甲方名称：

某网站

甲方需求：

- 抓取目标网站：58.com，ganji.com
- 抓取数据内容：

(1) 58同城、赶集网的招聘数据；

数据字段如下：

<1> 58.com相关数据字段

i. 工作职位信息列标题

'省份', '城市', '分类1', '分类2', '分类3', '标题', '招聘企业', '招聘企业ID', '薪资待遇', '学历要求', '招聘职位', '工作年限', '工作地址', '福利待遇', '联系电话', '职位描述', 'Scrape_url'

ii. 招聘企业（公司）信息列标题

'企业ID', '企业名称', '公司资质', '公司性质', '公司行业', '公司规模', '联系人', '联系电话', '邮箱', '企业网址', '公司地址', '公司介绍', 'Scrape_url'

iii. 求职简历信息列标题

'省份', '城市', '分类1', '分类2', '姓名', '性别', '年龄', '申请职位', '认证方式', '更新时间', '基本情况', '求职意向', '自我介绍', '个人特长', '学历教育', '在校情况', '获得证书', '语言能力', '工作经验', '项目经验', '专业技能', 'Scrape_url'

<2> gnaji.com相关数据字段

i. 工作职位信息列标题

'省份', '城市', '分类1', '分类2', '分类3', '标题', '招聘企业', '招聘企业ID', '职位名称', '月薪', '最低学历', '工作经验', '年龄', '招聘人数', '联系电话', '工作地点', '公司福利', '职位描述', 'Scrape_url'

ii. 招聘企业（公司）信息列标题

'公司ID', '公司名称', '公司规模', '公司行业', '公司类型', '联系人', '联系电话', '公司地址', '公司网站', '岗位福利', '公司介绍', 'Scrape_url'

iii. 求职简历信息列标题

'省份', '城市', '分类1', '分类2', '分类3', '姓名', '性别', '年龄', '认证方式', '摘要', '更新时间', '期望职位', '期望月薪', '期望地区', '最高学历', '工作年限', '籍贯', '个人特长', '工作经验', '教育经历', '自我描述', '语言技能', '专业技能', '证书奖项', '项目/培训经验', 'Scrape_url'

(2) 58同城所有城市的"家政服务 > 维修"类商户数据，具体包括"家电/房屋/家具/电脑/手机/空调/数码"类的商户数据；

数据字段如下：

'province'(省份), 'city'(城市), 'category1'(一级分类), 'category2'(二级分类), 'category3'(三级分类), 'title'(标题), 'publish_date'(发布/更新日期), 'shop_id'(商家ID), 'shop_name'(商家名称), 'address'(商家地址), 'phone'(联系电话), 'vip_years'(会员年限), 'evaluation_number'(累计评价), 'reservation_number'(预约总数), 'service_area'(服务区域), 'service_class'(服务类别), 'brand'(品牌), 'door-to-door_service'(是否上门), 'contacts'(联系人), 'description'(详情描述), 'scrape_url'(详情页原始链接)

技术难点：

- 数据量大，采集周期较长，对代理IP的需求较多；
- 网站有反采集策略，访问频率过高会被网站封IP；

- 产品列表页有页数限制，不能看到所有页；
- 详情页有多种模板，数据项样式不统一；

实现方案：

- 使用多机器、多进程相结合的方法，提高下载和数据（CPU密集型操作）提取速度；
- 通过大量稳定高匿HTTP代理IP轮换发出请求，并严格控制每个IP的两次访问间隔，以有效防止请求被网站拦截；
- 通过"区域"筛选器，尽可能多的让列表页页数减少；
- 分别处理不同模板的数据提取规则；
