# PikaPika简历爬虫项目说明

舜飞后端 陈锦瀚 2016.5.17

## 一．项目代号

**PikaPika**

## 二．需求说明

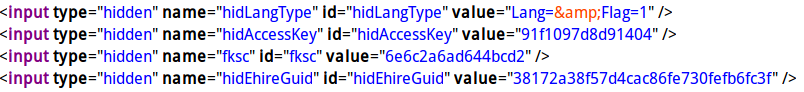
定时根据若干关键字对51jobs进行简历抓取，简历数据进行保存，方便HR以Web端离线形式进行检索/查看。

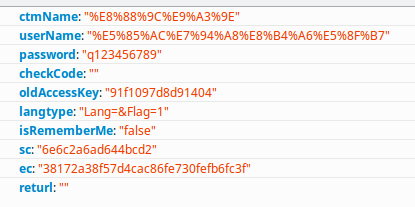
## 三．搜索流程

1.登陆入口：<http://ehire.51job.com/MainLogin.aspx>

登陆携带的字段包括隐藏字段，如下：

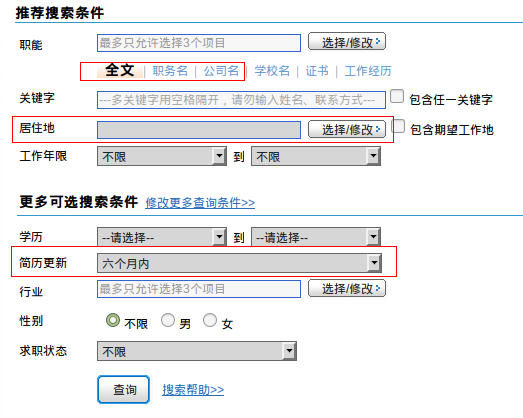




1. 登陆完毕后本地生成cookie

访问搜索常规入口：

<http://ehire.51job.com/Candidate/SearchResumeIndex.aspx>



不如直接页：

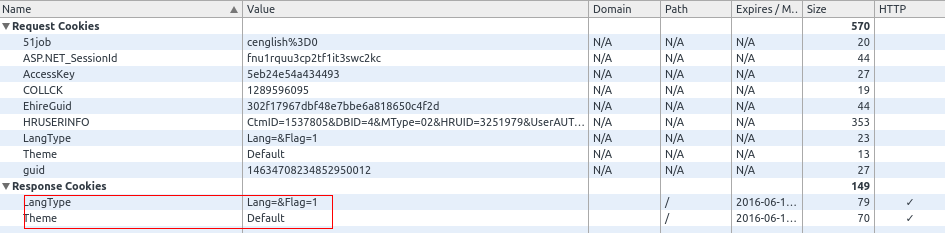
<http://ehire.51job.com/Candidate/SearchResume.aspx>

显示简历列表

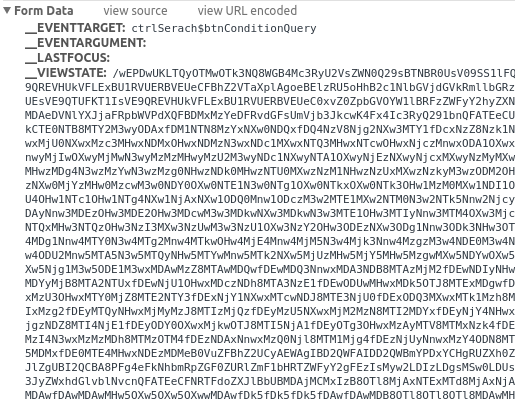
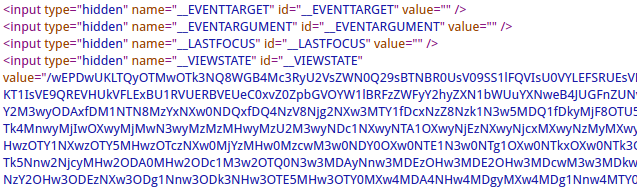


检索字段：**关键字【分类】 居住地 简历更新**

请求情况和相应情况如下（貌似永恒不变，棒极了！）：

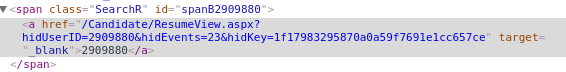


但是隐藏字段很多，需要逐项测试



1. 爬取所有简历详细信息

搜索列表附带简历详细链接如下





根据相关字段进行数据存储。

## 四．设计分析

编码使用Go语言，数据库自行设计。关键字表/简历表/简历详情表/关键字-简历表

两种方式抓取：

1. 提前进行关键字设置，关键字数据表一张，进行定时爬取，存入Mysql数据库，每次抓取直接覆盖之前即可，如果不同关键字爬取到同一个人，则简历-关键字表增加一记录。Web端查看时根据已有关键字进行检索，无过多选择，不能进行关键字设置，关键字下的简历则可以检索。

考虑数据量过大，可通过关键字ID取余数进行简历表/简历详情表分表

1. 实时抓取，当关键字不满足时，可设置关键字实时抓取，直接显示，不存入数据库，并且可选择是否加入定时抓取计划。

## 五．编码实现

1.模拟加载登陆页面链接，提取隐藏字段，输入用户名/验证码等模拟登陆，保存cookie。

2.模拟加载搜索最终页，更新cookie，提取隐藏字段，构造参数。

3.接2，url携带参数进行模拟加载搜索最终页，得到搜索列表，更新cookie。搜索列表数据信息可并发进行存入数据库。

4.接3，提取搜索列表简历链接，进行模拟加载简历详情页，对数据进行提取存入数据库。

5.Web端显示数据增删查改即可。

## 六．难点

1.登陆需突破验证码

2.分析服务器端发送了哪些加密字段并隐藏发送url

3.Go语言并发抓取如何分配任务，如何储存cookie，动态更新cookie

4.离线网站前端界面如何设计。

5.定时抓取是手动run还是自动设置run。

手动run两种，命令行run，其次web端点击启动。