

甲方名称:

厦门大学经济学院

甲方需求:

- 抓取目标网站: 中国裁判文书网 (<http://www.court.gov.cn/>)
- 抓取数据内容: 定制采集“中国裁判文书网”2017年1月1日前的所有文书
- 数据字段如下:
'case_type'(案件类型), 'document_type'(文书类型), 'referee_date'(裁判日期),
'keyword'(关键词), 'case_name'(案件名称), 'trial_procedure'(审判程序),
'case_number'(案号), 'court_name'(法院名称), 'document_id'(文书ID),
'document_content'(文书全文), 'document_tail'(文尾), 'document_url'(文书详情页链接)
- 结果数据格式: csv格式+MySQL格式

技术难点:

- 数据量非常大, 采集周期较长, 对代理IP的需求较多;
- 网站有反采集策略, 容易出现验证码;
- 搜索结果有访问页数限制, 每个搜索最多只能看到前500条, 不能看到所有页;
- 文书详情页有多种模板, 数据项样式不统一;

实现方案:

- 启动多个进程, 每个进程负责采集一个类型(刑事案件、民事案件、行政案件、赔偿案件、执行案件等)的文书;
- 进程中采用多线程结构, 并使用两个多线程函数来分别完成文档搜索和文档详情页的采集;
- 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;
- 先获取“按关键词筛选”的数据项, 再结合高级搜索选项<案件类型+文书类型+裁判日期>和关键词来进行搜索, 其中裁判日期的范围以天为单位; 这样可以最大限度地获取尽可能多的文书数据;
- 分别处理不同模板的数据提取规则;
- 前期使用MongoDB来存储结果数据, 后期再导出为需要的数据格式;



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
