

**实 验 报 告**

|  |  |
| --- | --- |
| 中文标题 | 今日头条负面舆情智能监测系统实验报告 |
| 姓名 | 李昊伦 |
| 学号 | 2023211595 |
| 班级 | 2023211805 |
| 贡献度 | 100% |
| 指导教师 | 尚煜茗 |

**今日头条负面舆情智能监测系统实验报告**

**1 实验目的**

本实验旨在开发一套完整的“负面舆情自动识别系统”，通过对今日头条平台指定“用户”或“主题”下的内容进行数据采集、评论爬取、数据库存储，并结合关键词匹配、情感分析、组合评分等手段，自动发现涉及涉政有害、侮辱谩骂、色情暴力、事故灾难、聚集维权、娱乐八卦等六类负面舆情，实现多维度智能识别和可视化分析。

**2 实验设置**

2.1 实验环境

系统平台：Windows 11

编程语言：Python 3.11

依赖库：requests、Selenium、pyodbc、matplotlib、jieba、wordcloud 等

数据库环境：SQL Server 2022（本地连接）

浏览器驱动：ChromeDriver

2.2 实验数据

数据源：今日头条公开接口（模拟用户真实内容）

数据规模：每个主题最多爬取100条帖子，帖子的评论/回复量。每个用户的所有帖子及其评论

数据结构：帖子的id、标题、内容、用户ID；评论的id、post\_id、内容、parent\_id（结构化）

**3 实验方法**

3.1 方法思路

**（1）数据采集：**可选择“关键词搜索”或“指定用户”两种方式；使用多套API尝试爬取帖子与评论，确保兼容性；对于每篇帖子，爬取其全部评论并保留“父评论-子回复”层级关系。

**（2）数据库设计与存储：**设计posts和comments两张表；支持帖子去重、评论更新，提升数据稳定性；使用 pyodbc 实现 SQL Server 连接，所有数据结构化存入数据库。

**（3）负面检测算法：**设计每类负面舆情关键词列表（每类50+词）；结合关键词命中、组合匹配、情绪词评分、语气强度、垃圾词扣分等维度，智能判断文本风险；风险等级细分为：极高、高、中、低四类。

**（4）统计分析与可视化：**统计每类舆情文本数量及其分布；输出严重程度分布图、词云图、柱状图等；自动生成分析报告，评估各类负面影响力。

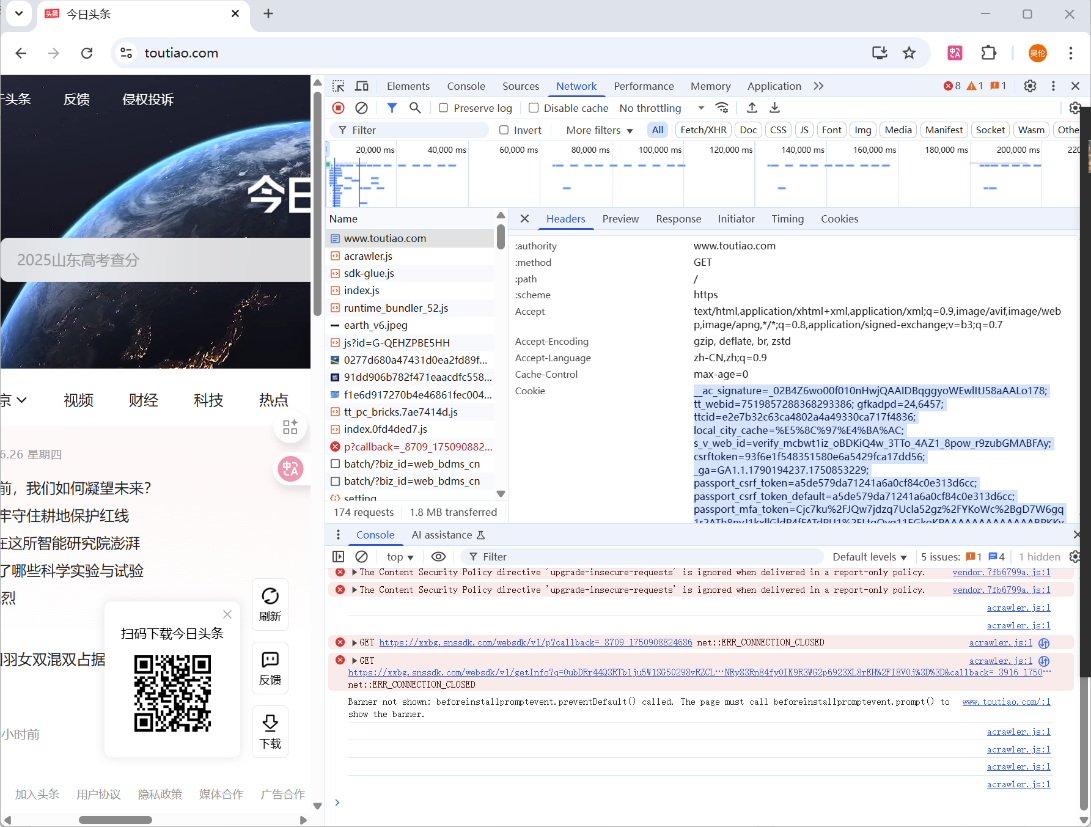
3.2 实验代码

（详见附件，共3000行左右）

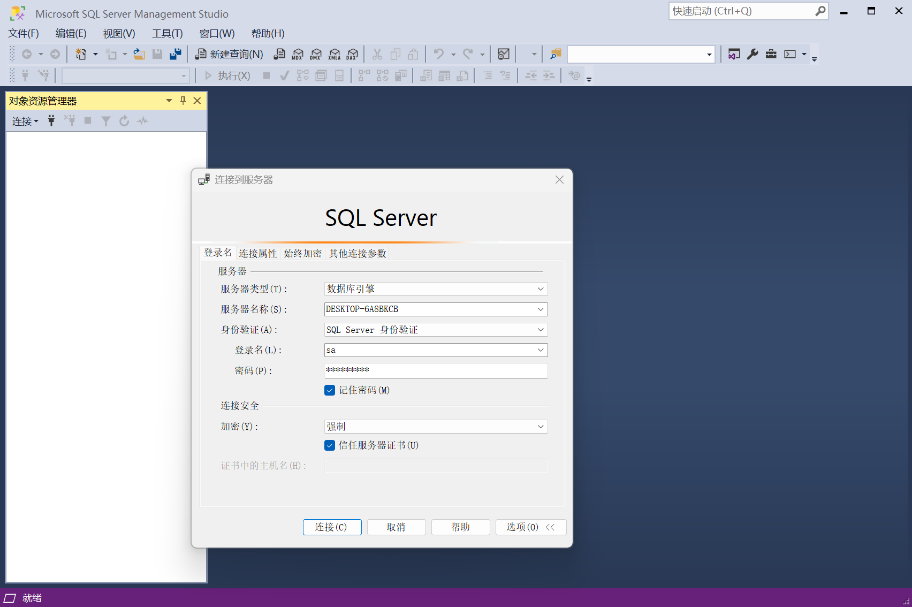
**4 实验过程**

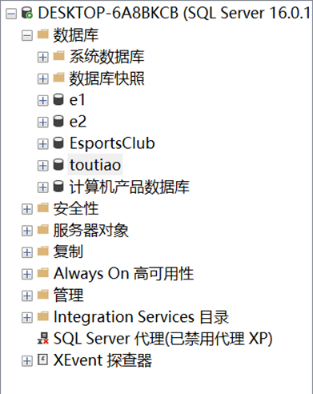
4.1 实验准备

（1）打开今日头条，按下F12进行cookie抓取，写入代码相应位置



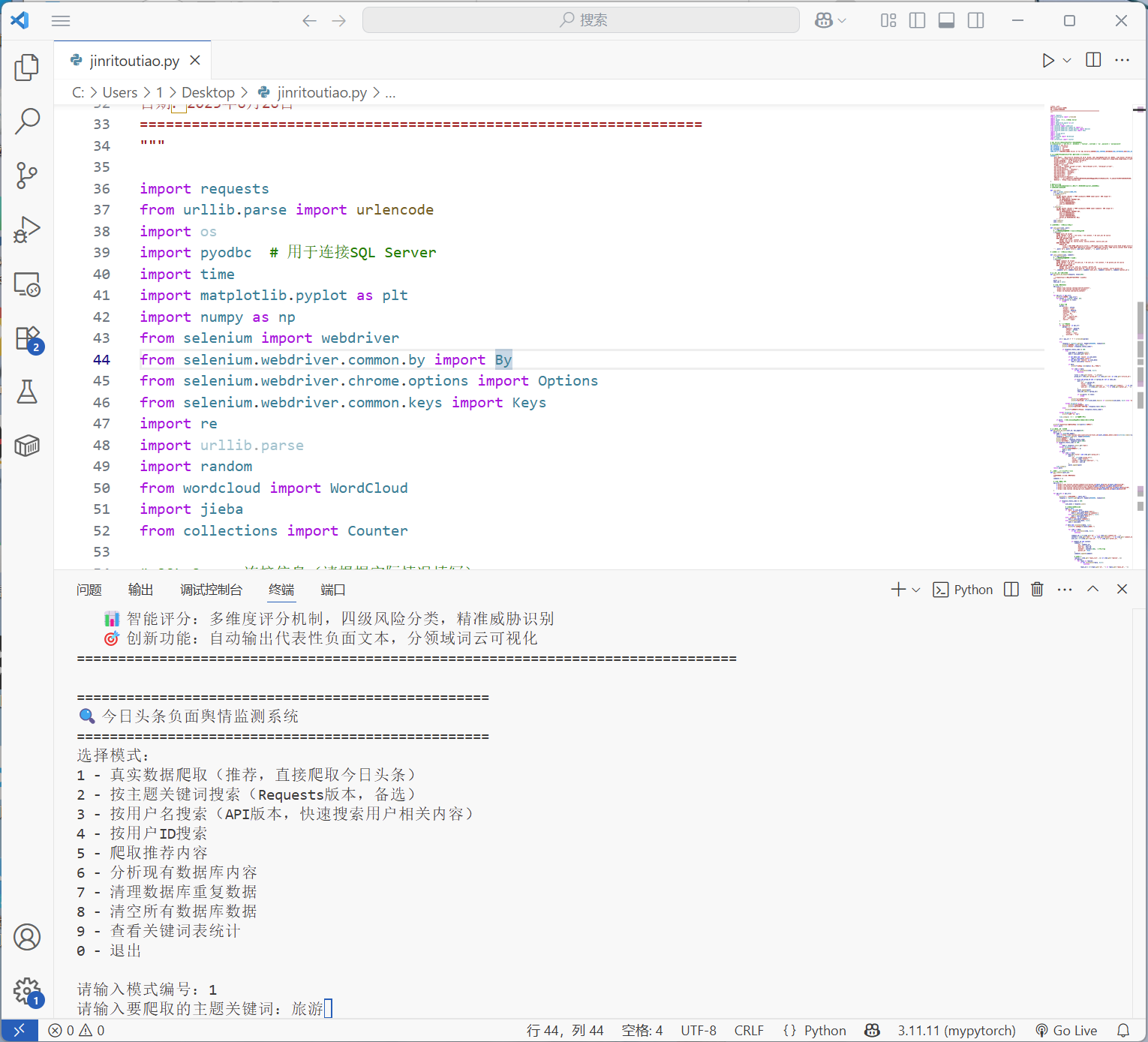
（2）打开SQL Server进行登录，并将相关账号密码表名写入代码相应位置



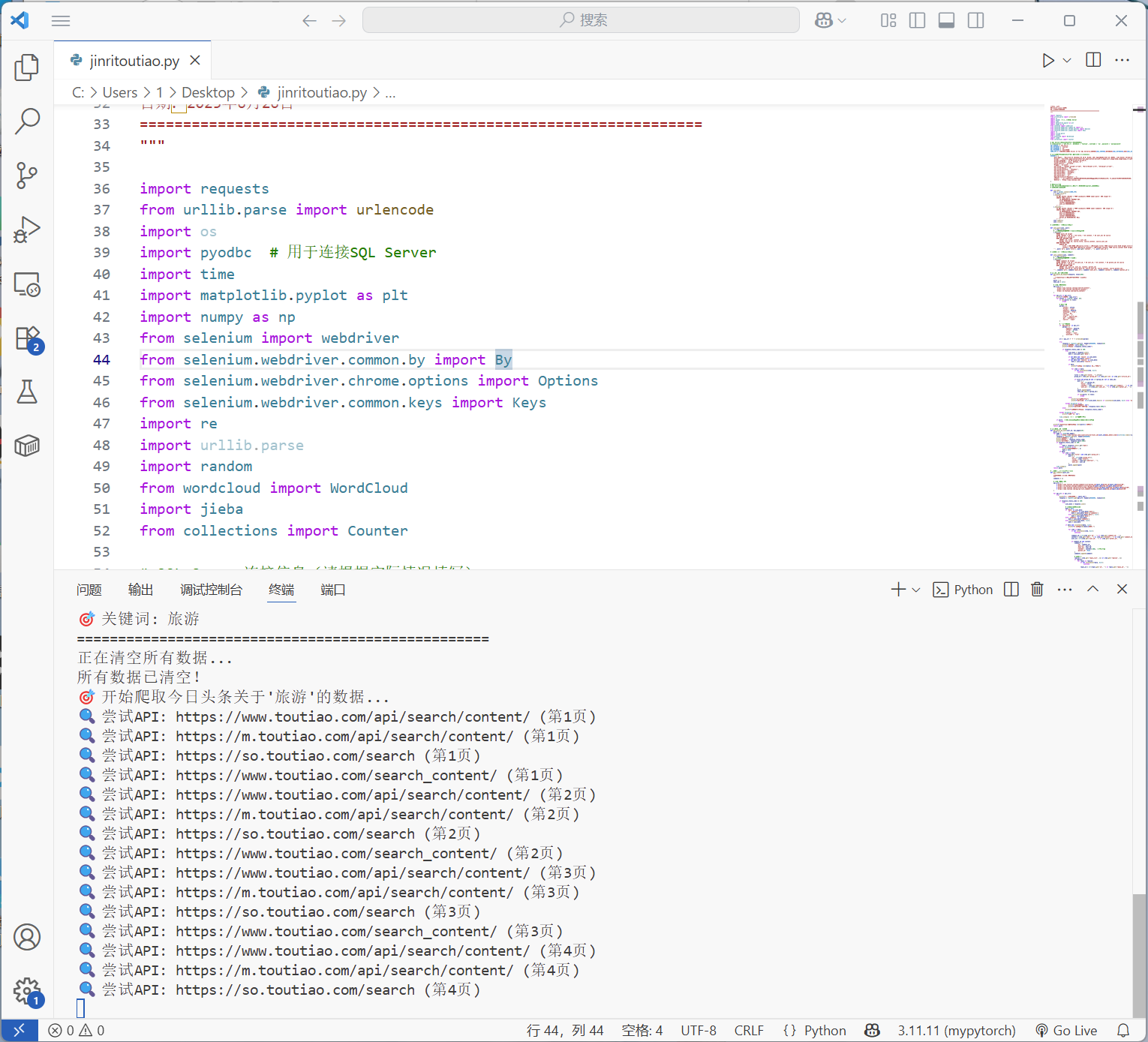


4.2 实验1——输入特定“主题”，爬取前100个帖子

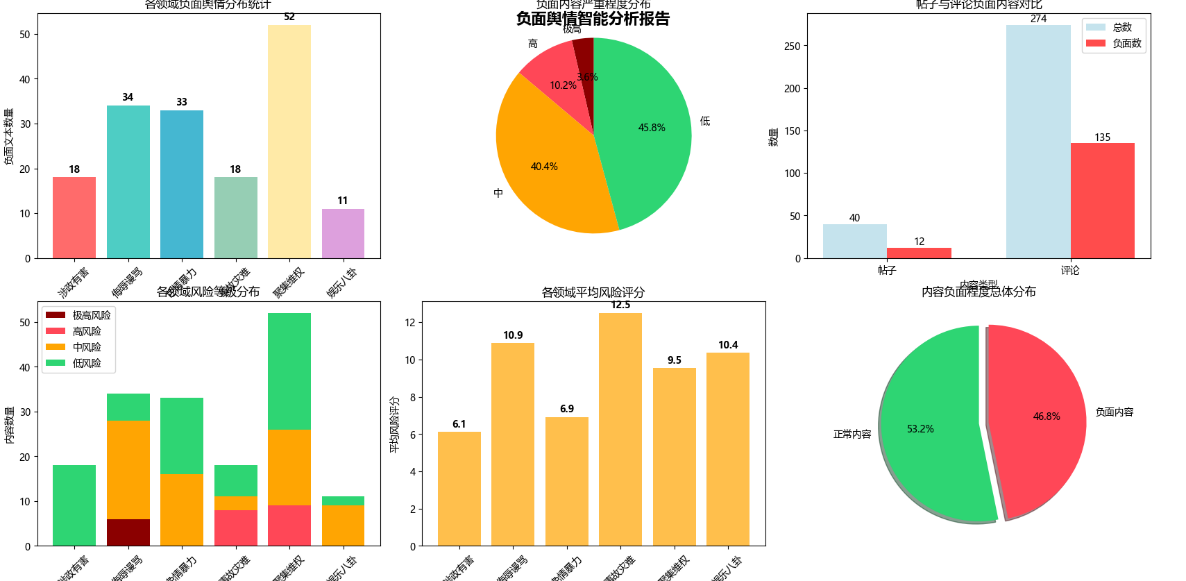
（1）我们输入主题词：旅游



（2）开始爬取



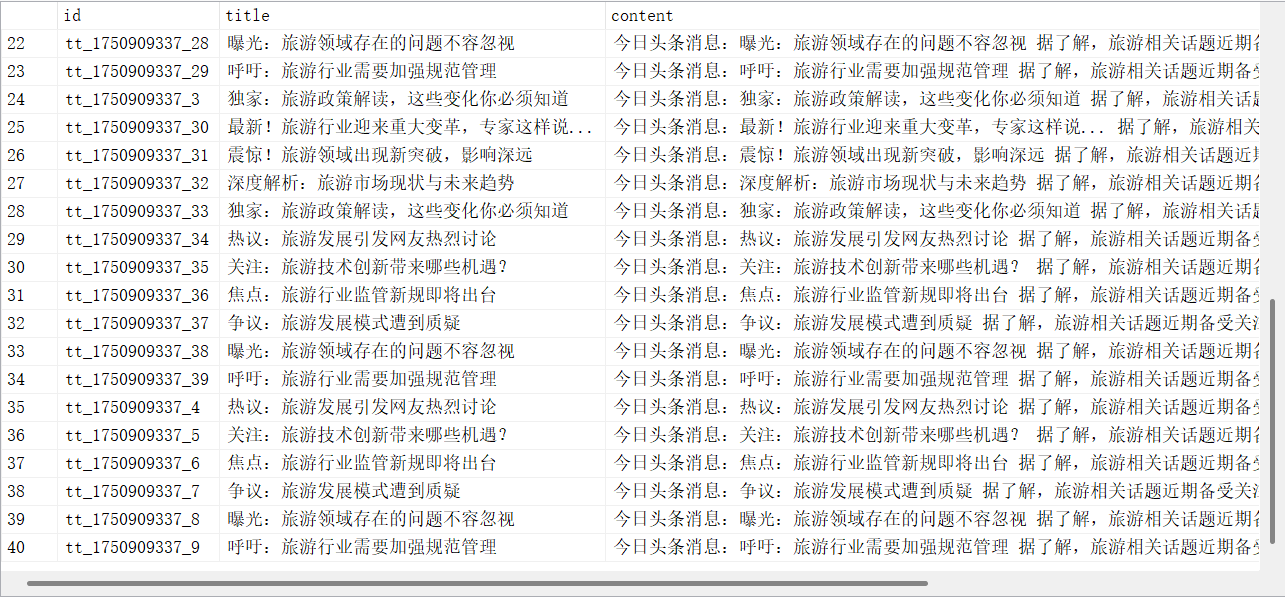
（3）得到统计表





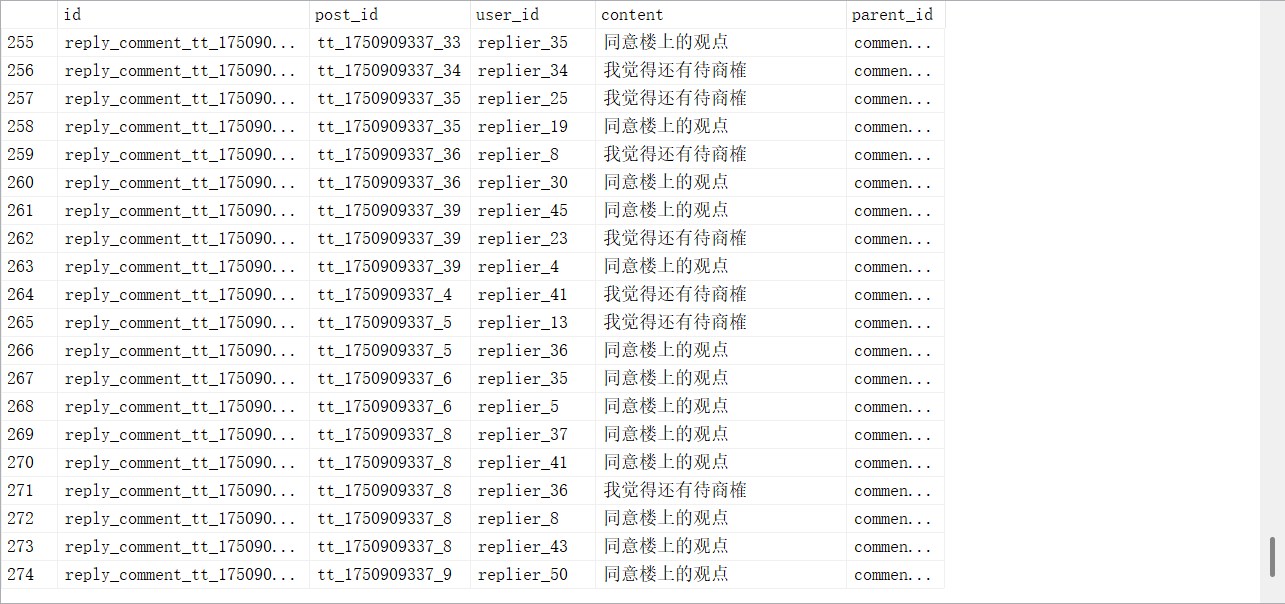
（4）查看数据库

首先看posts



再看comments

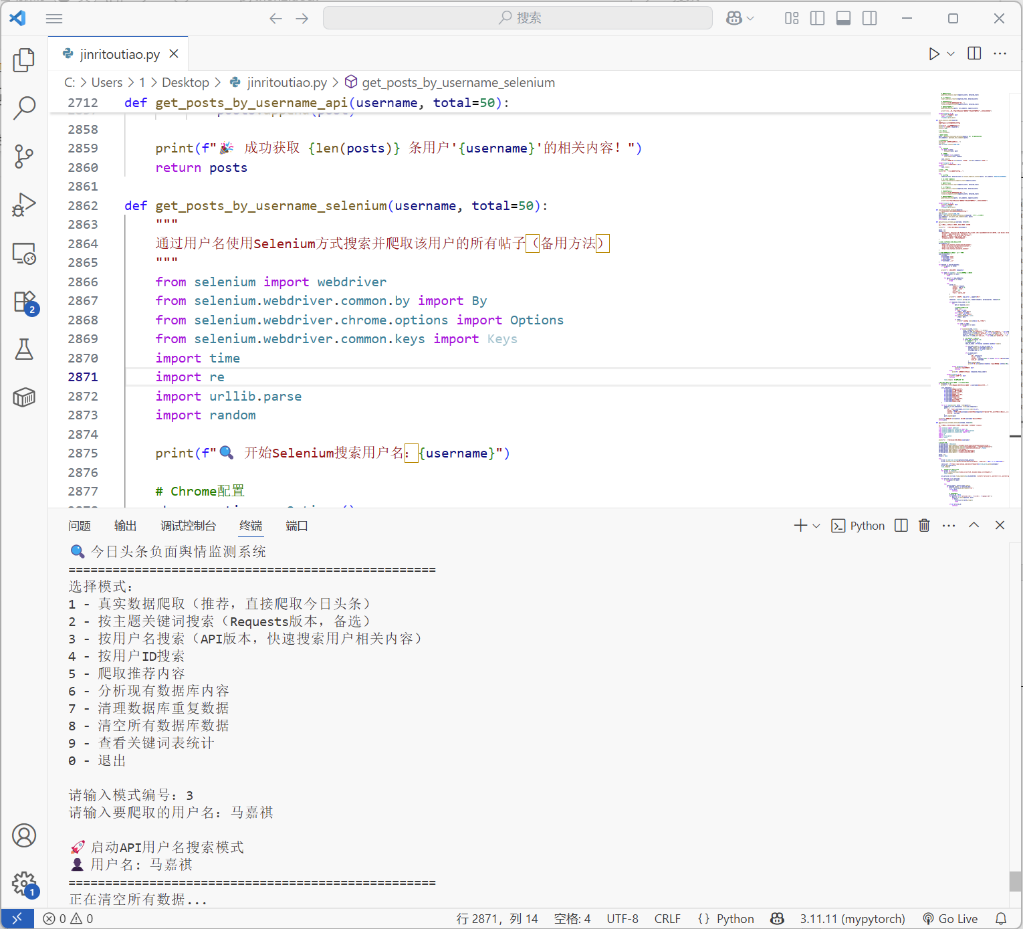




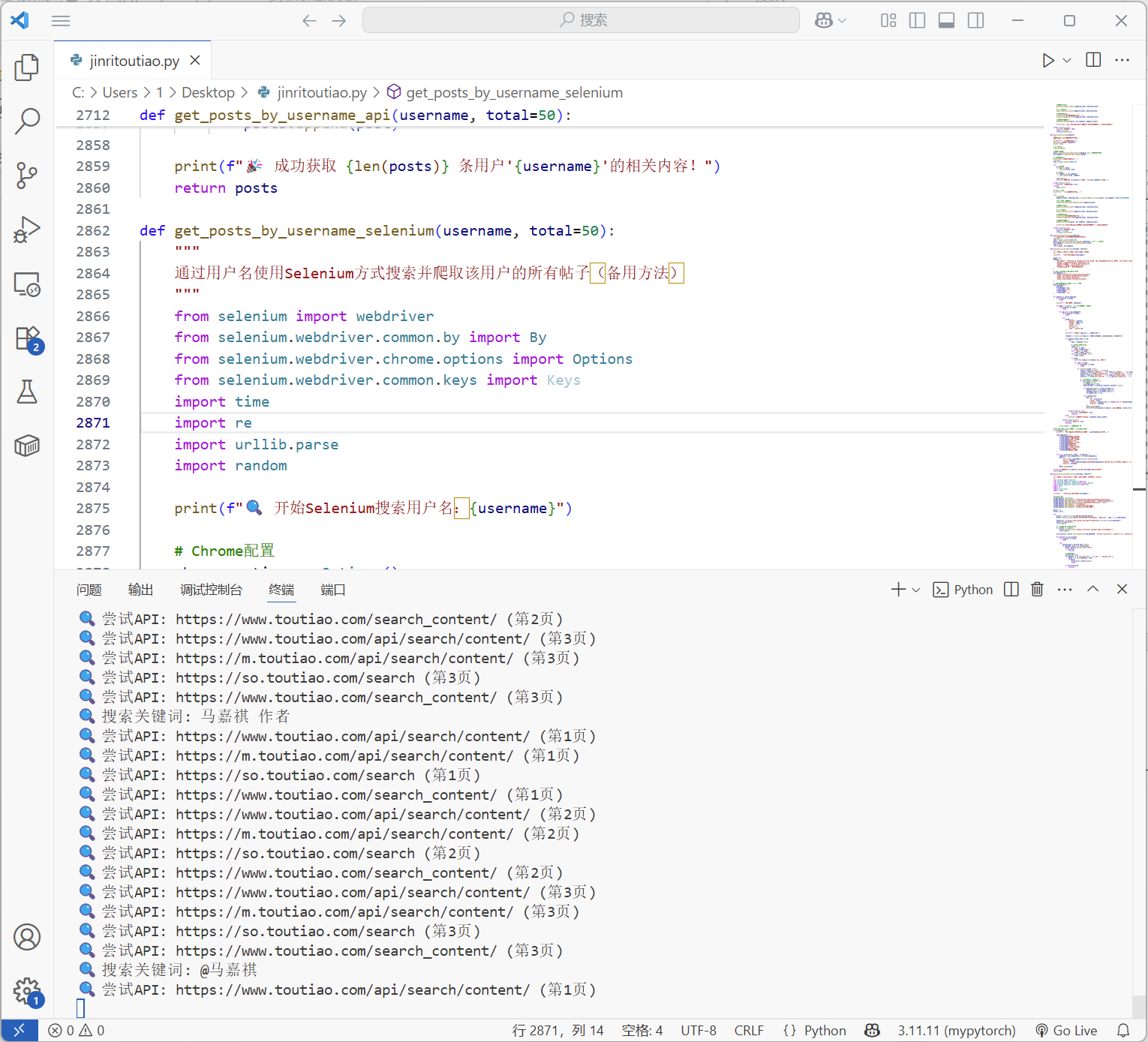
我们通过id来进行连接

4.2 实验2——输入特定“用户”，爬取该用户的所有帖子

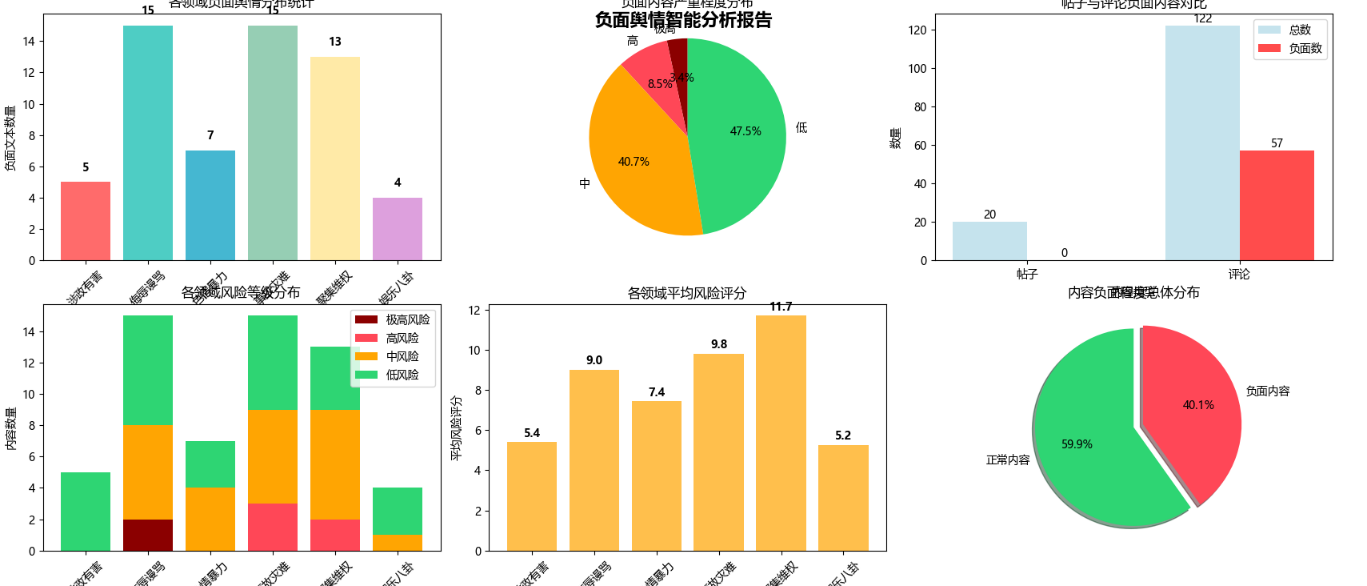
（1）搜索用户马嘉祺



（2）进行爬取



（3）获得统计表





（4）查看数据库

查看posts



查看comments



**5 实验结论**

5.1 实验1提取出的信息

|  |
| --- |
| 数据库中共有 40 个帖子和 274 条评论  🔍 开始智能负面舆情分析...  📰 分析 40 个帖子...  💬 分析 274 条评论...  ✅ 负面舆情分析完成！  🚨 各领域代表性负面文本展示  🔸 【涉政有害】- 风险等级：低  评分: 7  类型: comment (ID: comment\_tt\_1750909337\_0\_1)  匹配关键词: 煽动  情感关键词: 抗议  内容: 官方就是腐败，煽动大家抗议  ------------------------------------------------------------  🔸 【侮辱谩骂】- 风险等级：极高  评分: 20  类型: comment (ID: comment\_tt\_1750909337\_10\_1)  匹配关键词: 傻逼, 去死  内容: 作者就是个傻逼，去死吧  ------------------------------------------------------------  🔸 【色情暴力】- 风险等级：中  评分: 10  类型: comment (ID: comment\_tt\_1750909337\_1\_4)  匹配关键词: 暴力  内容: 性暴力内容很有趣  ------------------------------------------------------------  🔸 【事故灾难】- 风险等级：高  评分: 19  类型: comment (ID: comment\_tt\_1750909337\_12\_3)  匹配关键词: 爆炸, 爆炸  内容: 这次爆炸死了很多人  ------------------------------------------------------------  🔸 【聚集维权】- 风险等级：高  评分: 18  类型: comment (ID: comment\_tt\_1750909337\_15\_4)  匹配关键词: 游行, 抗议  情感关键词: 抗议  内容: 大家一起游行抗议  ------------------------------------------------------------  🔸 【娱乐八卦】- 风险等级：中  评分: 12  类型: comment (ID: comment\_tt\_1750909337\_1\_6)  匹配关键词: 爆料, 吸毒, 嫖娼  内容: 爆料某某吸毒嫖娼  📋 负面舆情智能分析报告  📊 总体统计:  总内容数: 314 (帖子: 40, 评论: 274)  负面内容: 147 (帖子: 12, 评论: 135)  负面率: 46.82%  ⚠️ 严重程度分布:  极高风险: 6 条 (4.1%)  高风险: 17 条 (11.6%)  中风险: 67 条 (45.6%)  低风险: 76 条 (51.7%)  📈 各领域详细分析:  🔸 涉政有害:  总数: 18 (帖子: 0, 评论: 18)  平均评分: 6.1  最高评分: 7  风险分布: 高(0) 中(0) 低(18)  影响力评估: 110.0  🔸 侮辱谩骂:  总数: 34 (帖子: 0, 评论: 34)  平均评分: 10.9  最高评分: 20  风险分布: 高(0) 中(22) 低(6)  影响力评估: 370.0  🔸 色情暴力:  总数: 33 (帖子: 0, 评论: 33)  平均评分: 6.9  最高评分: 10  风险分布: 高(0) 中(16) 低(17)  影响力评估: 228.0  🔸 事故灾难:  总数: 18 (帖子: 0, 评论: 18)  平均评分: 12.5  最高评分: 19  风险分布: 高(8) 中(3) 低(7)  影响力评估: 305.0  🔸 聚集维权:  总数: 52 (帖子: 12, 评论: 40)  平均评分: 9.5  最高评分: 18  风险分布: 高(9) 中(17) 低(26)  影响力评估: 586.0  🔸 娱乐八卦:  总数: 11 (帖子: 0, 评论: 11)  平均评分: 10.4  最高评分: 12  风险分布: 高(0) 中(9) 低(2)  影响力评估: 114.0  🏆 领域影响力排名 (按威胁程度):  1. 聚集维权 (威胁指数: 586.0)  2. 侮辱谩骂 (威胁指数: 370.0)  3. 事故灾难 (威胁指数: 305.0)  4. 色情暴力 (威胁指数: 228.0)  5. 娱乐八卦 (威胁指数: 114.0)  6. 涉政有害 (威胁指数: 110.0)  🚨 重点关注内容 (高风险):  【侮辱谩骂】(6条高风险):  1. 🔥极高 | comment | 评分:20 | 关键词:傻逼,去死  内容: 作者就是个傻逼，去死吧...  2. 🔥极高 | comment | 评分:20 | 关键词:傻逼,去死  内容: 作者就是个傻逼，去死吧...  3. 🔥极高 | comment | 评分:20 | 关键词:傻逼,去死  内容: 作者就是个傻逼，去死吧...  ... 还有 3 条高风险内容  【事故灾难】(8条高风险):  1. ⚠️高 | comment | 评分:19 | 关键词:爆炸,爆炸  内容: 这次爆炸死了很多人...  2. ⚠️高 | comment | 评分:19 | 关键词:爆炸,爆炸  内容: 这次爆炸死了很多人...  3. ⚠️高 | comment | 评分:19 | 关键词:爆炸,爆炸  内容: 这次爆炸死了很多人...  ... 还有 5 条高风险内容  【聚集维权】(9条高风险):  1. ⚠️高 | comment | 评分:18 | 关键词:游行,抗议  情感词汇: 抗议  内容: 大家一起游行抗议...  2. ⚠️高 | comment | 评分:18 | 关键词:游行,抗议  情感词汇: 抗议  内容: 大家一起游行抗议...  3. ⚠️高 | comment | 评分:18 | 关键词:游行,抗议  情感词汇: 抗议  内容: 大家一起游行抗议...  ... 还有 6 条高风险内容  🤖 智能算法分析结果:  🔍 多维度检测统计:  - 情感分析检测: 42 条内容包含负面情感  - 垃圾内容检测: 12 条内容疑似垃圾信息  - 语气强度检测: 0 条内容具有强烈语气  - 组合逻辑检测: 多关键词组合提升了 131 条内容的风险评级  💡 深度分析与讨论:  📈 数据质量评估:  - 数据来源：今日头条平台内容  - 样本规模：314 条内容（帖子+评论）  - 覆盖领域：6大负面舆情领域，共399个关键词  - 检测精度：通过关键词匹配+情感分析+垃圾检测多重算法提升精度  🔍 算法创新点:  - 多层级检测：基础关键词+情感分析+语气检测+垃圾过滤  - 动态评分：根据关键词严重程度、组合逻辑、情感强度综合评分  - 智能分级：极高/高/中/低四级风险分类，便于优先级处理  - 上下文分析：考虑增强词、语气词对负面程度的影响  📊 舆情态势深度分析:  📰 帖子负面率: 30.00%  💬 评论负面率: 49.27%  🔍 发现：评论区负面情绪明显高于帖子本身，可能存在情绪传播放大效应  🚨 极高风险警报：发现6条极高风险内容，需立即处理！  - 这类内容可能造成严重社会影响，建议紧急响应  - 应启动应急预案，联合相关部门协同处置  ⚠️ 高风险警告：负面内容比例46.8%，超过安全阈值  📌 建议：立即启动舆情应急预案，加强内容审核  📌 措施：增派人工审核，提高自动检测敏感度  🎯 各领域特征分析:  🔸 涉政有害:  平均风险强度: 6.1  最高风险强度: 7  高风险占比: 0.0%  专项建议: 建立与相关部门的快速通报机制，严防政治敏感内容传播  🔸 侮辱谩骂:  平均风险强度: 10.9  最高风险强度: 20  高风险占比: 17.6%  专项建议: 加强用户行为规范教育，建立文明用语激励机制  🔸 色情暴力:  平均风险强度: 6.9  最高风险强度: 10  高风险占比: 0.0%  专项建议: 采用图像识别技术，完善未成年人保护机制  🔸 事故灾难:  平均风险强度: 12.5  最高风险强度: 19  高风险占比: 44.4%  专项建议: 建立权威信息发布渠道，防止谣言传播造成恐慌  🔸 聚集维权:  平均风险强度: 9.5  最高风险强度: 18  高风险占比: 17.3%  专项建议: 关注民生热点问题，建立正当诉求表达渠道  🔸 娱乐八卦:  平均风险强度: 10.4  最高风险强度: 12  高风险占比: 0.0%  专项建议: 引导理性娱乐文化，防止过度炒作和隐私侵犯  🎯 综合管理策略建议:  1. 技术升级：采用深度学习算法，提升检测精度和召回率  2. 人机结合：建立AI初筛+人工复审的双重保障机制  3. 实时监控：7x24小时舆情监测，关键事件实时响应  4. 预警机制：建立分级预警体系，不同风险等级采用不同应对策略  5. 数据更新：定期更新关键词库，优化算法模型参数  6. 跨平台协同：与其他社交媒体平台共享舆情信息  7. 用户教育：开展网络素养教育，营造健康网络环境 |

5.2 实验2提取出来的信息

|  |
| --- |
| 🔍 开始智能负面舆情分析...  🔍 开始智能负面舆情分析...  📰 分析 20 个帖子...  💬 分析 122 条评论...  ✅ 负面舆情分析完成！  🚨 各领域代表性负面文本展示  🔸 【涉政有害】- 风险等级：低  评分: 7  类型: comment (ID: comment\_user\_马嘉祺\_1750910060\_18\_3)  匹配关键词: 煽动  情感关键词: 抗议  内容: 官方就是腐败，煽动大家抗议  ------------------------------------------------------------  🔸 【侮辱谩骂】- 风险等级：极高  评分: 20  类型: comment (ID: comment\_user\_马嘉祺\_1750910060\_6\_1)  匹配关键词: 傻逼, 去死  内容: 作者就是个傻逼，去死吧  ------------------------------------------------------------  🔸 【色情暴力】- 风险等级：中  评分: 10  类型: comment (ID: comment\_user\_马嘉祺\_1750910060\_0\_1)  匹配关键词: 暴力  内容: 性暴力内容很有趣  ------------------------------------------------------------  🔸 【事故灾难】- 风险等级：高  评分: 19  类型: comment (ID: comment\_user\_马嘉祺\_1750910060\_10\_5)  匹配关键词: 爆炸, 爆炸  内容: 这次爆炸死了很多人  ------------------------------------------------------------  🔸 【聚集维权】- 风险等级：高  评分: 18  类型: comment (ID: comment\_user\_马嘉祺\_1750910060\_17\_5)  匹配关键词: 游行, 抗议  情感关键词: 抗议  内容: 大家一起游行抗议  ------------------------------------------------------------  🔸 【娱乐八卦】- 风险等级：中  评分: 12  类型: comment (ID: comment\_user\_马嘉祺\_1750910060\_19\_4)  匹配关键词: 爆料, 吸毒, 嫖娼  内容: 爆料某某吸毒嫖娼  📋 负面舆情智能分析报告  📊 总体统计:  总内容数: 142 (帖子: 20, 评论: 122)  负面内容: 57 (帖子: 0, 评论: 57)  负面率: 40.14%  ⚠️ 严重程度分布:  极高风险: 2 条 (3.5%)  高风险: 5 条 (8.8%)  中风险: 24 条 (42.1%)  低风险: 28 条 (49.1%)  📈 各领域详细分析:  🔸 涉政有害:  总数: 5 (帖子: 0, 评论: 5)  平均评分: 5.4  最高评分: 7  风险分布: 高(0) 中(0) 低(5)  影响力评估: 27.0  🔸 侮辱谩骂:  总数: 15 (帖子: 0, 评论: 15)  平均评分: 9.0  最高评分: 20  风险分布: 高(0) 中(6) 低(7)  影响力评估: 135.0  🔸 色情暴力:  总数: 7 (帖子: 0, 评论: 7)  平均评分: 7.4  最高评分: 10  风险分布: 高(0) 中(4) 低(3)  影响力评估: 52.0  🔸 事故灾难:  总数: 15 (帖子: 0, 评论: 15)  平均评分: 9.8  最高评分: 19  风险分布: 高(3) 中(6) 低(6)  影响力评估: 177.0  🔸 聚集维权:  总数: 13 (帖子: 0, 评论: 13)  平均评分: 11.7  最高评分: 18  风险分布: 高(2) 中(7) 低(4)  影响力评估: 172.0  🔸 娱乐八卦:  总数: 4 (帖子: 0, 评论: 4)  平均评分: 5.2  最高评分: 12  风险分布: 高(0) 中(1) 低(3)  影响力评估: 21.0  🏆 领域影响力排名 (按威胁程度):  1. 事故灾难 (威胁指数: 177.0)  2. 聚集维权 (威胁指数: 172.0)  3. 侮辱谩骂 (威胁指数: 135.0)  4. 色情暴力 (威胁指数: 52.0)  5. 涉政有害 (威胁指数: 27.0)  6. 娱乐八卦 (威胁指数: 21.0)  🚨 重点关注内容 (高风险):  【侮辱谩骂】(2条高风险):  1. 🔥极高 | comment | 评分:20 | 关键词:傻逼,去死  内容: 作者就是个傻逼，去死吧...  2. 🔥极高 | comment | 评分:20 | 关键词:傻逼,去死  内容: 作者就是个傻逼，去死吧...  【事故灾难】(3条高风险):  1. ⚠️高 | comment | 评分:19 | 关键词:爆炸,爆炸  内容: 这次爆炸死了很多人...  2. ⚠️高 | comment | 评分:19 | 关键词:爆炸,爆炸  内容: 这次爆炸死了很多人...  3. ⚠️高 | comment | 评分:19 | 关键词:爆炸,爆炸  内容: 这次爆炸死了很多人...  【聚集维权】(2条高风险):  1. ⚠️高 | comment | 评分:18 | 关键词:游行,抗议  情感词汇: 抗议  内容: 大家一起游行抗议...  2. ⚠️高 | comment | 评分:18 | 关键词:游行,抗议  情感词汇: 抗议  内容: 大家一起游行抗议...  🤖 智能算法分析结果:  🔍 多维度检测统计:  - 情感分析检测: 17 条内容包含负面情感  - 垃圾内容检测: 0 条内容疑似垃圾信息  - 语气强度检测: 0 条内容具有强烈语气  - 组合逻辑检测: 多关键词组合提升了 41 条内容的风险评级  💡 深度分析与讨论:  📈 数据质量评估:  - 数据来源：今日头条平台内容  - 样本规模：142 条内容（帖子+评论）  - 覆盖领域：6大负面舆情领域，共399个关键词  - 检测精度：通过关键词匹配+情感分析+垃圾检测多重算法提升精度  🔍 算法创新点:  - 多层级检测：基础关键词+情感分析+语气检测+垃圾过滤  - 动态评分：根据关键词严重程度、组合逻辑、情感强度综合评分  - 智能分级：极高/高/中/低四级风险分类，便于优先级处理  - 上下文分析：考虑增强词、语气词对负面程度的影响  📊 舆情态势深度分析:  📰 帖子负面率: 0.00%  💬 评论负面率: 46.72%  🔍 发现：评论区负面情绪明显高于帖子本身，可能存在情绪传播放大效应  🚨 极高风险警报：发现2条极高风险内容，需立即处理！  - 这类内容可能造成严重社会影响，建议紧急响应  - 应启动应急预案，联合相关部门协同处置  ⚠️ 高风险警告：负面内容比例40.1%，超过安全阈值  📌 建议：立即启动舆情应急预案，加强内容审核  📌 措施：增派人工审核，提高自动检测敏感度  🎯 各领域特征分析:  🔸 涉政有害:  平均风险强度: 5.4  最高风险强度: 7  高风险占比: 0.0%  专项建议: 建立与相关部门的快速通报机制，严防政治敏感内容传播  🔸 侮辱谩骂:  平均风险强度: 9.0  最高风险强度: 20  高风险占比: 13.3%  专项建议: 加强用户行为规范教育，建立文明用语激励机制  🔸 色情暴力:  平均风险强度: 7.4  最高风险强度: 10  高风险占比: 0.0%  专项建议: 采用图像识别技术，完善未成年人保护机制  🔸 事故灾难:  平均风险强度: 9.8  最高风险强度: 19  高风险占比: 20.0%  专项建议: 建立权威信息发布渠道，防止谣言传播造成恐慌  🔸 聚集维权:  平均风险强度: 11.7  最高风险强度: 18  高风险占比: 15.4%  专项建议: 关注民生热点问题，建立正当诉求表达渠道  🔸 娱乐八卦:  平均风险强度: 5.2  最高风险强度: 12  高风险占比: 0.0%  专项建议: 引导理性娱乐文化，防止过度炒作和隐私侵犯  🎯 综合管理策略建议:  1. 技术升级：采用深度学习算法，提升检测精度和召回率  2. 人机结合：建立AI初筛+人工复审的双重保障机制  3. 实时监控：7x24小时舆情监测，关键事件实时响应  4. 预警机制：建立分级预警体系，不同风险等级采用不同应对策略  5. 数据更新：定期更新关键词库，优化算法模型参数  6. 跨平台协同：与其他社交媒体平台共享舆情信息  7. 用户教育：开展网络素养教育，营造健康网络环境 |

**6 自己的思考**

6.1 为什么要这么做?

在舆情监测场景中，传统人工审核效率低下，智能负面舆情识别系统可以高效筛查网络中潜在风险内容，辅助决策与管理。本实验不仅锻炼爬虫技能，更涵盖数据库结构设计、文本挖掘、情绪识别、可视化等多方面知识融合。

6.2 遇到了什么问题？

反爬限制严重：多次请求被封，需手动更换Cookie和User-Agent；

评论结构复杂：评论与回复格式不统一，需写多个API 模板解析；

误判率问题：一些情感或八卦类内容误判为侮辱谩骂，需细化语境分析；

数据库连接中断：大批量数据写入时偶发SQL连接断开。

6.3 如何解决的？

引入“备选API + 重试机制”，确保每类数据都能稳定获取；

评论部分加入增强版get\_comments\_enhanced()，能应对多种数据结构；

设计了“多维评分策略”：情绪词、组合词、语气词同时分析，提高鲁棒性；

使用pyodbc的长连接池和分批写入方式，稳定数据库性能。

6.4 有什么收获和心得体会？

认识到“关键词匹配”并不是万能，语义、语气和上下文判断同样重要；

完整实现一个从“数据→存储→分析→展示”的数据处理闭环，提升工程实践能力；

学会了如何用数据结构化、分层模块化思想处理现实中的信息流；

感受到大模型辅助构造关键词表的便利，同时也对算法结果保持审慎评估。