甲方名称：

上海某信息技术公司（因保密协议限制无法公开具体信息）

甲方需求：

实现对微信公众号文章数据进行定制采集，并进行长期维护。

1. 采集目标：搜狗微信文章；
2. 通过搜狗搜索引擎，匹配指定关键字，对命中到的关键字文章进行采集，采集属性和微信公众号文章均保持一致；
3. 通过客户提供的数据接口获取搜索关键词，并将搜索采集到的结果数据通过客户提供的数据接口发送给客户；
4. 采集时间范围：采集当天及一天内（24小时内）发布的文章；
5. 采集文章属性应包括：标题、正文内容（含html）、发布时间、来源、公众号-账号名称、阅读量、点赞数、文章详情页链接；
6. 采集频率：每日采集，每两小时采集一次（因为有账号登录限制，每次搜索只能查看前100条结果，因此采用全天采集的方式，以尽可能获取更多的数据）；
7. 采集量：每天约10万条文章数据（搜索词约900多个）；

技术难点：

- 搜狗账号封锁很厉害，账号登录困难；
- HTTP请求量大，每天至少数十万；

实现方案：

- 使用双重队列管理，一个队列负责搜索页下载，另外一个队列负责详情页采集；同时每个队列配合以多线程处理，以加快处理速度；
- 使用SSDB做后端实现的CACHE来保存搜索到的结果文章队列，可以让多个服务器上的运行的采集进程以及不同时段启动的采集进程，可以共享搜索找到的文章队列，可以有效过滤重复的文章。

项目状态：

微信公众号文章数据采集于2016年10月上线运行，目前仍在稳定进行中。
