**输入内容：**

网金趋势在和讯中的微博（http://t.hexun.com/13436232/default.html）从这个网页中的源码提取所需的数据，这些网页会由田园同学抓取后提供。

之后这些网页原始HTML页面的结构，做出通用的过滤模板。

**输出内容：**

输出包括一条微博内容的全部属性，以及这条微博下对应的评论属性，详细内容见以下两个表。

webo\_table表存储微博的基本信息，其表结构如下所示。

|  |  |  |
| --- | --- | --- |
| 数据字段 | 数据类型 | 说明 |
| weibo\_id | int  (auto\_increment) | 为每条微博消息，自定义一个微博id，其能自增或者使用uuid，  现估计微博的数量是：8w/day |
| weibo\_content | varchar | 微博内容，长度在1-280个字节  当是转发的内容时，需要将转发和内容组合 |
| weibo\_topic | varchar | 从微博中抽取的内容，用于最热话题 |
| weibo\_author | varchar | 发布微博的作者名称 |
| come\_from | smallint | 0：新浪；1：腾讯；2：网易；3：搜狐；4：和讯财经 |
| weibo\_time | datetime | 将不同微博网站时间统一为yyyy-mm-dd hh:mm:ss的格式 |
| comment\_num | int | 评论数，用于最热微博 |
| zhuanfa\_num | int | 转发数，用于最热微博 |
| status | smallint | 微博消息状态：0：存在；1：标记删除 |
| raw\_id | bigint | 原始微博的id,主要用于微博评论更新 |

comment\_table表存储微博评论的基本信息，其表结构如下所示。

|  |  |  |
| --- | --- | --- |
| 数据字段 | 数据类型 | 说明 |
| comment\_id | unsigned int | 评论id |
| weibo\_id | unsigned int | 评论所属微博id |
| comment\_content | varchar | 评论的内容 |
| comment\_author | varchar | 作者名称 |
| comment\_time | datetime | 评论时间 |