**社会计算期末大作业**

**基于小红书“TikTok难民”事件的**

**爬虫与数据分析**

**目 录**

[**--基于“TikTok难民”涌入小红书时期的数据分析 1**](#_Toc11526)

[**一、研究主题选择——数据需求 1**](#_Toc26353)

[**二、数据获取 2**](#_Toc10902)

[1.爬虫编写 2](#_Toc3049)

[2.开源爬虫使用 4](#_Toc8672)

[3.数据说明 5](#_Toc13849)

[**三、数据清洗 6**](#_Toc11073)

[**四、LLM处理 7**](#_Toc22979)

[1.处理原理 7](#_Toc31232)

[2.分析维度 7](#_Toc21742)

[3.主题的提取 8](#_Toc5216)

[4.Prompt设计 9](#_Toc7114)

[5.函数介绍 10](#_Toc20733)

[6.工作流 11](#_Toc14592)

[7.运行结果&截图 11](#_Toc27735)

[**五、数据分析 12**](#_Toc18850)

[1.内容互动语言影响分析 12](#_Toc24446)

[2.内容类型偏好分析 13](#_Toc6565)

[3.关键时间对公众情绪分析 14](#_Toc10543)

[4.互助内容类型分析 15](#_Toc27904)

[5.跨文化情感共鸣分析 16](#_Toc1265)

[6.话题优质评价与话题声量演化可视化 17](#_Toc8932)

[7.词云高频词分析 20](#_Toc10217)

[8.中外用户评论特点分析 25](#_Toc2993)

[9.话题声量变化分析 26](#_Toc6528)

# 一、研究主题选择——数据需求

美国政府官员长期以来以“中国可能利用TikTok收集美国用户数据并进行监视”为借口，将TikTok运营问题政治化。TikTok母公司字节跳动明确表示，不会出售TikTok业务。2024年12月27日，美国当选总统特朗普向美国最高法院提出请求，要求暂停执行TikTok强制出售令。特朗普表示，希望法院在他于2025年1月20日正式就职后，给予他时间通过政治手段解决此争议。当地时间2025年1月10日，据《纽约时报》报道，美国最高法院正加速审理一项涉及强制出售或全美封禁短视频社交媒体平台TikTok的法案，并可能最早于13日这周作出裁决。根据国会2024年通过的法律，TikTok需在1月19日前与其母公司——字节跳动分离，否则将面临全面禁令。

在此政治背景下，面对可能被禁的TikTok，美国网民们提前为自己寻找其他线上的“安身之处”。中国的社交媒体软件“小红书”是本次事件中最大的受益者，大量美国TikTok用户选择小红书作为自己的社交媒体转移地，这些美国网友自嘲这是一种网络“移民”，称自己为“难民”。在这一趋势下，大量美国人或者其他欧美国家用户大量注册小红书，由于事件的突发性，小红书并没有做好用户隔离，使得中国大陆的“小红书”用户可以直接和新注册的国外用户进行直接对话，打破了长期以来中国大陆的简中互联网对外的直接交流，而不需要进行有门槛的翻墙。2025年1月14日，TikTokRefugee词条下显示有7万+笔记，小红书位居美国AppStore免费App排行榜第一名。谷歌趋势也显示，“小红书”在美国的搜索热度，在1月14日快速上涨，一度达到最高值100。2025年1月15日，中国网友打开小红书，被不少英文帖子霸屏。

本次分析目标就是分析在“小红书”上发生的“TikTok难民“事件。因此，需要爬取“小红书”上的笔记和评论数据。通过需要分析，决定获得在本次事件主要关键词下以下笔记数据、评论数据和博主数据。

“笔记ID”：小红书上的笔记唯一标识。

“笔记链接”：小红书笔记URL。

“博主ID”：小红书上博主的唯一标识。

“博主昵称”：小红书博主的个人昵称。

“笔记类型”：小红书笔记类型，有普通的文字图片型和视频型。

“笔记标题”：笔记的标题。

“笔记详情”：笔记的详细内容。

“点赞数”：笔记获得的点赞数。

“评论数”：笔记获得的评论数。

“收藏数”：笔记获得的收藏数。

“分享数”：笔记的分享数。

“视频链接”：视频类型的笔记的URL。

“发布时间”：笔记的发布时间。

“IP地址”：笔记发布时的IP地址

“图片列表”：笔记中图片的URL。

“标签列表”：笔记带有的标签

“检索关键词”：爬取笔记数据时获得该数据的检索关键词。

“评论ID”：小红书中评论的唯一标识。

“笔记ID”：该评论所属的笔记ID。

“评论内容”：评论的内容。

“博主ID”：发布该评论的博主ID。

“博主昵称”：发布该评论的博主昵称。

“IP地址”：发布评论的博主所属的IP地。

“子评论数”：该一级评论拥有的二级评论数。

“点赞数”：评论拥有的点赞数。

“评论时间”：评论发布的时间。

“博主性别”：博主主页介绍中的性别描述。

“头像”：博主头像的图片资源URL。

“简介”：博主主页的个人简介。

“IP地址”：博主的IP归属地。

“关注”：博主关注的博主数量。

“粉丝”：博主拥有的粉丝数量。

“获赞和收藏”：博主获得的获赞和收藏数量。

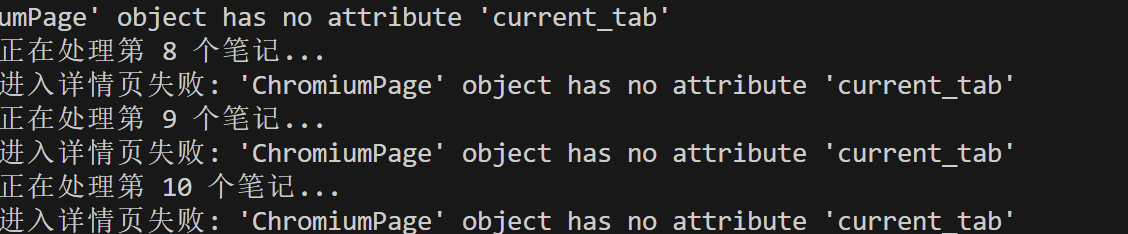
“标签列表”：博主拥有的标签。

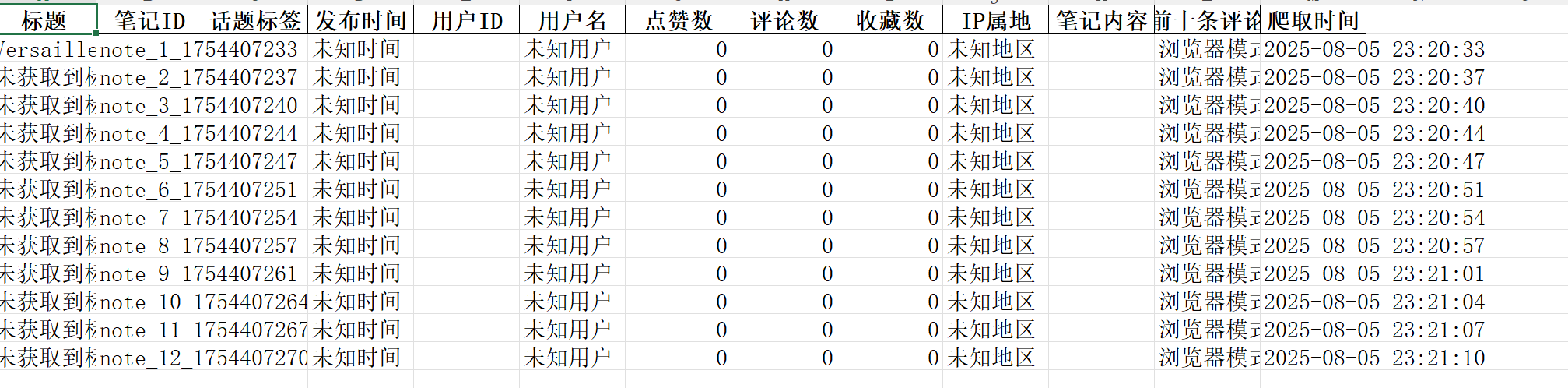
# 二、数据获取

## 1.爬虫编写

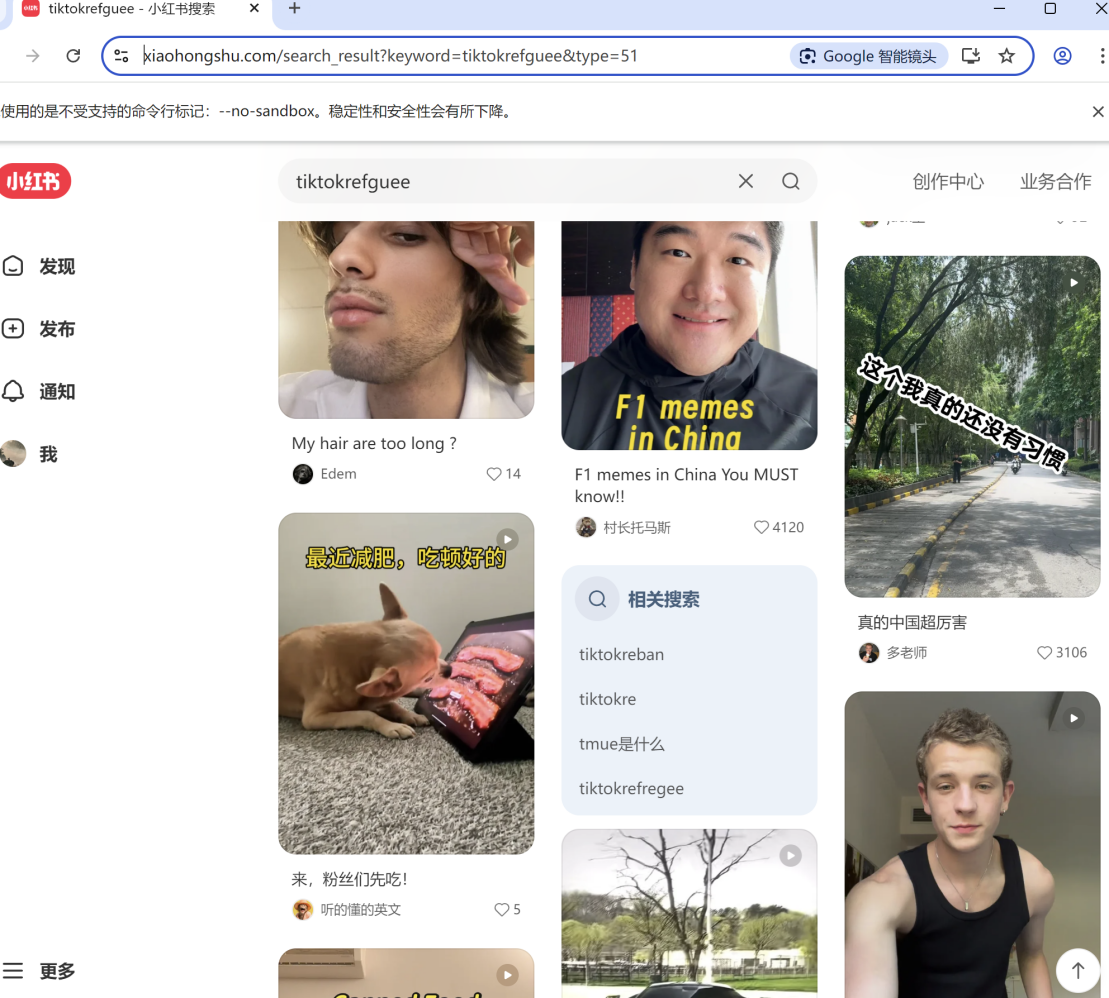
为了爬取到足够的笔记数据，利用Python编写爬虫程序进行数据获取。最开始的想法是编写一个简单的爬虫程序，利用requests库进行网络请求，和DrissonPage库进行浏览器操作，通过手动进行小红书网页的开发者模式获得网站登录后的Cookie，并进行分析网页数据，利用代码爬取html数据并进行解析，最后将数据存储为CSV文件。但是在程序运行时发现，小红书有反爬机制，导致利用程序自动化根本无法获得网页的响应数据。所以进行了调整。

为了应对小红书的反爬机制，选择对小红书的签名算法进行逆向，通过网络检索之前的项目，和一些简单分析，获得小红书得js签名，在新的代码中引入execjs库进行JavaScript签名处理，通过签名算法生成小红书API请求签名头，通过随机延迟模拟人工浏览行为。这次没有使用DrissonPage运行浏览器，而是直接使用requests库进行网络请求和Api处理。但最后可能因为小红书签名算法的更新和反爬机制的升级，所有的运行尝试中都只得到406报错，经过多次调试也不能成功获得数据。

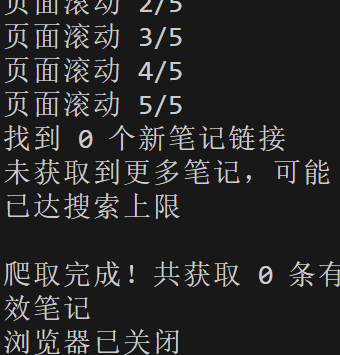




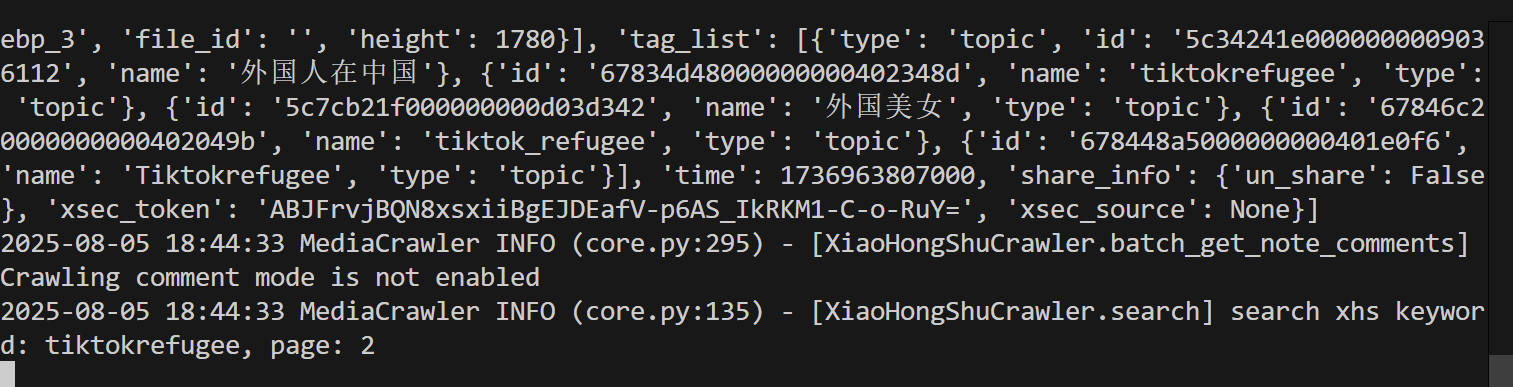
又尝试了只使用DrissonPage浏览器自动化进行数据获取的方式，绕过复杂的API签名机制，直接操作浏览器页面，但还是无法获得数据，通过浏览器自动化进行小红书笔记打开时，总是显示无法浏览笔记详情页，必须手机扫码查看笔记详情，并且并不是解锁网页，而是在手机端进行查看。结合API获取，在浏览器完成人工登录后获取到有效的Cookie，将Cookie同步到requests会话，首先在浏览器模拟搜索，检索需要的笔记关键词，通过正则表达式提取笔记ID，通过笔记ID进入笔记详情页，提取有效数据字段，但同理，进入笔记详情失败。



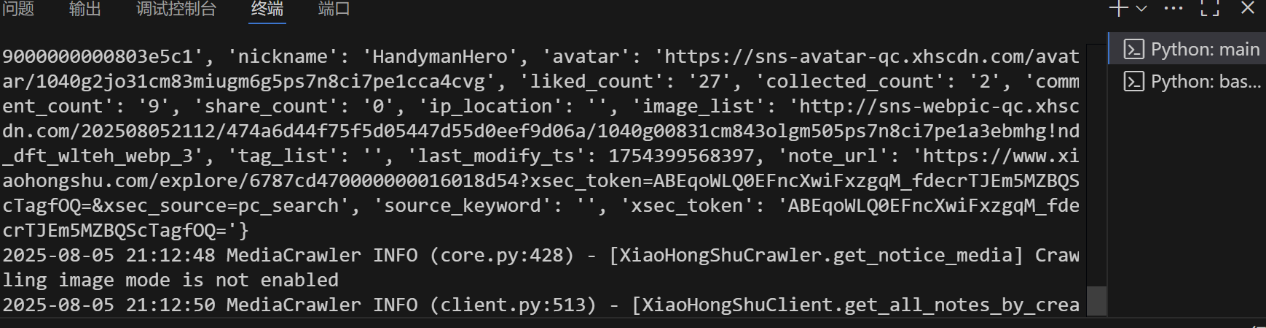
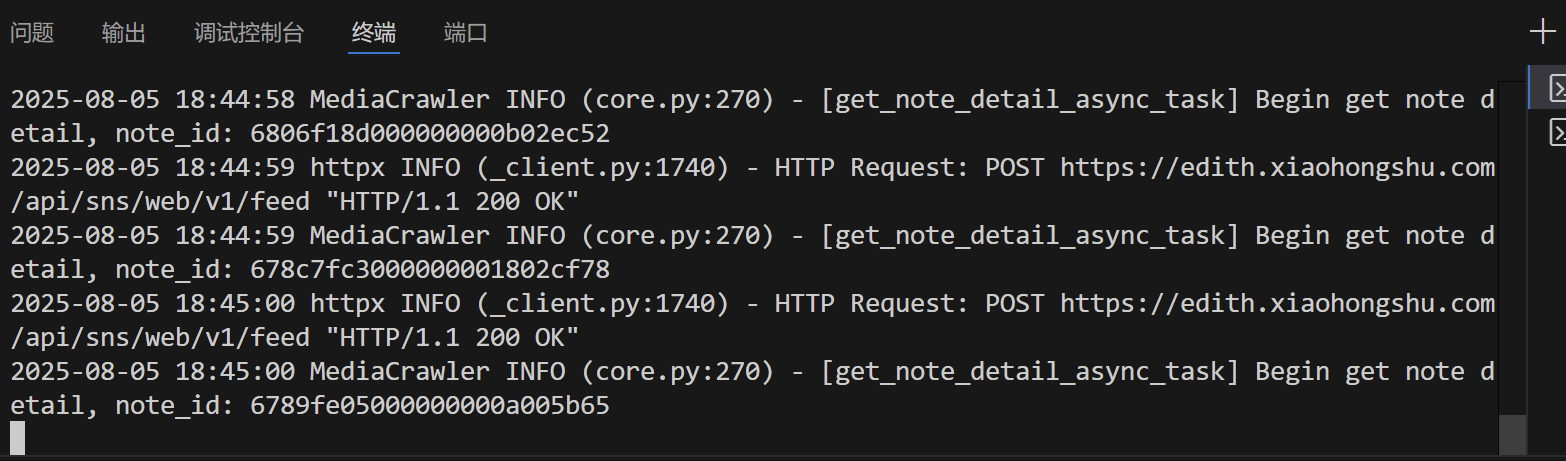
虽然通过人工获取“X-S”、“X-S-Common”用户代理等数据来模拟真实的浏览器环境，通过页面滚动和随机延迟来模拟真实用户行为，但是还是被小红书识别为爬虫，无法成功爬取数据。最后，为了获得小红书数据，不得不使用GitHub上开源的社交媒体爬虫项目MediaCrawler。



## 2.开源爬虫使用

MediaCrawler-自媒体平台爬虫项目在Github上已经获得34.9k星，已经是相当成熟的爬虫项目了，该项目基于Playwright浏览器自动化框架进行平台登录并保存登录状态，利用保持登录状态的浏览器上下文环境，通过JS表达式获取网站签名参数，而不需要人工对网站进行JS逆向，这是比较有技术难度的。作为一个成熟的项目，该爬虫项目由基础架构层，进行基本的爬虫实现、数据存储配置和代理管理；平台实现层，是小红书、抖音等自媒体平台爬虫的具体实现；工具层：负责浏览器启动，签名算法和数据的处理。该项目提供三种数据爬取类型，分别是根据关键词进行的搜索爬取、根据给定的笔记链接进行的详情爬取和根据给定的博主ID进行的博主爬取。

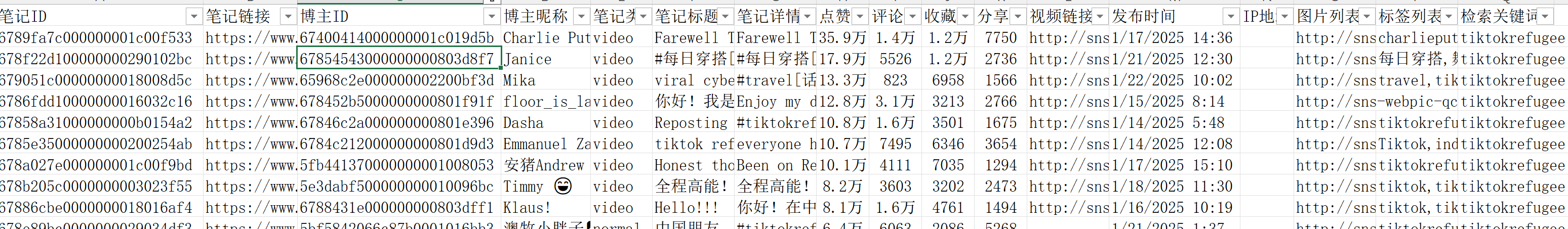
核心的爬取流程首先是初始化客户端，启动浏览器，注入反检测脚本应对反爬机制，创建客户端，人工进行账号登录。然后进行笔记搜索，通过关键词搜索笔记，获取笔记ID和链接，然后通过API和HTML解析两种方式获取笔记详情数据。小红书的签名算法的逆向获取是本项目的核心之一，通过获取动态参数生成签名。还有就是通过代理的方式建立浏览器客户端，使得浏览器的自动化代理程度高，模拟程度高。通过IP代理池进行多IP的轮换避免IP封禁。另一个核心是其反检测脚本的注入，避免了被小红书网站监测为自动化工具。



## 3.数据说明

通过MediaCrawler获得的数据并不是直接符合项目需求，所以人工对数据进行清洗。该项目获得的时间数据是13位毫秒级时间戳数据，并不符合直观数据需求，所以通过公式将其转化为“yyyy-mm-dd hh:mm:ss”时间格式。此外，数据中还有跟多不需要的数据，比如其为了在存储为数据库数据时避免数据重复建立的唯一标识主键“last-modify-ts”，数据被爬取时的时间戳，还有笔记博主的头像URL等，将这些无用的数据删除，然后依据笔记ID这一唯一标识，对数据进行筛选，删除重复数据，删除不同关键词检索并爬取到的重复数据。因为爬取时没有限制时间，所以有一些关键词还可能检索出2024年甚至更早的数据，这些数据肯定是不符合本次关于“TikTok难民”事件分析需求的。所以进行筛选后删除。最终得到有效数据。





# 三、数据清洗

在数据清洗阶段，我们重点关注了数据来源的相关性与规范性，其中检索词的选择对数据质量起到了决定性作用。由于数据爬取以检索词为核心线索，检索词的精准度直接影响所获帖子与研究主题的关联程度：当使用“TikTok refugee”相关精准检索词时，爬取结果多聚焦于“TikTok难民涌入小红书”这一核心议题，数据相关性较高；而若仅以“TikTok”为检索词，结果则会混入大量国外电商运营教学、外国短剧讨论等无关内容，这类数据不仅增加了后续处理的冗余度，还可能干扰情感分析或主题统计的准确性（所以在进行主题提取工作时暂时排除了该类帖子）。

此外，爬取过程中出现的空行也需进行系统性删除，避免在调用大语言模型进行标注或分析时因格式问题导致报错，确保数据格式的规范性。

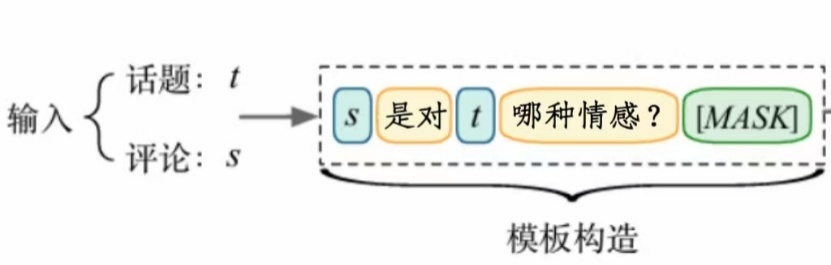
在后续主题标签归类时还进行了一些清洗筛选和特殊数据的标记处理。

# 四、LLM处理

## 1.处理原理

在对小红书平台中与“TikTok和难民涌入”相关帖子的评论进行情感分析时，我们采用**提示词工程**（Prompt Engineering）借助大模型完成标注工作，这一方式相比传统的人力标注具有显著优势。一方面，大模型能够高效处理大规模评论数据，在短时间内完成成百上千条评论的情感标签标注，大幅提升了研究效率，尤其适用于社交媒体中高频更新、数据量庞大的场景；另一方面，大模型的标注逻辑基于预设的统一规则，可减少因标注人员个体主观差异、疲劳状态或认知偏差导致的标注不一致问题，从而在一定程度上保证了标注结果的稳定性和可靠性。此外，这种方式还能降低人力成本，避免长期重复劳动带来的资源消耗，更符合实证研究中对数据处理效率与一致性的基本要求。

同时，我们并未直接对评论进行孤立的情感识别，而是使用**面相话题的评论情感识别**，预先提取的帖子主题词，并与评论内容相结合形成**“话题-评论”对**。这一做法的核心优势在于，主题词为评论提供了具体的语境背景，使情感识别更贴合评论所指向的具体话题，减少了脱离上下文导致的情感误判。例如，同一情感表达在不同主题下的含义可能存在差异，结合主题词的分析能更精准地捕捉这种差异。这一思路也与当前相关领域的研究实践一致——越来越多的研究表明，面向具体话题的评论情感识别能有效提升分析的针对性，为后续的统计分析，如不同主题下情感倾向的差异比较，提供更可靠的基础，使研究结论更具实际解释力。



## 2.分析维度

为精准且全面地对社交媒体评论进行情感分级，研究整合Ekman基本情绪理论、社会情绪理论及情绪维度模型，构建了一套逻辑严谨的分级体系。

首先，以为精准且全面地对社交媒体评论进行情感分级，研究整合Ekman基本情绪理论、社会情绪理论及情绪维度模型，构建了一套逻辑严谨的分级体系。

首先，以 **Ekman 基本情绪理论**中的快乐、悲伤、愤怒、恐惧、厌恶、惊讶为基础框架；纳入**社会情绪理论**中与群体传播密切相关且具有普遍性的“感动”；考虑到社交媒体中评论的交互特性，补充**互动性社会情绪**，包括驱动进一步交流的“疑问”以及作为正向互动反馈的“赞美”，以捕捉评论中的交流动态；针对评论中常见的比较场景，设置“对比”性情绪，以应对具体的**比较情境**。这10个共同构成基础核心情绪，是个体与群体共通的情感表达基石。

此外，还设有**中性**情绪，用于标注无明显情感倾向或情感表达模糊、中立的评论内容。

同时，借助**情绪维度模型**对上述所有情绪进行补充描述：通过愉悦度衡量情绪的正负倾向（如从悲伤到快乐）；通过唤醒度反映情绪的强度（如从平静到兴奋）；通过支配度体现态度倾向（如从反对到赞同），从而实现对社交媒体评论情感的多维度、精细化分级。

## 3.主题的提取

在话题提取环节，我们首先将爬取到的数百条帖子内容整合为文本集合，通过AI工具读取文本并统计主题词词频，以此初步归纳核心话题。这一过程与通过代码分词处理的逻辑本质相通，均以文本内容中的高频核心词为依据，结果具有一定的一致性。基于词频统计结果，我们将主题词初步划分为10类，并对每类主题的内涵进行了说明。随后，通过编写代码调用LLM，让模型根据帖子内容判断其所属主题类别，若不在 10 类范围内则标记为“不相关”。

代码详见【Python代码-附件一：topic.py】

| **类** | **释义** | **数目统计** |
| --- | --- | --- |
| tiktokrefugee/tiktok\_refugee /tiktok 难民 | 涵盖所有与 “TikTok 难民” 直接相关的标签和表述，是贯穿全文的绝对核心主题 | 100 |
| 猫税 /cattax/cat tax | 美国用户入驻小红书后形成的标志性友爱互动仪式，形成现象级文化梗 | 80 |
| 音乐 /music/ 歌曲分享 | 包括独立音乐、TikTok 热曲、翻唱、乐器演奏等内容，反映音乐成为难民群体与本土用户的重要交流媒介 | 35 |
| 取名/中文名/Chinese name | 外国用户寻求中文名、中国用户反向取英文名，体现跨文化身份认同的趣味互动 | 25 |
| 对账/信息差/物价对比 | 中美网友围绕房价、医疗、教育、税收等生活成本的“赛博对账”热潮，突出经济差异话题 | 20 |
| 宠物/狗/动物/animal | 猫税梗之外，单纯分享宠物日常的内容，强化“萌宠外交”属性 | 18 |
| 文化差异/语言学习/翻译 | 包括 “姐妹” 含义、英语试卷梗、中文名避雷指南等，聚焦文化碰撞细节 | 15 |
| 小红书/RedNote/平台生态 | 讨论小红书功能、教程、社区规则，或对比 TikTok 与小红书的差异 | 12 |
| 美国/中国/国籍身份 | 明确标注国籍的内容（如 “我来自美国”“德国网友提问”），体现地域标签的集中使用 | 10 |
| 搞笑/抽象/梗图 | 对“猫税”梗的二次创作、表情包、谐音梗等娱乐化内容 | 8 |

为避免遗漏重要话题，我们对“不相关”帖子中评论量较高的内容进行了二次核查，其他不相关话题下的评论则筛去，最后发现了两个值得关注的额外主题：其一为**“交友意向（friend）”**，这类帖子多为用户发布的交友内容，虽文字描述较简洁，但评论区互动活跃，属于高讨论度话题，却因文本信息量有限未被初始词频统计捕捉；其二与 “难民回归（remain）” 相关，部分帖子讨论TikTok运营机制改革后，原活跃于平台的“难民”用户或回归、或潜水，甚至有用户提问“难民是否都已离开”。这类帖子不仅评论量高，且参与讨论的外国用户占比显著——**即便这些用户未在小红书发布原创内容，仍保持着浏览和互动行为**，这一用户行为特征具有独特的研究价值。因此，我们将这两个主题补充至话题体系中，最终形成的评论集也主要围绕上述核心及补充主题展开分析。

## 4.Prompt设计

研究所采用的Prompt设计，在结构上涵盖了背景、身份、具体任务、传递参数及few-shot示例等关键要素，形成了一套逻辑完整且实用性强的提示框架。

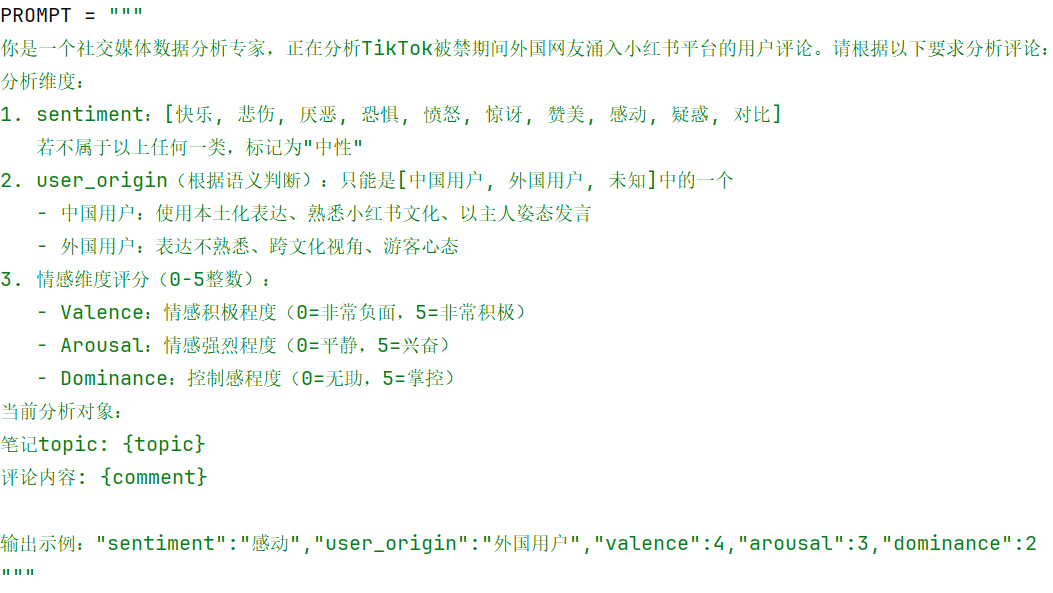
首先，明确的**背景说明**为大模型提供了必要的上下文，使其能够理解任务产生的场景和研究的核心对象，为后续任务奠定了认知基础。

其次，赋予大模型明确的**身份定位**，将其设定为专注于社交媒体内容分析的标注助手，使其能以符合研究需求的角色开展工作，在处理任务时保持一致的视角和专业度。

具体的**任务描述**则清晰界定了模型需要完成的操作，像对评论进行情感标签标注、判断帖子主题词所属类别等，让模型能够精准把握工作目标，避免出现任务偏离。

**传递参数**的设置为模型处理数据提供了灵活的依据，通过 { } 形式纳入主题词和评论的具体信息，使模型能自动读取Excel中的内容进行批处理。

而**few-shot**示例的加入，即提供具体的案例展示，为模型提供了直观的参考范例，帮助其更快理解任务规则和预期结果，有效减少了因理解偏差导致的错误，进一步保障了处理结果的准确性。



整体而言，这样的 Prompt 设计全面且合理，各要素相互配合，能够引导大模型高效、准确地完成相关分析任务。

最终代码详见【Python代码-附件二：LLM.py】

## 5.函数介绍

**（1）call\_api(topic, comment, timeout)**

功能：构造请求参数，调用API并返回结果

优化：添加timeout参数，10秒内无响应则终止请求

**（2）parse\_response(generated\_text)**

功能：解析API返回的文本，提取5个目标字段

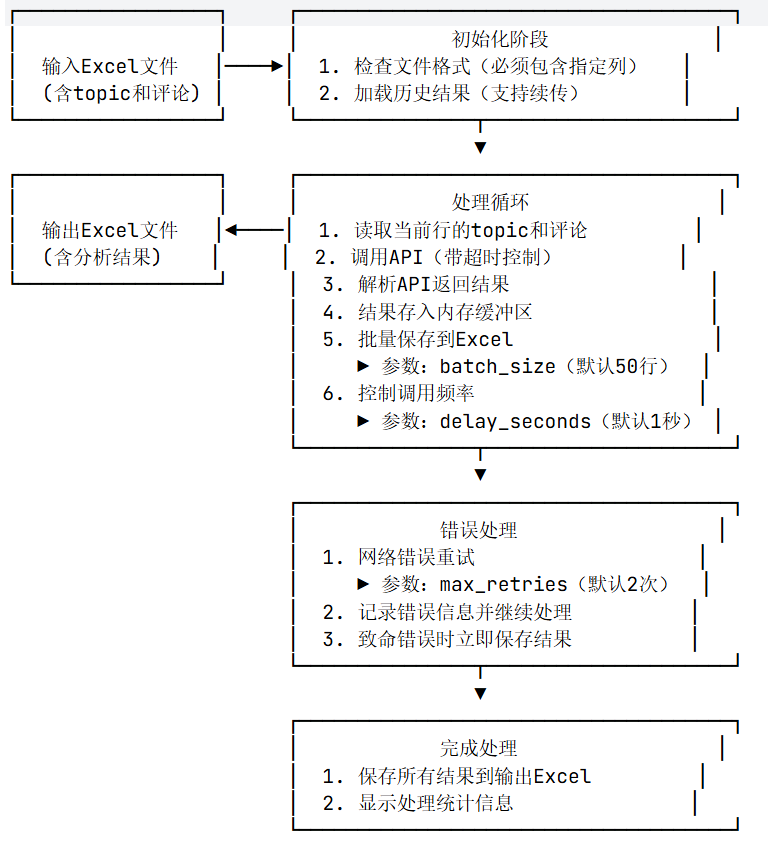
特点：找不到关键词时返回0，确保程序不崩溃

**（3）process\_excel(...)**

功能：协调整个处理流程，包括文件读写、进度显示、错误处理

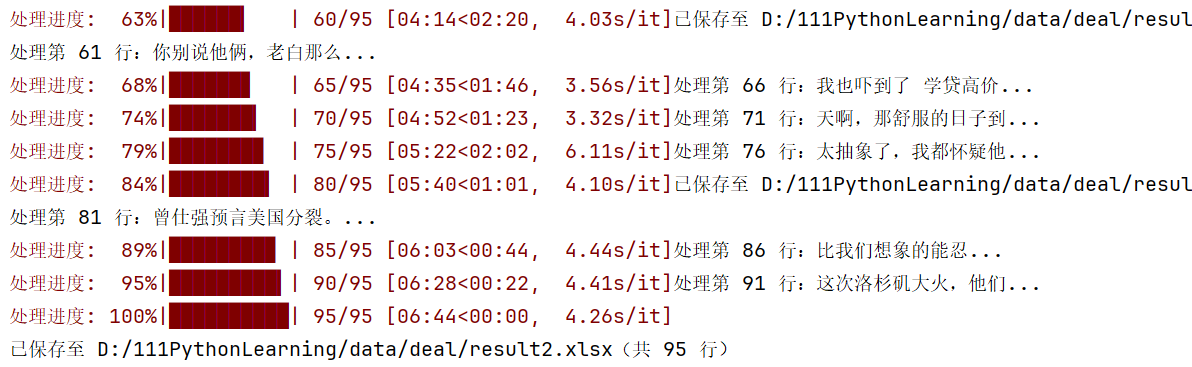
核心机制：内存缓冲区+批量保存，减少文件操作频率；实时打印进度，方便定位问题行

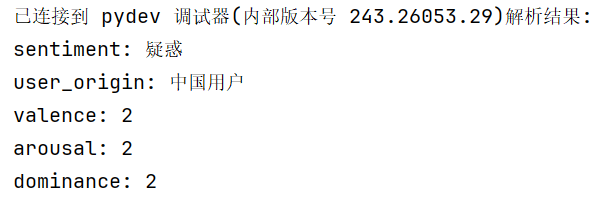
## 6.工作流



## 7.运行结果&截图

数据集处理：



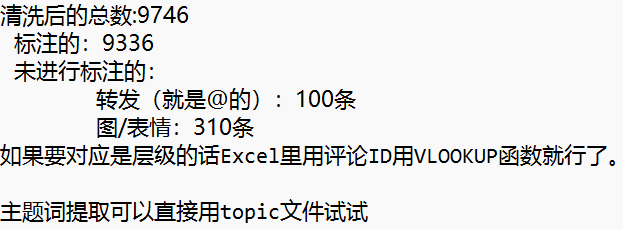
费用：单元测试：

先进行了小单元测试，确保处理的输出可以安全接入Excel，然后使用截断批次大规模流程处理。大概1k条一小时，分批跑了十多个小时，花了10元左右，有一些处理空值。最终数据集见评论update.xlsx的第三个sheet。

截图：



一些额外处理说明：

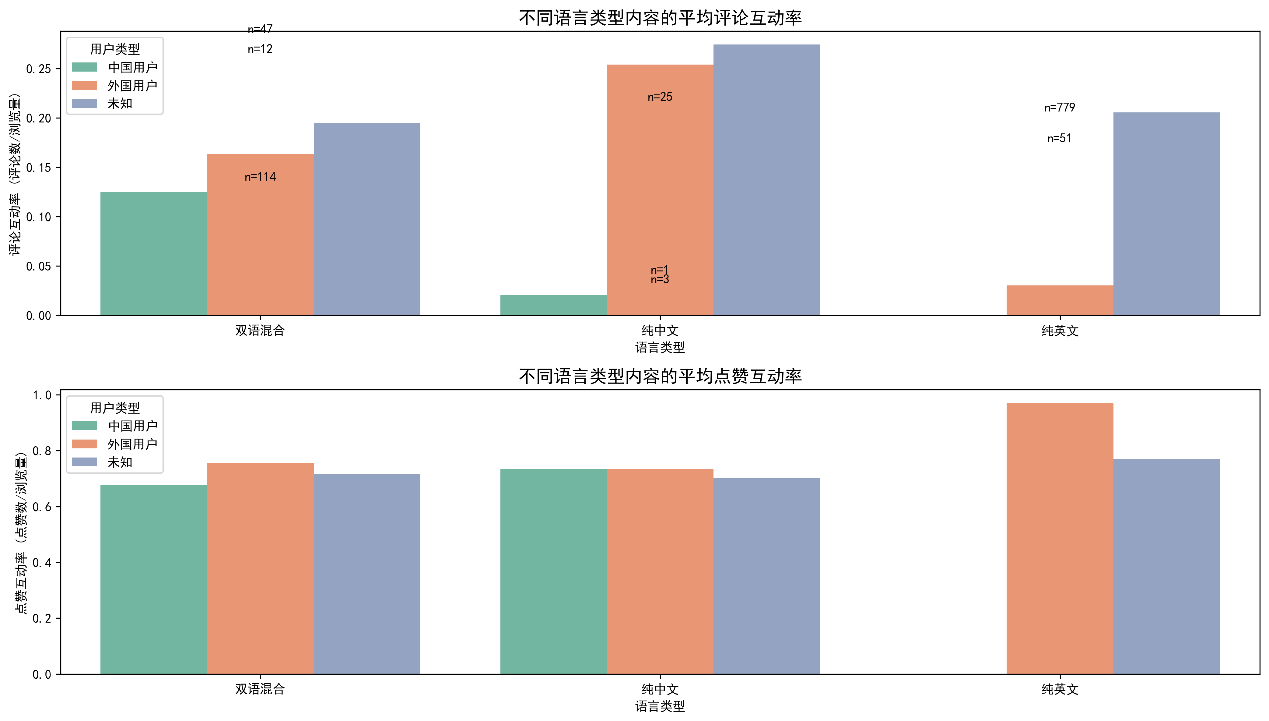


# 五、数据分析

## 1.内容互动语言影响分析

统计中外用户发布的纯英文、纯中文、双语混合内容的互动率，定义互动率为评论数/浏览量和点赞数/浏览量，由于获得的数据中没有获得笔记浏览量，定义浏览量=点赞数+收藏数+评论数，探究双语表达是否显著提升跨文化互动效率。

代码详见【Python代码-附件三：dataAnalysis.py】



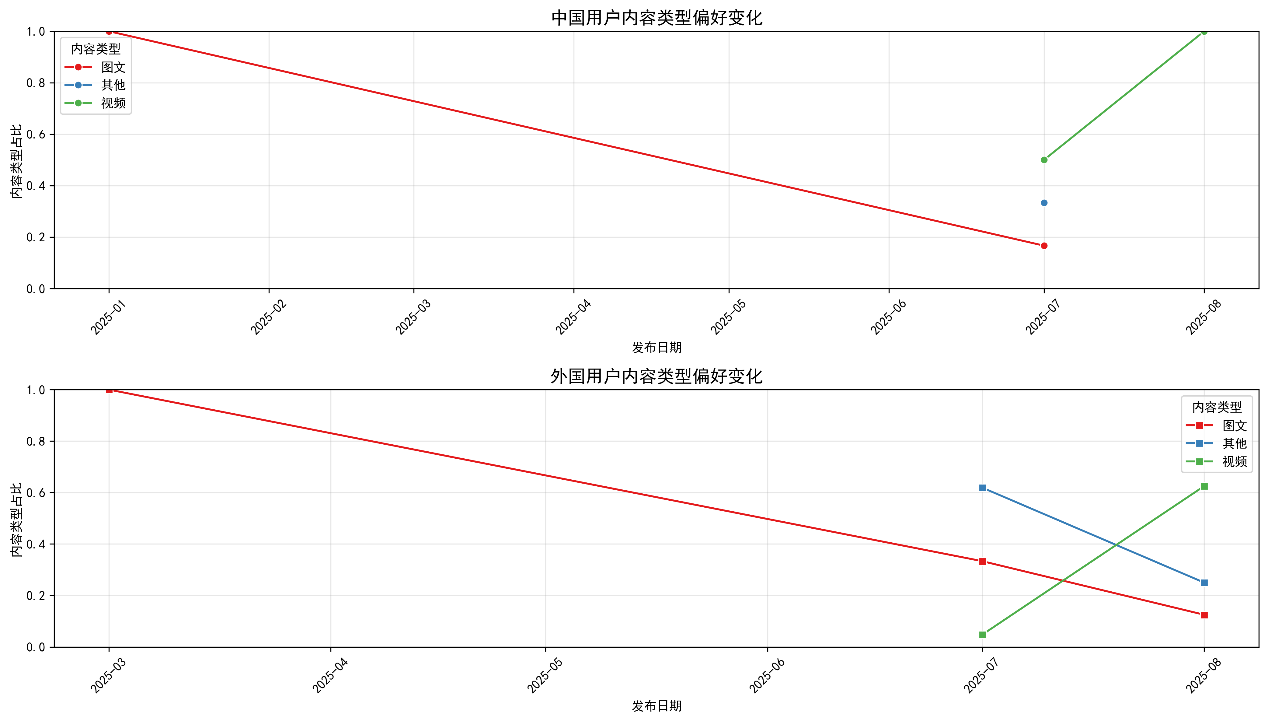
可视化结果显示，双语混合内容确实可以一定的提升跨文化互动效率。在评论互动率中，双语混合内容的评论互动率，中国用户≈0.12、外国用户≈0.16，显著高于纯中文内容中中国用户≈0.02、纯英文内容中外国用户≈0.03的互动率。说明双语表达能有效激发中外用户的评论互动，尤其对外国用户吸引力更强。双语混合内容的点赞互动率为中国用户≈0.7、外国用户≈0.75，与纯中文、纯英文内容差距不大，但整体处于较高水平。双语内容更易引发“深度互动+认可” 的组合行为。

从用户类型角度看，纯中文内容的评论互动率最低，只有约等于0.02，但双语混合内容的互动率中评论≈0.12、点赞≈0.7，接近外国用户水平。反映中国用户对双语内容的包容性，既理解中文语境，也能通过英文元素感知外国用户需求，在本次Tiktok难民涌入中国社交媒体事件中主动参与跨文化互动。

## 2.内容类型偏好分析

对比外国用户与中国用户的笔记类型占比，外国用户是否更倾向发布短视频作为从tiktok平台转进来的用户，是否延续 TikTok 使用习惯，而中国用户是否更喜欢在相同话题下发图文笔记。并且这一趋势是否随时间发生变化，“难民”逐渐融入小红书平台，更加适应平台主要特征。

代码详见【Python代码-附件三：dataAnalysis.py】



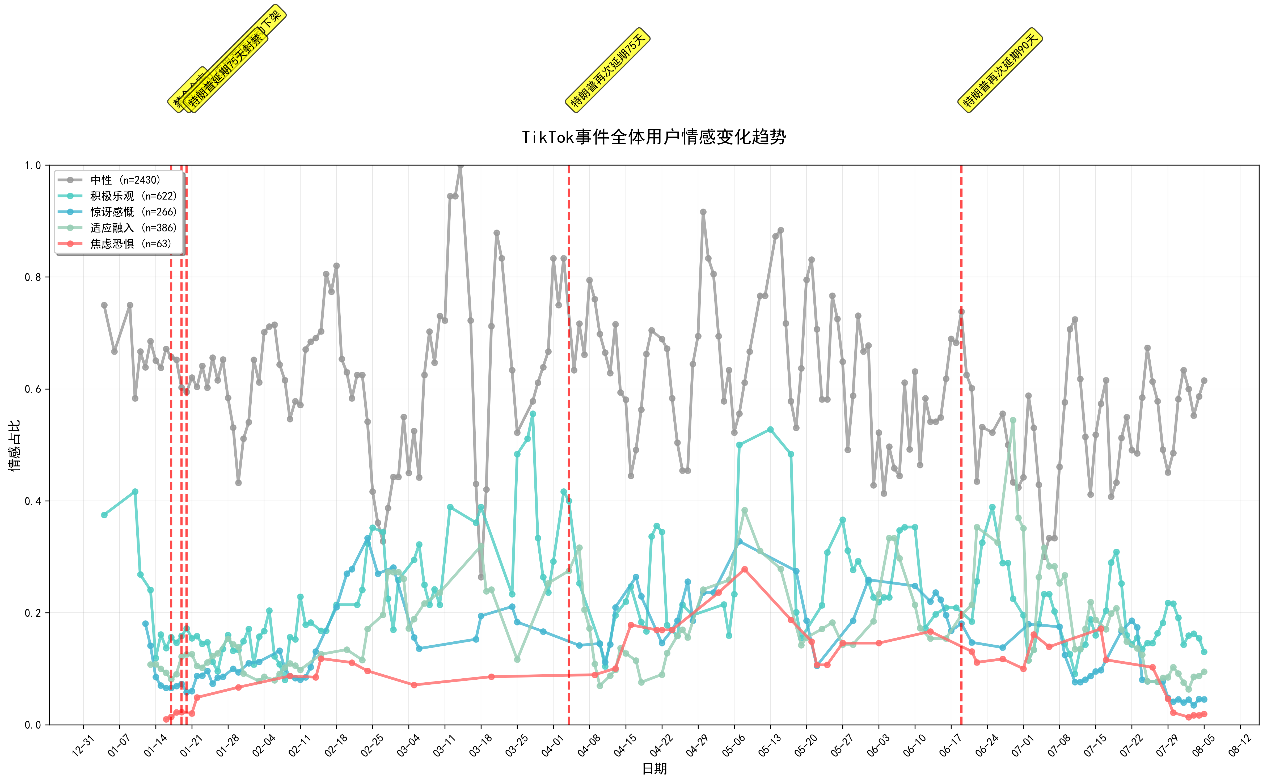
中国用户的笔记类型偏好趋势是图文主导到以视频类型为主，在“难民时间”前期，图文内容占比长期稳定在0.8以上，视频占比极低，说明中国用户在跨文化内容互动初期，延续小红书“图文社区”的传统创作习惯，更擅长用图文传递信息。7月视频占比迅速拉升至 0.5，可能是在“难民事件”后期，相关检索词能检索到的笔记数量减少，同时外国用户涌入后，中国用户为适配其“短视频”偏好，进行视频创作的数量增多，导致比例发生极大变化。

国外用户也并非如预期在初期进入小红书时依赖视频类型笔记，延续其在Tiktok平台的使用习惯，而是适应小红书平台特色，使用图文笔记类型的帖子比例很高。作为“难民”，其不得不被动适应小红书用户习惯。同样是在7月，视频类型比例逐渐上升，外国用户逐渐适应平台，更多回归自身习惯。中国用户前期“图文依赖”和外国用户“渐进视频转型”，本质是小红书本土生态与TikTok 外来习惯的碰撞，中国用户用熟悉的图文降低跨文化沟通成本，外国用户用视频延续内容创作惯性。

## 3.关键时间对公众情绪分析

按时间轴标记关键节点，比如今年时间的起点，一月十七日美国最高法院支持“不卖就禁”法案合宪，二十号特朗普延期禁令75天执行，分析节点前后的笔记和评论的情感分布变化，并且分析中国用户的情感变化是否同步波动。

代码详见【Python代码-附件三：dataAnalysis.py】



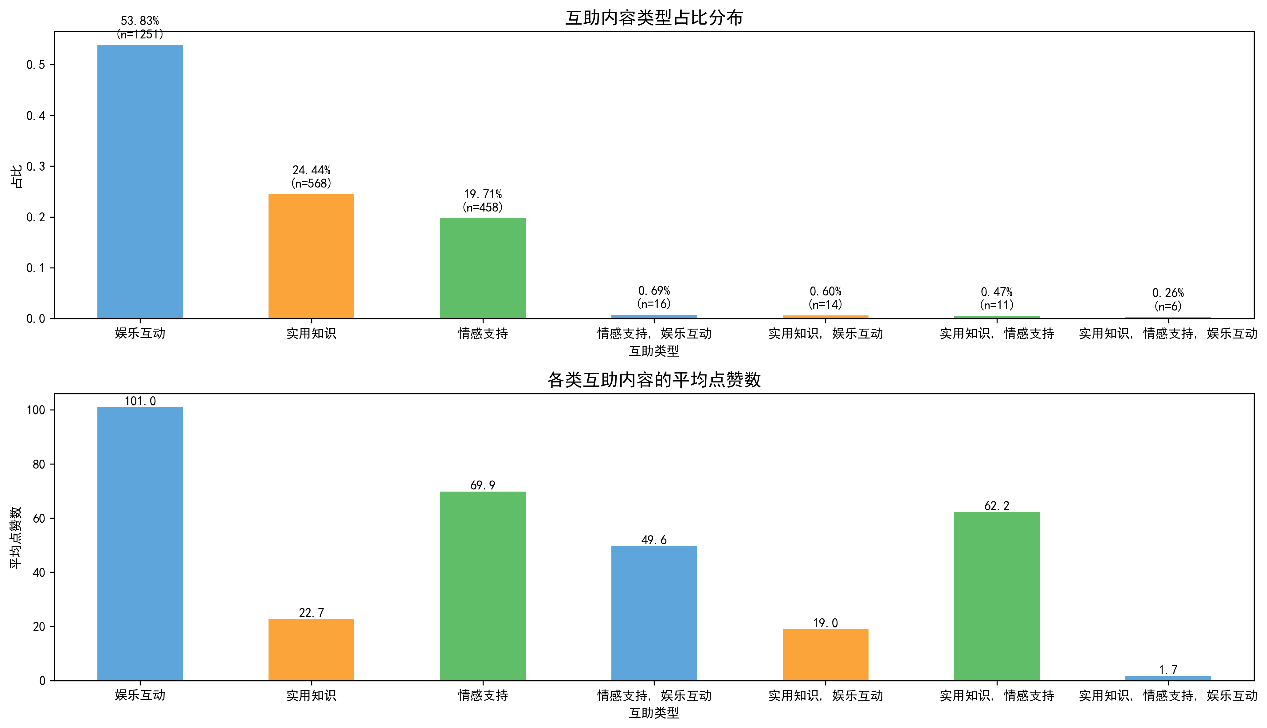
从图可清晰看到 TikTok 事件发展过程里，全体用户情感随关键事件波动的趋势。中性情感占比长期稳定在0.6-0.8，说明TikTok事件中，多数用户保持围观态度，未被单一情感主导。这符合跨文化事件传播规律，大众更多是信息关注者而非深度参与者。其他情感如积极乐观、惊讶感慨、适应融入、焦虑恐惧占比均低于0.4，且波动与关键事件强相关，说明Tiktok难民事件中，美国政策的关键事件对涌入小红书的难民用户会产生明显的情感影响。

在第一阶段，也就是美国最高法院判定针对Tiktok的“非卖即封”禁令合法时，“难民”开始涌入小红书，这时焦虑恐惧情绪占比从0飙升至0.15+，中性情绪占比从0.8跳水至0.6。在4月份特朗普对禁令第二次延期时，惊讶感慨情绪、适应融入情绪占比同步拉升，中性情绪再次跳水；焦虑恐惧情绪占比从0.1 回落至0.05。二次行政令触发复杂情感反应，既有对政策反复的惊讶感慨，也有用户开始主动适应，同时用户的恐惧下降，说明更多人意识到彻底封禁Tiktok的最坏情况发生的概率降低。在六月，特朗普再次延期90天，积极乐观情绪、适应融入情绪占比持续高位，焦虑恐惧占比跌至0.05以下，中性缓慢回升。三次行政令后，用户完成情感脱敏，部分人接受政策反复现实，积极寻找跨文化互动新路径。三次行政令时期均触发焦虑，但峰值逐次降低，说明用户对封禁Tiktok这一同类负面事件逐渐脱敏，焦虑强度下降。

## 4.互助内容类型分析

分类统计评论中的互助信息，将评论中的信息分为实用知识、情感支持、娱乐互动等几个类别，并分析哪种类型的互助获得更高点赞，即更符合这一特殊事件下用户的迫切需求。

代码详见【Python代码-附件三：dataAnalysis.py】



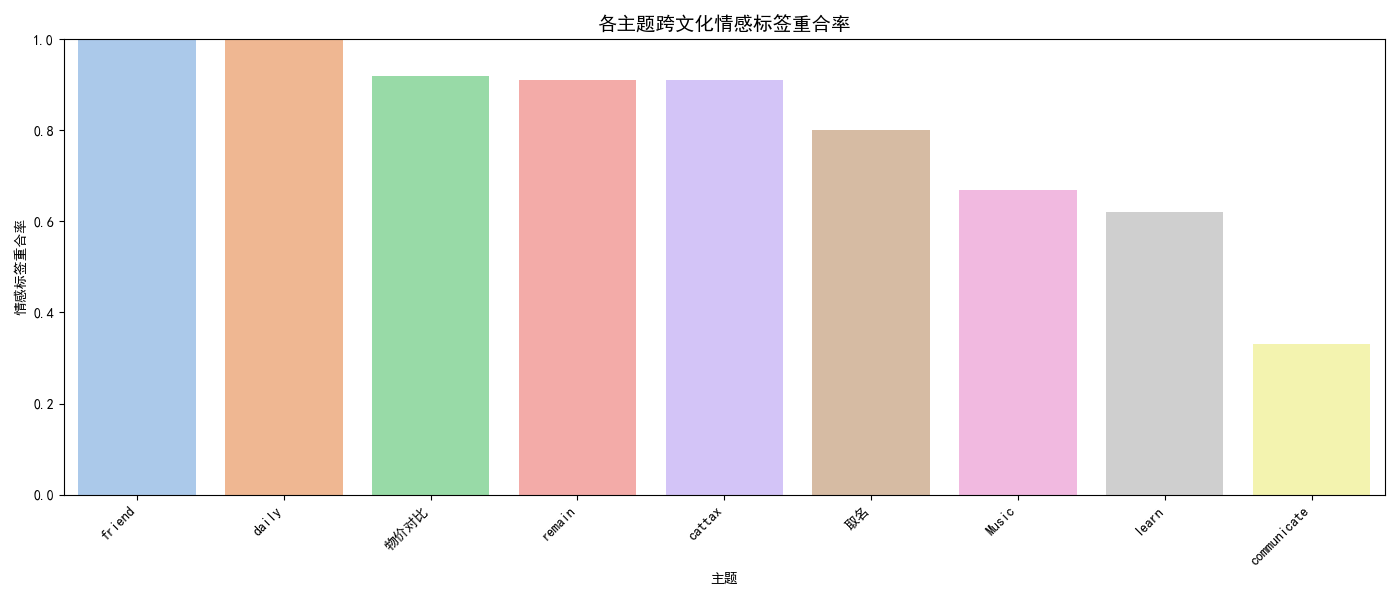
从需求优先级上来看，“娱乐互动”是跨文化互助的最大需求。娱乐互动类互助占比为53.83%，远高于实用知识的24.44%和情感支持的19.71%，占比绝对领先。说明跨文化社区中，用户更倾向用“轻松娱乐”打破陌生感，比如玩梗、分享搞笑等，成为中外用户初次互动的首选形式。娱乐互动无需专业知识、情感深度，简单的“猫税表情包”“文化搞笑对比”就能引发互动，契合跨文化场景中用户的心理，自然成为最主流的互助类型。娱乐互动平均点赞 101.0，是实用知识的4.4倍、情感支持的1.45倍。说明“纯粹娱乐”更易获得广泛认可，用户用点赞投票鼓励轻松互动。情感支持点赞69.9高于实用知识，反映跨文化场景中“情感共鸣”比“实用帮助”更稀缺、更具价值。对于复合类型，情感支持+娱乐互动的平均点赞为49.6，点赞数远高于“实用知识+娱乐互动”的19.0和“实用知识+情感支持”的62.2。说明“情感+娱乐”的组合，既保留互动轻松感，又叠加情感价值，是高效的跨文化互助公式。

## 5.跨文化情感共鸣分析

该分析基于通过LLM处理打分并打过标签的评论数据进行。现有数据已明确 12类核心主题、三维度评分（valence/arousal/dominance）及用户来源。对每类主题，计算中外用户情感标签的重合率，即中外用户共同提及的情感标签占比，指数越高说明情感认知越一致，及通过皮尔逊系数计算中外用户在愉悦度、唤醒度和支配度三维度评分上的相关性系数。重合率越高、相关性越强，认为主题的跨文化共鸣强度越高。对比高共鸣主题与低共鸣主题的内容属性，分析前者的共性特征。

各主题跨文化情感标签重合率：结果显示friend和daily为绝对的高共鸣主题，此外，物价对比和难民回归的共鸣指数也很高。说明“交友、日常分享、生活对比”是跨文化场景中情感共识最易达成的话题。这类话题贴近衣食住行、人际关系等普世生活经验，中外用户的情感认知高度一致，天然适配跨文化传播。learn、communicate重合率垫底，且中国用户样本量极少，外国用户样本也有限。反映“学习、深度沟通”类话题，因文化教育体系差异大、专业抽象内容多，情感共识难建立。

代码详见【Python代码-附件四：跨文化情感共鸣分析.py】



跨文化情感共鸣强度分析结果：愉悦度、唤醒度、支配度呈现“弱相关”特征。所有主题的情感维度相关性几乎趋近于0，甚至出现负相关。这说明情感标签重合率高，不代表中外用户的情感强度感知同步。communicate主题的支配度相关高达0.53，但样本量仅极少，且情感标签重合率0.33。说明小样本下的“强相关”无意义，反而暴露深度沟通类话题具有情感理解偏差大，但偶尔能达成强烈共识的矛盾。

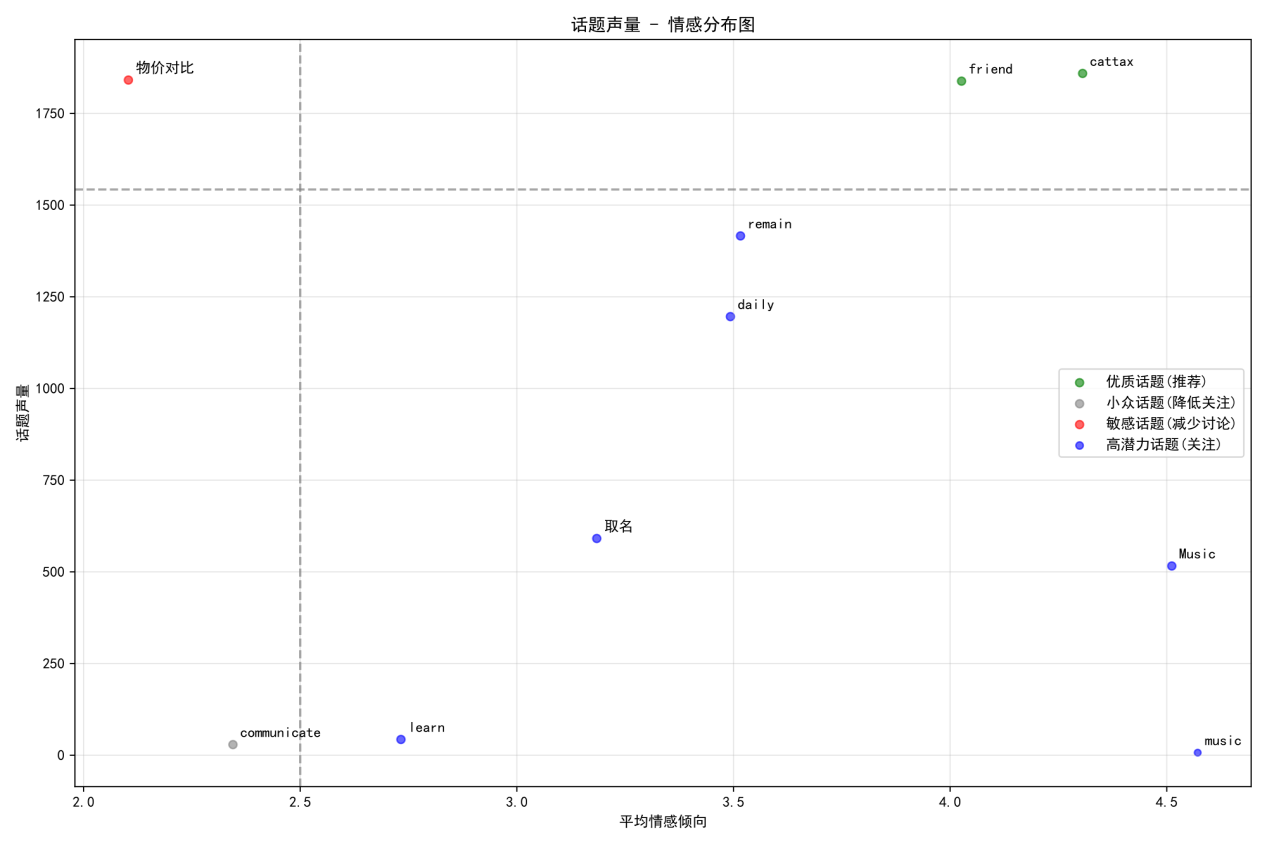
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | topic | 中国用户样本大小 | 外国用户样本大小 | 情感标签重合率 | 愉悦度相关性 | 唤醒度相关性 | 支配度相关性 |
| 4 | friend | 388 | 1073 | 1.0 | 0.03 | -0.02 | -0.02 |
| 3 | daily | 213 | 697 | 1.0 | -0.11 | -0.03 | -0.05 |
| 8 | 物价对比 | 503 | 697 | 0.92 | 0.04 | 0.02 | 0.01 |
| 6 | remain | 446 | 558 | 0.91 | -0.07 | 0.02 | 0.04 |
| 1 | cattax | 359 | 585 | 0.91 | 0.02 | -0.05 | 0.06 |
| 7 | 取名 | 122 | 131 | 0.80 | -0.07 | 0.04 | 0.01 |
| 0 | Music | 205 | 204 | 0.67 | 0.07 | -0.09 | 0.05 |
| 5 | learn | 12 | 22 | 0.62 | -0.38 | -0.16 | -0.12 |
| 2 | communicate | 5 | 15 | 0.33 | 0.21 | -0.56 | 0.53 |

## 6.话题优质评价与话题声量演化可视化

结合前文LLM提示词对数据“sentiment”列文字进行映射：5分组（快乐，赞美，感动），2.5分组（非积极非消极的中性情感），2分组（惊讶，疑惑，对比），0分组（悲伤，厌恶，恐惧，愤怒），计算各话题的总情感倾向。以2.5分作为阈值线，总情感倾向高于2.5的为积极情感话题，低于2.5的为消极情感话题。

在得到的各话题总情感倾向后，利用爬取到的各话题下评论总数计算该话题声量（Volume）即热度，可以绘制出声量-情感分布图，以衡量话题优质程度。声量排名在前70%视为高声量话题即热门话题。“高声量+积极感情”的话题是做优质话题应以予推荐，“低声量+积极情感”的话题作为高潜力话题关注，“高声量+消极情感”的话题应作为敏感话题减少讨论，“低声量+负面情感”的话题作为小众话题可以降低关注。基于对话题的评价，可以使用推流与限流机制以维护平台良好氛围。运行代码得到可视图表如下：

代码详见：【Python代码-附件五：话题筛选.py】

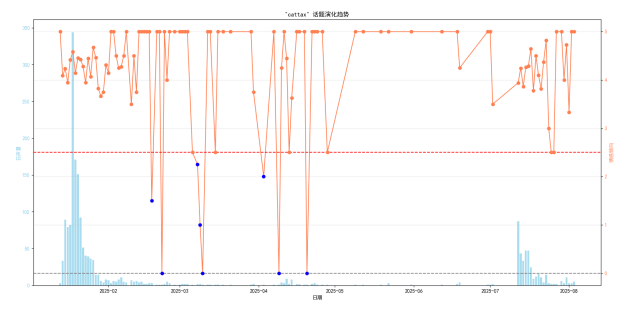
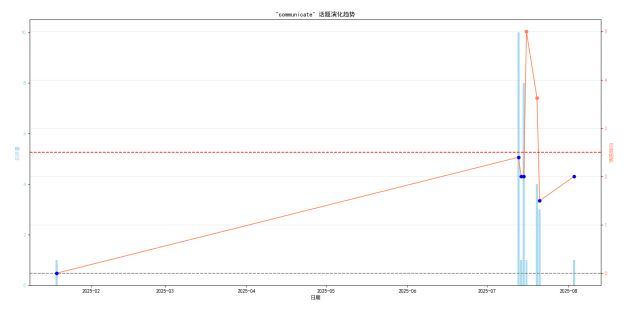
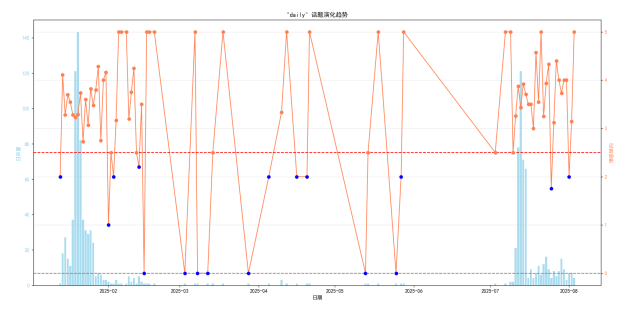
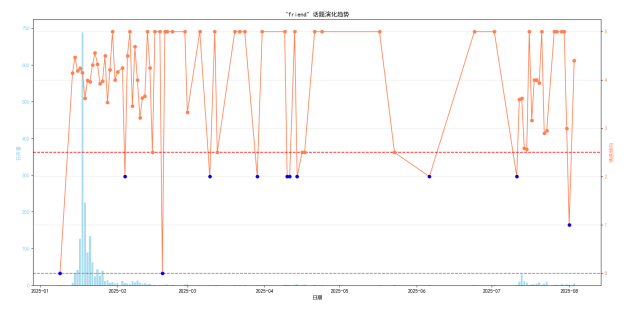
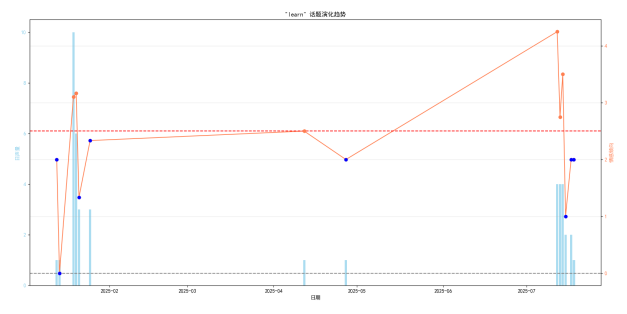
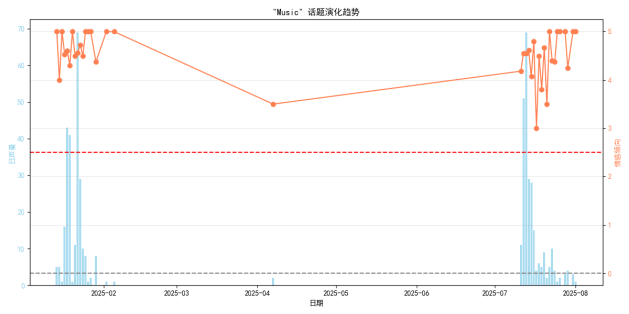
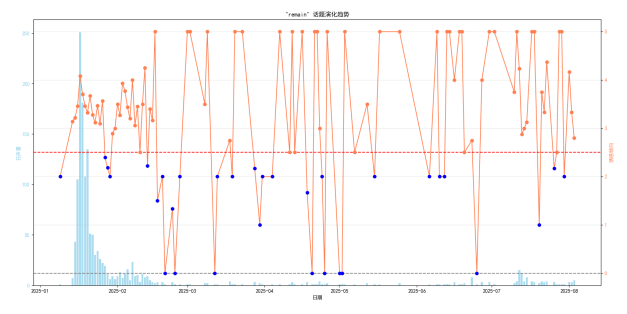
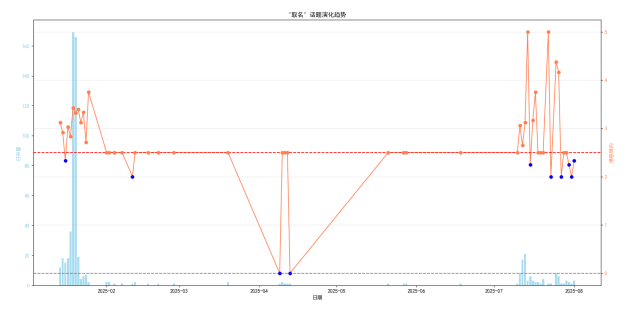
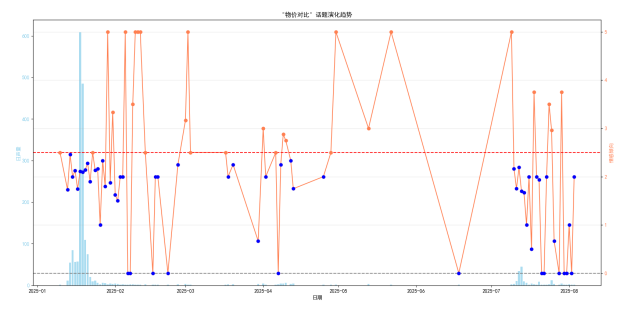


|  |  |  |  |
| --- | --- | --- | --- |
| 笔记topic | 分类 | 总声量 | 情感平均得分 |
| cattax | 优质话题(推荐) | 1859 | 4.305810 |
| 物价对比 | 敏感话题(减少讨论) | 1841 | 2.103476 |
| friend | 优质话题(推荐) | 1838 | 4.026387 |
| remain | 高潜力话题(关注) | 1416 | 3.516243 |
| daily | 高潜力话题(关注) | 1196 | 3.492893 |
| 取名 | 高潜力话题(关注) | 591 | 3.184433 |
| Music | 高潜力话题(关注) | 516 | 4.511628 |
| learn | 高潜力话题(关注) | 43 | 2.732558 |
| communicate | 小众话题(降低关注) | 29 | 2.344828 |

得到以上数据，平台应对“cattax”“friend”话题进行推流，对“物价对比”话题进行限流，同时可将“communicate”话题的资源适当调配到“remain”“daily”等高潜力话题，促使其声量增大。

统计各话题的热度与情感更迭，在验证上图结论的同时可以匹配现实时间节点，观察现实时事对用户和话题的活跃度的影响。

代码详见：【Python代码-附件五：话题筛选.py】



大多数话题都是在2025年1月17号附近日声量突然暴涨，应该是由1月13日TikTok禁令引发的用户迁移潮以及1月19日TikTok全面封禁禁令最后期限引发的二次讨论。可以看到，几乎所有话题都在寒假和暑假期间声量暴涨，这说明**学生可能为与本次TikTok难民潮活动时间主流群体**。若想提升话题的声度，可以开展助推学生活跃的活动迎合主流群体喜好，在提升平台总流量的同时也能够提高平台黏性。

## 7.词云高频词分析

限定result数据集中的“笔记topic”列，对“评论内容”列进行分词，小写化，删去停用词，统计词频。以词频排名在前10%的词为基础制作词云图，观察各个话题中的常用词。本部分对话题头部常用词分析排名也恰好能验证前文已获取到话题的总情感倾向。

代码详见：【Python代码-附件六：分词.py】

①“cattax”话题评论中的常用词词云图：



评论中最常出现的“可爱”“害羞”“喜欢”“吧唧”皆是表达大量喜爱与赞美的词汇，与上文所得"cattax"话题评价一致。而所出现的“偷笑”“飞吻”与“doge”则是表情符号，其强化了评论营造轻松愉快氛围的能力。本词云图能较好呈现“TikTok难民”涌入初期，外国用户可能通过分享猫咪内容来融入社区，中国用户则通过围绕猫咪的赞美接纳外来用户的行为特点。萌宠分享作为低门槛，高共鸣的内容载体极适合用于两种不同文化进行融合的场景。

“cattax”话题作为“高声量+积极情感”的优质话题，其高频词的出现可以辅助平台学习如何运营维护话题良好氛围。平台可以在需推流话题评论回复中添加人工智能回复类似词组的搭配，便能在提高声量的同时促进话题向积极情感倾向的发展。

②“物价对比”话题评论中的常用词词云图：



评论中常出现的核心词组有“我们”“你们”“他们”，三种人称代词的频繁使用体现出中外用户交流的热烈程度之高。尤其“美国”一词的出现隐含了中美对比的情景。

“种菜”“不能”“违法”“犯法”“可以”等次生词组也高频出现，反映“物价对比”话题下对美国“私人种菜受限” 这一争议性话题的激烈讨论。外国用户羡慕中国的种菜自由，而中国用户则庆幸本国政策，对比中体现出两国人民对于自身环境的反思与认同，这也反映出“物价对比”的本质其实是生活自由度的对比。

最后，“捂脸”“石化”“失望”“汗颜”等词也能体现出用户对于美国相关政策的震惊与不满，反映出用户对美国相关政策批判。负面情绪居多，与上文所得"物价对比"话题评价一致。

③“friend”话题评论中的常用词词云图：



评论中常出现“want”“friend”“外国”“foreign”等词组，直接体现出涌入的外国用户与本土中国用户强烈的交友需求。从“want”到“make”再到“talk”，用户由交友意愿发展到具体行动，直至“Wechat（微信）”被提及，用户希望从评论区的初步互动，延伸到更私密的社交平台。有趣的是，图中“学习”翻译”“中文”“英文”的频繁出现说明交友需求中同时伴随着语言学习需求。

当渴望文化交流与友好接触相碰撞时，多元的氛围为个体敞开心扉提供了充分保障，从而促使用户积极向外交流与提升平台黏性。

④“remain”话题评论中的常用词词云图：



“tiktok”在该话题下作为核心词限定了讨论背景，大量“tiktok”因政策原因来到小红书，在政策有新的变动时是“left”“back”还是“remain”引发了众多用户的讨论。使用最多的“share”一词展现了多数外国用户对小红书平台的定位——日常内容分享平台。

⑤“daily”话题评论中的常用词词云图：



评论中“我们”“他们”“中国”“美国”频繁使用可以看出daily话题下用户的交互方式为对话式。而“help”“thank”“how”“what” 等词，反映出双方日常交流中对帮助、知识的双向探索。

⑥“取名”话题评论中的常用词词云图：



“正宗”“地道”“本土化” 等词，体现外国用户在取中文名字时，渴望融入中国文化、让名字更符合中文表达习惯的诉求。而“捂脸”“黑化”“石化”“受不了”“可以”表示出取名过程中用户的情绪表现，剧烈的情绪波动提升“取名”活动的活跃度。

⑦“Music”话题评论中的常用词词云图：



Music话题下可以十分显著地体现出本次赛博难民文化碰撞交流的特点。TikTok难民涌入时，将“美国”“乡村音乐”“牛仔”等这类带有文化符号的西方音乐元素内容带到小红书，引发用户对“乡村音乐”这类西方曲风的关注。“原唱”“original”则指向对“原创音乐”“正版演绎”的重视，反映音乐创作与版权相关的交流。

同时“内娱”一词的出现说明中国用户也在对国内的音乐环境进行思考。“love”“beautiful”“nice”“great”等英文赞美词，与“喜欢”“好听”“赞”等中文表达并存，体现中外用户对音乐超越文化的共同喜爱。

⑧“learn”话题评论中的常用词词云图：



评论中“中文”“法国”“英语”等高频词反映出文化交流中语言学习成为了核心载体，在对彼此语言的相互学习用户实现了跨文化互动。“容易”“真的”“不是”“shock”等词反映出学习中的真实感受，降低了大家集体学习的枯燥感，加深了几个不同文化群体之间的共鸣。

⑨“communicate”话题评论中的常用词词云图：



“撒谎”“吹牛”“浮夸”“exaggerating”“lying” 等词高频出现体现跨文化交流中，用户对夸张说法的聚焦。这可以体现出用户对真实、客观交流的在意，也能折射出此刻双方进行语言磨合的行为。

## 8.中外用户评论特点分析

网络评价作为用户对话题看法的载体，能反应用户的性格特征。根据评价者的IP地址以及评价的愉悦度，唤醒度，支配度定量分析中外用户网络发言的行为特点。基于爬取的数据，可以直接计算评论愉悦度，唤醒度，支配度的三大统计数据。

代码详见【Python代码-附件七：评论特点分析.py】

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | valence | | | arousal | | | dominance | | |
| 用户类型 | 计数 | 平均值 | 标准差 | 计数 | 平均值 | 标准差 | 计数 | 平均值 | 标准差 |
| 中国用户 | 2670 | 3.257678 | 1.294817 | 2670 | 3.083146 | 0.71371 | 2670 | 2.704494 | 0.627772 |
| 外国用户 | 4698 | 3.242018 | 1.142947 | 4697 | 2.928678 | 0.749376 | 4698 | 2.181779 | 0.539826 |
| 未知 | 1821 | 2.859418 | 1.234132 | 1821 | 2.398682 | 1.116942 | 1821 | 2.312466 | 0.666798 |

愉悦度（valence）数据：

中国用户和外国用户的平均愉悦度较为接近，中国用户为3.258，外国用户为3.242，这表明中外用户在小红书平台上发言时整体的愉悦程度差异不大。但从标准差来看，中国用户的愉悦度标准差1.295略高于外国用户的1.143，说明中国用户在评论时愉悦度上的波动稍大，发言时的愉悦程度更为多样化，因个体差异产生的变化更多。但总体来看，先前未专门针对外国用户适配的小红书平台却能让短时间大量涌入的外国游客感到与中国用户同样的愉悦感，不难发现小红书运营团队的快速响应强劲实力，以及中国用户的热心好客的特点，最终使得外国用户宾至如归，快速融入到了社区中。

支配度（dominance）数据：

中国用户的平均支配度为2.704，显著高于外国用户的2.182，说明中国用户在网络发言中相对更倾向于表达具有一定支配性的观点。对其进行分析，这有可能是因为外国用户涌入小红书，多是发表询问帮助、遵从建议等较为被动的言语。而中国用户作为小红书平台的本地用户，会更容易以一种“东道主”的视角发表一些主动的，积极介入的观点，能反映出中国用户在小红书上热情好客的评论行为特点。但若是要归纳外国网友与中国网友在全球互联网平台上的评论特点，则还需对双方在其他网络平台的评价发言进行定量研究。

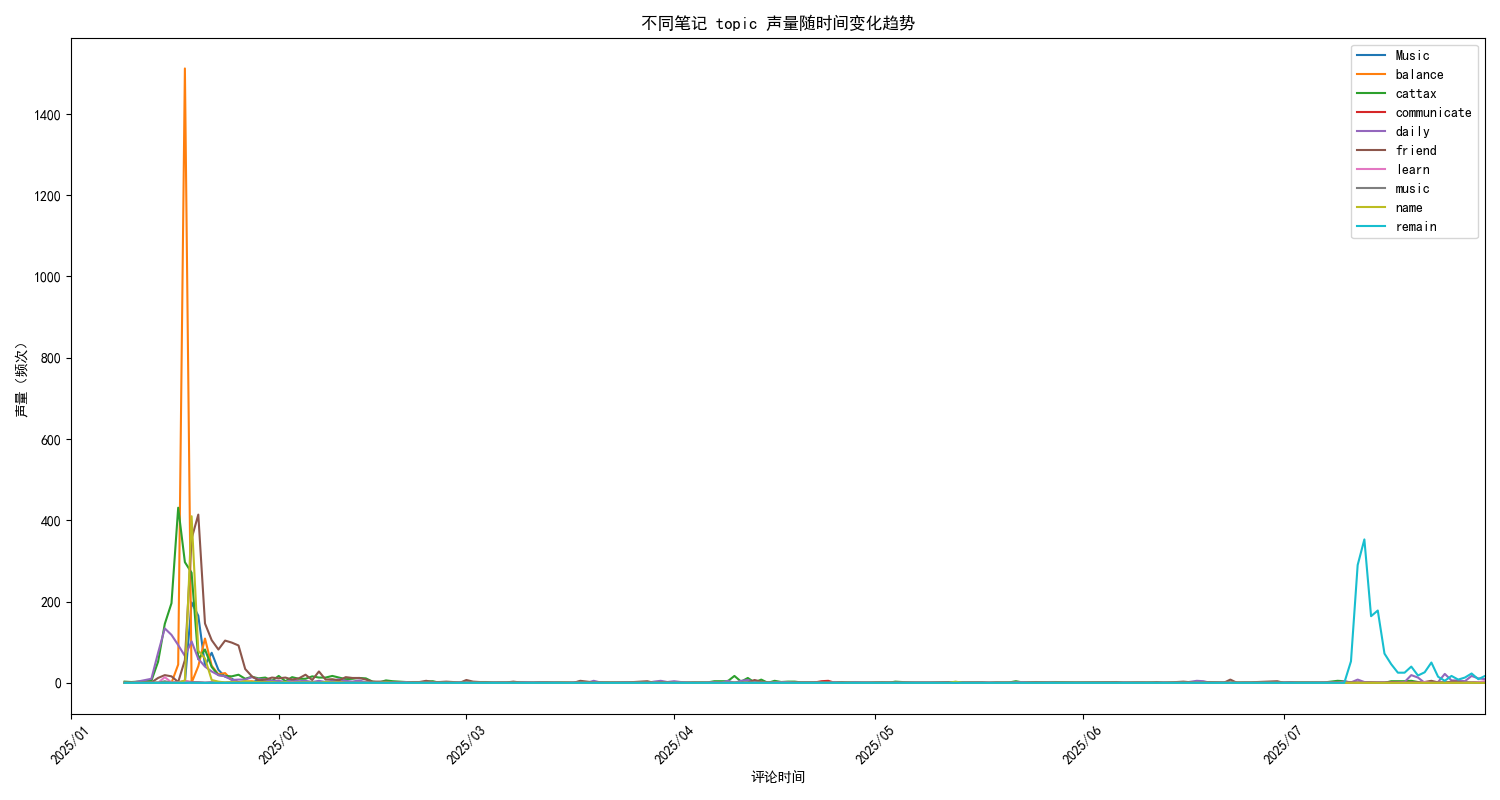
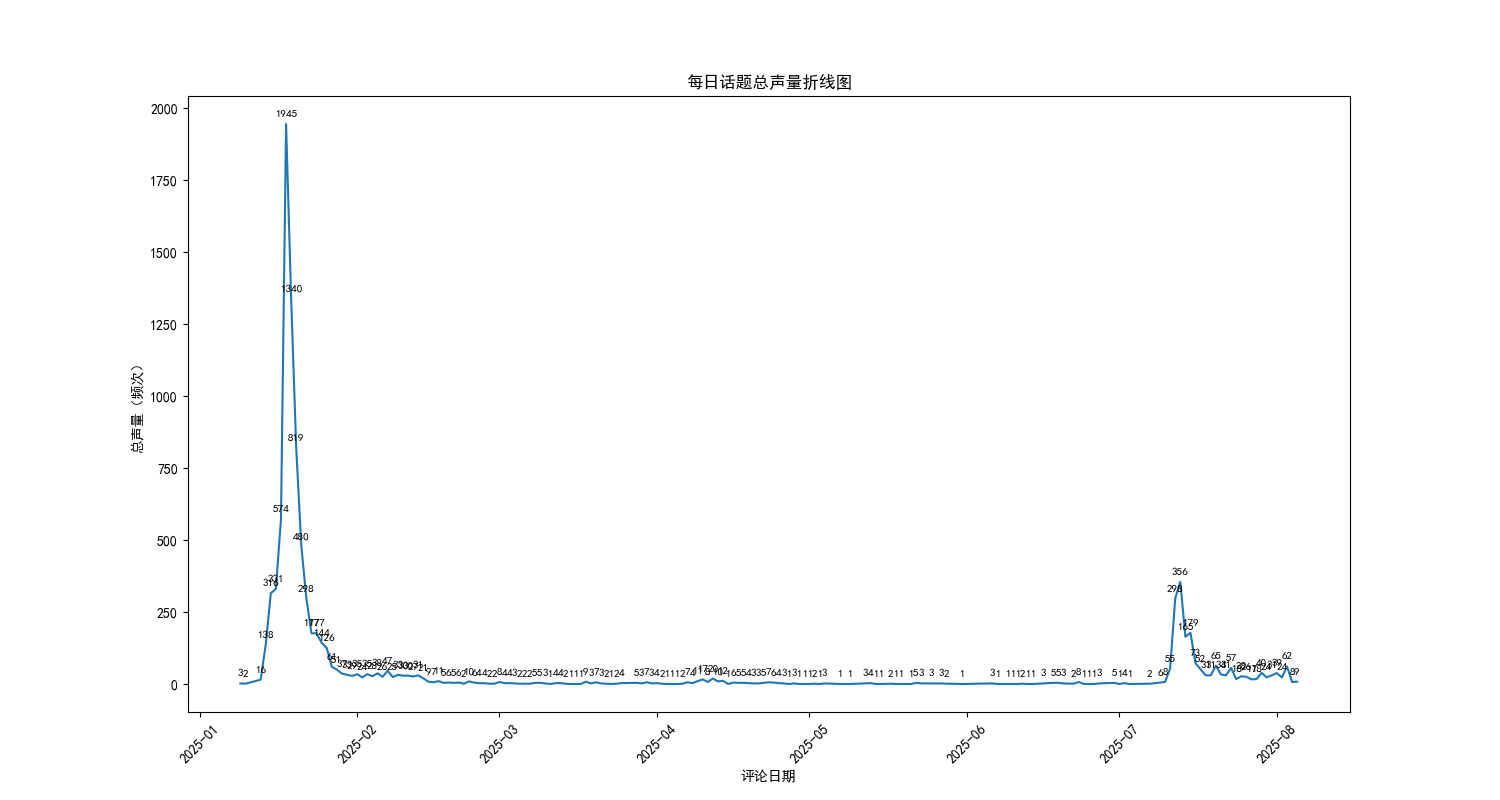
唤醒度（arousal）数据

中国用户的唤醒度均值为3.083，外国用户为2.929，中国用户略高于外国用户，这可能表示中国用户在网络发言时整体的兴奋或激动程度稍高一些。同时中国用户的唤醒度标准差1.140略高于外国用户的1.012，说明中国用户在唤醒度上的波动更大，发言时的兴奋状态更为多样化。

## 9.话题声量变化分析

按评论时间和笔记话题分组，统计每个组合下评论ID的数量（即声量），便能得到不同话题在不同时间的声量数据。进行可视化后，便可以进行分析：

代码详见【Python代码-附件八：话题声量分析.py】



可以看出Tiktok难民潮在小红书上大体引起了两拨热潮，第一次在2025年1月，第二次是在2025年7月。尤其是friend话题在1月份剧烈上升，可以是因为TikTok禁令迫使用户迁移，小红书短期内新增大量外国用户，大量新用户的“社交破冰”需求集中爆发。而小红书1月19日上线“中英双语翻译”功能，解决了跨语言交流障碍，让“friend” 相关的跨国互动更顺畅，间接推动话题持续发酵。

而“remain”话题讨论1月未爆发但7月突然爆发十分令人寻味。分析其背后原因，可能与小红书实施“人脸+身份证双认证”新规有关，这一规定直接冲击了赛博难民群体。新规要求所有发布内容的账号必须完成人脸+身份证认证，否则永久封号。但大量外国用户无中国身份证，面临“无法认证即失去账号”的困境。因此用户在评论区密集讨论，直接推高“remain”话题声量。