一、数据清洗：

共 9037 条评论

移除了 979 条完全重复的评论

移除了 139 条高度相似的评论

移除了 20 条广告评论

完成文本规范化处理

移除了 10 条空评论

移除了 5 条无意义短评

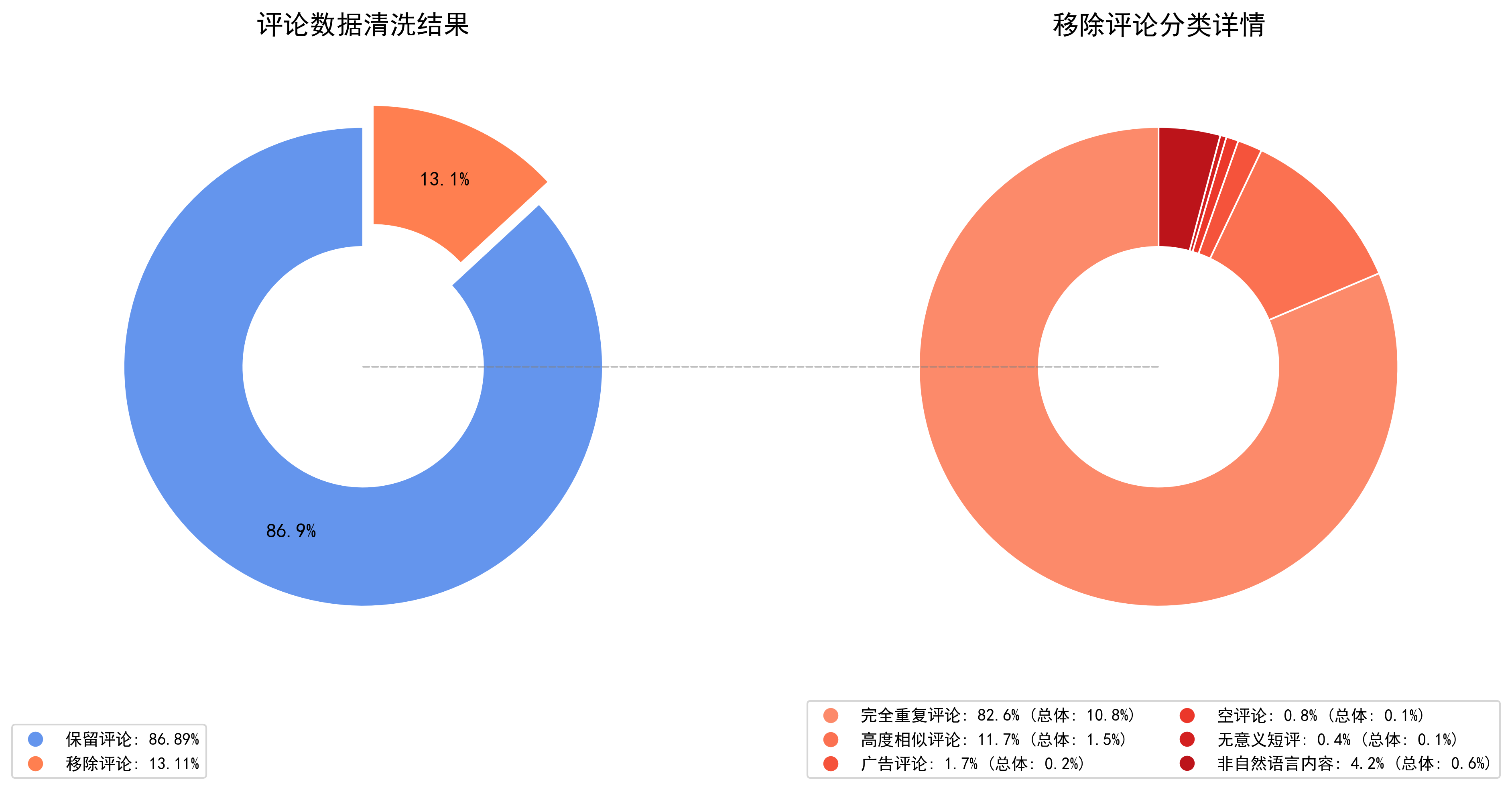
移除了 50 条非自然语言内容

总评论数: 9037

保留评论数: 7852

移除评论数: 1185

保留率: 86.89%



二、短评情感分析

大连理工情感词典

1. 长评情感分析

大模型api

1. 情感系数合成
2. 筛选出大模型认为有用的评论
3. 用Z\_score标准化函数进行标准化



i：维度（情感分数，故事讲述，角色表演，制作水平）

t：日期（1月29日-3月24日）

1. 利用熵权法得到每一个维度的权重（引入时间衰减后加权）

方法原理：

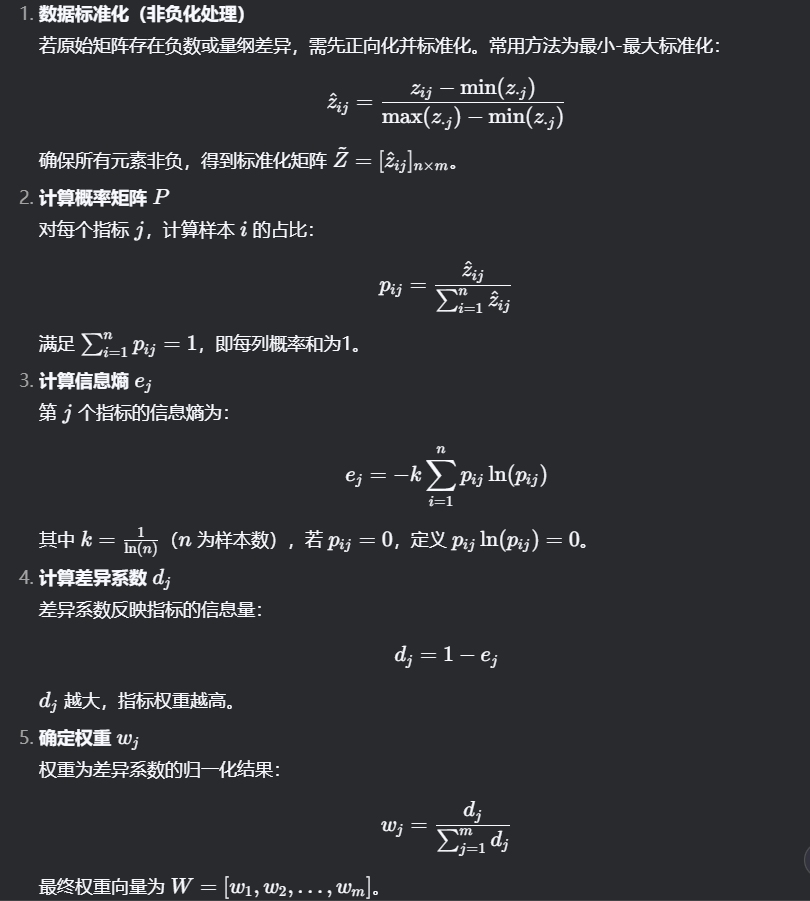
熵权法是一种基于信息熵的客观赋权方法，通过计算指标的变异程度（信息量）确定权重。信息熵越小，指标的变异程度越大，提供的信息量越多，权重越高；反之，熵值越大，权重越低。

参数：

信息熵  ：衡量指标数据分布的混乱程度。熵值越小，数据分布越不均衡，信息量越大。

差异系数：将熵值转化为正向指标，直接反映信息量大小。

权重 ：通过归一化差异系数，实现权重的客观分配。



Sentiment\_score:0.2981473871395439

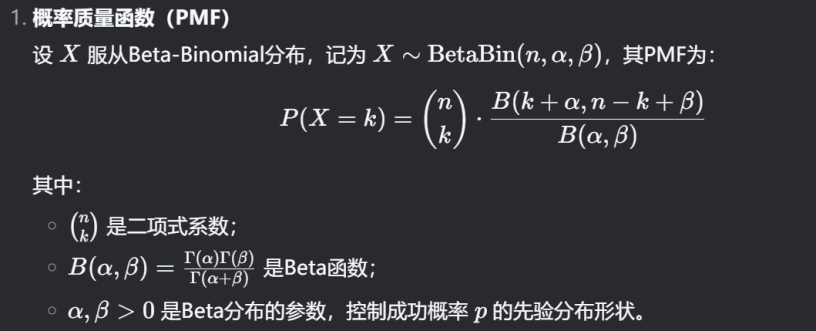
storytelling\_score:0.29770190721885587

Character\_performance score:0 2109536856534225

Production\_score:0.19319701998817784

1. 引入时间衰减

I.确定时间衰减函数：



Beta-Binomial分布是Binomial分布的扩展，通过将Binomial分布中的成功概率 p 替换为服从Beta分布的随机变量，从而引入参数不确定性。这种混合模型能够更好地描述具有过度离散性的数据，即观测方差显著大于Binomial分布的理论方差的情况。

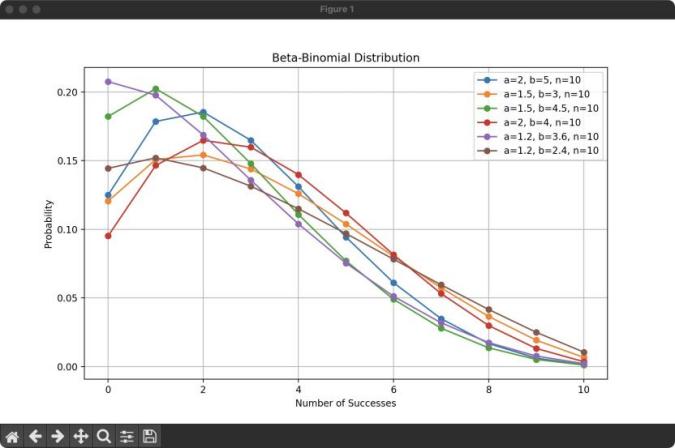
核心意义：离散化的自然延伸

Beta-Binomial分布通过以下方式扩展了Binomial分布：

引入参数不确定性：将固定的 p 替换为随机变量，更灵活地建模真实场景中的不确定性。

解决过度离散性：当观测数据的方差大于 np(1−p)时，Beta-Binomial分布通过调整参数 α,β拟合这种变异性。

贝叶斯框架的共轭性：Beta分布是Binomial分布的共轭先验，便于贝叶斯推断中后验分布的解析计算。



最佳参数选择：α=2，β=5

II.窗口选择（衰减函数应用时间范围，向前天数，即n的值）

一、核心原则

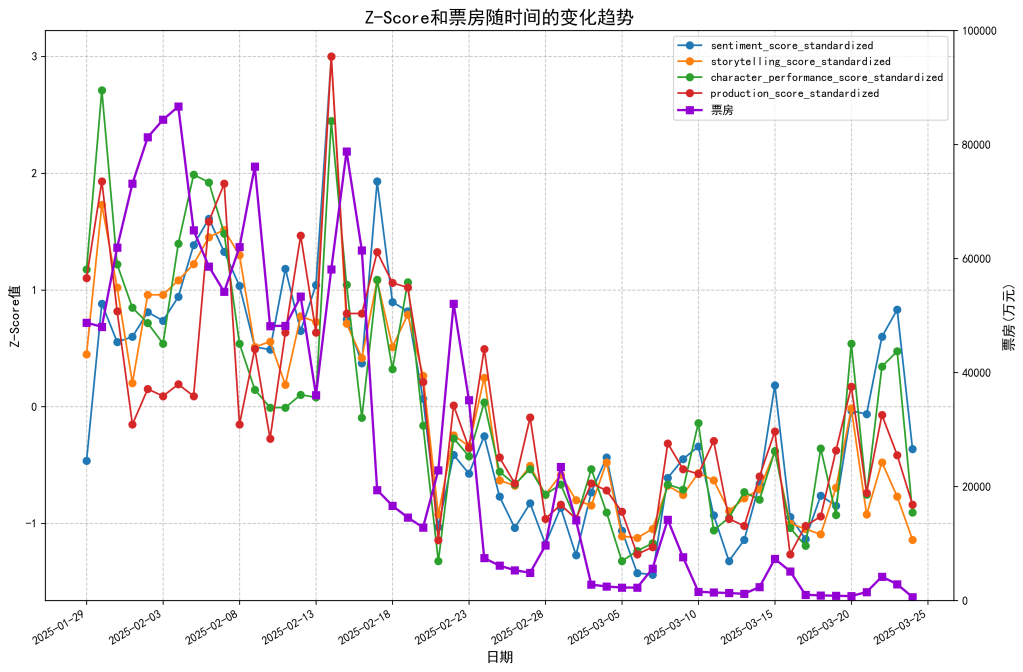
确定时间窗口时需平衡 实时性 与 稳定性：

短窗口：捕捉近期快速变化，但对噪声敏感；

长窗口：平滑波动，但可能滞后于真实趋势。

二. 数据分析

绘制时间序列图：观察数据的波动周期和衰减速度。



计算自相关性：通过自相关函数（ACF）和偏自相关函数（PACF）确定显著滞后阶数。

统计指标分析：计算不同窗口下数据的均值和方差，选择方差稳定且趋势明显的窗口。

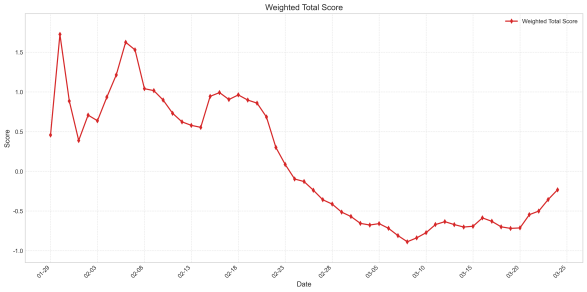
确定n=10

III.进行时间衰减

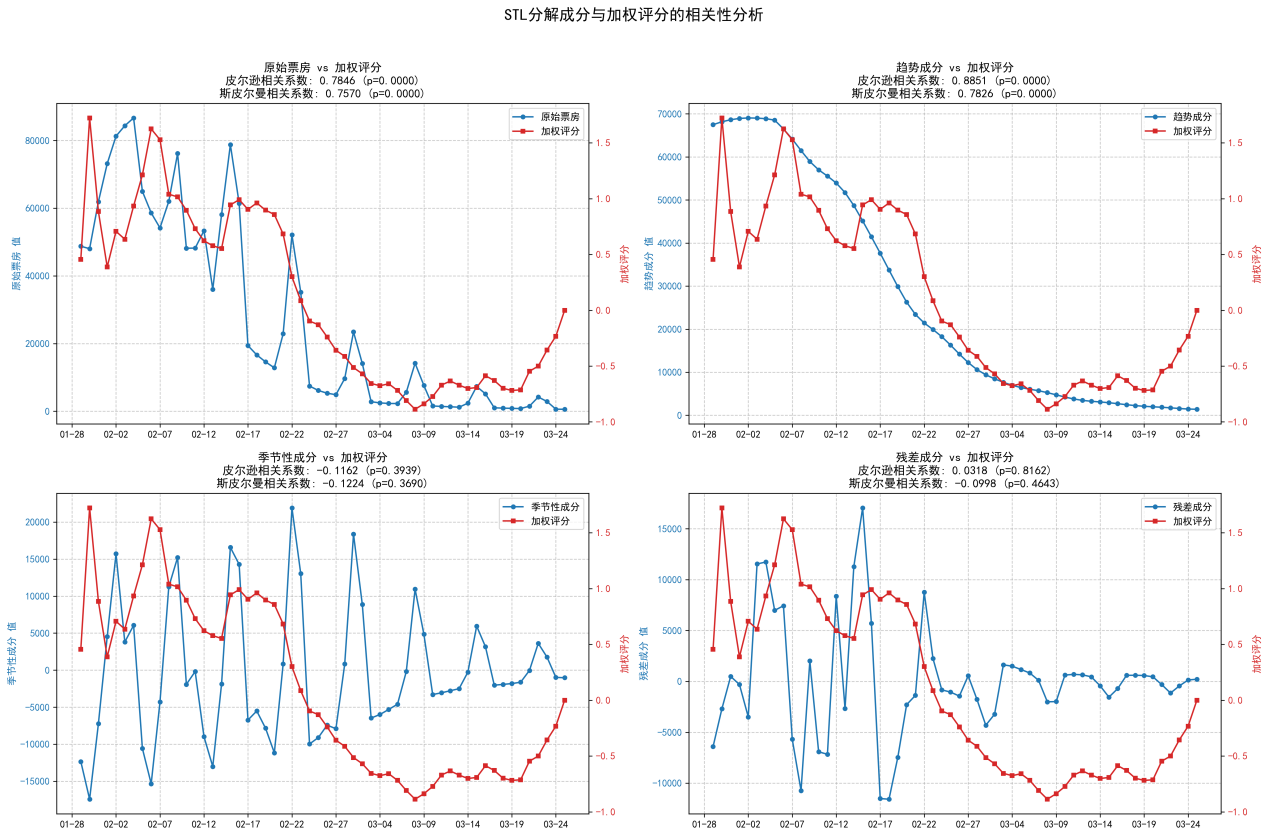


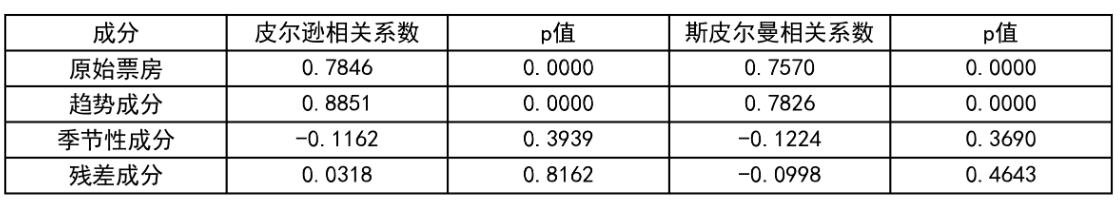
5.使用熵权法得出的权重对衰减后的值进行加权



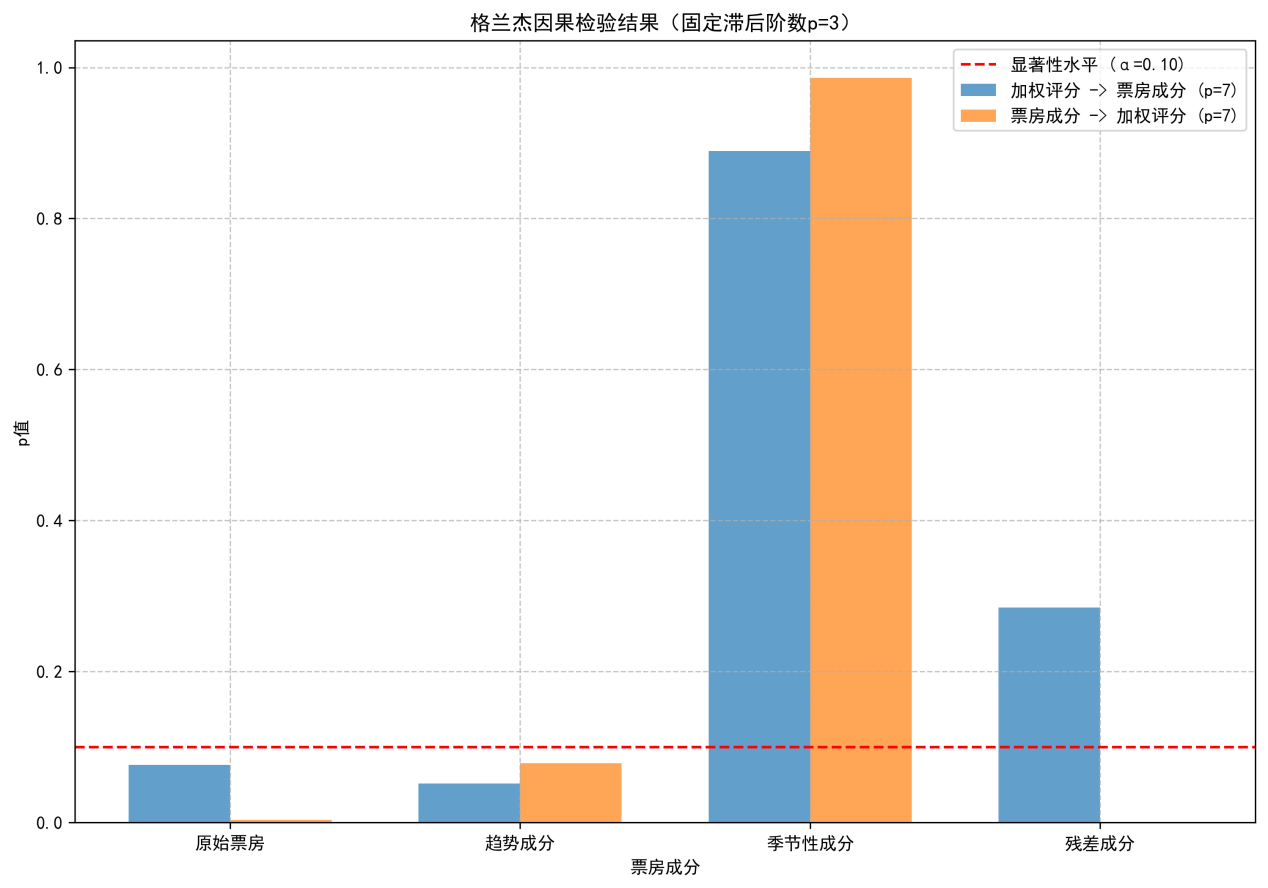


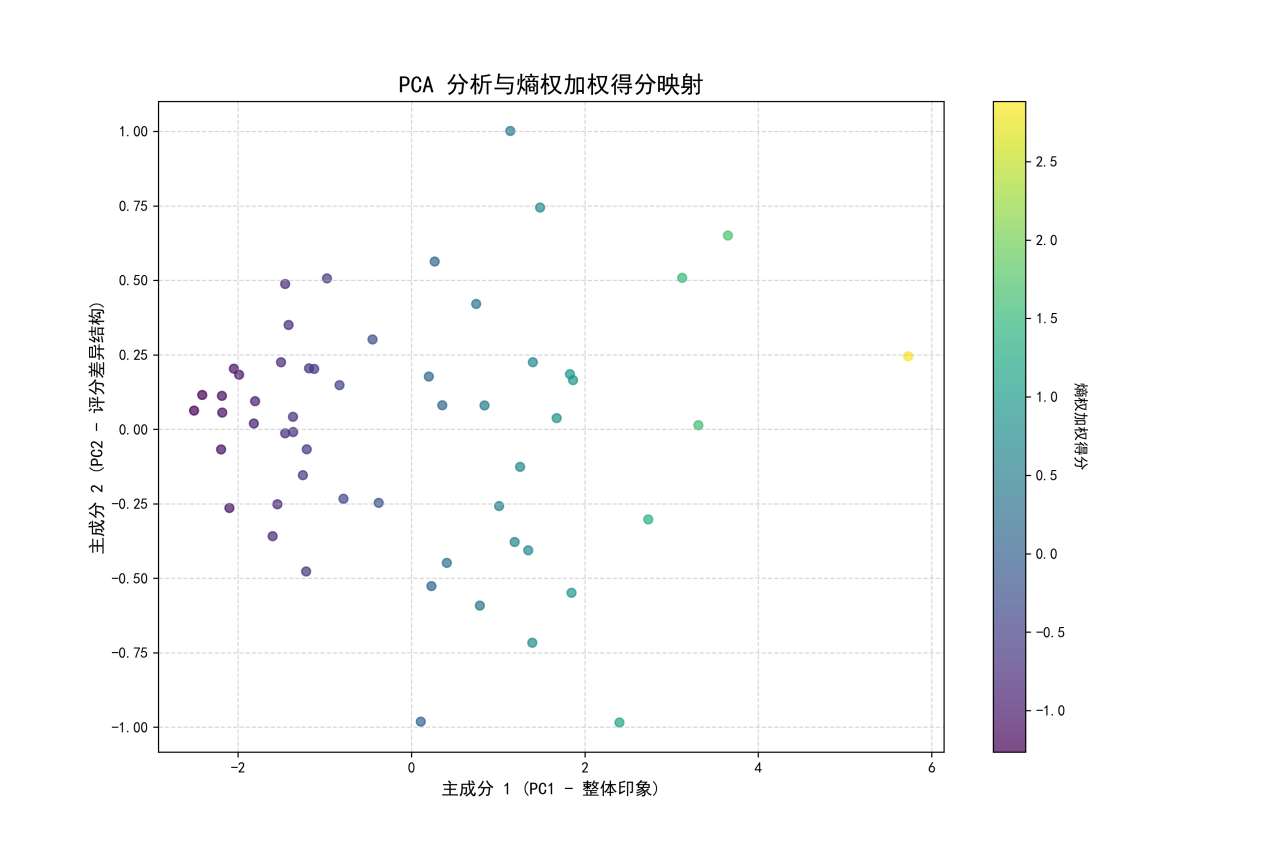
1. STL分解并计算相关系数





1. 格兰杰因果检验



1. 二维PCA + 加权情感映射  
   

 X轴 (PC1 - 主成分1): 代表了四个评分维度线性组合后的主要变化方向。根据解释，这通常可以理解为 “整体评价” 或 “综合印象”。得分越高（越靠右），表示在所有四个维度上的综合得分倾向于越高。

Y轴 (PC2 - 主成分2): 代表了与 PC1 正交的第二大变化方向。根据解释，这可以理解为 “评分结构差异” 或“维度侧重”。它揭示了除了整体高低之外，评分在不同维度间存在的特定模式或对比关系（例如，故事性强但制作一般的项 vs. 制作精良但故事性弱的项）。

颜色深浅 (熵权加权得分): 点的颜色强度直接反映了根据熵权法计算出的加权总分。颜色越深，表示加权得分越高，意味着根据权重（情感分: 0.2981, 故事: 0.2977, 角色: 0.2109, 制作: 0.1932），该项获得了更高的综合评价，尤其是在情感和故事讲述方面得分较高，观众评价更积极。颜色越浅，表示加权得分越低，观众评价相对偏低。

密集区域: 点密集的区域可能代表了具有相似评价特征（相似的整体得分水平和评分结构）的样本集合。

稀疏区域/离群点:远离主要群体的点代表了评价较为独特或极端的样本。

样本点主要集中在第二象限，右下象限 (PC1 > 0, PC2 < 0): 此区域的点也具有较高的整体评价 (PC1)，但在评分结构 (PC2) 上呈现出与右上象限不同的模式。

深色点: 整体评价高、结构不同于右上、加权得分高的样本。

浅色点: 整体评价高、结构不同于右上、但加权得分相对不高的样本。

颜色深浅（加权得分/情感倾向）:

深色区域: 重点关注颜色最深的那些点。它们代表了观众评价最高（根据熵权加权）的样本。观察它们主要分布在哪个或哪些象限，可以了解高评价样本的主要特征（是普遍高分，还是有特定结构的高分）。由于情感分和故事讲述权重最高，深色点很可能在这两项上得分突出。

浅色区域: 颜色最浅的点代表加权得分最低的样本。它们的位置则反映了低分样本的特征。这些样本可能在情感和故事讲述上得分较低。

1. 词云

