关于微博爬虫的功能说明

一．功能概述

通过关键字和日期进行筛选，爬取符合过滤条件的微博信息，包括话题、大 V、关键字三个方面的内容。

注：以下所有功能均包括：

如果需要增加追踪微博转发层数，可以依次迭代。

可以定位部分水军和僵尸粉。

二．功能细述

2.1 过滤条件

日期

式样：20180522

规则：设置起始时间和终止时间

关键词

枚举目标微博中出现的关键词

2.2 结果数据结构

2.2.1 关键词爬取

关键词爬取是通过在微博全网中搜索关键词，获取相关目标微博的爬虫模块，该模块目前可以爬取 2 层微博数据。

第一层微博

微博信息

微博内容

作者主页地址

评论数

转发数

点赞数

发表日期

Tag(该微博的标志字段)

评论信息
评论人 id
评论人主页地址
评论内容
评论所属微博内容
评论所属微博日期
Tag

转发信息
转发人 id
转发人主页地址
转发评论内容
转发评论所属的微博内容
转发日期
Tag

第二层微博

第二层微博是转发后的微博信息，即转发的转发。
微博信息
转发人主页地址
转发评论内容
转发微博的转发数
转发微博的评论数
转发微博的点赞数
转发日期
被转发微博发表日期
微博关键字（用于搜索转发微博时使用）
转发评论关键字（用于搜索转发微博时使用）
Tag

2.2.2 大 V 微博爬取

大 V 微博爬取是通过限定关键字和时间对指定大 V 的微博进行过滤和筛选，目前可支持爬取两层微博，追踪层数可增加，大 V 个数理论上不限，同时可以对评论的粉丝信息进行收集。

第一层微博

微博信息
微博内容
作者主页地址
评论数
转发数

点赞数
发文日期
Tag(该微博的标志字段)

评论信息
评论人 id
评论人主页地址
评论内容
评论所属微博内容
评论所属微博日期
Tag

转发信息
转发人 id
转发人主页地址
转发评论内容
转发评论所属的微博内容
转发日期
Tag

评论粉丝信息
粉丝主页地址
粉丝关注的忍术
粉丝的粉丝数
粉丝的微博数
粉丝的性别
粉丝的所在地（如果有）
该粉丝所属微博的内容
Tag

第二层微博

第二层微博是转发后的微博信息，及转发的转发。
微博信息
转发人主页地址
转发评论内容
转发微博的转发数
转发微博的评论数
转发微博的点赞数
转发日期
被转发微博发表日期
微博关键字（用于搜索转发微博时使用）
转发评论关键字（用于搜索转发微博时使用）
Tag

2.2.3 话题微博爬取

通过对微博的话题的进行爬取，获取所有话题微博的相关信息。本功能目前只支持一层微博的爬取。

话题信息

话题被阅读的次数
话题讨论次数
话题参与人数
Tag

微博信息

微博内容
作者主页地址
评论数
转发数
点赞数
发文日期
Tag

三．其他项目信息

稳定性

目前测试过连续爬取 10 个小时左右的任务，但是长期爬取是否可行无法确保，新浪可能会做出相应的措施，若被新浪发觉，理论上可以通过跟换微博账号的方法进行规避。

建议频率

1. 编写脚本每天定时爬取，只抓取前一天的微博内容，每天迭代。
2. 隔一段时间启动脚本，爬取某个时间段内的消息。

功能迭代

目前微博爬虫可支持如上的方向及相关字段的获取，如本项目具有一定的实际意义，并需要在实验室中投入使用，业务方可根据实际需求提前 3-5 个工作日提出具体要求，这样在截止日获取数据的风险较低。

后续如果有长期稳定的需求改动可以及时提出，由本人配合修改。

结果文件

目前支持 excel 和 json 两种格式的文件输出。

其他相关爬虫项目

微信公众号可以爬取公众号文章及点赞阅读等基础的数据，评论信息由于腾讯的控制是无法获取的（开发进度：暂时阻塞）。

百度贴吧可以爬取相应吧内的帖子，帖子可以由起始时间和关键字来过滤（开发进度：80%）。