

基于LDA主题建模的文学语料主题演化分析

——以简·奥斯汀六部小说为例

苏 乐

上海海事大学外国语学院，上海

收稿日期：2026年6月16日；录用日期：2026年7月31日；发布日期：2026年8月14日

摘 要

为厘清简·奥斯汀创作主题的阶段性变化，并建立文本词汇特征与时代社会环境的关联，本文选取其6部长篇小说构建小型英文语料库，采用LDA主题建模开展量化文本分析，并结合19世纪初英国文学发展脉络完成综合解读。量化结果显示：奥斯汀早期创作文本高频词汇集中于情感、社交婚约相关表述，文本整体侧重婚恋关系中的情感与礼仪探讨；成熟期作品高频出现财产、生存处境类词汇，文本更侧重个体婚恋选择背后的现实约束。分作品独立建模能够捕捉同一创作阶段下不同文本的微观词汇差异。统计层面呈现的主题阶段性分化，主要源于当时英国财产继承、婚姻财产相关法律对女性生存选择的硬性约束，社会财富流动变化仅作为长期时代背景存在。本研究构建了“量化数据 + 文学解读 + 社会语境”的融合框架，搭建“量化词汇统计 + 文本细读 + 时代制度语境”的综合分析路径，验证了LDA模型在小规模文学语料库中作为统计发现工具的辅助性应用价值，为计算语言学与文学研究的交叉提供了可复用的分析路径。

关键词

简·奥斯汀，LDA主题建模，主题演变，量化分析，语料库语言学

Topic Evolution in Literary Corpus Based on LDA Topic Modeling

—A Case Study of Six Jane Austen Novels

Le Su

School of Foreign Languages, Shanghai Maritime University, Shanghai

Received: June 16, 2026; accepted: July 31, 2026; published: August 14, 2026

Abstract

To clarify the phased shifts in Jane Austen's creative themes and establish correlations between textual lexical features and the social context of her era, this paper constructs a small English corpus consisting of her six full-length novels and conducts quantitative textual analysis via LDA topic modeling, followed by an integrated interpretation against the literary context of early nineteenth-century Britain. Quantitative findings indicate that high-frequency lexical items in Austen's early works revolve around emotions, social etiquette and marital engagements, with narratives centering on emotional bonds and ritual norms within romantic relationships. Her mature writings, by contrast, feature frequent terms pertaining to property and living circumstances, placing greater emphasis on practical restrictions underlying individuals' marital decisions. Independent modeling of individual novels further captures micro-level lexical disparities within the same creative phase. Statistically evident thematic segmentation primarily stems from rigid legal constraints imposed by British laws governing property inheritance and spousal assets on women's life choices, while shifts in social wealth circulation merely function as a long-term historical backdrop. This study establishes an integrated analytical framework combining quantitative data, literary interpretation and social context, alongside a comprehensive research path integrating lexical statistics, close reading and institutional context of the age. It verifies the auxiliary value of the LDA model as a statistical auxiliary tool for small-scale literary corpora, and delivers a replicable analytical workflow for interdisciplinary research bridging computational linguistics and literary studies.

Keywords

Jane Austen, LDA Topic Modeling, Thematic Evolution, Quantitative Analysis, Corpus Linguistics

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

简·奥斯汀聚焦乡绅婚恋与日常社交，文本是考察 18 至 19 世纪英国个体生存的典型材料，长期受到学界持续关注。奥斯汀的创作生涯(1795~1815)恰好横跨英国工业革命的发展阶段，社会财富分配逐步调整，不同乡绅群体因财富观念产生分化，深刻改变乡绅群体的生活选择，女性受相关法律约束缺乏独立财产处置权，婚恋直接决定终身生活走向，这一现实背景为其创作主题演变提供基础，也值得系统梳理创作主题随时间产生的变化规律。

从文学史的内部脉络来看，奥斯汀的创作处于小说文体转型节点：18 世纪感伤、哥特小说盛行，19 世纪写实叙事逐步兴起。奥斯汀承接感伤文学对情感、道德的关注，以生活化反讽叙事摆脱哥特作品夸张文风，为后续写实小说发展铺垫。同期作家创作同样呈现由情感书写转向现实探讨的走向，奥斯汀的主题变化与整体文学发展脉络相契合。因此，其作品主题的演变既是个人创作成熟的体现，也呼应英国小说从情感书写向日常写实过渡的整体走向。

梳理现有研究成果可见，学界对奥斯汀作品的解读已形成多元维度：在婚恋主题研究层面，Bailey [1] 从法律视角铺垫历史语境，Manimozhi [2] 与国内何志燕、陈剑雯 [3] 等聚焦婚恋与经济的关联，黄福奎 [4] 虽观察到奥斯汀后期婚恋观念存在变化，但全部研究结论均依靠主观文本细读得出，缺少客观词汇数据佐证。在研究方法层面，数字人文理念推动下，Wijitsopon [5]、Fischer-Starcke [6] 借助语料库工具分析奥

斯汀文本语言风格，Rha 与 Silver [7]采用主题建模横向对比 19 世纪女性创作者文本，但现有量化研究多聚焦静态文本风格、跨作家横向对比，缺少针对奥斯汀创作时间线的纵向主题演化分析，也未搭建量化数据与时代制度、文学脉络相结合的综合分析框架。

现有研究仅依靠定性判断奥斯汀主题变化，割裂词汇数据、文本内涵与时代背景的关联，难以厘清主题转变的多重动因。基于此，本文选取奥斯汀六部完整长篇小说，采用分层 LDA 主题建模开展量化分析，核心研究问题为：第一，依托 LDA 量化统计结果，奥斯汀早期、成熟期文本的词汇分布具备何种特征，两类文本是否存在可区分的统计层面主题差异；第二，同一创作阶段下不同小说的高频词汇存在哪些微观区别，词汇差异如何体现创作者叙事侧重的多样性；第三，奥斯汀文本词汇分布呈现的阶段性变化，与 18~19 世纪英国财产、婚姻法律制度、小说文体转型存在何种关联，主题转向的核心驱动因素是什么。本研究依托 LDA 建模客观呈现创作宏观变化与单篇差异，搭建“量化统计 - 文本细读 - 制度与文学语境”融合分析框架，为经典文学的计算语言学研究提供新思路。

2. 研究方法步骤

2.1. 语料选取与文本预处理

本研究选取简·奥斯汀 6 部长篇英文小说，按初稿创作时间划分为早期(1795~1798)、成熟期(1811~1815)两组，分组及预处理后词汇总量见表 1。原始文本取自 Project Gutenberg，仅保留正文，批量清除章节标题、换行空格完成基础清洗。

统一标准化预处理流程：全部字母小写、清除标点数字、分词、词形还原；统计全局词频，人工构建专属停用词表，剔除称谓、叙事动词、程度虚词；过滤全文频次 ≤ 2 的低频词，生成建模可用干净词表。

Table 1. Basic information of Jane Austen’s corpus
表 1. 简·奥斯汀语料基本信息

创作阶段	作品名称	创作年份	词汇量(预处理后)
早期	《理智与情感》	1795	18,173
	《傲慢与偏见》	1797	18,912
	《诺桑觉寺》	1798	10,802
成熟期	《曼斯菲尔德庄园》	1811	24,195
成熟期	《爱玛》	1815	22,077
成熟期	《劝导》	1815	11,399

2.2. 分层 LDA 主题建模设置

- 建模分为阶段合并建模、单篇独立建模两层，统一参数： $\alpha = 0.2$ ， $\eta = 0.8$ ，迭代 1500 次。
- (1) 阶段合并建模：同阶段文本词表合并为总文档，设置 $k = 2$ ，输出分阶段核心主题词汇；
 - (2) 单篇独立建模：遍历 $k = 2 \sim 5$ ，依靠一致性、困惑度双指标筛选最优主题数 $k = 2$ ，输出单篇词汇权重与主题分布。

2.3. 模型固有局限前置说明

LDA 依托词袋模型仅统计词汇共现频次，忽略句法、反讽、戏仿等修辞意图，对奥斯汀文本存在显著分析偏差。以《诺桑觉寺》哥特戏仿叙事为例，模型仅依据字面词汇频次统计，无法区分词汇的戏谑使用与正面叙事，易误判文本核心导向。本文统一界定：LDA 词频、权重仅为文本表层词汇分布统计线

索，所有文本主题解读均为结合细读的人工推论，量化数据仅作辅助参考，不能独立证明叙事观点。

整体分析流程如图 1 所示，整体逻辑为小说文本获取、文本清洗预处理、停用词筛选、双层 LDA 建模、模型效度检验，最后结合文本、制度、文学语境综合解读。

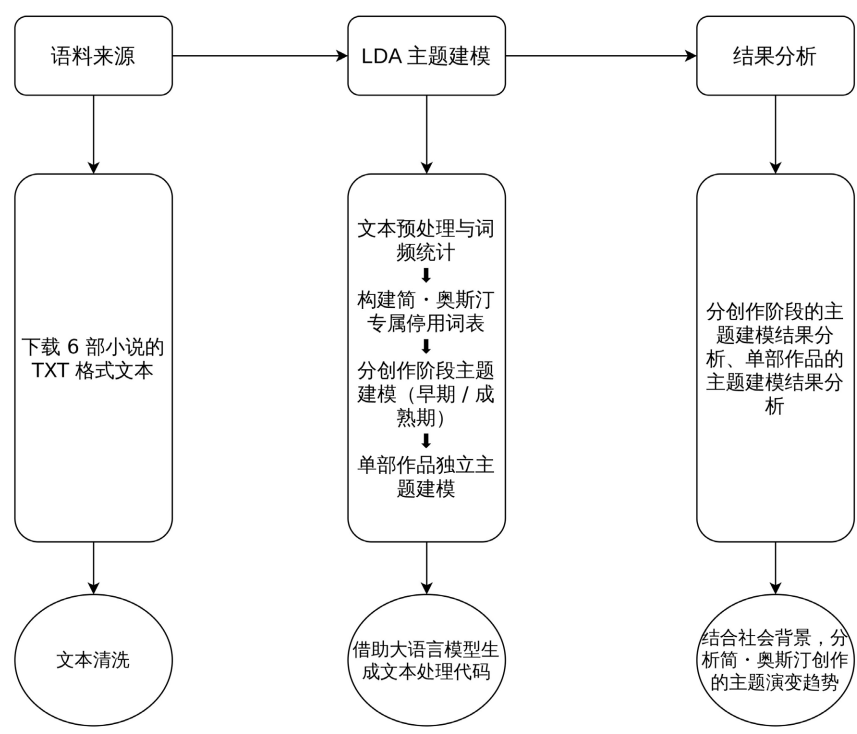


Figure 1. Analytical framework
图 1. 分析框架

3. LDA 建模结果与分析

3.1. 模型有效性验证

为验证第二步模型主题数选择的合理性，本研究遍历主题数 $k=2\sim5$ 并计算量化评估指标，一致性和困惑度，寻找一致性高、困惑度低的最佳主题数量，模型计算结果如表 2 所示。

如图 2 所示， $k=2$ 时一致性得分达到 0.2509，困惑度为-7.6508，综合判定为最优主题数。尽管整体一致性得分低于 0.4，该现象在小型文学语料建模中具备普遍性，根源为样本总量少、文学文本词汇灵活多变，与标准化新闻、专利文本词汇聚集规律存在差异。因此本文效度判断不单纯依靠量化阈值，重点对照主题词汇与小说叙事内容的匹配度完成人工校验。

Table 2. LDA model evaluation metrics with different topic numbers (k)
表 2. 不同主题数下的 LDA 模型评估指标(k)

主题数量(k)	一致性得分	困惑度	评估结果
2	0.2509	-7.6508	最优(最高/最低)
3	0.2471	-7.6369	次优
4	0.2384	-7.6369	较差
5	0.2393	-7.6372	较差

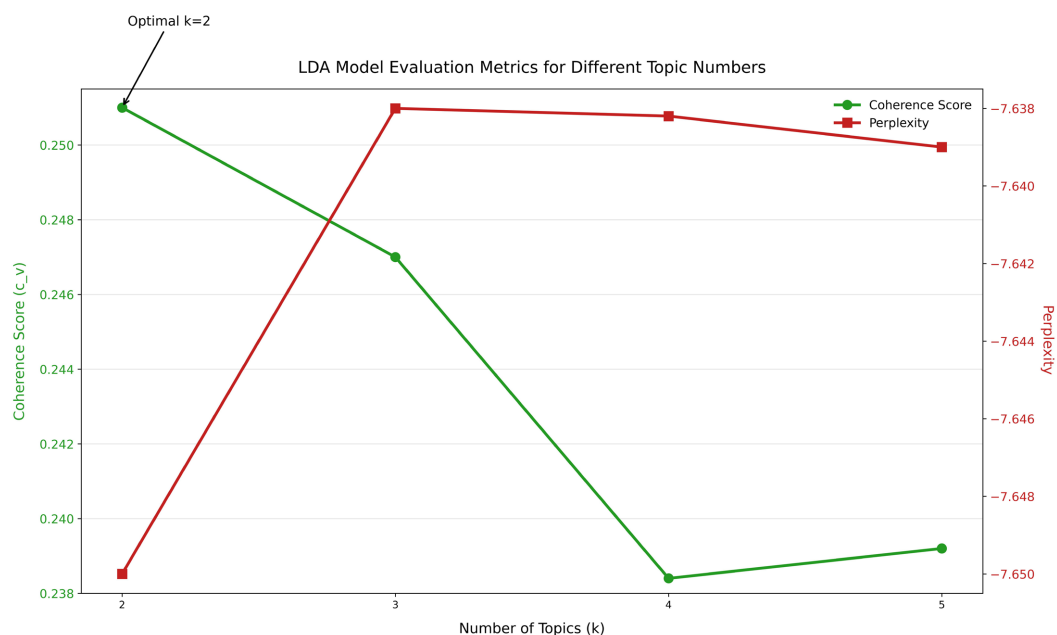


Figure 2. LDA model evaluation metrics
图 2. LDA 模型评估指标变化

3.2. 宏观阶段主题演变特征

阶段主题建模结果清晰呈现出奥斯汀早期与成熟期作品的主题边界，具体结果如表 3 所示，以及早期和成熟期的词云图见图 3 和图 4：

Table 3. Core vocabulary and weights of stage-specific topics
表 3. 阶段主题的核心词汇及权重

创作阶段	主题编号	核心词汇(权重)	核心主题特征
早期	主题 1	family (0.006), pleasure (0.004), love (0.004), opinion (0.004), affection (0.003), illustration (0.003), gentleman (0.003), engagement (0.003)	情感、社交、婚约类词汇集中
成熟期	主题 2	family (0.005), love (0.005), pleasure (0.005), idea (0.004), deal (0.003), doubt (0.003), circumstance (0.002), interest (0.002)	新增财产、处境、自我反思类词汇

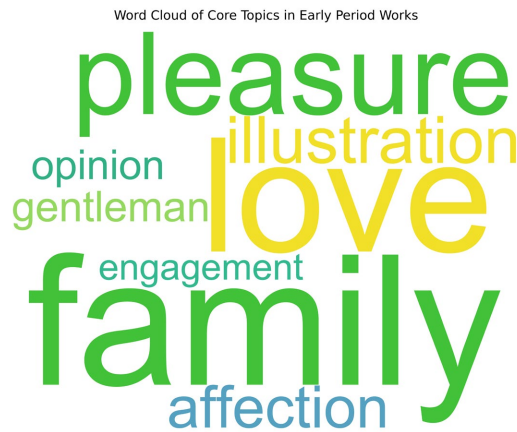


Figure 3. Word cloud of early period topics
图 3. 早期主题词云

早期文本高频出现 affection、illustration、engagement 等词汇，illustration 频次偏高仅代表文本对话、观点辩驳场景数量更多，结合细读可推论《傲慢与偏见》男女主多次争辩社交偏见是该词汇高频的典型场景；engagement 对应《理智与情感》埃莉诺的婚约故事，体现早期叙事侧重情感与婚恋契约。

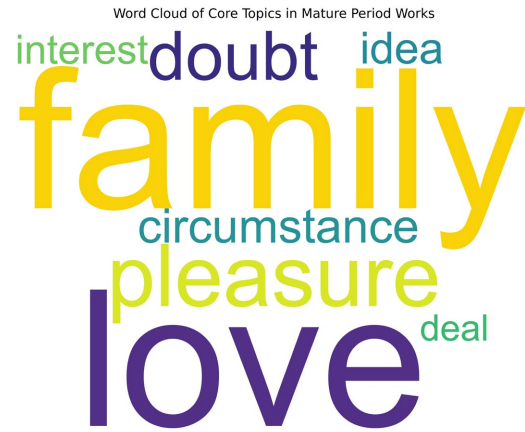


Figure 4. Word cloud of mature period topics
图 4. 成熟期主题词云

成熟期新增 deal、price、circumstance 等词汇，词汇分布线索结合文本可推论：《曼斯菲尔德庄园》大量书写财产婚配权衡，《劝导》侧重人物基于现实境遇做出婚恋抉择。分类数据显示六部小说可被完整划分至对应阶段，从统计层面印证创作叙事存在清晰的阶段性偏移。

3.3. 微观单部作品主题差异

基于最优主题数 $k=2$ 的单部作品建模结果，6 部作品在核心主题框架下展现出显著的微观差异，单篇建模提取各文本前十高频词，具体结果如表 4 所示：

Table 4. Top 10 feature words and frequencies of each work
表 4. 各作品 top 10 特征词及频次

创作阶段	作品名称	Top 10 特征词(频次)	核心特征
早期	《理智与情感》	affection (90), family (85), love (79), opinion (76), engagement (76), person (69), pleasure (68), comfort (64), life (63), behaviour (63)	情感、婚约词汇数值突出
	《傲慢与偏见》	family (159), illustration (153), pleasure (98), love (94), gentleman (76), opinion (75), netherfield (73), happiness (72), character (67), marriage (66)	家庭、辩论相关词汇频次最高
	《诺桑觉寺》	family (65), pleasure (52), life (45), love (44), gentleman (38), happiness (37), account (35), head (33), opinion (33), horse (33)	整体词频偏低，叙事轻松日常
成熟期	《曼斯菲尔德庄园》	family (139), love (132), price (122), pleasure (120), comfort (98), opinion (96), happiness (86), idea (80), end (78), deal (76)	财产相关词汇显著偏高
	《爱玛》	love (122), pleasure (120), idea (111), doubt (104), deal (90), family (84), opinion (77), person (77), life (75), happiness (74)	自我怀疑、反思类词汇突出
	《劝导》	family (85), life (57), character (51), idea (50), interest (45), deal (45), gentleman (43), circumstance (43), love (43), course (42)	侧重人生境遇与现实权衡

从表 4 的词汇统计结果能够看出三部早期小说存在明显的频次差异，但词频高低仅为表层数据线索，

文本内涵仍需结合文本细读推导。《理智与情感》affection、engagement 数值相对突出，可推论小说叙事更侧重真挚情感与婚约承诺的书写；《傲慢与偏见》family、illustration 出现频次更高，能够辅助说明作品大量围绕家庭关系、观点辩论展开叙事；《诺桑觉寺》整体高频词数值偏低、分布零散，这一现象与文本篇幅有限、通篇运用反讽戏仿手法密切相关。本文统计显示，abbey、mystery、horror 等哥特标志性词汇在该书中总频次远低于 family、pleasure 等日常用词。由于 LDA 采用词袋统计逻辑，仅能计算词汇出现次数，无法识别词语背后戏谑解构的修辞用意，单纯依靠建模结果极易误判文本创作立场，该实例充分体现模型在修辞识别上的短板，解读奥斯汀这类反讽文本必须辅以人工细读校正。

成熟期各部文本的词汇倾向同样存在区分度，全部词汇特征仅作为叙事倾向参考线索。《曼斯菲尔德庄园》price、deal 高频出现，可推论作品反复探讨财产条件对婚配的束缚；《爱玛》love、idea、doubt 等词汇数值靠前，对应主人公不断反思自我、调整婚恋观念的成长主线；《劝导》circumstance、interest 集中分布，从统计层面能够佐证小说聚焦中年人基于现实条件做出婚恋抉择的叙事特点。

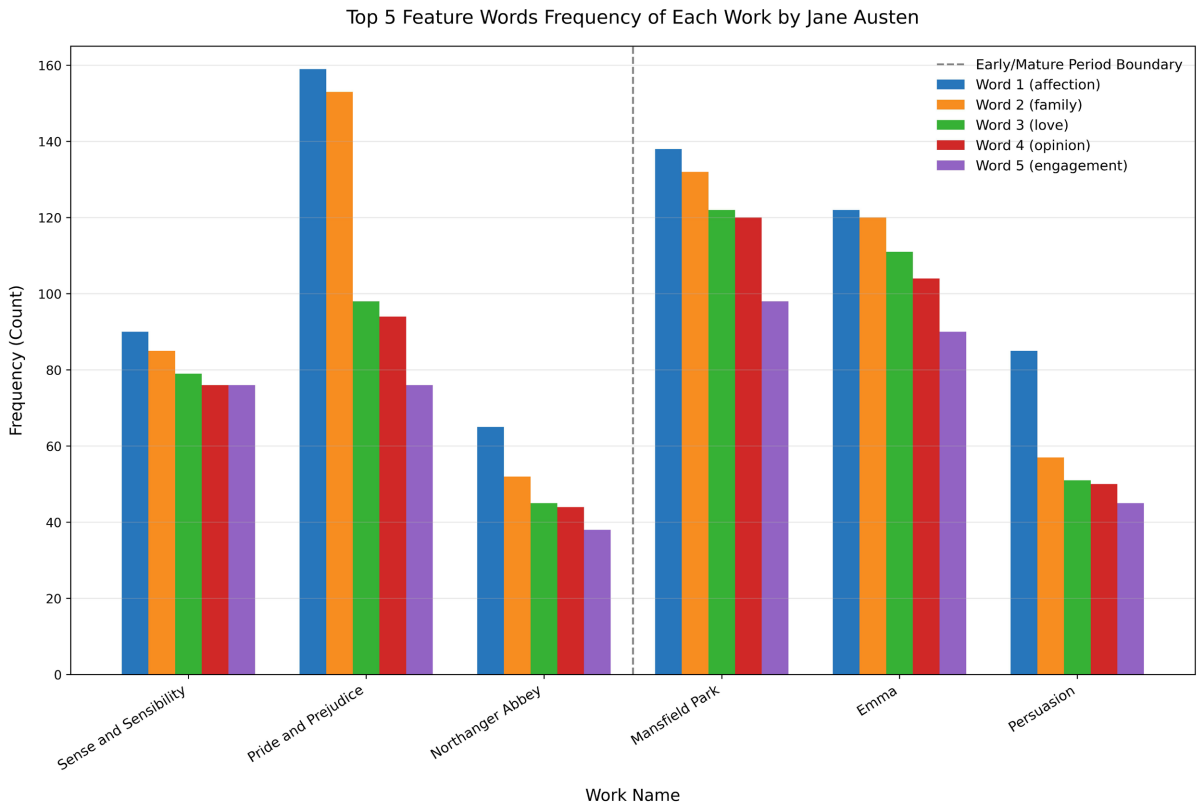


Figure 5. Top 5 feature words frequency
图 5. 各作品前 5 特征词频次对比

图 5 展示六部小说前五高频词对比，family 在《傲慢与偏见》中数值最高，仅能作为该作偏重家庭社交书写的统计线索；price 在《曼斯菲尔德庄园》频次突出，可辅助佐证文本频繁探讨财产对婚恋的影响。成熟期各文本词汇各有侧重，《爱玛》聚焦自我认知，《劝导》着眼中年现实抉择，量化数据直观展现奥斯汀单篇叙事的差异化设计。

4. 主题演化讨论

简·奥斯汀的创作生涯(1795~1815)横跨英国社会转型阶段，工业发展仅缓慢改变乡村财富格局，当

时固定的民事法律制度才是约束人物婚恋、生存选择的核心因素。在财产继承层面，长子继承制与限定继承制限定土地仅由男性承袭，《傲慢与偏见》班纳特家族便是典型案例；婚姻领域通行夫妻一体原则，女性婚后财产全部归丈夫所有，19 世纪中后期相关法案才放宽女性财产权限。这套制度决定婚姻是女性获取稳定生活的核心途径，文本叙事重心的变化可从制度约束、文学思潮两条线索解读。

从社会制度层面来看，1795~1798 年汉普郡当地土地格局稳定，情感共鸣与社会礼仪仍是婚恋主流评判标准，早期文本 affection、gentleman 高频分布。结合词汇线索与文本可推论，《理智与情感》侧重纯粹爱恋与婚约坚守，《傲慢与偏见》围绕身份偏见展开大量对话辩论。1811 年后财富流动差异扩大，财产矛盾凸显，deal、price、circumstance 成为成熟期高频词汇，《曼斯菲尔德庄园》书写财产干预婚配，《爱玛》反思盲目婚恋撮合，《劝导》刻画成年人基于现实条件的理性相爱。

从文学思潮演变来看，奥斯汀的创作转变契合感伤主义向写实小说过渡的整体脉络，且本文 LDA 建模结果可与文体特征形成双向印证。18 世纪感伤文学普遍高频出现 love、affection 等情感词汇，与奥斯汀早期词汇分布完全匹配；19 世纪写实创作增多财产、处境类表述，对应成熟期词汇变化。横向对比拉德克利夫、埃奇沃思同期作品能够发现，由纯粹情爱书写转向现实生存探讨是当时小说整体创作趋势。奥斯汀全程以道德德性为评判标尺，早期赞颂真挚情感，成熟期倡导理性抉择，两类叙事均完成对功利化婚恋的反思。

综合来看，奥斯汀创作主题的阶段性变化并非单一因素作用的结果，外部刚性法律制度划定了女性生存选择的边界，内部文学文体发展又推动叙事视角持续下沉，二者相互交织，共同促成文本词汇与叙事重心由纯粹情感向现实生存问题的偏移，也让其小说成为映照 19 世纪初普通人婚恋生存状态的典型文学样本。

5. 结论

本文依托分层 LDA 建模对奥斯汀六部小说开展量化统计，结合 19 世纪初英国财产、婚姻法律与小说文体发展脉络完成综合分析。量化词汇分布能够将奥斯汀创作清晰划分为两大阶段，早期文本高频出现 affection、gentleman、engagement 等情感、婚约相关词汇，叙事重心偏向婚恋内部的情感与礼仪探讨；成熟期文本新增 deal、price、circumstance 等指向财产、生存处境的词汇，文本更侧重书写婚恋选择背后的现实生存约束，这种阶段性分化的核心成因是当时约束女性财产权益的各类法律制度，社会财富流动仅作为长期宏观时代背景。同一创作阶段内部的文本也存在明显微观词汇差异，早期作品分别侧重纯粹情感表达、社交观念辩驳与轻松日常喜剧叙事，成熟期作品则围绕财产婚配捆绑、人物自我认知、中年人生现实权衡形成差异化书写，量化数据直观印证奥斯汀创作叙事的多元特质。

本文搭建的分层建模搭配多重语境校验的分析流程，证实 LDA 模型可作为小型经典文学语料的辅助统计工具，整套分析方法具备良好迁移性，能够为其他叙事文本的主题演化对比研究提供可复用的计算语言学分析思路。

本研究存在一定局限：LDA 词袋模型无法识别反讽、戏仿修辞，量化结果仅能作为文本细读的辅助材料；研究语料仅选取六部完整长篇，未纳入残稿与书信，分析素材存在局限，且仅单一使用 LDA 模型，量化维度较为单薄。后续研究可融合情感、句法分析工具完善量化模型，扩充奥斯汀书信、残稿语料，并开展同期作家对比与译本主题分析，进一步拓展相关研究维度。

参考文献

- [1] Bailey, M. (2015) The Marriage Law of Jane Austen's World. *Persuasions: The Jane Austen Journal On-Line*, 36. <https://jasna.org/publications-2/persuasions-online/vol36no1/bailey/>
- [2] Manimozhi, V. (2020) Marriage Money and Society in the Novels of Jane Austen. *Malaya Journal of Matematik*, S, 2366-2368.

-
- [3] 何志燕, 陈剑雯. 建筑在金钱基础上的婚姻与爱情——浅析简·奥斯汀文学作品的主题[J]. 江西师范大学学报(哲学社会科学版), 2008, 41(6): 81-84.
- [4] 黄福奎. 简·奥斯汀后期作品中婚恋观变化刍探[J]. 英语广场, 2017(8): 19-20.
- [5] Wijitsopon, R. (2013) A Corpus-Based Study of the Style in Jane Austen's Novels. *Manusya*, **16**, 41-64. <https://doi.org/10.1163/26659077-01601003>
- [6] Fischer-Starcke, B. (2009) Keywords and Frequent Phrases of Jane Austen's *Pride and Prejudice*: A Corpus-Stylistic Analysis. *International Journal of Corpus Linguistics*, **14**, 492-523. <https://doi.org/10.1075/ijcl.14.4.03fis>
- [7] Rha, L. and Silver, S. (2021) Topic Modeling and Analysis: Comparing the Most Common Topics in 19th-Century Novels Written by Female Writers. *Aresty Rutgers Undergraduate Research Journal*, **1**, 1-7. <https://doi.org/10.14713/arestyrurj.v1i3.172>