# 拉勾网反爬系统说明

### 基本说明

拉勾网与2019年1月10日升级反爬策略（1月9号可以正常获取数据）。受到影响的接口为https://www.lagou.com/gongsi/${cityCode}-${rongziCode}-${industryCode}-0.json和https://www.lagou.com/gongsi/searchPosition.json

这两个接口主要用于获取公司导航信息和公司发布岗位信息。

### **反爬/加密方式**

访问加密接口主要需要2个cookie，分别是user\_trace\_token和LGRID。这两个参数由接口https://a.lagou.com/collect?v=1&\_v=j31&a=327242320&t=pageview&\_s=1&dl=https%3A%2F%2Fwww.lagou.com%2F%3F\_from\_mid%3D1&ul=zh-cn&de=UTF-8&dt=%E6%8B%89%E5%8B%BE%E7%BD%91-%E4%B8%93%E4%B8%9A%E7%9A%84%E4%BA%92%E8%81%94%E7%BD%91%E6%8B%9B%E8%81%98%E5%B9%B3%E5%8F%B0\_%E6%89%BE%E5%B7%A5%E4%BD%9C\_%E6%8B%9B%E8%81%98\_%E4%BA%BA%E6%89%8D%E7%BD%91\_%E6%B1%82%E8%81%8C&sd=24-bit&sr=1366x768&vp=563x609&je=0&\_u=MEAAAAQBK~&jid=839549805&cid=489375350.1548057431&tid=UA-41268416-1&\_r=1&z=1408632925返回。而collect这个请求及其相关参数是由https://a.lagou.com/js/a.js生成的。a.js经过混淆加密，而且其和浏览器进行多次交互，因此无法使用nodejs或是其它js引擎对其进行代码层面的模拟（主要是无法模拟浏览器里面的window对象以及对window对象的相关属性设置）。另外，获取到的cookie和申请该cookie的IP地址进行绑定，也就是说cookie无法共享使用。

### **解决方案**

**3.1 headless方案**

使用phantomjs或者selenium等服务器端 JavaScript API来完全模拟用户行为，并获取相关cookie。

* 优点：不需要关注加解密等繁琐的细节，可见即可得；
* 缺点：需要完全启动一个webdriver才可以获取相关cookie，效率较低。
* 结论：而且由于使用付费代理，且cookie不能共享且需求方需要每日更新，因此该方案不可用。

**3.2 破解加密过程和相关算法**

经分析a.js文件，代码层面破解难度较大，破解时间不定（无法完全估计工期），因此该方案不可用。

**3.3 采集移动端（m站）**

* 优点：拉勾网的移动端采集相对容易，虽然也需要cookie，但是是直接返回结果，不需要和加密文件交互（user\_trace\_token和LGRID）。
* 缺点：移动端无法通过多个条件进行组合查询。这有可能导致7个大城市的相关数据采集不全。而且移动端数据没有web版的详细。
* 结论：在没有更好的方案的情况下可以采集移动端

**3.4 绕过反爬策略**

经过调研分析发现，移动端cookie和web端cookie通用，那么可以通过访问移动端拿到相关cookie，然后用2个主要cookie（user\_trace\_token和LGRID）来访问web端加密接口。

* 优点：可以获取到web端的全部数据；代码改动代价较小；
* 缺点：如果拉勾升级策略，使得移动端和web端cookie不通用，那么该方案就会失效。
* 结论：当前采用该方案