舆情事件的Granger因果检验软件V1.0

说明书

摘要

企业通常会在舆情论坛注册账号，推广产品或服务，但是舆情氛围可能被恶意用户以大量发表垃圾评论的方式扭转，影响了客户群体对真实商品或服务质量的认知。本软件通过BERT-3模型、聚类等方法归纳文本代表的舆情事件，通过Granger因果检验分析用户受到谁的影响、对谁产生了影响。本软件帮助企业更快地定位舆情事件传播的链路，保障舆情论坛能够真实、客观地反映企业产品或服务的质量。

# 软件描述

博客社区、电子商城、社交媒体等每个用户都能发表言论、转发、评论、点赞的网络空间被称为“舆情论坛”。在现代社会，企业通常会在舆情论坛注册账号，推广产品或服务；当企业宣传新活动时，会在舆情论坛引发一段时间的讨论。但是，一些恶意用户（不满意服务的客户或竞争对手）可能雇佣大量的垃圾账户或机器人，通过大量发表负面评论，引发舆情转向。这种行为影响了客户群体对真实商品或服务质量的认知，所以企业需要找到诱导舆情的用户，采取维护自身权益的措施。

本软件通过分析舆情论坛上的文本，将帖子按表达内容聚合成舆情事件；再根据用户参与舆情事件的时间序列，通过Granger因果检验，分析用户受到谁的影响、对谁产生了影响。在文本处理方面，软件通过BERT-3自然语言处理模型提取文本的特征，接着通过PCA模型（主成分分析）和DBSCAN模型将评论聚类为数个舆情事件。在影响力评价方面，软件根据每种互动计数的信息熵加权计算帖子的影响力。然后，软件将帖子代表的舆情事件和影响力，归纳为用户的舆情表达的时间序列。最后，软件通过Granger因果检验，在选定的舆情事件中，分析一个用户与其他用户在舆情表达方面的因果关系，从而帮助企业更快确定舆情发展的链路。

本软件能够帮助企业了解舆情论坛中事件的传播规律，保障舆情论坛能够真实、客观地反映企业产品或服务的质量。值得注意的是，Granger因果检验是统计学上判断时间序列是否互相影响的方法，不能成为认定社会学意义上因果关系的证据，因此本软件计算的结论仍需人工分析后才能得到应用。

# 软件功能

软件的分析流程如图 1所示。其中绿色部分代表用户输入的原始数据，蓝色部分（平行四边形）代表计算过程，黄色部分（矩形）代表中间过程的数据。

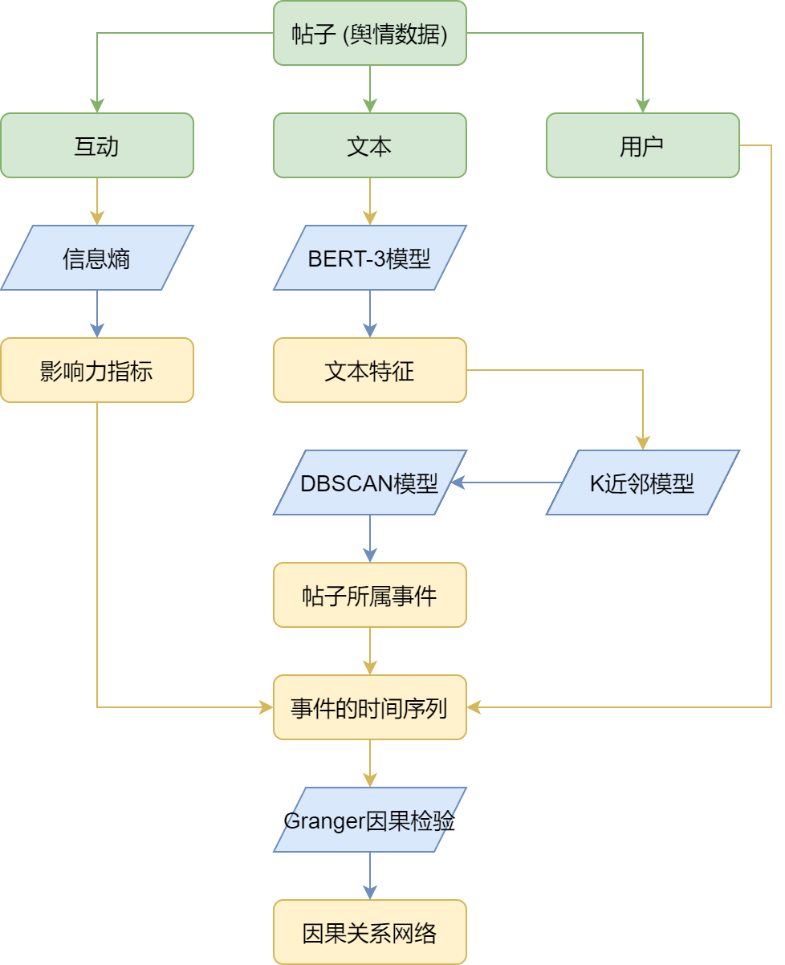


图 1 软件的分析流程

帖子可以通过舆情论坛提供的工具导出，通过网络爬虫获得，也可以通过手工方式收集。专门地，新浪微博的帖子推荐使用<https://github.com/dataabc/weibo-search>工具采集，但不是必须的。帖子的结构是一个Microsoft Excel格式的二维表格，其中必须包含1个代表用户身份（唯一序号）的字段，1个代表帖子正文的字段，不定数量的代表互动计数（转发、点赞、评论等）的字段。

在用户通过表单标注这些字段后，软件内部分为2个主线，分别处理后合成“事件的时间序列”数据。

## 分析互动

所有帖子的互动指标是一个的矩阵，其中是帖子的数量，是互动指标的数量。矩阵的每一列代表单个互动指标。软件先对矩阵作对数变换，接着**按列**作归一化变换，然后将矩阵**按行**以信息熵加权。每个指标的信息熵如公式所示。

帖子的影响力指标是一个的向量，代表帖子引起其他用户的参与的能力。

## 分析文本

所有帖子的文本是一个长度为的字符串列表。BERT-3模型是一个已经训练好的神经网络模型（pooled模式），接受字符串列表作为输入，输出的矩阵，表示文本特征。通过BERT-3模型，字符串列表被转化成矩阵，其中每行代表一个字符串的特征。

软件需要将舆情文本聚类成事件，所以需要降维。在主成分分析的转换后，特征向量的每个维度相互独立，且按照解释原数据的方差从大到小排列。软件使用主成分分析（PCA模型）将特征向量投影到正交空间上，绘制“解释方差比例折线图”，横轴是维度的排名（记作x），纵轴是前x个维度解释原数据的方差，占所有维度解释原数据的方差的比例。用户通过观察“解释方差比例折线图”，在保留维度的数量和保留信息的比例之间作出权衡，填写保留维度的数量。

软件按所填数量降维，然后**按列**作标准化变换，作为聚类的样本空间。软件采用DBSCAN算法聚类；该算法需要确定minPts和Eps参数的数值，具体含义如下：

1. 算法随机选择一个样本点开始，如果它的半径为Eps的邻域中有不少于minPts个点，则将其和其邻域半径中的点标记为一类，否则它成为孤立点。
2. 如果这个点不是孤立点，则算法遍历其邻域中的点，重复步骤（1），直至从这个点出发，通过邻域半径可桥接的所有点，都被标记为一类。
3. 算法从未被标记的样本点中随机选择一个开始，重复步骤（1，2），此时将样本点标记为另一个类。直至所有的样本点都被标记为某一类或孤立点，聚类结束。

由于不具备数据科学背景的用户不能正确设定minPts和Eps，软件通过一些经验法则自动设定参数。minPts被设定为样本维度的2倍[[1]](#footnote-1)。软件计算每个点到其他样本点的Chebyshev距离，按从小到大排列，取第minPts个距离作为这个点的K近邻距离（K=minPts）。然后，软件将每个点按K近邻距离从小到大排列，绘制“K近邻距离折线图”。用户观察图像，选择曲线的下凸函数的拐点，对应的纵坐标即是Eps参数的取值。

进行低维空间的聚类时，曲线几乎始终是下凸的，所以通常应选择下凸函数的拐点，这是最合适的聚类方式；在高维空间中，曲线包含了多段下凸函数，每个拐点有不同的含义，所以可选择任意一个拐点，或拐点之间的任意一点。在文本特征聚类中，每个拐点代表了不同层次的相似，一般距离从小到大分别是语言层面的相似、语义层面的相似、话题层面的相似、语义风格的相似，用户可根据分析偏好选择。如果为分析舆情事件，选择话题、语义风格层面的相似即可。

待用户设定Eps参数后，软件执行DBSCAN算法，输出每个帖子的类别编号，代表这个帖子所属的舆情事件。所有的孤立点作为编号为-1的舆情事件，代表帖子内容不属于任何事件。

## Granger因果检验

设定时间序列的采样次数，软件将最早和最晚的发帖时间之间的时间段，等距离划分为个区间（第1个区间表示不在时间段内，为应对时间错误而保留，通常为空）。软件生成一个的全零张量，其中是发帖者（舆情社区的用户）的数量，是聚类得到的舆情事件的数量（包含事件-1）。

对于每个帖子，找到所属发帖者的序号、所属时间区间、所属话题，在张量的对应位置加上影响力指标，在发帖的时间区间之后的每个时间区间，加上，其中是自然对数的底数，是影响力衰减倍率（由用户设定），该时间区间距离发帖的时间区间，间隔了个时间区间。于是，软件获得了每个发帖者关于每个事件的影响力的时间序列，简称为事件的时间序列。

软件比较两两用户关于指定事件的时间序列，对于给定的“最大滞后阶”参数，计算Granger因果检验统计量F和显著性水平p. 一般来说，如果在第个滞后阶上，则说明一个用户对另一个存在影响（因果关系是有方向的），影响传播用时个时间区间。Granger因果检验使用线性的Pairwise 因果判据[[2]](#footnote-2),[[3]](#footnote-3)，有2个基本假设：

1. 因先于果；
2. 结果序列的未来值，有特定的信息取决于原因序列。

根据以上基本假设，推定对存在影响的条件如公式所示。其中表示在时间以前的所有信息；则表示仅在时间以前的信息。

严格来说，Granger因果需要进行差分，使得时间序列平稳。考虑舆情传播的现实意义，时间序列代表舆情的影响力变化，所以软件未对时间序列作平稳性检验和差分。

软件将统计量F和p输出给用户。

# 软件运行环境

## 操作系统

服务端：Linux发行版的任意操作系统；Windows Server 2016以上版本。服务端计算机必须拥有NVIDIA GeForce GTX 1080Ti以上计算能力的显卡，4Mb/s以上带宽的网络连接。

客户端：Google Chrome / Microsoft Edge 89以上版本浏览器支持的任意操作系统；iOS, iPadOS 13.0以上版本；macOS 10.15以上版本。

## 软件开发工具

JetBrains PyCharm 2021.3；MobaXTerm Personal 20.6

## 依赖软件

服务端：

Python 3.9语言及源码中列出的函数库；Bootstrap 5.0 CDN；BERT模型：

<https://tfhub.dev/tensorflow/bert_zh_preprocess/3>

<https://tfhub.dev/tensorflow/bert_zh_L-12_H-768_A-12/3>

客户端：

Google Chrome / Microsoft Edge 89以上版本浏览器

（新浪微博数据源）数据采集工具<https://github.com/dataabc/weibo-search>

## 开发语言

Python 3.9, HTML/CSS, JavaScript

# 软件特点

（1）本软件的引导流程易于使用

传统地，Granger因果检验由具有统计学背景的分析人员，使用专业统计软件实现。本软件在舆情分析这一特定场景，建立了一个完整的引导分析流程，使得非专业用户跟随着本软件的引导，就能正确完成分析。软件具有网页形式的图形界面，每一步作为一个页面，绘制了作为分析依据的统计图，提示用户如何设定最优的参数，如何解读分析结果。另外，软件也具有一定的自由度，用户可以根据偏好调整参数。

（2）本软件使用技术成熟的标准模型

本软件的作者具有数据科学背景，通过阅读学术文献，确定了截至软件开发完成日期的最流行的模型，分析结果符合规范并且能保证科学性。在提取文本特征时使用BERT-3模型，在因果检验中使用Granger因果检验标准，使用DBSCAN的最优的参数设定方法。

（3）本软件易于部署和维护

本软件以云计算为基础，不需要用户安装软件，所需部署和维护全部是服务端的工作。软件使用Django函数库实现自动化管理数据库和网站：

* 数据库：本软件源码中已包含自动生成SQLite3数据库的脚本，不需要人工创建数据库。
* 网站请求：本软件自身提供解析静态文件和前后端交互的程序，不需要特定的运行环境。
* 算法：本软件由模块化算法组装，容易添加或删除组件，方便用户修改功能或仅安装选定的功能。

软件内部也有清晰的模块化结构，主要使用自动化脚本，方便在尽量不中断服务的情况下更新升级。

# 软件设计与实现

## 软件设计

本软件使用Django函数库解析网络请求，对用户的每个操作，基本流程如下：

1. 软件收到用户的HTTPS请求，与路由表匹配，根据请求的路径分配到不同的视图函数执行。
2. 视图函数大致分为5个部分：验证输入的格式是否正确，从数据库获取运算所需数据，执行计算过程，将数据写入数据库，将人类可读的结果返回到上下文字典（本质是一个Python字典类型的变量）。
3. HTML渲染器读取静态网页模板，将上下文字典中的内容渲染到网页上，作为HTTPS的返回请求发送。

每个视图函数关联一种操作，操作数据库的一个或多个数据表，是实现软件功能的基本单位。

舆情事件的Granger因果检验软件运行后，用户通过浏览器访问<https://example.com/main>地址（根据部署地址确定），可以与软件交互：

1. 提供登录组件，供输入用户名和密码登录；提供“注册”页面，供登记用户名和密码，软件发送确认注册邮件，用户点击邮件上的链接注册。
2. 在首页上放置“采集数据”和“分析数据”按钮，“采集数据”的下拉列表包含了不同的数据采集工具的链接，页面上有采集数据的指导。“分析数据”按钮链接到任务列表，提供新建、删除、修改、查看功能。
3. 用户新建一个任务，进入引导流程的第一步；如果用户中途退出任务，再次进入时，点击任务名称，回到已完成的最后一个步骤。
4. 用户上传舆情数据集、填写分析任务的名称，软件将数据集解析为pandas.DataFrame格式，创建任务并将数据集保存在数据库中，跳转到下个步骤。
5. 用户通过表单勾选用户编号、互动指标、帖子文本字段，软件保存信息到数据库中，跳转到下个步骤。
6. 软件提供描述性统计图，用户确认数据分布正确，启动BERT-3模型，然后可以离开软件，软件在后台运行。软件连续运行BERT-3模型、PCA算法。
7. 当软件执行耗时计算时，任务被标记为繁忙，用户无法操作任务，但可以在其他任务上工作；每个任务单独具有1个线程，不同任务可以并行计算，但受服务器运算能力的限制。
8. 软件显示“解释方差比例折现图”，用户观察图像，设定保留的维度数量。提交后，软件保存信息并完成降维、计算K近邻距离，跳转到下个步骤。
9. 软件显示“K近邻距离折线图”，用户填写Eps参数，软件保存信息并完成聚类、计算帖子的影响力指标，跳转到下个步骤。
10. 软件显示“影响力指标分布图”，用户填写时间序列的采样次数、Granger因果检验的最大滞后阶，软件保存信息。
11. 软件显示搜索舆情社区的用户编号的表单，显示Granger因果检验的参数设置表单；后者的“用户编号”栏目的选项来自于前者的搜索结果。用户填写表单，软件在同一个页面异步返回结果。
12. Granger因果检验的结果（一个用户与其他用户的因果关系）按照最大滞后阶上的p值从小到大排列（p值越小代表影响越显著），用户获取分析结果。

## 工作模块

软件的数据结构如图 2所示，计算步骤如图 3所示。



图 2 软件的数据结构

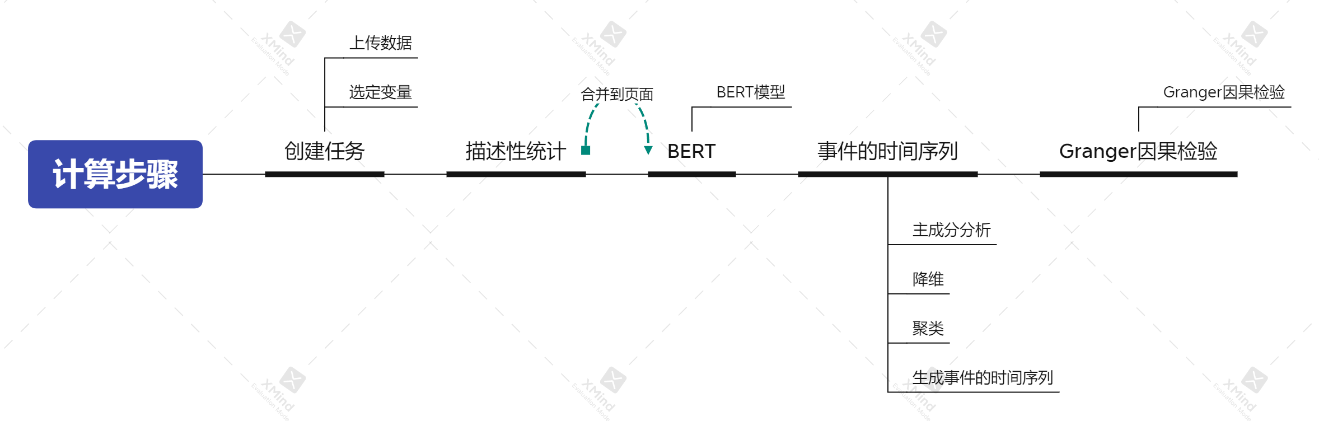


图 3 软件的计算步骤

所有的分析流程模块对任务管理模块有依赖关系，每个分析模块对计算步骤上位于前面的模块有依赖关系。

**用户登录：**本模块负责用户的登录和注册功能，继承自Django函数库的Authorize模块，具有向用户发送注册确认邮件的功能。

**任务管理：**本模块负责管理所有的分析任务，每个任务具有独立的数据、模型、结果；本模块管理每个分析任务的繁忙状态，收集错误提示，并保持数据集及相关信息。本模块在分析流程中，具有创建任务、接收数据集、解析变量的功能。

**描述性统计：**本模块负责绘制描述性统计图，包括用户发帖数量-用户数量的分段折线图、每日帖子量的柱状图。BERT-3模型没有交互操作，与描述性统计模块合并使用1个页面。本模块具有“描述性统计图”数据库表。

**BERT：**本模块负责调用BERT-3模型，保存运算结果。每个任务提取的文本特征，在本模块管理的数据库表对应一条记录。本模块的启动按钮在与“描述性统计”模块共用的页面上，完成分析后直接跳转到“事件的时间序列”模块。

**事件的时间序列：**本模块负责分析帖子的影响力和所属舆情事件，包含3个数据库表格：降维模型、影响力模型、事件的时间序列。对应的多个分析步骤如图 3所示，过程中所有的中间变量、模型、参数存储在3个数据库表格中。

**Granger因果检验：**输入事件的时间序列，本模块负责计算Granger因果关系，和存储上一次分析所固定的参数，包含1个数据表格。

## 主要源码文件的作用

**manage.py**: 软件运行的主函数，是启动和管理网络服务的主体。

**sina\_event\_chain\_django\_cn/urls.py**: 路由表，负责分流网页请求。

**sina\_event\_chain\_django\_cn/asgi.py, sina\_event\_chain\_django\_cn/wsgi.py**: 由Django函数库自动生成，包含在ASGI和WSGI服务器上运行的兼容性程序源码。

**sina\_event\_chain\_django\_cn/settings.py**: 全局设置，包含时区、语言、静态文件存储路径、媒体文件存储路径、密码规则、软件部署密码、模块列表等所有全局变量。

**templates**: 静态网页模板，包含软件界面的静态部分。动态部分由视图函数返回的上下文字典渲染。

其余的my\_login, task\_manager, statistics\_figure, bert, events\_time\_series, granger\_causality分别对应用户登录、任务管理、描述性统计、BERT-3模型、事件的时间序列、Granger因果检验，每个文件夹的结构非常相似，内部文件的功能是：

**migrations**: 数据库迁移记录，用以比较源码和数据库的差异；如果因为版本更新，软件检测到源码被修改，会自动和migrations文件夹中的记录比较，确认自身的版本位置。如果管理员执行更新，软件会按序号生成新的数据库迁移记录，使得源码定义的数据库结构和事实上的数据库结构一致。

**views.py**: 包含了视图函数，函数名称链接到路由表。

**models.py**: 定义了数据库结构，包括表名、字段名、字段类型、限制条件（例如文本字段的最大长度，整数字段的最大和最小值，文件储存地址，别名，日期字段的自动写入现在时间）。

**admin.py**: 数据库表格在管理员后台的显示格式，包括显示哪些字段，修改时使用哪种表单结构，是否分组筛选等。

**apps.py**: 供sina\_event\_chain\_django\_cn/settings.py全局设置中，模块列表读取模块的入口。

# 软件使用说明

## 管理员安装过程

（1）管理员获得程序源码后，在软件根目录下创建token文件夹，在文件夹中创建django\_secret\_key文本文件，写入一个52位的随机字符串（包含英文大小写和数字），作为网络通信密钥。创建smtp.json函数（如果需要使用邮件注册功能），填入发送注册邮件的SMTP邮件服务信息，格式为：

{

"host": "example.com",

"port": 465,

"username": "registration@example.com",

"password": "anypassword"

}

（2）在服务器上通过Anaconda 3创建Python虚拟环境（如果服务器专供本软件使用，不需要创建虚拟环境），进入程序源码的根目录，使用以下命令部署软件：

pip install -r requirements

python manage.py migrate

这是一个最大的运行环境，但不是每个函数包都是必要的。

（3）在软件根目录创建bert\_models文件夹，下载BERT-3模型，网址分别是：

<https://tfhub.dev/tensorflow/bert_zh_preprocess/3>

<https://tfhub.dev/tensorflow/bert_zh_L-12_H-768_A-12/3>

解压后以文件夹bert\_zh\_preprocess\_3和文件夹bert\_zh\_L-12\_H-768\_A-12\_3的形式保存在bert\_models目录下。

（4）管理员使用以下命令创建最高权限的账号：

python manage.py createsuperuser

根据命令行提示，输入用户名和密码（其它非必填项）。

（5）管理员使用以下命令启动程序：

python manage.py runserver 0.0.0.0:8000

其中8000是端口号，可以自定义；但是IP地址必须填写0.0.0.0或127.0.0.1，不能修改为其它值。如果软件是联网供用户使用的，IP地址必须填写0.0.0.0

（6）管理员进入后台，依次在数据库的各个表中添加记录。

管理员后台的首页如图 4所示。导航栏的右侧是返回软件前端和修改密码的按钮；页面主体部分的左栏是数据库中的每个数据表（在每个模块的admin.py文件中注册的），带有底纹的是模块名称，模块名称下的是数据表名；页面主体部分的右栏是当前管理员账号的操作记录。

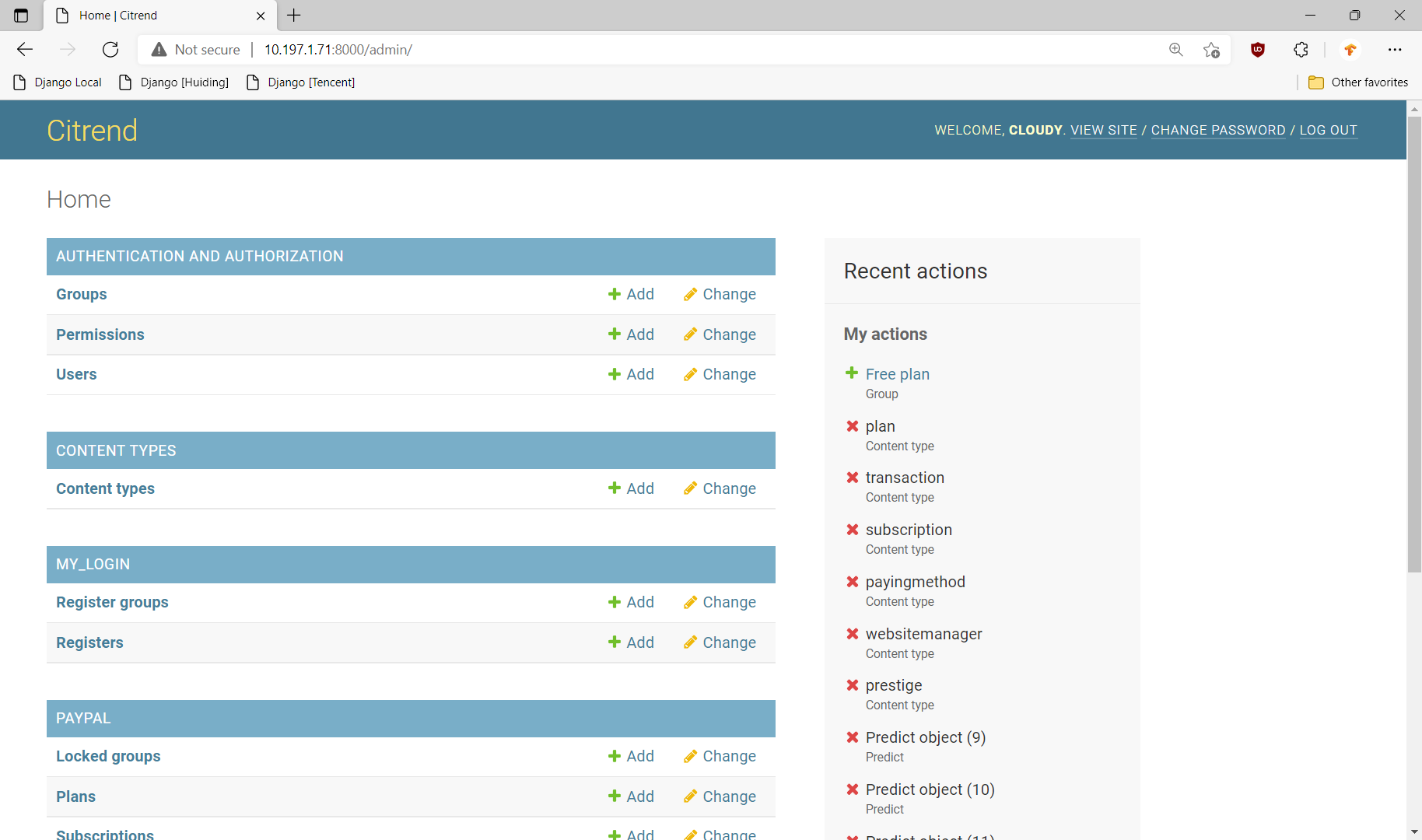


图 4 管理员后台的首页

管理员在部署软件时的操作如下：在Groups表中添加用户组，为用户组设置权限，用户组的权限在Permissions表中预览。在Register groups中添加可自由注册的用户组（一般来说，添加一个用户注册时默认赋予的用户组）。在Locked groups中添加需要付费才能进入的用户组，在Plans中添加对应的订阅计划。

完成操作后，退出管理员账号，完成部署。

## 用户使用过程

（1）软件的首页如图 5所示。页面分为导航栏和主体部分，导航栏上的Clixove徽标链接到首页，“我的分析”链接到分析任务的列表，“帮助”链接到软件作者的信息（外部链接），点击“登入”弹出登录表单。

主体部分包含了软件标题，“采集数据”和“分析数据”按钮，点击“采集数据”的下拉菜单，能够链接到不同的数据采集工具（外部链接），点击“分析数据”跳转到任务列表。

接着是软件简介、分析流程图、模型的介绍，都是静态文本。



图 5 软件首页

（2）用户首次访问软件，在导航栏的右侧点击“登录”按钮，如图 6所示；如果没有账号，点击“不是用户？注册”按钮，进入注册页面。

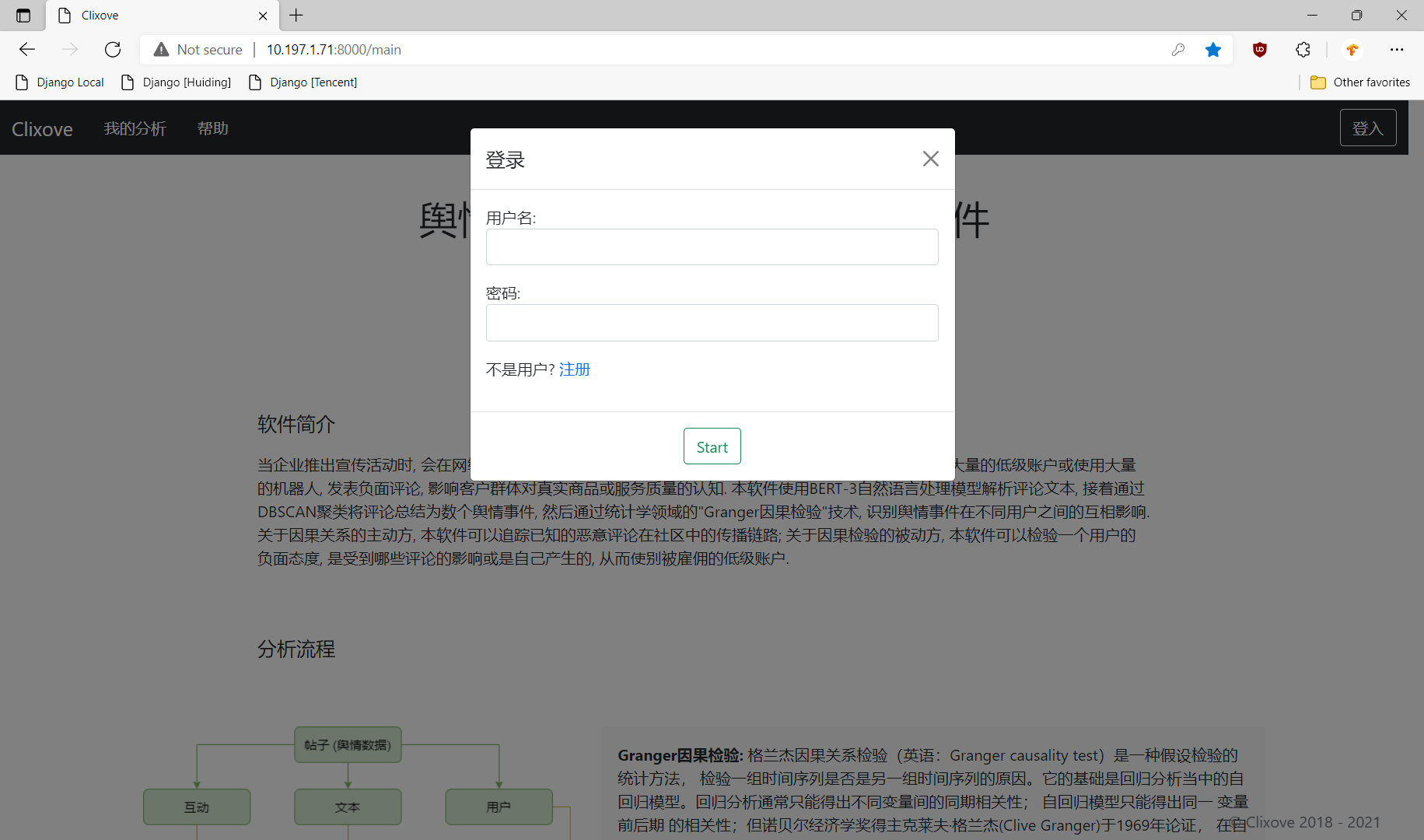


图 6 用户的登录页面

（3）任务列表如图 7所示。第1张表格列出了所有的分析任务，点击“新建分析”按钮新建一个任务，点击“删除”按钮删除一个任务，点击任务的名称进入离开时的分析流程。第2张表格列出了所有分析任务中，异步分析（用户离开）时产生的错误信息，方便用户查看到离开时发生的异常情况。

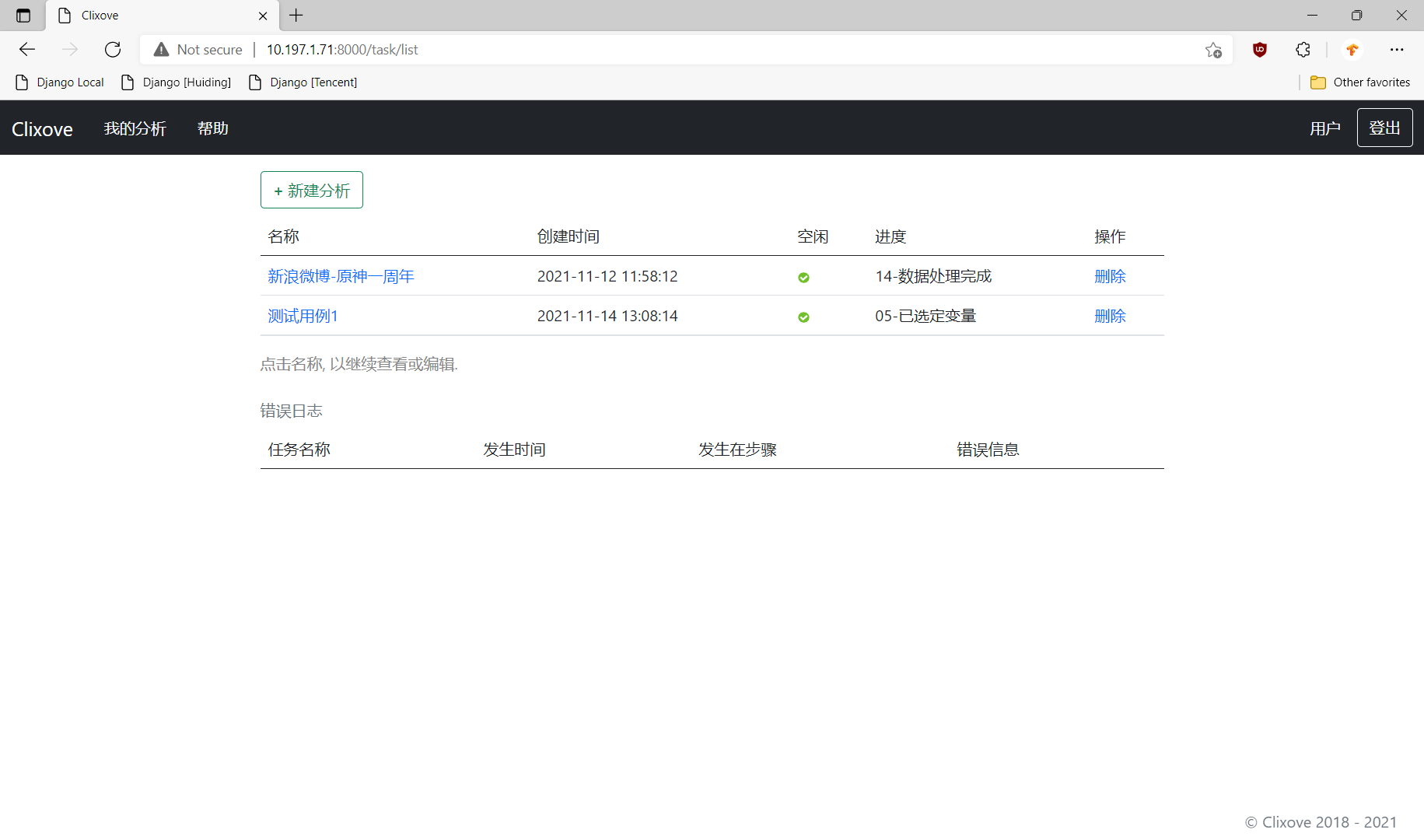


图 7 分析任务列表

（4）点击“新建分析”按钮，如图 8所示。用户填写任务的名称，上传舆情数据集，点击“提交”按钮进入下一个步骤。

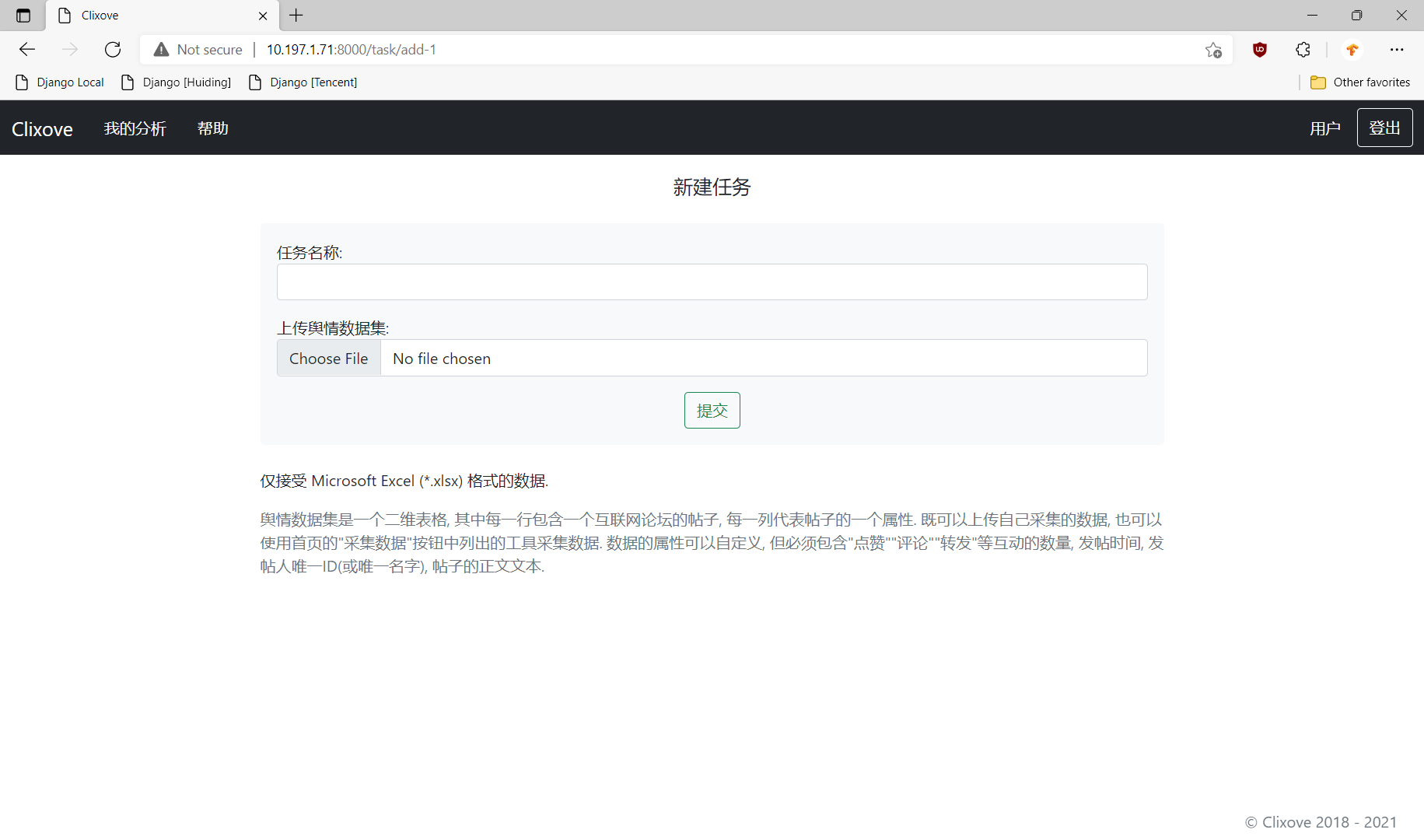


图 8 新建任务

（5）上一步跳转后，页面如图 9所示。用户依次填写不同的变量，点击“提交”按钮跳转到下一个步骤。软件在数据库中记录需要分析的变量名称。

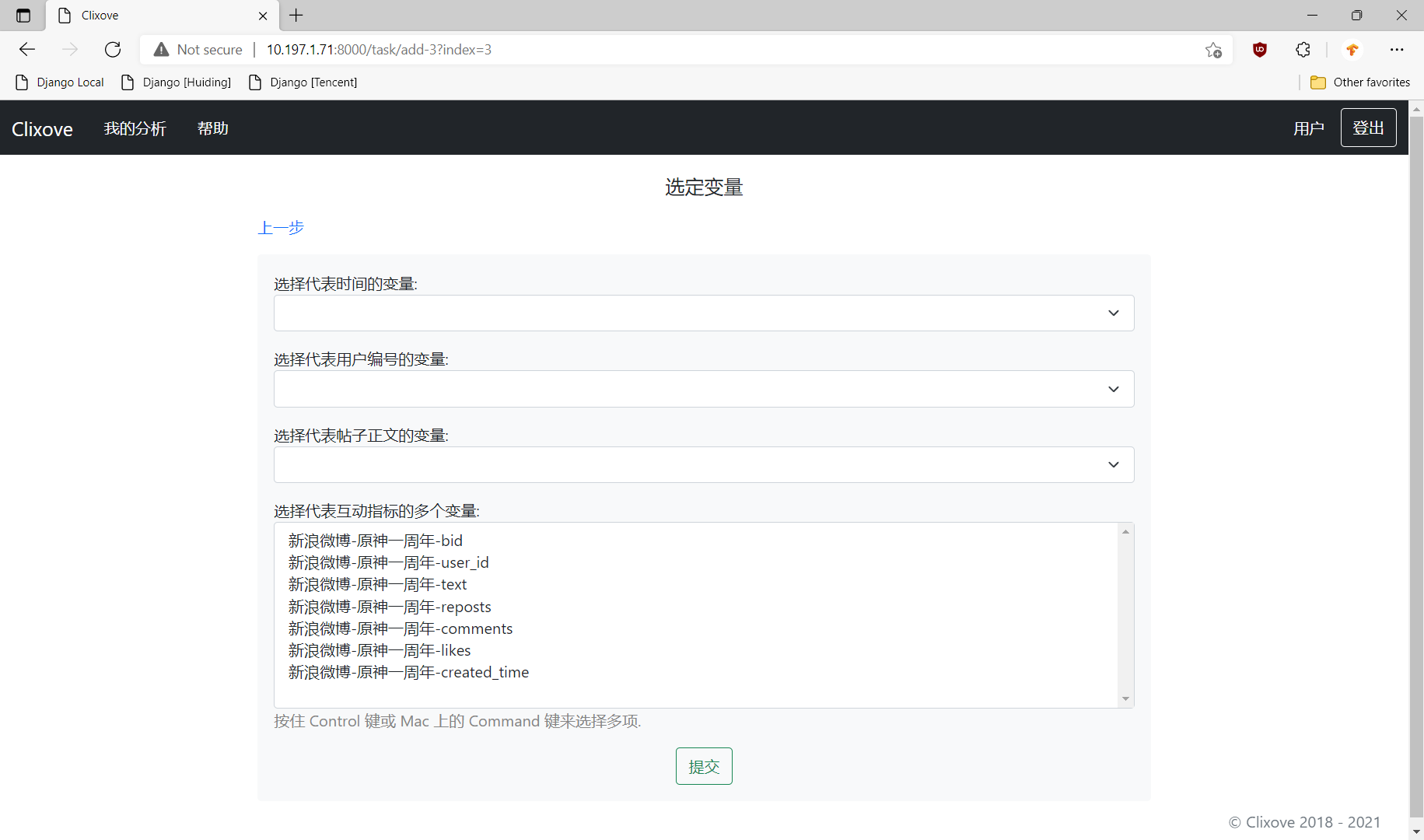


图 9 选定变量

（6）上一步跳转后，页面如图 10所示。页面上的两幅图分别是描述性统计图，用户观察图像，确定采集到的数据是自己所需要的，然后点击“启动BERT-3模型”。运行需要数分钟到数小时不等的时间（根据帖子数量而不同），软件异步执行这个过程，而在用户点击按钮后，直接将页面跳转回到任务列表。用户稍后应点击任务名称，彼时将会自动跳转到下一步骤的页面。软件后端依次运行BERT-3模型和PCA算法，将结果保存到数据库中，将发生的错误登记到异步错误日志。

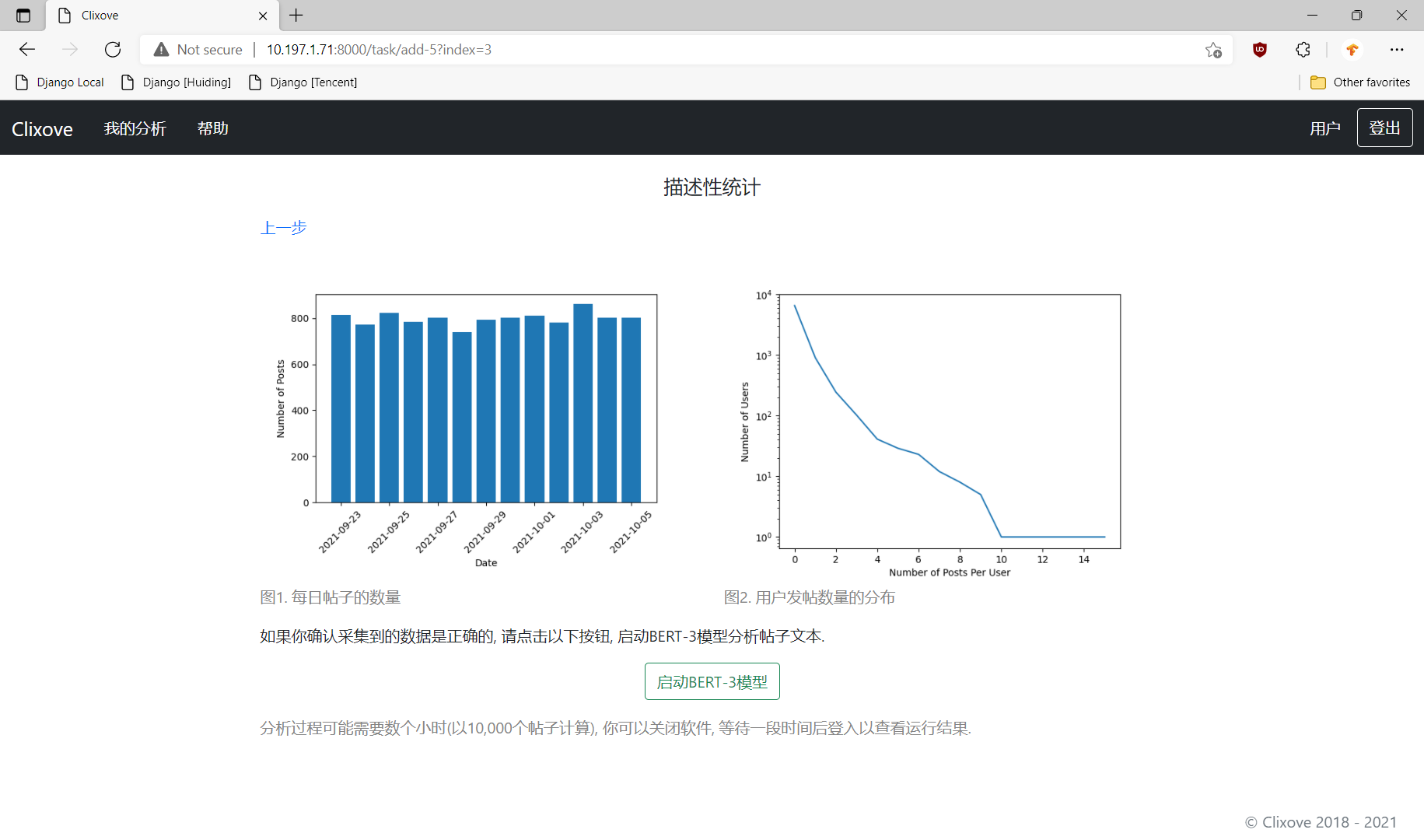


图 10 描述性统计和启动BERT-3模型

（7）用户返回软件时，进入如图 11所示的页面。页面上显示“解释方差比例图”，用户可以在“调整横轴坐标范围”中调整图片显示的范围，点击“重绘”更新图片，以不同的横轴缩放比例观察图像。用户填写“保留维度的数量”，点击“降维”，软件执行降维，然后跳转到下一个步骤。

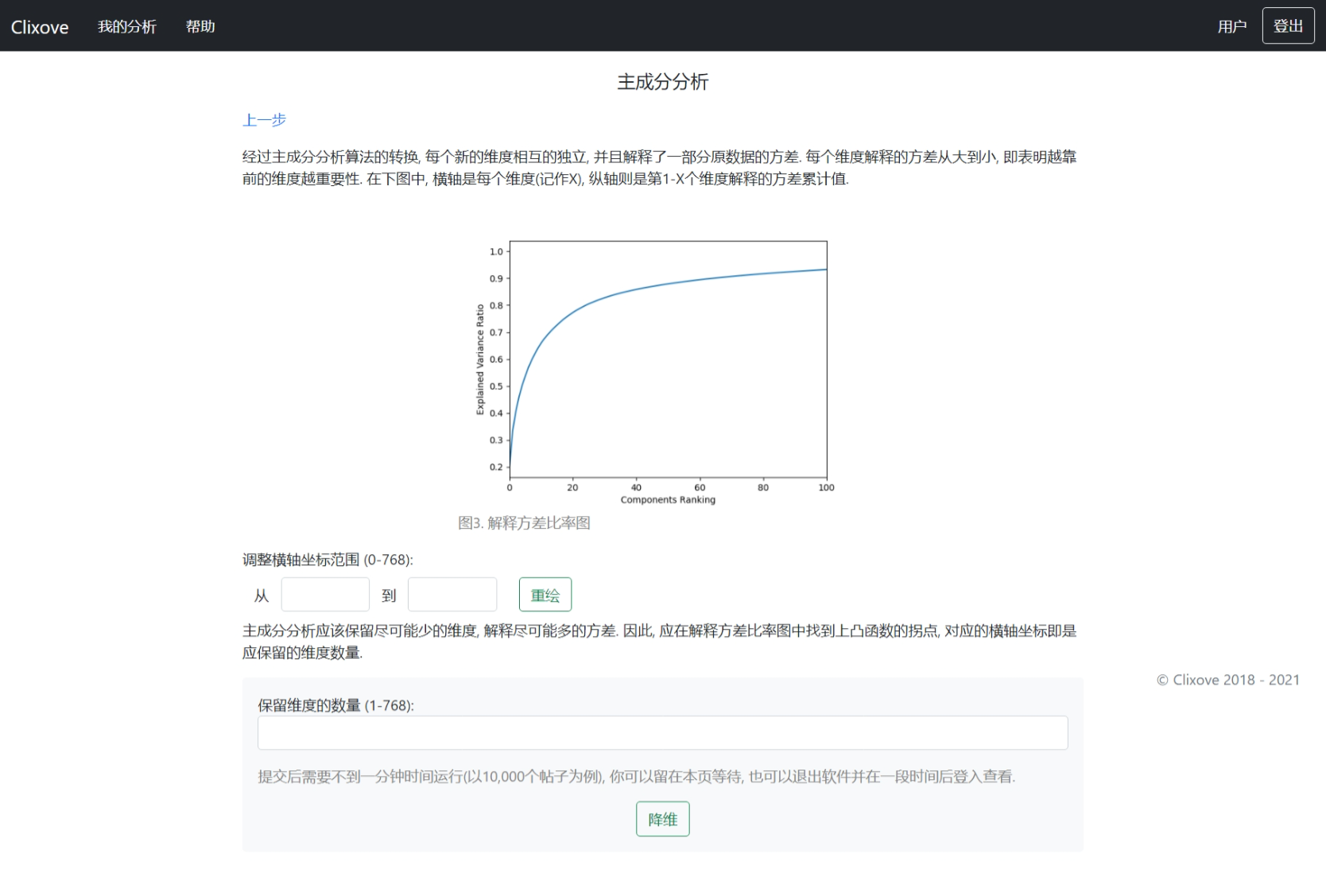


图 11 主成分分析

（8）上一步跳转后，页面如图 12所示。页面上显示了K近邻距离折线图，用户以此为依据确定Eps参数，填写到“Epsilon”的空格中。软件执行聚类算法，然后跳转到下一个步骤。

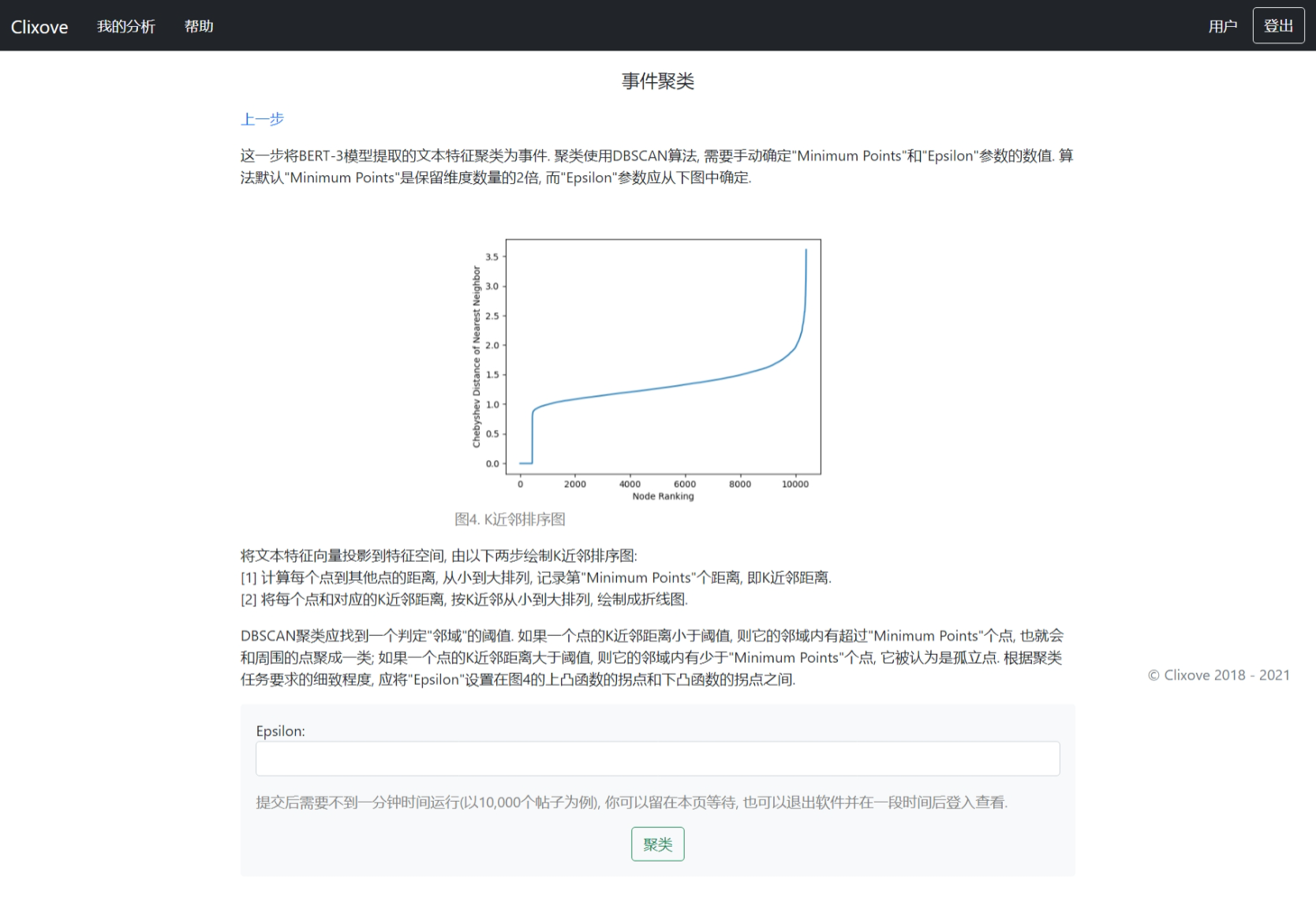


图 12 DBSCAN算法

（9）上一步跳转后，页面如图 13所示。页面上列出了每个舆情事件和对应的帖子数量，用户点击“查看代表性言论”，软件显示每个舆情事件对应的，文本特征最接近聚类中心的10个帖子正文，如图 14所示（代表性言论从数据集抽取，与本软件无关）。

页面上还显示了影响力指标的频率分布直方图。用户填写“等间隔采样次数”和“影响力衰减率”参数，点击“生成舆情的时间序列”按钮，软件生成事件的时间序列，然后跳转到下一个步骤。



图 13 事件的时间序列

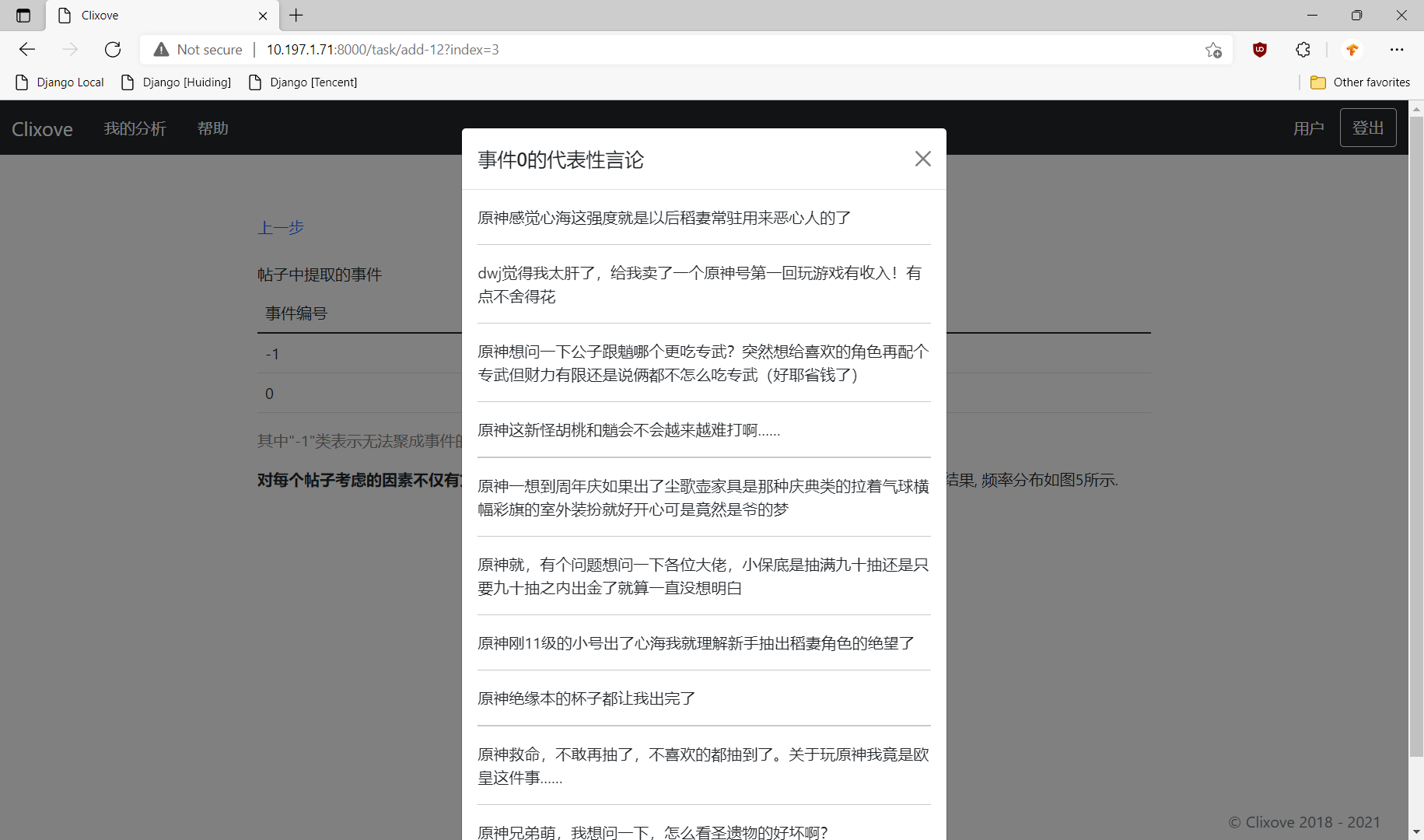


图 14 事件的代表性言论

（10）上一步跳转后，页面如图 15所示。用户首先搜索用户编号，点击“搜索”按钮后软件生成“设置Granger … 参数”表单的内容，如图 16所示。



图 15 Granger因果检验：搜索用户



图 16 Granger因果检验：运行模型

（11）用户填写表单内容，点击“检验”。软件异步执行Granger检验结果，然后在页面上更新检验结果，如图 17所示。

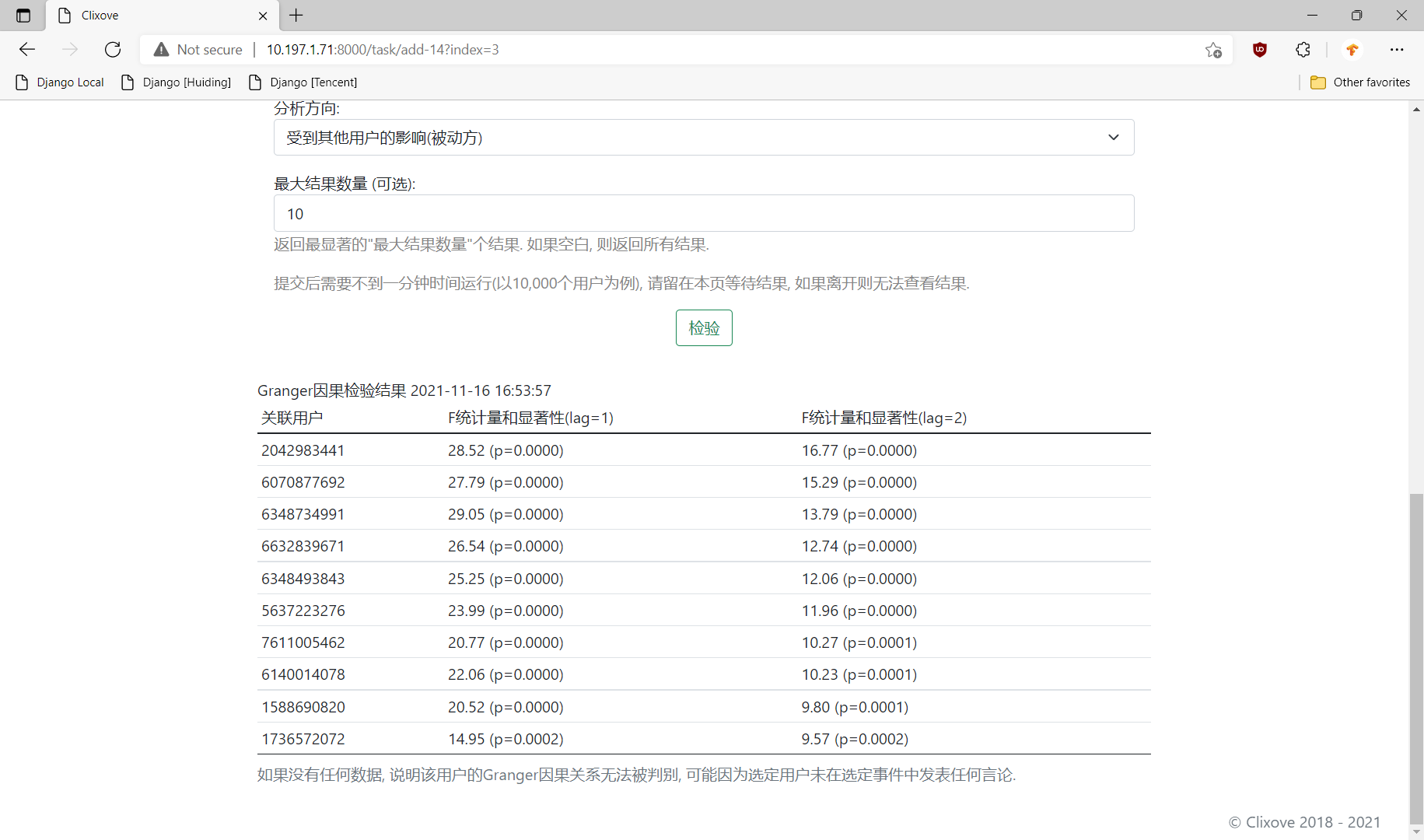


图 17 Granger因果检验：计算结果

至此，用户获取了所需结果，分析流程结束。用户离开软件并重新返回时，可以重复（10，11）步骤，从一个发帖者开始，不断分析舆情事件的因果关系链路。

1. Sander, Jörg, et al. "Density-based clustering in spatial databases: The algorithm gdbscan and its applications." *Data mining and knowledge discovery* 2.2 (1998): 169-194. [↑](#footnote-ref-1)
2. Eichler, Michael (2012). ["Causal Inference in Time Series Analysis"](http://researchers-sbe.unimaas.nl/michaeleichler/wp-content/uploads/sites/31/2014/02/causalstatistics.pdf) (PDF). In Berzuini, Carlo (ed.). *Causality : statistical perspectives and applications* (3rd ed.). Hoboken, N.J.: Wiley. pp. 327–352. [ISBN](https://en.wikipedia.org/wiki/ISBN_(identifier)) [978-0470665565](https://en.wikipedia.org/wiki/Special:BookSources/978-0470665565). [↑](#footnote-ref-2)
3. Granger, C.W.J. (1980). "Testing for causality: A personal viewpoint". *Journal of Economic Dynamics and Control*. **2**: 329–352. [doi](https://en.wikipedia.org/wiki/Doi_(identifier)):[10.1016/0165-1889(80)90069-X](https://doi.org/10.1016%2F0165-1889%2880%2990069-X). [↑](#footnote-ref-3)