

## 甲方名称:

某网站

## 甲方需求:

- 抓取目标网站: 菁优网 (<http://www.jyeoo.com/>)
- 抓取数据内容: 初中、高中、小学的所有学科, 按教材、年级、课本、章节的组题数据; 试题中的图片文件要存储为本地文件;
- 数据字段如下:  
'学科', '教材版本', '年级学期', '课本', '章节', '试题内容', '试题出处', '解析', '组卷', '真题', '难度', 'Scrape\_url'(本条数据对应的原始网页url)
- 结果数据格式: csv格式+html格式

## 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;
- 部分试题需要账号登录才能看到解析内容;

## 实现方案:

- 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;
- 为了保持试题内容原有的结构和样式, '试题内容' 和 '解析'部分保留原始数据的html格式;
- 结果数据文件, CSV格式文件是所有试题的合集, HTML格式文件是一个题目对应一个.html文件;



**西安鲲之鹏网络信息技术有限公司**

**选择我们, 所有数据都是你的!**



公司名称: 西安鲲之鹏网络信息技术有限公司

网 址: <http://www.site-digger.com/>

地 址: 陕西省西安市雁塔区长安中路99号长安文化综合大厦11821室

邮 编: 710061

联系电话: 029 - 87553281

手 机: 13571845363 齐工

13389148466 周工

客 服 QQ: 1649677458 或 312602670

邮 箱: [hello@site-digger.com](mailto:hello@site-digger.com)