（可能会是）应用语言学及英语教学的期末论文

医患关系语料库 关键点阐述

语料来源是微博，设置好时间区间，微博里的关键词是“医患关系”

历时性分析方法：

1. 将 发布时间 转换为 datetime 格式并提取年月
2. 聚合：每月评论数 + 正负面比例
3. 画图：时间轴上情绪趋势线

情感分析的单位为整条评论（评论内容里的所有内容）

情感得分标准和解读：



关键词检索方案：合成一个语料库进行统计

目前采用：思路1：全评论词频统计（推荐用于全局可视化）

操作：将所有评论内容拼接、分词、统计词频

适合生成词云、top N 高频词、看整体语料场域

缺点：可能混入与主题无关的高频词（如“今天”“真的”）

解决办法：

加入你之前打的“是否保留”筛选 → 只统计保留评论词频

可再细化：只统计“直接相关=True”的词频（更精准）

可能会采取的思路2:

思路2：关键词窗口/句子内词频（推荐用于深度分析）

操作：

找到所有含“关键词”的评论

把“关键词所在的句子”或“关键词前后±5词”切出

对这些局部进行分词和词频统计

目的：

分析“医患关系”这个词常常搭配哪些形容词、动词

看“医生”附近常出现“温柔”还是“吼我”

方法建议：

使用 jieba 进行分词

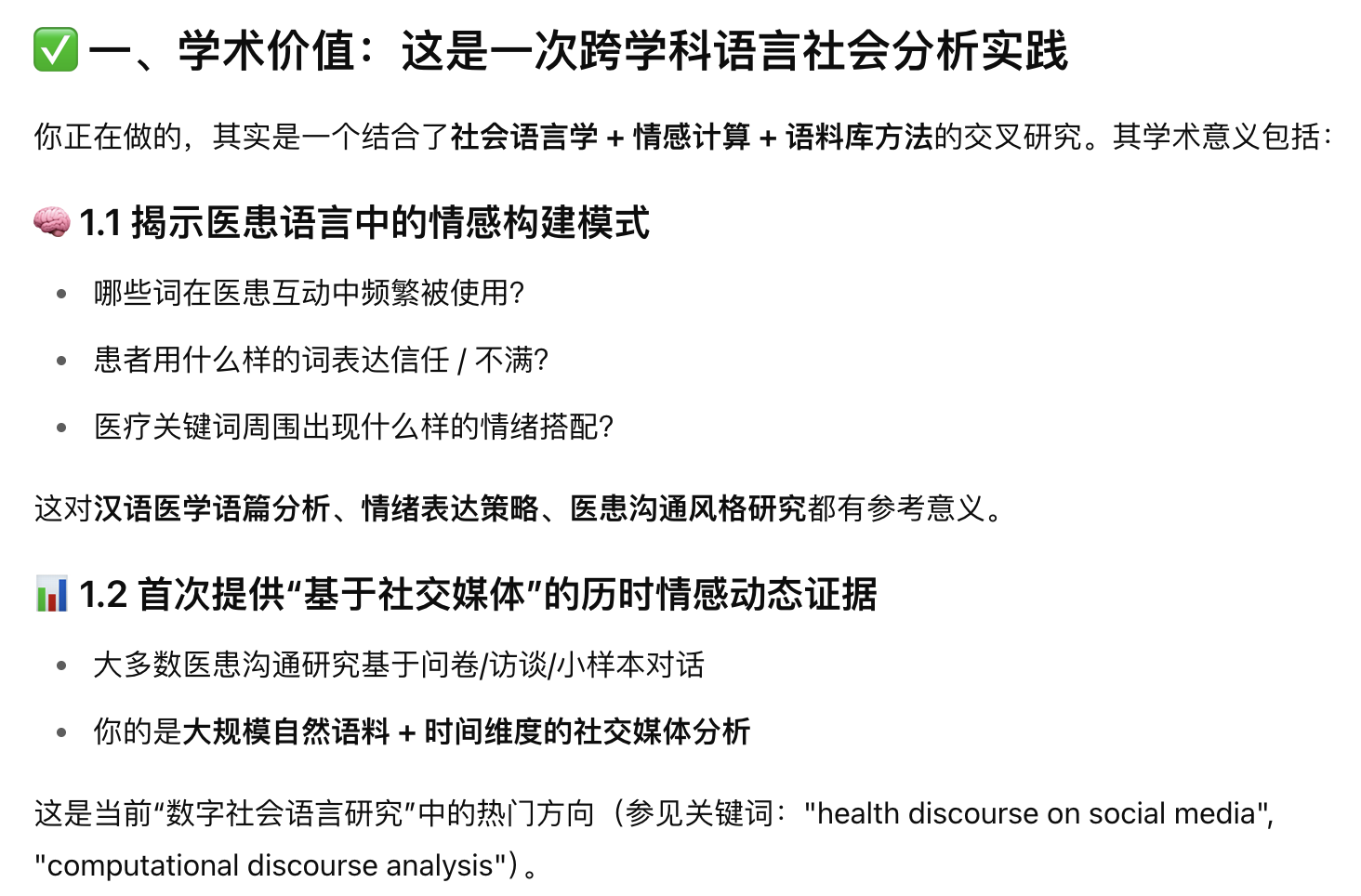
用 re 提取句子或窗口

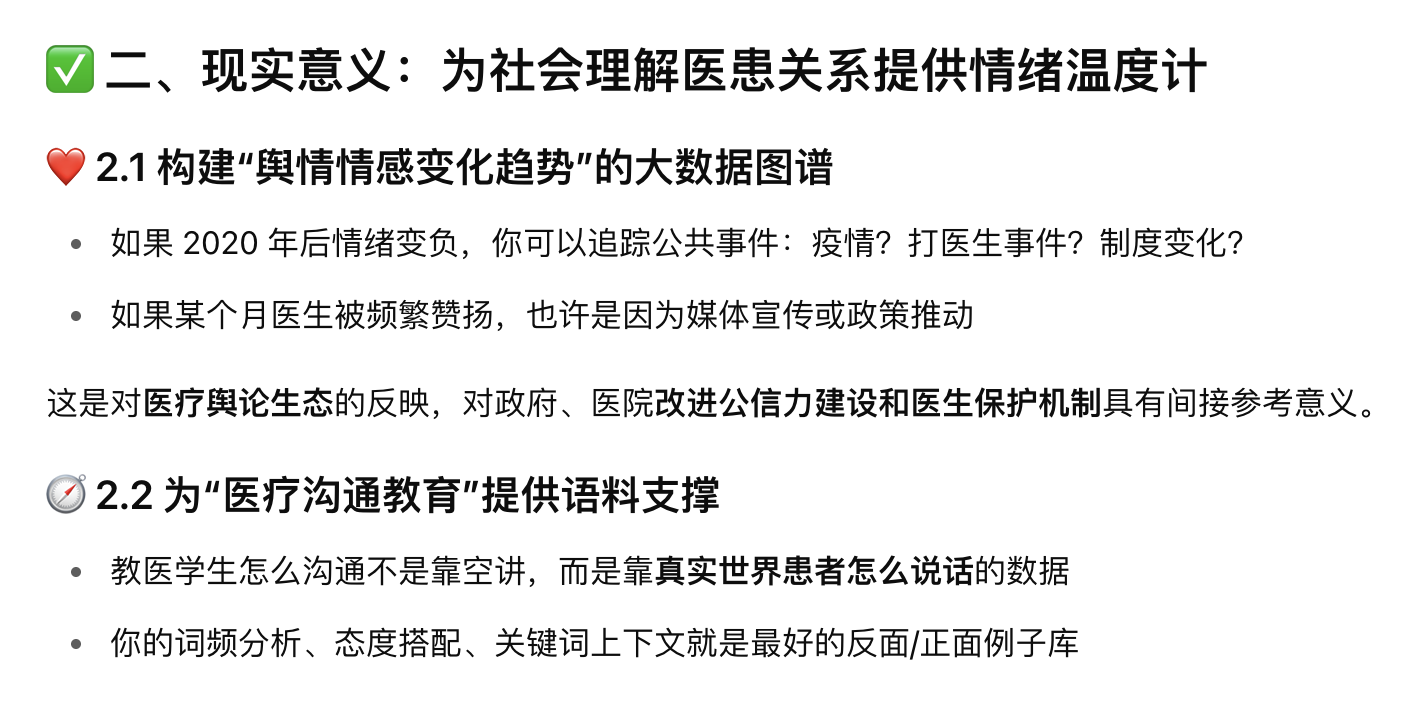
研究意义

“我研究出人们在微博中关于医患关系的情感变化，然后呢？”——这个问题的本质是：

“我的研究成果能为社会、语言研究、医疗制度带来什么价值？”

所以我们来帮你从学术价值 + 实际意义 + 研究定位三个角度梳理你的研究意义。







然后用r进行情感分析得分的可视化

时间变化可视化说明：

效果说明

蓝线：每日情感得分趋势

橙点：当日评论数（越多越大）

虚线红线：LOESS 平滑趋势

背景使用 theme\_minimal() 提高可读性

搭配贡献频率的那个可视化：

·蓝色节点大小 = 在网中的重要程度  
• 橙色边越粗 = 共现越频繁  
• 自动排斥文字重叠，整体清爽高级

可以根据bubbles里的结果来看具体做了什么，这只是个初步的稿子，语料肯定得重新搜集，比如设置了时间范围之后还需要让帖子根据时间顺序排序，不要漏掉很明显的时间段防止数据出现无法解释的现象。

GPT里面的名称为医患关系语料库构建。

以上内容都已经有初步的代码，之后主要是更换语料库，然后根据实际需要对代码进行微调。

其他没啥想说了，注意文件名称吧，因为我改了不少。

目前已经有一个pdf，里面内容蛮全面，不过有几点需要更改：

LDA可以作为后续补充研究方法看看需不需要加上

预训练 Transformer 分类器暂不考虑使用

“未来研究可以结合媒体报道数据做多元回归验证”

用TF\_IDF分析词频和词云

可以先按感情分组，然后分别做高频词统计 & 共现分析。丰富讨论：正面语境中的关注点 vs 负面语境中的指责点。

### ****针对您的RQ2，如何使用TF-IDF和LDA？****

**您的目标是追踪议题随时间的变化，LDA是完成这个任务最理想的工具。** 正确的操作流程应该是这样的（两步法）：

**第一步：在【全部语料】上训练一个统一的LDA模型，以发现稳定的主题结构。**

1. **数据：** 使用您所有的微博数据（2019-2025）。
2. **目的：** 从这个长时段、大规模的语料中，让模型学习并总结出关于“医患关系”讨论中，反复出现的、具有稳定语义结构的核心议题。例如，模型可能会发现15个核心议题，您通过查看每个议题下的高频词，可以给它们命名，如Topic\_1: 医生执业安全, Topic\_2: 医疗费用与医保, Topic\_3: 患者就医体验, Topic\_4: 医德与感谢……
3. **为什么要在全部语料上训练？** 因为这样可以保证您获得一个统一的、贯穿所有时间段的议题分析框架。如果您对每一年都单独训练一个LDA模型，那么2020年的“主题1”和2021年的“主题1”可能指的是完全不同的东西，无法进行历时比较。

**第二步：分析这些已发现的议题在【不同时间段】的流行度变化。**

1. **数据切分：** 将您的微博语料按时间切片（例如，按年或按季度）。
2. **计算主题分布：** 对于每个时间片（比如“2020年”的所有微博），计算这些微博在第一步发现的15个核心议题上的平均分布。也就是说，计算出在2020年，人们的讨论中有百分之多少是关于Topic\_1: 医生执业安全的，有多少是关于Topic\_2: 医疗费用与医保的，以此类推。
3. **可视化：** 将计算出的每个议题在不同时间段的流行度（占比或平均权重）绘制成折线图或堆叠面积图。这样您就可以非常直观地看到，哪个议题在哪个时间段是讨论的热点。

**TF-IDF可以作为补充：**

您也可以使用TF-IDF来作为一种补充验证。例如，当您发现2020年“医生执业安全”这个议题的流行度激增时，您可以再计算一下2020年语料相对于其他年份语料的TF-IDF高分词。您很可能会发现，“陶勇”、“暴力”、“砍伤”等词的TF-IDF值在2020年特别高，这就能与LDA的结果相互印证。

### ****研究问题 (Research Questions)****

本研究旨在通过对2019年1月至2025年5月期间的微博语料进行分析，回答以下三个核心问题：

**RQ1: “医患关系”公众话语的情感基调是什么？ (What is the foundational sentiment polarity of the discourse?)**

* **中文表述：** 在2019年至2025年的微博平台中，关于“医患关系”的公众话语其整体情感极性（正面、负面、中性）分布呈现何种特征？
* **English Formulation:** What is the overall distribution of sentiment polarities (positive, negative, neutral) within the public discourse on 'doctor-patient relationships' on Weibo from 2019 to 2025?

**说明：** 这个问题是您研究的起点和宏观背景，用描述性统计回答，为后续的深入分析设定基调。

**RQ2: 公众对“医患关系”的情感与关注焦点如何随时间演变？ (How do public sentiment and focus evolve over time?)**

这个问题可以进一步分解为两个相互关联的子问题，共同描绘出动态变化的全貌：

**RQ2a (情感演变):**

* + **中文表述：** 在2019年至2025年期间，微博上关于“医患关系”的公众情感倾向如何随时间动态演变？哪些关键性的社会事件（如公共卫生事件、恶性伤医事件等）与情感的显著波动存在关联？
  + **English Formulation:** How does the public sentiment towards 'doctor-patient relationships' on Weibo dynamically evolve over the period of 2019–2025, and what correlations exist between significant sentiment fluctuations and key real-world events (e.g., public health crises, major violent incidents against doctors)?

**RQ2b (议题演变):**

* + **中文表述：** 伴随着情感的演变，公众在不同年份（2019-2025）讨论“医患关系”时，其核心议题（Topics）发生了怎样的变迁？哪些议题的关注度呈现出上升、下降或周期性变化的趋势？
  + **English Formulation:** Concurrent with the sentiment evolution, what are the shifts in the core topics of public discourse surrounding 'doctor-patient relationships' across different years from 2019 to 2025? Which topics demonstrate a rising, declining, or cyclical trend in public attention?

**说明：** RQ2是您研究的核心，它不仅追踪了情感的“波动”，还通过议题的变迁去探索波动背后的“原因”，完整地讲述了“演变”的故事。

**RQ3: 不同情感表达背后隐藏着哪些议题和语言特征？ (What topics and features characterize different sentiments?)**

* **中文表述：** 在整体语料中，表达不同情感极性（正面、负面、中性）的微博文本，其在核心议题（Topics）和特征词汇（Lexical Features）上存在哪些显著差异？
* **English Formulation:** Across the entire corpus, what are the significant differences in core topics and characteristic lexical features among Weibo posts that express positive, negative, and neutral sentiments towards 'doctor-patient relationships'?

**说明：** 这个问题深入到话语内部，通过对比分析，揭示出构成不同情感的“内容成分”，回答了“人们在表达正面/负面情绪时，到底在谈论什么以及如何谈论”的问题。

**如何在3分钟报告中呈现这些问题？**

您可以这样说：

“大家好，我的研究聚焦于医患关系的社交媒体话语。具体而言，我希望通过对2019至2025年的海量微博数据进行分析，回答以下三个层层递进的问题：

**第一，** 整体来看，公众对医患关系的讨论究竟抱持着怎样的情感基调？ **第二，** 在这六年多的时间里，公众的情感和他们关注的议题焦点是如何动态演变的？尤其是，像疫情、伤医事件这样的社会事件，是如何影响这种演变的？ **第三，** 当人们表达正面或负面情绪时，他们具体在讨论哪些议题，又使用了哪些特征词汇？

通过回答这些问题，本研究旨在为理解当代中国的医患关系提供一个数据驱动的、历时性的舆情洞察。”

### ****如何构建“语料库概况与描述性统计”****

您可以将这个部分分为两个方面：**总体概况**和**历时分布**。

**1. 语料库总体概况 (Overall Corpus Profile)**

这部分回答：“我的数据集有多大？总体互动情况如何？”

* **帖子总量：** 首先报告您收集到的微博总数。
  + 例如：“本研究共收集2019年1月至2025年5月期间，包含关键词‘医患关系’的微博共计XX万条。”
* **总体互动量：** 汇报整个语料库的总转发数、总评论数、总点赞数。
  + 例如：“整体来看，这批语料共获得了XX次转发、XX条评论和XX次点赞，反映了公众对该议题的高度关注。”
* **总体情感分布（您原先的RQ1）：** 在这里呈现您对全部数据进行情感分析后的宏观结果。
  + 例如：“在情感倾向方面，整体语料呈现出...的分布特征，其中负面情感占比XX%，正面情感占比XX%，中性情感占比XX%。”（用饼图或柱状图展示）

**2. 核心指标的历时分布 (Diachronic Distribution of Key Metrics)**

这部分回答：“在长达六年半的时间里，关于‘医患关系’的讨论热度和参与度是如何变化的？” 这正是您刚才提到的“汇报每年或者每月的帖子总数，转赞评数量”。

**讨论热度（帖子数量）的历时变化：**

* + **操作：** 以**月**为单位，统计每个月的微博发帖数量。
  + **可视化：** 绘制一张折线图或柱状图，X轴为时间（2019.01 - 2025.05），Y轴为帖子数量。
  + **意义：** 这张图能非常直观地展示出“医患关系”这个话题在哪些时期是公共舆论的焦点，讨论热度何时出现高峰和低谷。**这张图为您后续的情感波动分析提供了最重要的背景板。**

**公众参与度（转赞评数量）的历时变化：**

* + **操作：** 以**月**为单位，统计每个月的总互动量（转发数 + 评论数 + 点赞数之和）。
  + **可视化：** 同样绘制一张折线图或柱状图。
  + **意义：** 有时候，帖子的数量不一定多，但互动量可能非常高，这代表某些帖子引发了极强的公众共鸣或争议。这张图可以从“参与度”的维度补充“热度”分析。

### ****单页PPT设计方案：“研究背景与文献回顾”****

这页PPT的核心目标是：**在60-75秒内，清晰地告诉评委您为什么要做这个研究，以及您的研究与前人研究的关系。**

**PPT页面布局建议：**

* **标题：** 研究背景与文献回顾
* **内容结构：** 分为上下两个部分，“研究动机”和“文献基础”。

#### ****第一部分：研究动机 (Motivation)****

* **现实挑战：** 医患关系是持续的社会焦点，相关纠纷与事件引发强烈公众情绪。
* **数据机遇：** 微博已成为该议题的核心舆论场，为研究提供了海量、动态的自然语言数据。
* **研究缺口：** 现有研究多为静态、小样本分析，**缺乏对长时段、大规模舆论生态的动态追踪。**

#### ****第二部分：文献基础 (Literature Foundation)****

* **情感基调：** 文献表明公众情绪以**负面为主**，尤其在冲突事件中 (Chen et al., 2024; Lu et al., 2024; Zheng et al., 2025)。
* **情境差异：** 但在非冲突领域（如儿科日常）或事件不同阶段，情感亦呈现**复杂性与波动性** (Song et al., 2022; Zheng et al., 2025)。
* **研究视角：** 既有研究多聚焦于**单一事件**的话语分析 (Tu & Liu, 2015; Zheng et al., 2025) 或**特定领域**的舆情监测 (Song et al., 2022)。

#### ****本页结论/过渡句 (在PPT底部或作为讲解的收尾)****

**因此，对“医患关系”这一宏观议题进行长周期、跨事件的历时性情感与议题演变分析，尚有明确的研究空间。**

### ****为开题报告重新设计的PPT结构与逻辑（3分钟版）****

**总时长控制：** 3分钟，大约4-5张核心内容PPT，每张PPT讲解不超过45秒。

**Slide 1: 封面页 (Title Slide)**

* **标题：** 基于微博语料的医患关系情感表达和演变研究
* **副标题（可选）：** 开题报告
* **基本信息：** 报告人、导师、日期

**Slide 2: 研究背景与问题提出 (Background & Problem Statement)**

* **(约40秒)**
* **核心内容：**
  1. **现实背景：** 简要说明“医患关系”是长期存在的社会焦点议题，具有重要的现实意义。
  2. **数据背景：** 指出微博等社交媒体已成为公众讨论该议题的主要舆论场，产生了海量的、动态的自然语言数据。
  3. **研究缺口（Gap）：** 点明现有相关研究多依赖问卷、访谈等传统方法，样本小、时效性弱，**缺乏对长时段、大规模、自然状态下公众情绪和议题焦点的动态追踪。**
* **本页目标：** 用1-2句话清晰地告诉老师们，您发现了一个有待填补的研究空白。引出您的研究问题。

**Slide 3: 研究问题 (Research Questions)**

* **(约45秒)**
* **核心内容：**
  + 将我们之前提炼的三个核心研究问题（RQs）清晰地列出。
  + **RQ1: 整体画像：** 在2019-2025年间，微博上的“医患关系”话语呈现出怎样的总体讨论规模、参与度及情感倾向特征？
  + **RQ2: 历时演变：** 在此期间，公众的情感倾向和核心议题焦点如何随时间动态演变？这些演变与关键社会事件（如疫情、伤医事件）是否存在关联？
  + **RQ3: 话语特征：** 表达不同情感（正面/负面）的话语在议题和典型词汇上存在哪些显著差异？
* **本页目标：** 明确、具体地展示您要研究的**三个核心问题**。这三问层层递进，体现了您思考的深度。

**Slide 4: 研究设计与技术路径 (Research Design & Methodology)**

* **(约50秒)**
* **核心内容：**
  + 用一张流程图或简洁的列表，展示您将**如何**回答上述三个问题。
  + **1. 数据来源 (Data Source):**
    - 平台：新浪微博
    - 关键词：“医患关系”
    - 时间跨度：2019年1月 - 2025年5月
    - **（强调一句“数据采集工作已初步完成”，这会大大增加可行性分）**
  + **2. 研究方法 (Methods):**
    - 语料库语言学、情感分析、LDA主题模型。
  + **3. 分析路径 (Analytical Path):**
    - **第一步：描述性统计** - 分析语料的总体规模、互动量和情感分布，回答RQ1。
    - **第二步：历时性分析** - 绘制月度情感波动曲线和年度议题演变图，回答RQ2。
    - **第三步：对比性分析** - 对比正面/负面情感语料库的关键词和主题，回答RQ3。
* **本页目标：** 向老师们证明，您有一个清晰、严谨、可操作的研究方案。

**Slide 5: 预期工作与进度安排 (Work Plan & Timeline)**

* **(约30秒)**
* **核心内容：**
  + 这页是展示您研究**可行性**的直接证据。
  + 用一个简单的甘特图或时间轴，列出后续的主要工作阶段和预计完成时间。
    - **第一阶段：数据深度清洗与预处理** (预计X月-X月)
    - **第二阶段：模型训练与数据分析** (预计X月-X月)
    - **第三阶段：结果整理与论文撰写** (预计X月-X月)
    - **第四阶段：论文修改与完善** (预计X月-X月)
* **本页目标：** 表明您对后续工作有清晰的规划，确保能够按时完成。

**Slide 6: 结束页 (Thank You Slide)**

* **汇报完毕**
* **请各位老师批评指正**

**总结一下这个新结构：**

这个结构完全**聚焦于“计划”**，逻辑线非常清晰：**为什么要做（背景） -> 要做什么（问题） -> 要怎么做（方法） -> 能否做完（进度）**。

开始生成语料库描述性统计报告...

================================================================================

成功加载文件，共 93548 条预处理后的微博。

--- 1. 语料库总体概览 ---

微博总数: 93,548 条

时间跨度: 2019-01-01 到 2025-05-31

总点赞数: 1,136,197,233.0

总转发数: 790,079.0

总评论数: 846,972.0

总互动量: 1,137,834,284.0

平均每条微博互动量: 12163.11

--- 2. 讨论热度与参与度的历时分布 ---

各年度发帖数量:

年份

2019.0 15389

2020.0 26029

2021.0 13377

2022.0 13745

2023.0 11306

2024.0 10967

2025.0 2621

dtype: int64

各年度总互动量:

年份

2019.0 73650343.0

2020.0 735096877.0

2021.0 73531444.0

2022.0 201602805.0

2023.0 28648402.0

2024.0 16462651.0

2025.0 8841762.0

Name: 总互动量, dtype: float64

/Users/fafaya/Desktop/医患关系研究/描述统计.py:74: FutureWarning: 'M' is deprecated and will be removed in a future version, please use 'ME' instead.

posts\_per\_month = df\_monthly.resample('M').size()

讨论热度最高峰出现在: 2020年03月, 该月共发帖 3,711 条。

--- 3. 内容特征分析 ---

平均每条微博（清洗后）长度: 192.08 字符

总互动量最高的10条微博:

- 互动量: 4,710,288.0 (2020-02-01) | 内容: 今天核酸检测结果阳性，尘埃落定，终于确诊了 2武汉·武汉市中心医院...

- 互动量: 4,710,288.0 (2020-02-09) | 内容: 看李医生以前的微博，觉得他是一名认真负责的医生，也是一个可爱的会发发牢骚爱吃鸡腿的小哥。希望李医生的...

- 互动量: 4,710,282.0 (2020-03-22) | 内容: 今天又有伤医事件了，不是说医患关系会好吗？虽然我知道不会，但是我还是侥幸相信。。。 有一次梦碎了。。...

- 互动量: 2,352,497.0 (2020-06-08) | 内容: 如果能重启2020希望每个人都能珍惜当下的生活，希望没有战争，希望所有的孩子都有学上，希望医患关系不...

- 互动量: 2,352,497.0 (2020-06-08) | 内容: #重启2020你会做什么# 很高兴受邀参加由@央视新闻 @新世相 共同发起的「重启2020」特别计划...

- 互动量: 2,168,404.0 (2021-01-10) | 内容: 因为看到同事被抬上救护车我被拼多多开除了 L王太虚wray的微博视频...

- 互动量: 2,168,404.0 (2021-01-11) | 内容: 预计劳资关系将替代医患关系成为2021年度的热点矛盾之一。...

- 互动量: 1,912,253.0 (2021-08-14) | 内容: #余生请多指教MV# 遇见你之前，余生很长，漫漫时光，无处安放；遇见你之后，余生很短，一闹一笑，并肩...

- 互动量: 1,912,253.0 (2021-08-14) | 内容: 这医患关系住个院也能有对象国家什么时候分我一个呜呜呜...

- 互动量: 1,912,235.0 (2022-01-05) | 内容: 有的时候真的不能怪医患关系紧张，就这样，不紧张才怪呢！...

