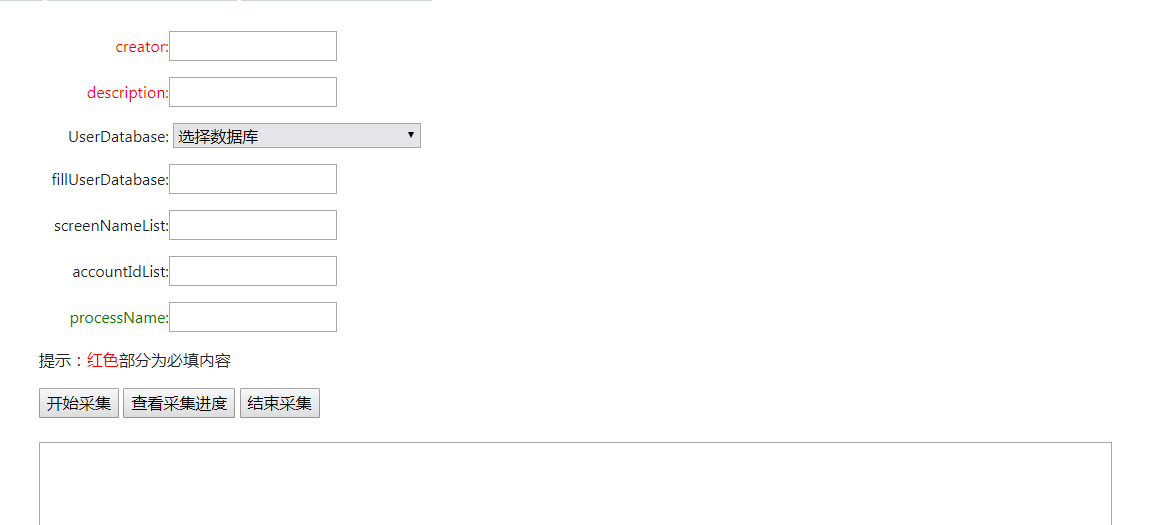
采集检索系统

一共5个模块：1）用户信息采集模块；2）用户推文采集模块；3）关键词推文采集模块；4）状态监控模块；5）检索查询模块。

数据库中集合有Prcoess（存储采集进程的信息）、DbAndColName（存储集合名与类型）。

## 用户信息采集模块



|  |  |
| --- | --- |
| Creator | 采集进程的创建人 |
| description | 对这个采集任务的描述 |
| UserDatabase | 选择存储的数据库 |
| FillUserDatabase | 自己填写数据库 |
| ScreenNameList | 推特的屏显名，用逗号隔开 |
| AccountIdList | 推特用户ID，用逗号隔开 |

采集用户用户详细信息时，直接将用户的信息与朋友、粉丝ID一起采完，不要选择是否采集朋友、粉丝ID。

## 用户推文采集模块

除了上面用户信息采集表格中的以外，还有如下：

|  |  |
| --- | --- |
| HistoryOrRealtime | 采集用户实时推文还是采集历史推文 |
| CaptureNum | 采集推文数量 |

## 关键词推文采集模块

除了上面用户信息采集表格中的以外，还有如下：

|  |  |
| --- | --- |
| Keywords | 采集推文的关键词 |
| HistoryOrRealtime | 采集实时关键词推文还是采集历史推文 |

## 状态监控模块

有4个模块，全部是实时更新，1）代理模块（有哪些代理正常，哪些异常）；2）服务器负载模块（包括CPU，内存，磁盘带宽信息）；3）正在运行的采集进程（包括创建者，进程描述，采集进度(这个需要讨论)）；4）点击进程查看，可以显示当前采集的最新的3条数据。

## 检索查询模块

包括3个模块（3个页面），1）原始信息查询（输入ID或者screen\_name获取完整的信息）；2）关键词匹配查询；3）关键词扩展查询

1. 原始信息查询

功能：只用于显示1条推文或者一个推特用户的完整信息。

输入：

|  |  |
| --- | --- |
| Database | 选择数据库 |
| ScreenName | 用户屏显名 |
| AccountID | 用户ID |
| TweetID | 推文ID |

输出：一条完整的信息。

1. 关键词匹配查询

在进行这个功能前，先执行ES建表，（还是每个mongodb都建？效率问题），用户数据库建立description,推文数据库建立text。

用户数据库返回信息：

|  |  |
| --- | --- |
| ScreenName | 用户屏显名 |
| AccountID | 用户ID |
| description | 用户描述 |

推文数据库返回信息：

|  |  |
| --- | --- |
| TweetID | 推文ID |
| CreatedAt | 创建时间 |
| Text | 用户描述 |

1. 关键词扩展查询

输入一些关键词，返回更多相关关键词的推文。（不慌）