 返回数据类型：html
* Headers设置：需要UA和Referer；
* 其它说明：无

###### **网站抓取流程图**



###### **其它说明**

**4.1 去重方式**

根据需求，该网站属于copy拉勾展示信息，每天一次，因此抓取不需要去重。

**4.2 开发语言**

Nodejs开发，并将进程托管至PM2。

**4.3 测试结果**

* 开始抓取时间：2018-12-08 08:00:00
* 抓取结束时间：2018-12-09 14:03:00
* 耗时：30h
* 数据量统计：
  + 公司信息：204185条；生成文件大小256M；
  + 职位信息：384045条；生成文件大小507M；
* 结论：七个大城市数据：
  + 武汉：公司数据8881条；岗位信息9078条；
  + 上海：公司数据35221条；岗位信息12272条；
  + 深圳：公司数据31727条；岗位信息11834条；
  + 广州：公司数据23133条；岗位信息41019条；
  + 成都：公司数据15161条；岗位信息17444条；
  + 杭州：公司数据19473条：岗位信息32499条；
  + 北京：公司数据48563条；岗位信息137146条；

可见这7个城市的公司信息占所有公司信息量的89%；岗位信息占到所有岗位信息的68%。另外像郑州（公司1000多家，岗位1873）等数据量都相对较小。

另外还有许多网站的公司并不发布招聘岗位。