

媒体融合与传播国家重点实验室

媒体大数据资源系统使用说明

大数据中心

2020 年 1 月 12 日

一、简介

本系统基于 MongoDB 分布式存储系统和 Elastic Research 全文检索系统搭建，在此基础上开发了一个简单的用户界面，提供数据检索与下载服务。

数据资源情况：

（1）微博数据

6.85 亿（2012 年至今）

（2）微信数据

3.46 亿（2014 年至今）

（3）网媒数据

0.46 亿（2012 年至今）

（4）实时热点（3 小时更新）

微博 top10

微信 top10

网媒 top10

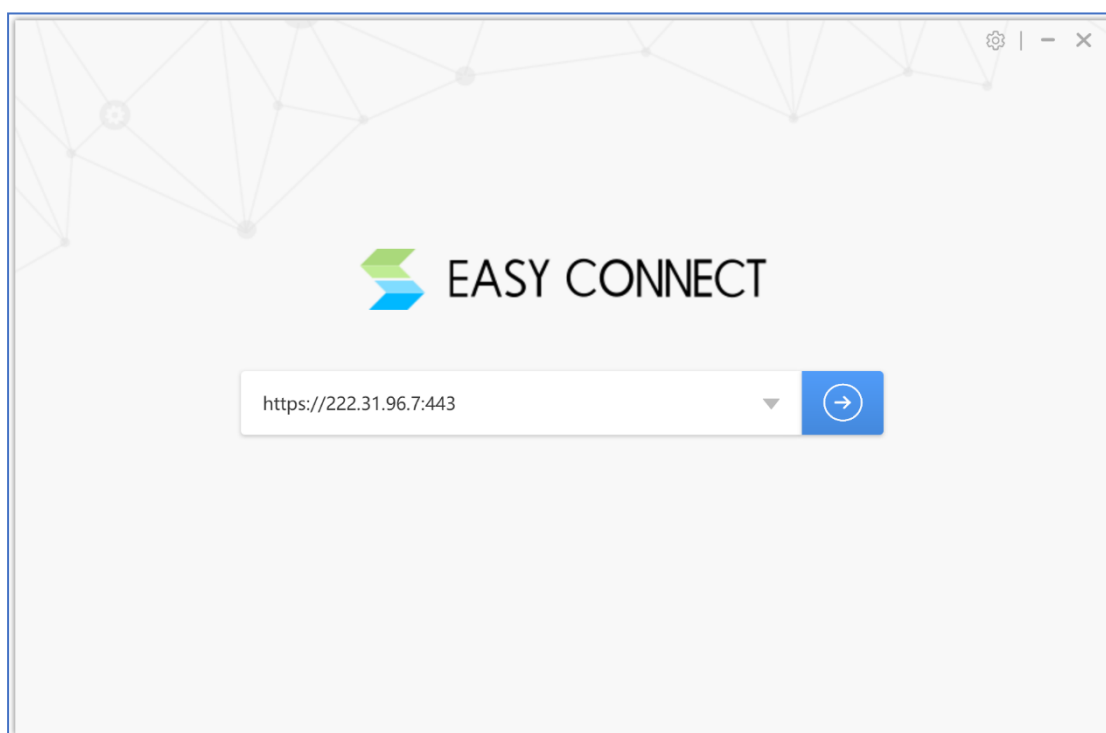
目前提供数据检索非常简陋，如果需要比较准确的数据进行分析，仍需要人工操作。

Jupyter 数据服务是一些直接操作数据库的简单实例，其中大数据分析的案例需要的操作时间会非常长，如果需要进行分析，仍需要根据具体情况开发专门的系统才实用。

二、访问与登录

本系统支持通过浏览器访问与登录。目前，支持基于校园网访问和校外 VPN 访问，支持校园账号统一认证。

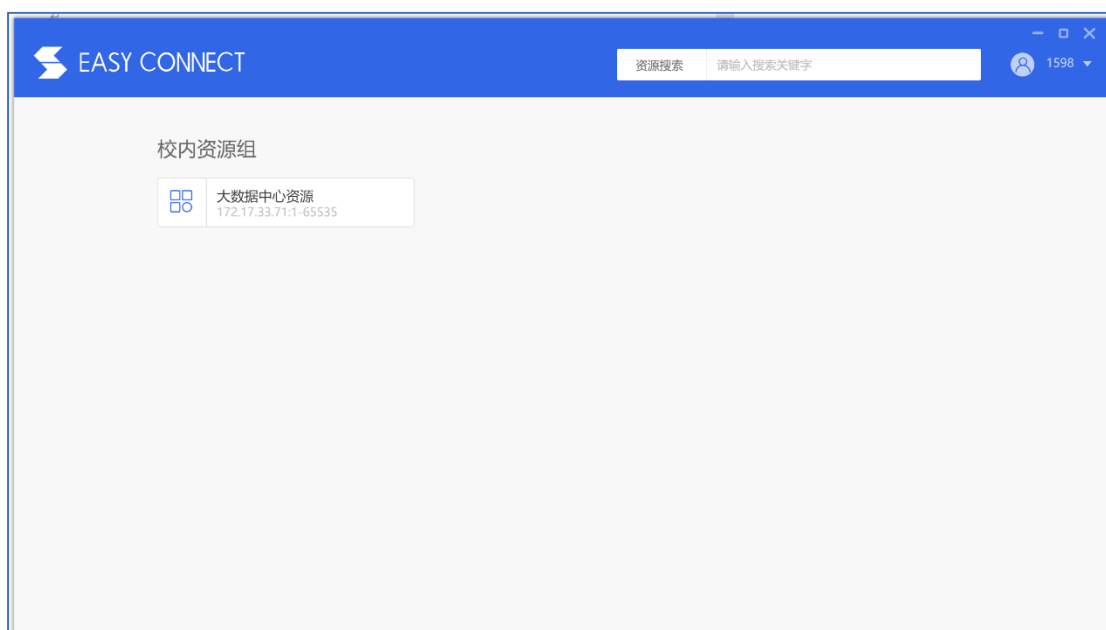
VPN 访问需要安装 Easy Connect 客户端，安装后打开客户端，输入服务链接：<https://222.31.96.7:443>，见下图：



输入服务链接后将出现以下界面：

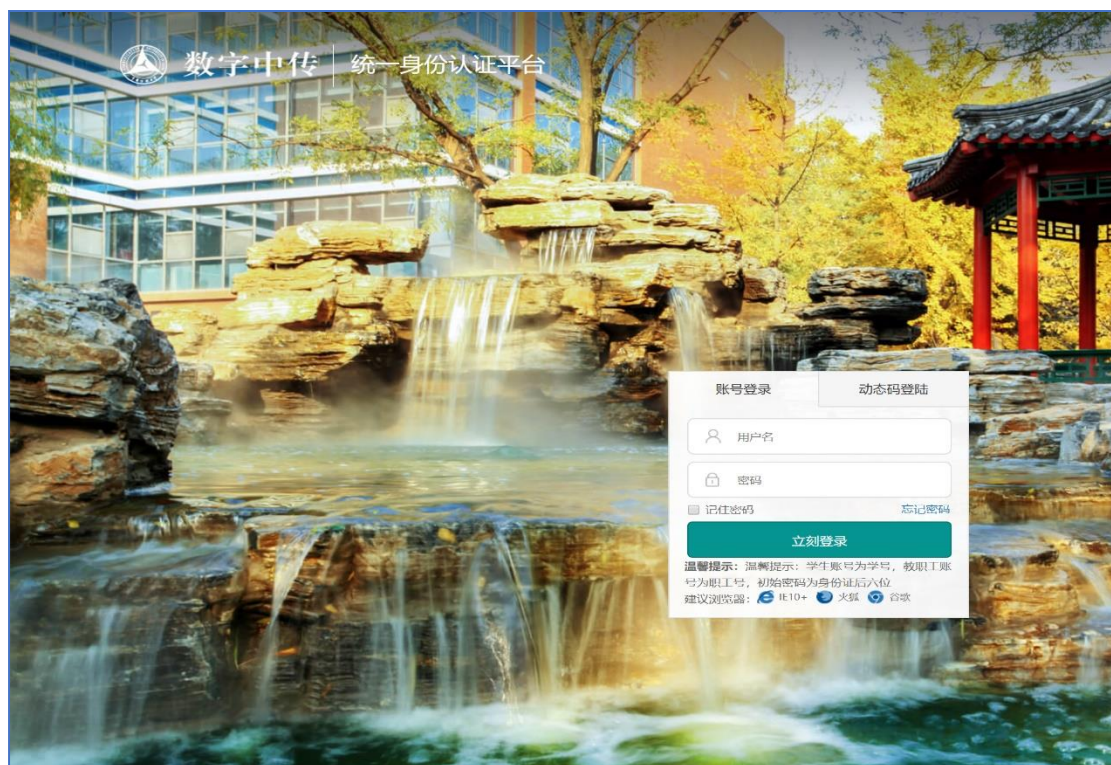


输入校园账号的用户名和密码登录系统，出现以下界面：



然后，在浏览器地址栏中输入系统地址 <http://mdes.cuc.edu.cn:8083> 即可进入平台。

如果在校园网登录，则直接在浏览器地址栏中输入系统地址 <http://mdes.cuc.edu.cn:8083>，回车后出现校园网统一登录界面；



输入校园账号的用户名和密码登录系统。

系统主页如下图：

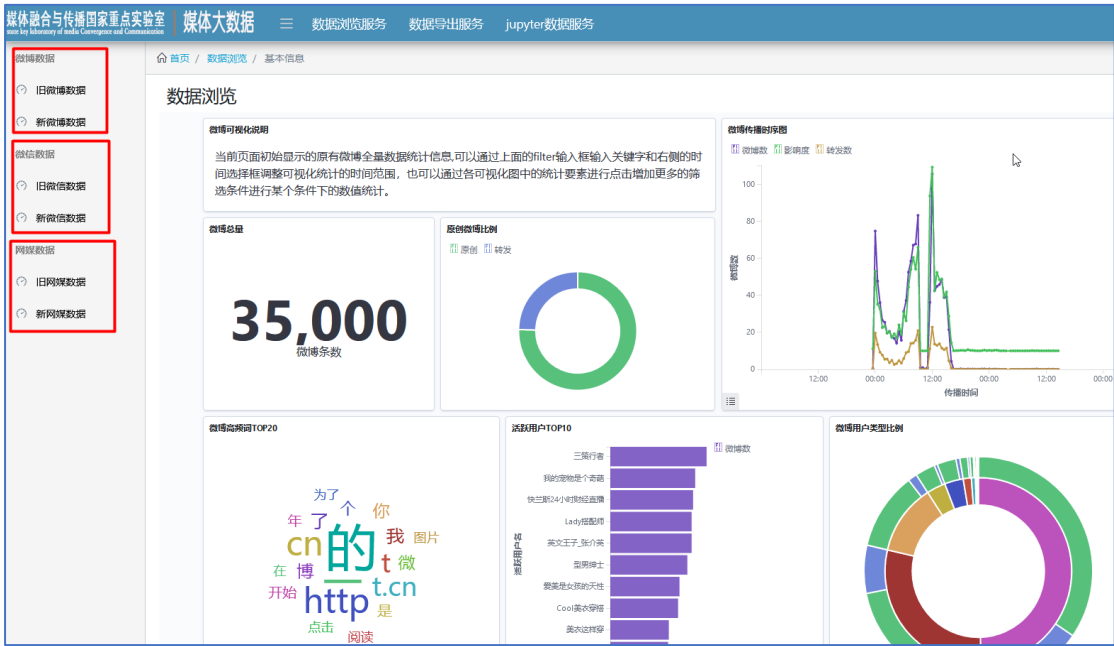
媒体融合与传播国家重点实验室 媒体大数据					
数据浏览服务		数据导出服务	jupyter数据服务	数据操作服务	设置 退出
基本信息					
数据浏览 / 数据浏览 / 基本信息					
数据集名称	数据库名称	数据库	数据表数	最后同步时间	说明
数据浏览	旧微博数据(2019年2月之前)	NetWork	20	2020-01-03 22:50:01	
数据浏览	新微博数据	NewWeibo	11	2020-01-03 22:50:05	
微信数据	旧微信数据(2019年2月之前)	wechat	14	2020-01-05 11:03:42	
微信数据	新微信数据	wechat	14	2020-01-05 11:03:49	
网络媒体数据	旧网络媒体数据(2019年2月之前)	mediaspider	98	2020-01-05 11:04:03	
网络媒体数据	新网络媒体数据	NewMedia	4	2020-01-13 09:15:50	
CUC Data Center © 2019 CUC.EDU.CN. Powered by CUC DataLabs.					

三、数据浏览功能

在主页中直接点击“数据浏览服务”：

媒体融合与传播国家重点实验室 媒体大数据					
数据浏览服务		数据导出服务	jupyter数据服务	数据操作服务	设置 退出
基本信息					
数据浏览 / 数据浏览 / 基本信息					
数据集名称	数据库名称	数据库	数据表数	最后同步时间	说明
数据浏览	旧微博数据(2019年2月之前)	NetWork	20	2020-01-03 22:50:01	
数据浏览	新微博数据	NewWeibo	11	2020-01-03 22:50:05	
微信数据	旧微信数据(2019年2月之前)	wechat	14	2020-01-05 11:03:42	
微信数据	新微信数据	wechat	14	2020-01-05 11:03:49	
网络媒体数据	旧网络媒体数据(2019年2月之前)	mediaspider	98	2020-01-05 11:04:03	
网络媒体数据	新网络媒体数据	NewMedia	1	2020-01-05 11:04:05	

将会出现如下界面：



用户可根据需要点击不同数据进行浏览查看，其中三大部分的旧数据是指：2019 年 2 月之前的数据，新数据是指 2019 年 2 月之后的数据。现在以新微博数据为例子，进行展示。新旧数据主要区别是数据格式变化。

点击新微博数据将会呈现如下所示：

The screenshot shows the '新微博数据' (New Weibo Data) table. The table has columns for '数据表' (Data Table), '说明' (Description), '字段数' (Number of Fields), '记录数' (Number of Records), '最后更新时间' (Last Update Time), and '操作状态' (Operation Status). The data is organized into a table with 10 rows of data.

数据表	说明	字段数	记录数	最后更新时间	操作状态
incomplete_text_201907		32	5484529	2020-01-03 22:50:04	standby
incomplete_text_201903		31	5573626	2020-01-03 22:50:04	standby
incomplete_text_201904		31	5600606	2020-01-03 22:50:04	standby
incomplete_text_201910		53	6751168	2020-01-03 22:50:04	standby
incomplete_text_201902		31	4674534	2020-01-03 22:50:04	standby
incomplete_text_201911		35	8218686	2020-01-03 22:50:05	standby
incomplete_text_201906		41	5105158	2020-01-03 22:50:05	standby
incomplete_text_201908		58	152858	2020-01-03 22:50:05	standby
incomplete_text_201909		34	5600735	2020-01-03 22:50:05	standby
incomplete_text_201905		31	5706918	2020-01-03 22:50:05	standby
incomplete_text_201912		54	8534689	2020-01-03 22:50:05	standby

用户根据需要选择任一月份放大镜进行数据的浏览查看：



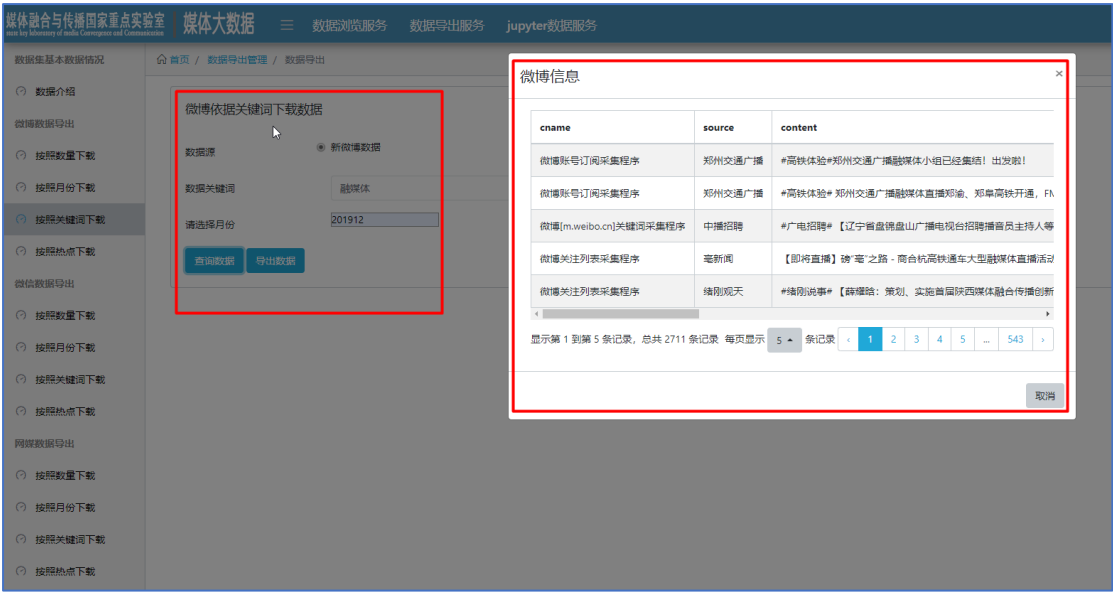
四、数据导出功能

点击主页“数据导出服务”：

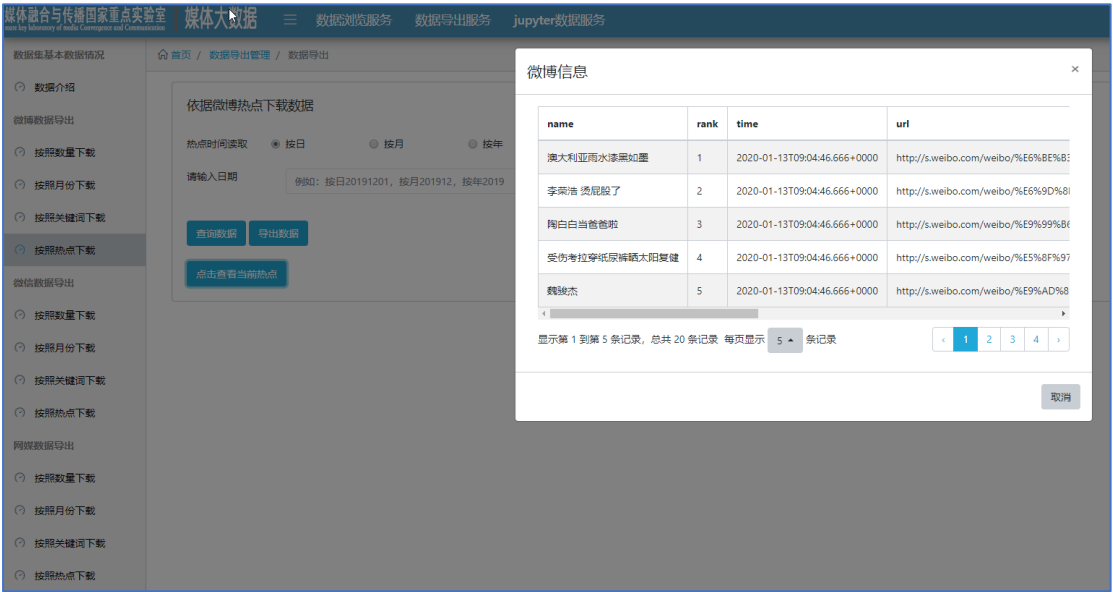


会出现如下图所示：

按照关键词进行检索后下载

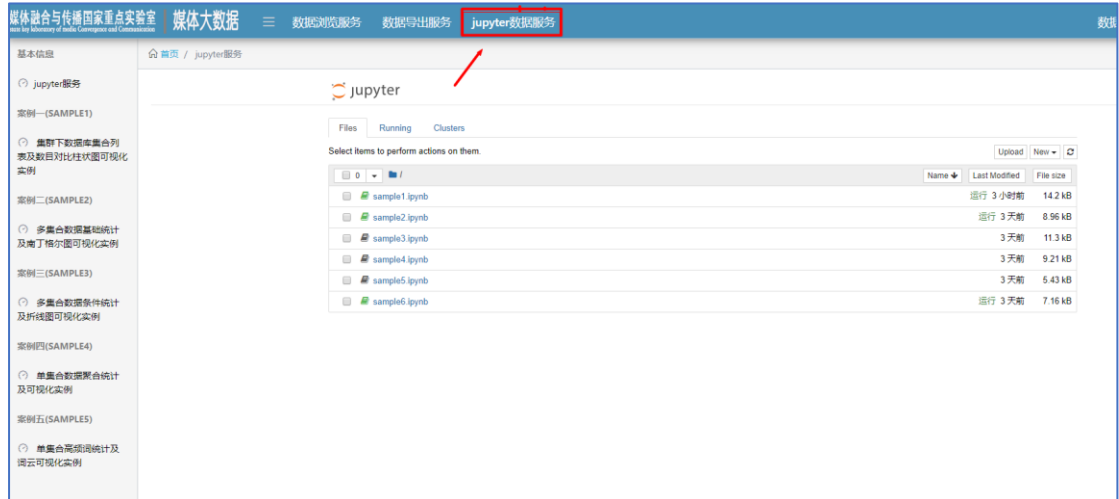


依据热点进行下载:

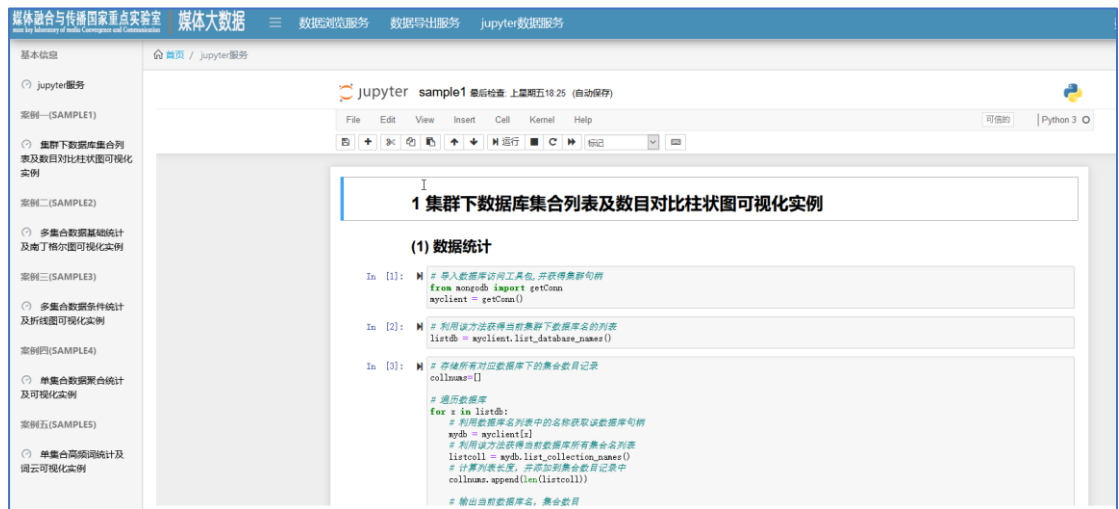


五、Jupyter 数据服务

点击“Jupyter 数据服务”:



随意点击一个案例：



可呈现可视化图形：

