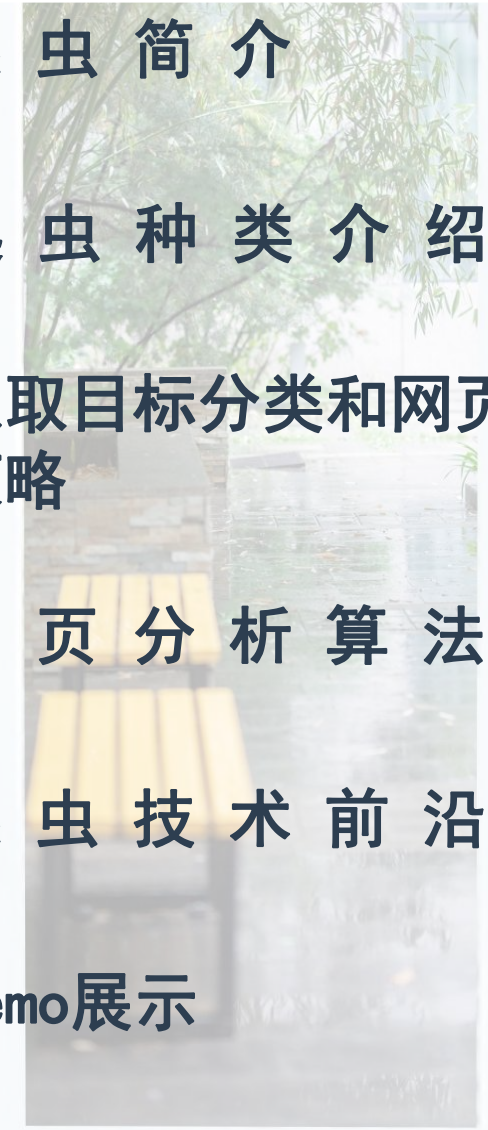
爬虫

第五组：李梓石，赵宇航，闫敏行，李若雪，郝思然

北京理工大學

目录



德以明理
學以精工

PART 01

爬虫简介

主讲人：郝思然

1.1 产生背景

- 随着网络的迅速发展，万维网成为大量信息的载体，如何有效地提取并利用这些信息成为一个巨大的挑战。搜索引擎，例如传统的通用搜索引擎AltaVista，Yahoo!和Google等，作为一个辅助人们检索信息的工具成为用户访问万维网的入口和指南。但是，这些通用性搜索引擎也存在着一定的局限性。

1.2 爬虫的定义

何为爬虫？

网络爬虫（又称为网页蜘蛛，网络机器人，更经常的称为网页追逐者），是一种按照一定的规则，自动的抓取万维网信息的程序或者脚本。

VS

为何叫爬虫？

由于专门用于检索信息的“机器人"程序像蜘蛛一样在网络间爬来爬去，因此，搜索引擎的“机器人”程序就被称为“蜘蛛"程序，即爬虫。

1.3 发展历史

1989年

Tim Berners-Lee发明的万维网，引入三个重要技术。

统一资源定位器(URL)，我们通过它来访问我们想看的网站；

内嵌的超链接，让我们可以在网页之间导航，例如产品详情页；网页不仅包含**文本**，还包括**图像、音频、视频**和**软件组件**。

1990年

第一个**网络浏览器**由Tim Berners-Lee发明，被称为**WorldWide网页(无空间)**，以WWW项目命名。在网络出现一年后，人们有了一条途径去浏览它并与之互动。

1991年第一个**网页服务器**和第一个http://网页页面。

1.3 发展历史

1993年

6月第一台网页机器人——**万维网漫游器**，用来测量网页的大小。

12月首个基于爬虫的网络搜索引擎 — **Jump Station**。**Jump Station**带来了新的飞跃。它是第一个依靠网络机器人的**WWW搜索引擎**。

2000年

（**API**表示应用程序编程接口） 2000年，Salesforce和eBay推出了自己的API，程序员可以用它访问并下载一些公开数据。

网页API通过收集网站提供的数据，为开发人员提供了一种更友好的网络爬虫方式。

1.3 发展历史

2004 年

2004年，**Beautiful Soup**发布。它被认为是用于网络爬虫的最复杂和最 先进的库，也是当今常见和流行的方法 之一。

2005-2006年

网络抓取软件的可视化，2006 年，Stefan Andresen和他的 Kapow 软件发布了**网页集成平 台6.0版本**，它允许用户轻松简 单的选择网页内容，并将这些 数据构造成可用的excel文件或 数据库。

1.4 爬虫过程

爬取过程：

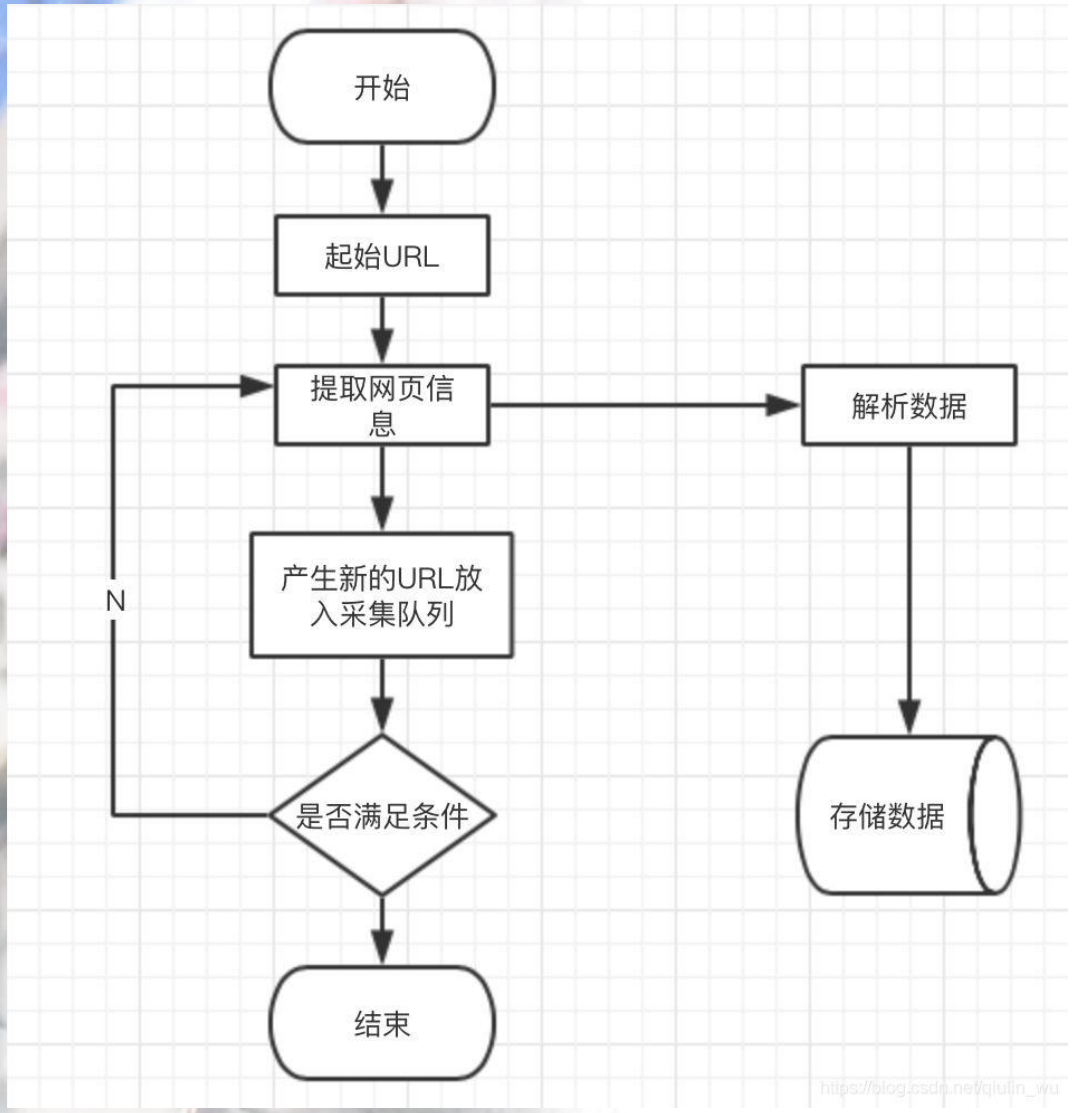
（1） 分析目标网站

（2） 发起请求

（3） 获取响应内容

（4） 解析内容

（5） 保存数据



1.5 面临的问题

- 截止到 2007 年底，Internet 上网页数量超出 160 亿个，研究表明接近 30%的页面是重复的
- 动态页面的存在：客户端、服务器端脚本语言的应用使得指向相同 Web 信息的 URL 数量呈指数级增长。
- 爬虫需要在单位时间内尽可能多的获取高质量页面

1.6 爬虫的种类

通用网络爬虫

deep web爬虫

聚焦网络爬虫

增量式爬虫

PART 02

爬虫种类介绍

主讲人：李若雪

2.1.1 通用网络爬虫

通用网络爬虫

通用网络爬虫又称**全网爬虫**。爬行对象从一些种子URL扩充到整个Web，主要为门户站点搜索引擎和大型Web服务提供商采集数据。这类网络爬虫的爬行范围和数量巨大，对于爬行速度和存储空间要求较高，对于爬行页面的顺序要求相对较低，同时由于待刷新的页面太多，通常采用并行工作方式，但需要较长时间才能刷新一次页面。虽然存在一定缺陷

，通用网络爬虫适用于为搜索引擎**搜索广泛的主题**，有较强的应用价值。

2.1.2 通用网络爬虫

深度优先策略是按照深度由低到高的顺序，依次访问下一级网页链接，直到不能再深入为止。爬虫在完成一个爬行分支后返回到上一链接节点进一步搜索其它链接。当所有链接遍历完后，爬行任务结束。

优势

- 适合垂直搜索或站内搜索

劣势

- 爬行页面内容层次较深的站点时会造成资源的巨大浪费

2.1.2 通用网络爬虫

广度优先策略是按照网页内容目录层次深浅来爬行页面，处于较浅目录层次的页面首先被爬行。当同一层次中的页面爬行完毕后，爬虫再深入下一层继续爬行。

优势

- 能够有效控制页面的爬行深度，避免遇到一个无穷深层分支时无法结束爬行的问题
- 实现方便
- 无需存储大量中间节点

劣势

- 需较长时间才能爬行到目录层次较深的页面

2.2.1 聚焦网络爬虫

聚焦网络爬虫（Focused Crawler），又称**主题网络爬虫**（Topical Crawler），是指选择性地爬行那些跟主题有相关性内容的网络爬虫。和通用爬虫相比，聚焦爬虫只需要爬行与主题相关的页面，极大地节省了资源，还可以很好地满足一些特定人群对特定领域信息的需求。

2.2.2 聚焦网络爬虫优势

优势1

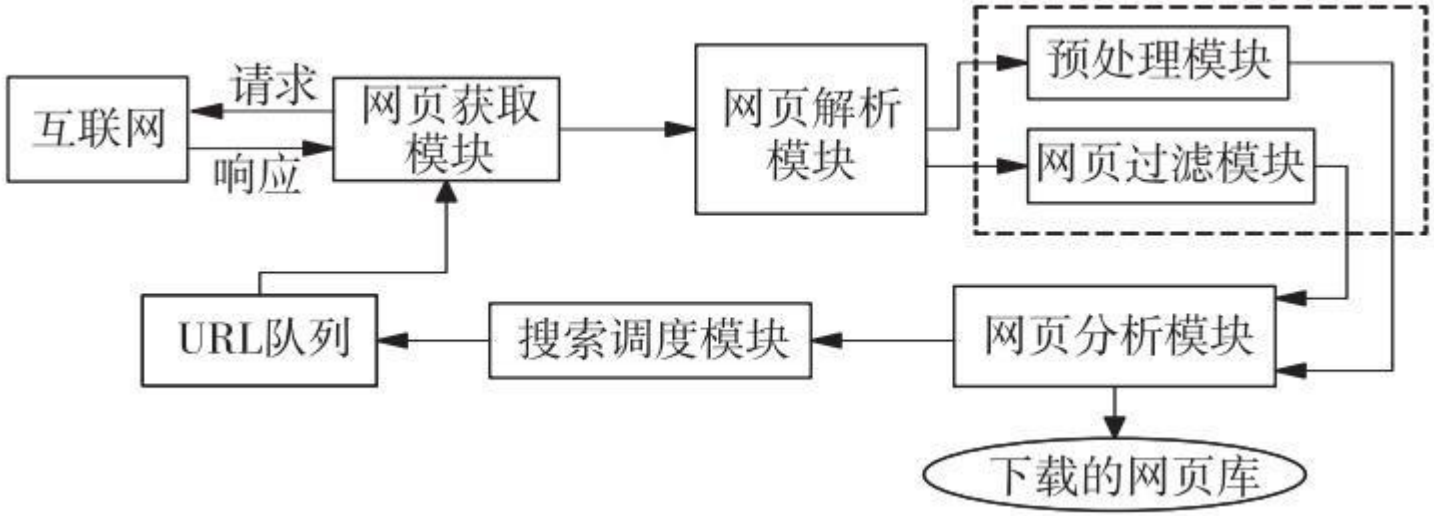
相比通用爬虫只能提供粗略的信息，主题爬虫主题明确且系统能够精准地获取有效信息。

VS

优势2

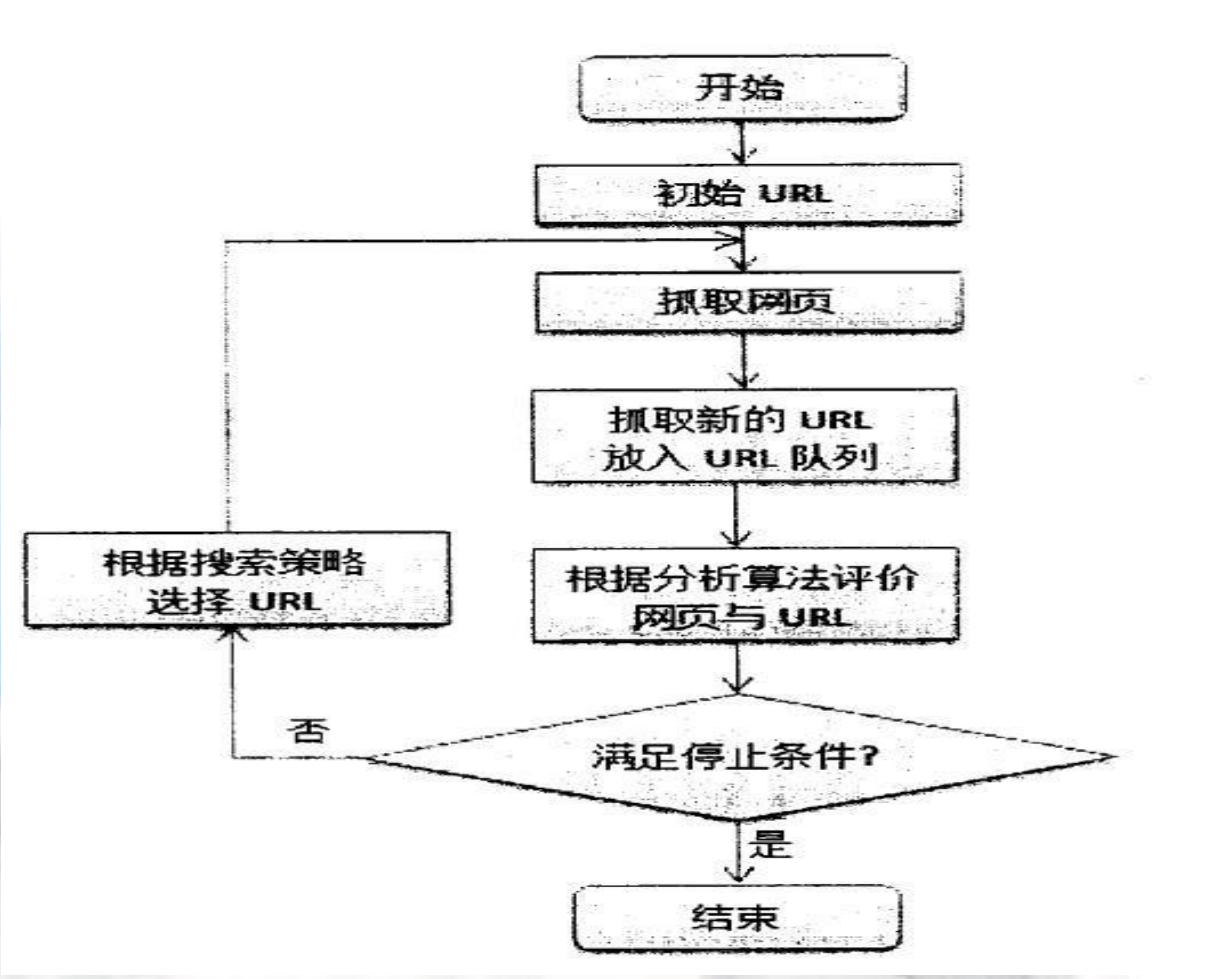
主题爬虫在存储网页URL时需要判断该URL与主题的相关性，尽可能地筛选出与主题相关的页面。

2.2.3 聚焦爬虫模块



聚焦爬虫模块图

2.2.4 聚焦爬虫过程



聚焦爬虫过程图

2.3.1 增量式网络爬虫

增量式网络爬虫是指对已下载网页采取增量式更新和只爬行新产生的或者已经发生变化网页的爬虫，它能够在一定程度上保证所爬行的页面是尽可能新的页面。 和周期性爬行和刷新页面的网络爬虫相比，增量式爬虫只会在需要的时候爬行新产生或发生更新的页面，并不重新下载没有发生变化的页面，可有效减少数据下载量，及时更新已爬行的网页，减小时间和空间上的耗费，但是增加了爬行算法的复杂度和实现难度。

2.3.2 增量式网络爬虫目标

目标1：保持本地页面集中存储的页面为最新页面

（1）统一更新法

（2）个体更新法

（3）基于分类的更新法

VS

目标2：提高本地页面集中页面的质量：

广度优先策略、PageRank 优先策略等。

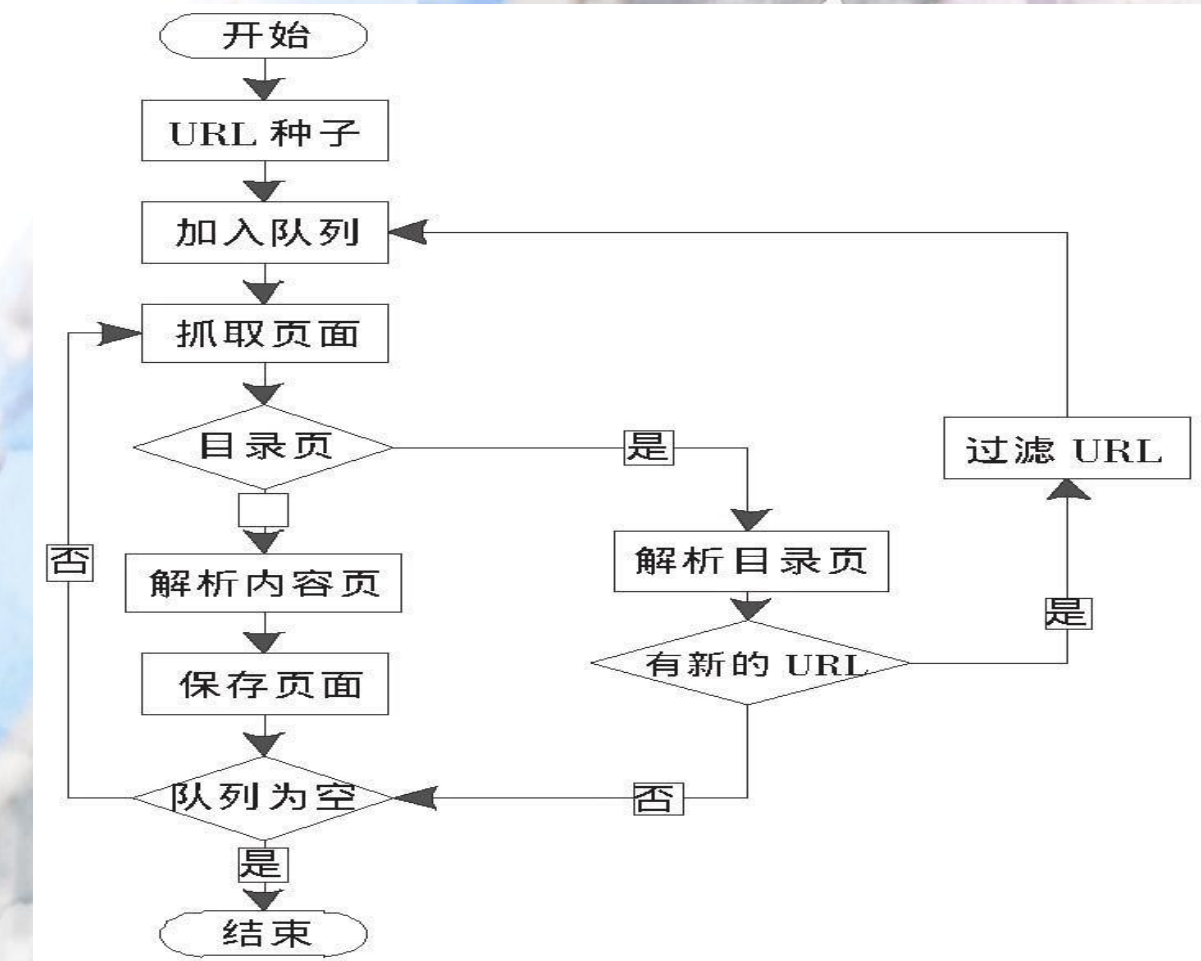
2.3.3 增量式爬虫的思路

在发送请求之前判断这个 URL 是否曾爬取过（适合不断有新页面的网站）

在解析内容后判断这部分内容是否曾爬取过（适合页面内容定时更新的网站）

写入存储介质时判断内容是否已存在于介质中（最大限度达到去重的目的）

2.3.4 增量式爬虫过程



增量式爬虫过程图

2.4.1 Deep web 爬虫

Deep web

Web 页面按存在方式可以分为**表层网页**（Surface Web）和**深层网页**（Deep Web，也称 Invisible Web Pages 或 Hidden Web）。表层网页是指传统搜索引擎可以索引的页面，以超链接可以到达的静态网页为主构成的 Web 页面。Deep Web 是那些大部分内容不能通过静态链接获取的、隐藏在搜索表单后的，只有用户提交一些关键词才能获得的 Web 页面。

2.4.2 Deep web 爬虫

Deep web 爬虫关键步骤

- 自动查找深层Web入口点。
- Form 建模。
- 查询选择。
- 表单提交。
- 学习爬行路径。

2.4.3 Deep web 爬虫六个基本模块

- 基本功能模块： 爬行控制器、解析器、表单分析器、表单处理器、响应分析器、LVS 控制器。
- 内部数据结构：URL 列表、LVS 表。

2.4.4 Deep web 爬虫表单填写

基于领域知识的表单填写

此方法一般会维持一个本体库，通过语义分析来选取合适的关键词填写表单。

VS

基于网页结构分析的表单填写

此方法一般无领域知识或仅有有限的领域知识，将网页表单表示成 DOM 树，从中提取表单各字段值。

PART 03

开发库及框架

主讲人：赵宇航

3.1 常用库

页面爬取

实现Http请求操作

- **urllib**
- **requests**
- selenium
- aiohttp

页面分析

从网页中提取信息

- **lxml**
- **beautifulsoup**
- pyquery

存储库

与数据库交互

- pymysql
- pymongo

3.1.1 urllib

request

HTTP请求模块，可以用来模拟发送请求。只需要给库方法传入URL以及额外的参数，就可以模拟在浏览器里输入网址然后回车一样。

error

异常处理模块，如果出现请求错误，我们可以捕获这些异常，然后进行重试或其他操作以保证程序不会意外终止。

parse

工具模块，提供了许多URL处理方法，比如拆分、解析、合并等。

robotparser

网站识别模块，主要是用来识别网站的robots.txt文件，然后判断哪些网站可以爬，哪些网站不能爬。

3.1.2 requests

requests是python 的第三方库，它是对 urllib的进一步封装， 因此在使用上显得更 加的便捷。

3.1.3 lxml

XPath，全称 XML Path Language，即 XML 路径语言，最初是用来搜寻 XML 文档的，但同样适用于
HTML 文档的搜索。所以在做爬虫时完全可以使用 XPath 做相应的信息抽取。

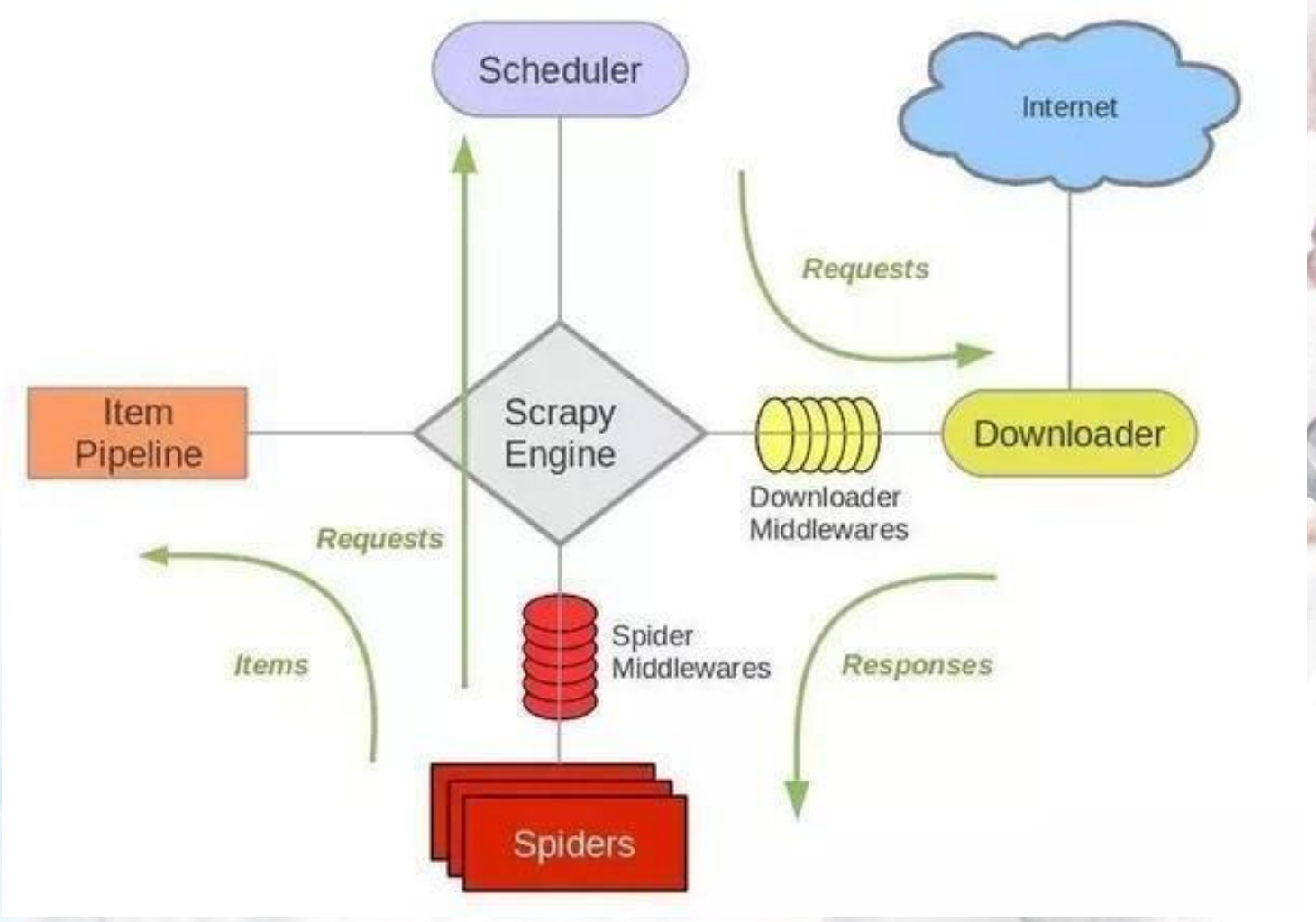
3.1.4 Beautiful Soup

Beautiful Soup将复杂HTML或 XML文档转换成一个复杂的树 形结构（文档树），树上每个 节点都是一个Python对象。所 有对象可以归纳为 4 种类型：**Tag，NavigableString，BeautifulSoup ,Comment** 。

3.2 Scrapy框架

- 是一个快速、高层次的屏幕抓取和web抓取框架，用于抓取web站点并从页面中提取结构化的数据。
- Scrapy吸引人的地方在于它是一个框架，任何人都可以根据需求方便的修改。它也提供了多种类型爬虫的基类，如BaseSpider、sitemap爬虫等。

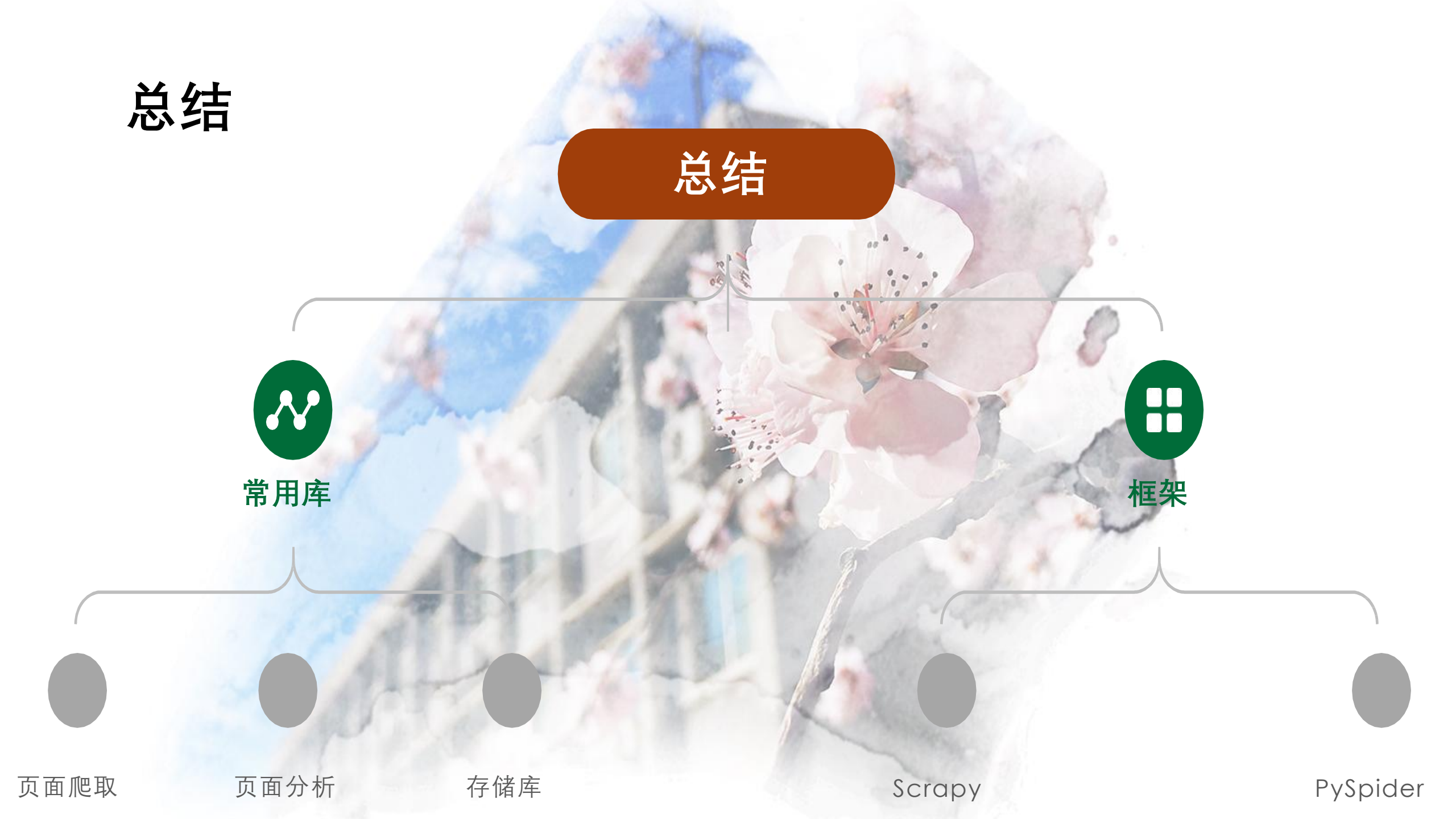
3.2.1 Scrapy框架



3.2.2 Pyspider框架

Pyspider是一个带有强大的Web UI、脚本编辑器、任务监控器、项目管理器以及结果处理器的框架。它 支持多种数据库后端、多种消息队列和Javascript 渲染页面采集。

总结



PART 04

网页分析算法

主讲人：郝思然 李若雪

4.1 抓取目标分类

- 基于目标网页特征

（1）预先给定的初始抓取种子样本；

（2）预先给定的网页分类目录和与分类目录对应的种子样本，如Yahoo!分类结构等

（3）通过用户行为确定的抓取目标样例

- 基于目标数据模式

- 基于领域概念

4.1.1 广度优先搜索

- 广度优先搜索策略是指在抓取过程中，在完成当前层次的搜索后，才进行下一层次的搜索。该算法的设计和实现相对简单。在目前为覆盖尽可能多的网页，一般使用广度优先搜索方法。

4.1.2 最佳优先搜索

- 最佳优先搜索策略按照一定的网页分析算法，预测候选URL与目标网页的相似度，或与主题的相关性，并选取评价最好的一个或几个URL进行抓取。它只访问经过网页分析算法预测为“有用”的网页。

4.1.3 深度优先搜索

- 深度优先搜索策略从起始网页开始，选择一个URL进入，分析这个网页中的URL，选择一个再进入。如此一个链接一个链接地抓取下去，直到处理完一条路线之后再处理下一条路线。深度优先策略设计较为简单。

4.2 拓扑分析算法

- 基于网页之间的链接，通过已知的网页或数据，来对与其有直接或间接链接关系的对象（可以是网页或网站等）作出评价的算法。又分为网页粒度、网站粒度和网页块粒度这三种。

4.2.1 网页粒度的分析算法

- PageRank和HITS算法是最常见的链接分析算法，两者都是通过对网页间链接度的递归和规范化计算，得到每个网页的重要度评价。

4.2.2 网站粒度的分析算法

- 网站粒度的资源发现和管理策略也比网页粒度的更简单有效。网站粒度的爬虫抓取的关键之处在于站点的划分和站点等级（SiteRank）的计算。SiteRank的计算方法与PageRank类似，但是需要对网站之间的链接作一定程度抽象，并在一定的模型下计算链接的权重。

4.2.3 网页块粒度的分析算法

- 在一个页面中，往往含有多个指向其他页面的链接，这些链接中只有一部分是指向主题相关网页的，或根据网页的链接锚文本表明其具有较高重要性。

4.3 网页内容分析算法

- 基于网页内容的分析算法指的是利用网页内容（文本、数据等资源）特征进行的网页评价。网页的内容从原来的以超文本为主，发展到后来动态页面数据为主，后者的数据量约为直接可见页面数据的400~500倍。

4.3.1 纯文本分类与聚类算法

- 很大程度上借用了文本检索的技术。文本分析算法可以快速有效的对网页进行分类和聚类，但是由于忽略了网页间和网页内部的结构信息，很少单独使用。

4.3.2 超文本分类和聚类算法

- 根据网页链接网页的相关类型对网页进行分类，依靠相关联的网页推测该网页的类型。

PART 05

爬虫技术前沿

主讲人：赵宇航

5.1 谷歌翻译爬虫

基于Gecko浏览器内核的谷歌翻译爬虫

此方法模拟浏览器加载网页，完成用户输入，触发执行脚本，最终获得目标数据。应用上述方法，设计并实现了面向谷歌翻译的专用爬虫，能够采用“多次少取”的方式解决大规模语料的自动翻译问题。

Google 翻译

5.2 网络舆情监测

网络舆情监测

采用网络爬虫技术从百度指数获取某一“热门事件”的数据，并对这些数据进行预处理，进而建立网络舆情的 Logistic 微分方程模型。结合已有数据，采用智能算法确定微分方程解中的 3 个关键参数；最后应用于网络舆情预测。

5.3 暗网爬虫

常规爬虫无法索引的这些数据内容，由暗网爬虫爬取

所谓暗网，是指目前搜索引擎爬虫按照常规方式很难抓取到的互联 网页面。

5.4.1　反爬虫

为什么要反爬?

1、爬虫占总PV比例较高，浪费服务器资源

2、公司可免费查询的资源被批量抓走，丧失竞争力

5.4.2　反爬虫

我们在反什么样的爬虫？

1、黄牛恶意竞争

2、没人去停止的失控爬虫

3、同行竞争对手

4、网站点击欺诈

5.4.3 反爬虫

怎样反爬?

1.User-Agent控制请求

User-Agent中可以携带一串用户设备信息的字符串，包括浏览器、操作系统等信息。我们可以通过在服务器设置user-agent白名单，只有符合条件的user-agent才能访问服务器。它的缺点就是很容易被爬虫程序伪造头部信息，进而被破解掉。

2.session访问限制

session是用户请求服务器的凭证，网络爬虫往往通过携带正常用户session信息的方式，模拟正常用户请求服务器。因此，我们同样可以根据短时间内的访问量的大小判断是否为爬虫程序，将疑似爬虫程序的用户的session加入黑名单。

3.蜘蛛陷阱

蜘蛛陷阱通过引导爬虫程序陷入无限循环的陷阱，消耗爬虫程序的资源，导致其崩溃而无法继续爬取数据。此方法的缺点就是会新增许多浪费资源的文件和目录，而且对正常网站排名有影响，会造成搜索引擎的爬虫程序也无法爬取信息，进而导致在搜索引擎的网站排名靠后。

5.4.3 反爬虫

怎样反爬?

4.IP限制

我们可以在服务器设置一个阈值，将短时间内访问量大的IP地址加入黑名单，禁止其访问，以达到反爬虫的目的。

5.验证码

在用户登录或访问某些重要信息时可以使用验证码来阻挡爬虫程序。验证码分为图片验证码、短信验证码、数值计算验证码、滑动验证码、图案标记验证码等。

6.数据加密

前端请求服务器前，将请求参数、user-agent、cookie等参数进行加密，用加密后的数据请求服务器，这样的话网络爬虫程序不知道我们的加密规则，就无法进行模拟请求我们的服务器。但是，这种方式的加密算法是写在js代码里的，很容易被用户找到并且破解。

7.对 Cookie 进行限制

用户向访问网站发送 Request 时,数 据中会包含特定的 Cookie 数据, 网站将会通过对Cookie值的验证来判断该用户操作是爬虫脚本还是真实的用户,当用户第二次及第 三次打开网页访问无 Cookie数据时,则说明该操作为爬虫脚本。

5.4.4 反爬虫

爬虫技术发展 >> 反爬虫保护隐私 >> 针对反爬的爬虫技术

- 爬虫这种自动化技术的确为人类互联网带来了许多好处,但 是同样的,滥用爬虫技术也会有很多坏处,水能载舟亦能覆舟,因此,我们要在法律规制和自身行为规范下学会正确使用这种技术, 才能最大化的发挥其优势,避免造成对互联网环境的危害。

PART 06

demo展示

主讲人：赵宇航 闫敏行 李梓石

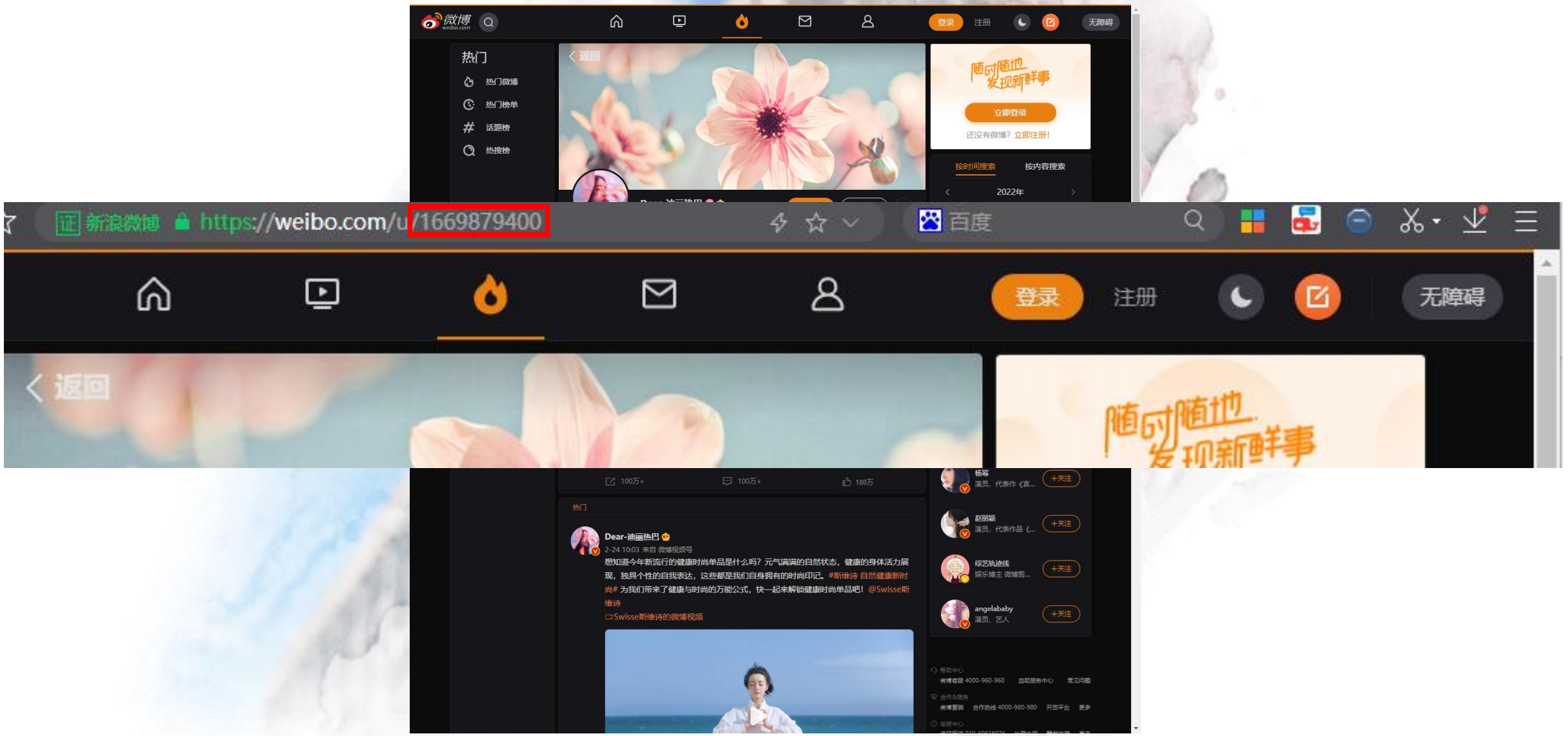
6. 微博爬取

输出

用户信息

- 用户id：微博用户id，如"1669879400"
- 用户昵称：微博用户昵称，如"Dear-迪丽热巴"
- 性别：微博用户性别
- 生日：用户出生日期
- 所在地：用户所在地
- 教育经历：用户上学时学校的名字
- 公司：用户所属公司名字
- 阳光信用：用户的阳光信用
- 微博注册时间：用户微博注册日期
- 微博数：用户的全部微博数（转发微博+原创微博）
- 粉丝数：用户的粉丝数
- 关注数：用户关注的微博数量
- 简介：用户简介
- 主页地址：微博移动版主页url，如https://m.weibo.cn/u/1669879400?uid=1669879400&luicode=10000011&lfid=1005051669879400
- 头像url：用户头像url
- 高清头像url：用户高清头像url
- 微博等级：用户微博等级
- 会员等级：微博会员用户等级，普通用户该等级为0
- 是否认证：用户是否认证，为布尔类型
- 认证类型：用户认证类型，如个人认证、企业认证、政府认证等
- 认证信息：为认证用户特有，用户信息栏显示的认证信息

6. 微博爬取


微博
登录
注册
无障碍
热门
热门微博
热门榜单
话题榜
热搜榜
返回
随时随地发现新鲜事
立即登录
还没有微博？立即注册！
按时间搜索
按内容搜索
2022年
新浪微博
https://weibo.com/u/1669879400
百度
登录
注册
无障碍
返回
随时随地发现新鲜事
100万+
100万+
180万
热门
Dear-迪丽热巴
2-24 10:03 来自 微博视频号
想知道今年新流行的健康时尚单品是什么吗？元气满满的自然状态，健康的身体活力展现，独具个性的自我表达，这些都是我们自身拥有的时尚印记。#斯维诗 自然健康新时尚# 为我们带来了健康与时尚的万能公式，快一起来解锁健康时尚单品吧！@Swisse斯维诗
Swisse斯维诗的微博视频
杨幂
演员，代表作《宫...
赵丽颖
演员，代表作品《...
综艺轨迹线
娱乐博主 微博剪...
angelababy
演员、艺人
+关注

6. 微博爬取

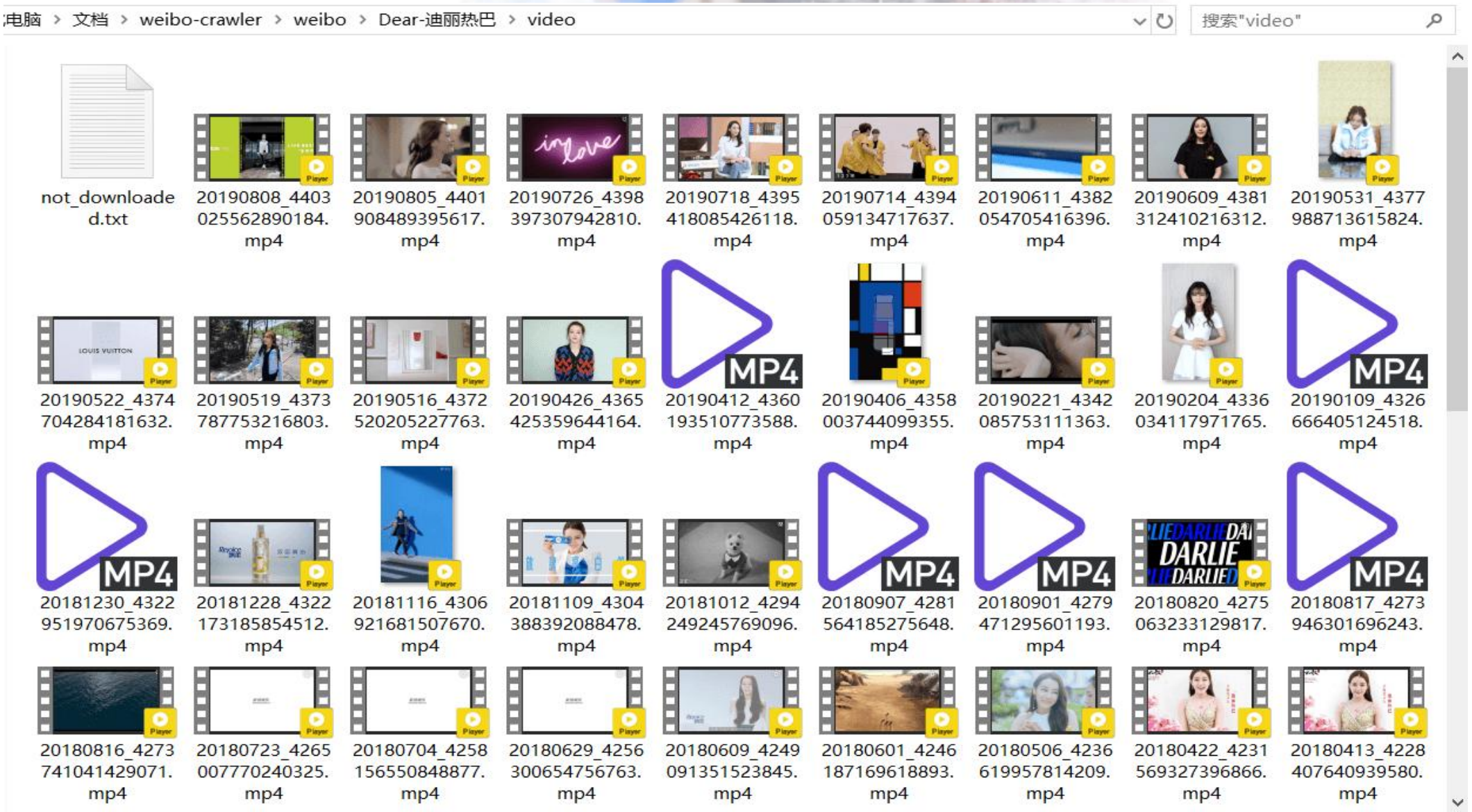

```
{
    "user_id_list": ["1887344341"],
    "filter": 1,
    "remove_html_tag": 1,
    "since_date": "2022-02-28",
    "start_page": 1,
    "write_mode": ["csv"],
    "original_pic_download": 1,
    "retweet_pic_download": 0,
    "original_video_download": 1,
    "retweet_video_download": 0,
    "download_comment":1,
```

6. 微博爬取

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	正文	原始图片url	视频url	位置	日期	工具	点赞数	评论数	转发数	话题	@用户	是否原创	源用户id	源用户昵称	源微博id	源微博正文
2	4403025562890180	各位潮搭特长生这个夏天可别让你的衣橱闲着来@a		http://gslb.n		2019/8/8		392131	405981	1000000	生来好动	adidasneo	TRUE				
3	4402632736773690	这个节日你们过就好了 不用管我	https://wx4.s			2019/8/7		1195699	855179	1000000			TRUE				
4	4402308588283590	#国模之美#祝Vogue生日快乐！♡				2019/8/6		274659	37659	40791	国模之美		FALSE	1711599694	angelica张宇	4402254083394030	#国模之美#
5	4401908489395610	#忘不掉的柔顺发香#如果我藏起来了？你们能找到我		http://gslb.n		2019/8/5		389636	299710	1000000	忘不掉的	飘柔Rejoic	TRUE				
6	4401638393359860	#五星红旗有14亿护旗手#我是护旗手！♡CN				2019/8/4		421590	167791	212229	五星红旗有		FALSE	2656274875	央视新闻	4401559745706760	#五星红旗有
7	4399837476445810	让我们一起#致敬中国军人#!				2019/7/30		376104	222407	266558	致敬中国军		FALSE	2656274875	央视新闻	4399830736230380	【#致敬中国
8	4399823278807940	支持小源新歌～源				2019/7/30		503255	24765	27180			FALSE	2812335943	TFBOYS-王源	4399692954085210	七年 我还是
9	4398397307942810	推开反转门，穿上叛逆香气，可甜可酷，你爱的样子		http://f.us.si		2019/7/26		423375	359801	1000000	YSL反转巴		TRUE				
10	4398075302999500	联合国可持续发展目标（SDG）目标2是消除饥饿，				2019/7/25		232620	152912	196433	联合国在华	联合国粮农	FALSE	5262686616	联合国粮农组	4397874005444600	@联合国粮农
11	4397614382669660	保护生态环境，促进农业发展，让我们和@联合国粮				2019/7/24		268106	159326	389701	联合国在华	联合国粮农	FALSE	6184877673	NicholasRos	4397556371211120	#联合国在华
12	4396056907479200	很高兴成为百雀羚的彩妆代言人！东方草本能量亦妆				2019/7/20		393811	315106	1000000		百雀羚	FALSE	2037441407	百雀羚	4396053971301850	官宣！她是
13	4395418085426110	#和爱豆一起读书# 强国一代，#榜样阅读# 我在@中		http://f.us.si		2019/7/18		330342	196139	924267	和爱豆一起	中国青年报	TRUE				
14	4394059134717630	#瓶盖挑战# 有些瓶盖，一旦错过就还在#极限挑战#		http://f.us.si		2019/7/14		467585	225030	1000000	瓶盖挑战,		TRUE				
15	4393551061952470	@路易威登 经典中不断创新的LV记录成长的LV爱了	https://wx2.s			2019/7/13		555749	384218	1000000	路易威登	路易威登	TRUE				
16	4392834054029370	很高兴成为MISS SIXTY亚太区代言人！@MissSixty 高	https://wx2.s			2019/7/11		494507	257404	1000000	MISSION S	MissSixty	TRUE				
17	4389261824190130	不忘初心，牢记使命！祝福党，祝福祖国！CNCN				2019/7/1		360568	151124	157328			FALSE	2803301701	人民日报	4389138709375150	【今天，中
18	4387591560200900	哥哥起飞了哦～大家晚安🎵	https://wx3.s			2019/6/27		929783	573930	1000000			TRUE				
19	4384485850075610	希望伤亡不要再增加了！愿平安🙏🙏🙏				2019/6/18		330030	40636	49044			FALSE	2803301701	人民日报	4384465759882170	【#宜宾地震
20	4384122253963000	#617腾讯影业年度发布会# 嫦娥姐姐终于要来见阿丝	https://wx1.s			2019/6/17		357134	101609	414871	617腾讯影		TRUE				
21	4383183661430860	搬完这块砖，就可以吃好吃的了	https://wx3.s			2019/6/14		620242	442456	1000000			TRUE				
22	4383180406740600	大家一起电影院约起来啊～～～				2019/6/14		200660	10880	11475			FALSE	2007341923	代斯daisy	4383031864995270	写在《秦明·
23	4382054705416390	#不潮不出街#悠闲下午时光～Dear-迪丽热巴的秒拍		http://f.us.si		2019/6/11		391705	213204	339442	不潮不出街		TRUE				
24	4381312410216310	带话题词#我的身边有非遗#发布微博，分享你知道的		http://f.us.si		2019/6/9		300056	169151	883820	我的身边有		TRUE				
25	4381207217392720	100多天，说长不长，说短不短的，特别开心可以陪	https://wx4.s			2019/6/9		625187	346221	1000000			TRUE				
26	4380517519047350	#比耶季##高考加油#大家高考加油鸭！！！	https://wx3.s			2019/6/7		756704	365680	1000000	比耶季,高		TRUE				
27	4379825714272340	#蓝天保卫战我是行动者# 我承诺使用环保购物袋代	https://wx4.s			2019/6/5		397493	25608	337653	蓝天保卫战		TRUE				
28	4379188905591760	这一年，做自己。不要停下努力的步伐。	https://wx2.s			2019/6/3		1331611	1000000	1000000			TRUE				
29	4378893399854470	今天是我的生日，来祝福我吧!				2019/6/3	生日云	723284	109236	690128			TRUE				
30	4378514142544190	27岁❤第一次有这么多人陪我一起过生日🎂我真的	https://wx1.s			2019/6/1		1145119	314980	1000000			TRUE				
31	4378432861003770	#迪丽热巴生日会# 热·爱终于实现了承诺，紧张～				2019/6/1	一直推	256045	58967	14784	迪丽热巴生		TRUE				
32	4378408970314840	看到你们在乖乖排队偷拍了一下一会儿见阿丝儿们	https://wx2.s			2019/6/1		561340	237139	1000000			TRUE				
33	4377988713615820	#极限挑战#差不多就行了 咋害没完没了呢 Dear-迪		http://f.us.si		2019/5/31		375330	71786	842474	极限挑战		TRUE				

1669879400

6. 微博爬取


电脑 › 文档 › weibo-crawler › weibo › Dear-迪丽热巴 › video
搜索"video"
not_downloaded.txt
20190808_4403025562890184.mp4
20190805_4401908489395617.mp4
20190726_4398397307942810.mp4
20190718_4395418085426118.mp4
20190714_4394059134717637.mp4
20190611_4382054705416396.mp4
20190609_4381312410216312.mp4
20190531_4377988713615824.mp4
20190522_4374704284181632.mp4
20190519_4373787753216803.mp4
20190516_4372520205227763.mp4
20190426_4365425359644164.mp4
20190412_4360193510773588.mp4
20190406_4358003744099355.mp4
20190221_4342085753111363.mp4
20190204_4336034117971765.mp4
20190109_4326666405124518.mp4
20181230_4322951970675369.mp4
20181228_4322173185854512.mp4
20181116_4306921681507670.mp4
20181109_4304388392088478.mp4
20181012_4294249245769096.mp4
20180907_4281564185275648.mp4
20180901_4279471295601193.mp4
20180820_4275063233129817.mp4
20180817_4273946301696243.mp4
20180816_4273741041429071.mp4
20180723_4265007770240325.mp4
20180704_4258156550848877.mp4
20180629_4256300654756763.mp4
20180609_4249091351523845.mp4
20180601_4246187169618893.mp4
20180506_4236619957814209.mp4
20180422_4231569327396866.mp4
20180413_4228407640939580.mp4

感谢您的观看

北京理工大学