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