LDA的操作与运用

2nd Study 2025.04.03

前期准备

1.安装PyCharm

如果尚未安装PyCharm，请下载并安装，见群内教程

2.将文件复制到桌面

在开始之前，确保您的Python文件(.py)和相关数据源已复制到桌面上

请复制 数据1.CSV 和 LDA代码.py 到桌面

直接打开Python文件

1.找到桌面上的Python文件

1. 在桌面上找到.py文件

2.右键点击文件

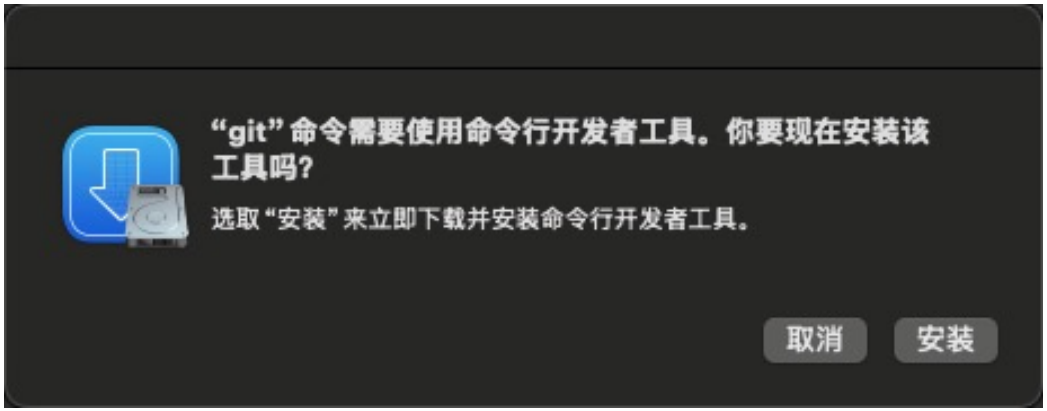
1. 右键点击该文件，将出现上下文菜单
2. 选择"打开方式"或"Open with"选项

3.选择PyCharm

1. 在应用程序列表中选择PyCharm
2. 如果PyCharm不在列表中，请选择"选择其他应用"并浏览找到PyCharm可执行文件

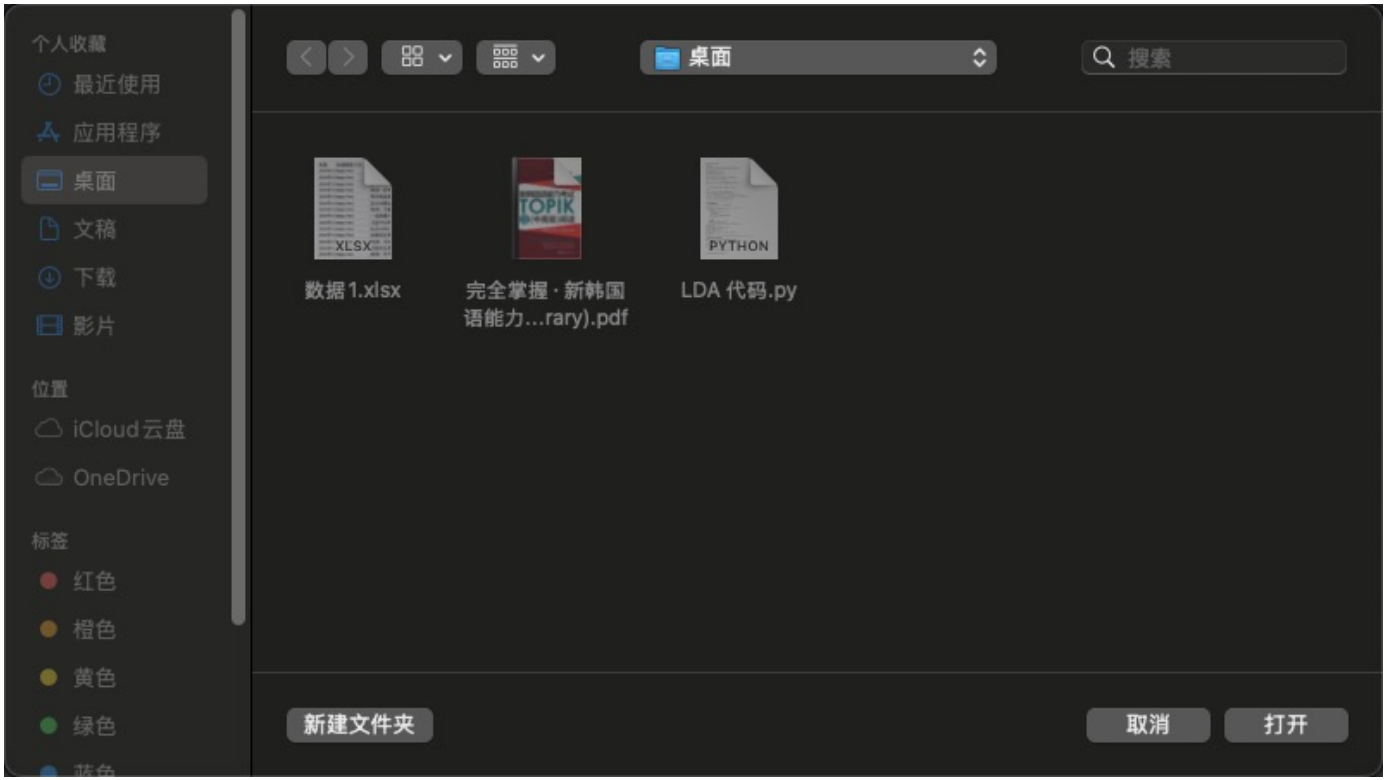
4.文件将在PyCharm中打开

1. PyCharm将启动并打开Python文件



⬅️Mac提示请点击取消

⬅️请点击在项目中打开



- 打开之后选择一个路径保存（不要选择桌面），日后所有相关的都在这个地方，可以理解为代码的保存库

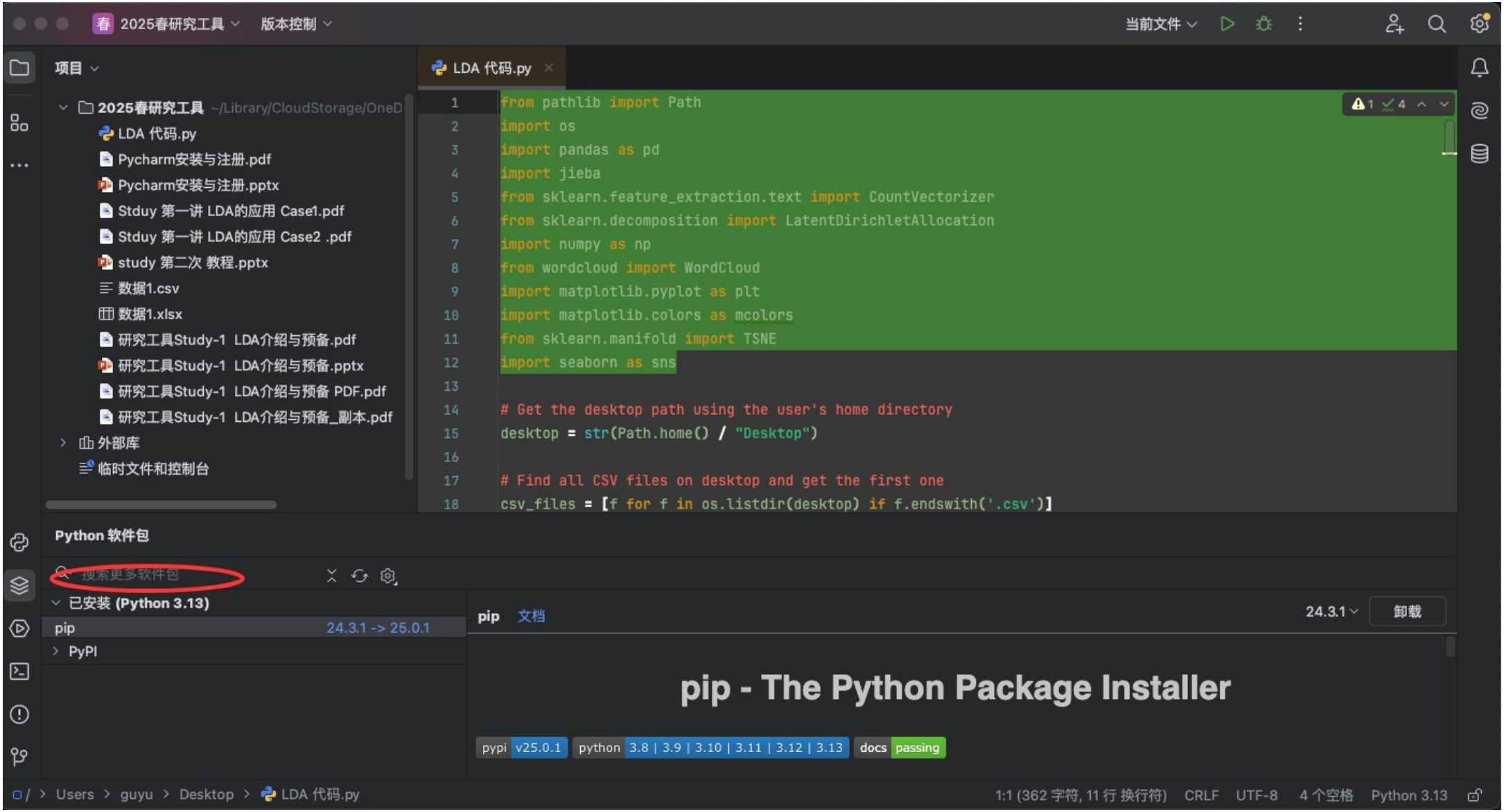
```python
from pathlib import Path
import os
import pandas as pd
import jieba
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
import numpy as np
from wordcloud import WordCloud
import matplotlib.pyplot as plt
import matplotlib.colors as mcolors
from sklearn.manifold import TSNE
import seaborn as sns

```

- from pathlib import Path # 用于处理文件路径，是比os.path更现代的路径处理方式
- import os # 提供与操作系统交互的功能
- import pandas as pd # 用于数据分析和处理表格数据
- import jieba # 中文分词库，用于将中文文本分割成单词
- from sklearn.feature_extraction.text import CountVectorizer # 将文本转换为词频矩阵
- from sklearn.decomposition import LatentDirichletAllocation # LDA主题模型，用于发现文本中的主题
- import numpy as np # 科学计算库，提供高效的数组操作
- from wordcloud import WordCloud # 生成词云图
- import matplotlib.pyplot as plt # 绘图库，用于数据可视化
- import matplotlib.colors as mcolors # matplotlib的颜色模块
- from sklearn.manifold import TSNE # 用于降维可视化
- import seaborn as sns # 基于matplotlib的高级可视化库

import **jieba** ⬅️**Python包**

Python包管理概念

- **pip**: Python的标准包管理工具
- **包/库**: 别人写好的代码模块，可直接导入使用
- **requirements.txt**: 项目依赖文件，列出所有需要的库
- **虚拟环境**: 隔离的Python环境，避免包版本冲突


2025春研究工具
版本控制
当前文件
项目
2025春研究工具 ~/Library/CloudStorage/OneD
LDA 代码.py
Pycharm安装与注册.pdf
Pycharm安装与注册.pptx
Stduy 第一讲 LDA的应用 Case1.pdf
Stduy 第一讲 LDA的应用 Case2 .pdf
study 第二次 教程.pptx
数据1.csv
数据1.xlsx
研究工具Study-1 LDA介绍与预备.pdf
研究工具Study-1 LDA介绍与预备.pptx
研究工具Study-1 LDA介绍与预备 PDF.pdf
研究工具Study-1 LDA介绍与预备_副本.pdf
外部库
临时文件和控制台
LDA 代码.py
from pathlib import Path
import os
import pandas as pd
import jieba
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.decomposition import LatentDirichletAllocation
import numpy as np
from wordcloud import WordCloud
import matplotlib.pyplot as plt
import matplotlib.colors as mcolors
from sklearn.manifold import TSNE
import seaborn as sns
# Get the desktop path using the user's home directory
desktop = str(Path.home() / "Desktop")
# Find all CSV files on desktop and get the first one
csv_files = [f for f in os.listdir(desktop) if f.endswith('.csv')]
Python 软件包
搜索更多软件包
已安装 (Python 3.13)
pip 24.3.1 -> 25.0.1
PyPI
pip 文档
24.3.1
卸载
pip - The Python Package Installer
pypi v25.0.1
python 3.8 | 3.9 | 3.10 | 3.11 | 3.12 | 3.13
docs passing
/ > Users > guyu > Desktop > LDA 代码.py
1:1 (362 字符, 11 行 换行符)
CRLF
UTF-8
4 个空格
Python 3.13

初学者可以使用PyCharm界面

1.点击菜单 **File > Settings** (Windows/Linux) 或 **PyCharm > Preferences** (macOS)

2.导航至 **Project: [项目名] > Python Interpreter**

3.点击 **+** 按钮打开包安装对话框

4.搜索并选择要安装的包

5.点击 **Install Package** 按钮

numpy pandas matplotlib seaborn jieba scikit-learn wordcloud

有基础的使用终端
在PyCharm底部工具栏找到 Terminal 标签页
点击打开内置终端
输入pip安装命令

- # 基础科学计算和数据处理库
- pip install numpy pandas matplotlib seaborn

- # 中文分词库
- pip install jieba

- # 机器学习库
- pip install scikit-learn

- # 词云生成库
- pip install wordcloud

代码以及实现功能讲解

```python
# 查找桌面上的CSV文件并加载第一个
csv_files = [f for f in os.listdir(desktop) if f.endswith('.c
file_path = os.path.join(desktop, csv_files[0])

# 处理不同编码
try:
    df = pd.read_csv(file_path, encoding='utf-8-sig')
except:
    df = pd.read_csv(file_path, encoding='gbk')

# 使用jieba进行中文分词
df['cut'] = df['摘要'].apply(lambda x: ' '.join(jieba.lcut(str

# 将文本转换为词频矩阵
vectorizer = CountVectorizer(max_df=0.95, min_df=2)
tf = vectorizer.fit_transform(df['cut'])
```

预处理

【摘要】所在的列，
所以要读取的列名应该
【摘要】

你也可以DIY成别的

LDA 与 困惑度

python Copy

```python
# 计算不同主题数量(2-10)的困惑度
n_topics_range = range(2, 11)
perplexities = []

for n_topics in n_topics_range:
    lda_model = LatentDirichletAllocation(
        n_components=n_topics,
        random_state=0,
        max_iter=10,
        learning_method='online'
    )
    lda_model.fit(tf)
    perplexity = lda_model.perplexity(tf)
    perplexities.append(perplexity)

# 基于最小困惑度找到最佳主题数量
optimal_topics = n_topics_range[np.argmin(perplexities)]
```

困惑度（Perplexity）是评估主题模型（如LDA）质量的重要指标，它可以帮助我们确定最佳的主题数量。

困惑度的定义

困惑度本质上是对模型预测能力的衡量，它源自信息论，表示为：

Perplexity = exp(-平均对数似然)

- 在LDA主题模型中，较低的困惑度值表示模型对文档的拟合程度更好。

困惑度越低，这个主题合理。
但是注意：困惑度如果单调递增则可以认定无最佳主题数，可以人为筛选或引用其他指标

```python
# 计算不同主题数量(2-10)的困惑度
n_topics_range = range(2, 11)
perplexities = []

for n_topics in n_topics_range:
    lda_model = LatentDirichletAllocation(
        n_components=n_topics,
        random_state=0,
        max_iter=10,
        learning_method='online'
    )
    lda_model.fit(tf)
    perplexity = lda_model.perplexity(tf)
    perplexities.append(perplexity)

# 基于最小困惑度找到最佳主题数量
optimal_topics = n_topics_range[np.argmin(perplexities)]
```

基于最小困惑度找到最佳主题数量

```python
# 使用最佳主题数量训练最终LDA模型
lda = LatentDirichletAllocation(
    n_components=optimal_topics,
    random_state=0,
    max_iter=50
)

# 将文档转换到主题空间
doc_topics = lda.fit_transform(tf)
```

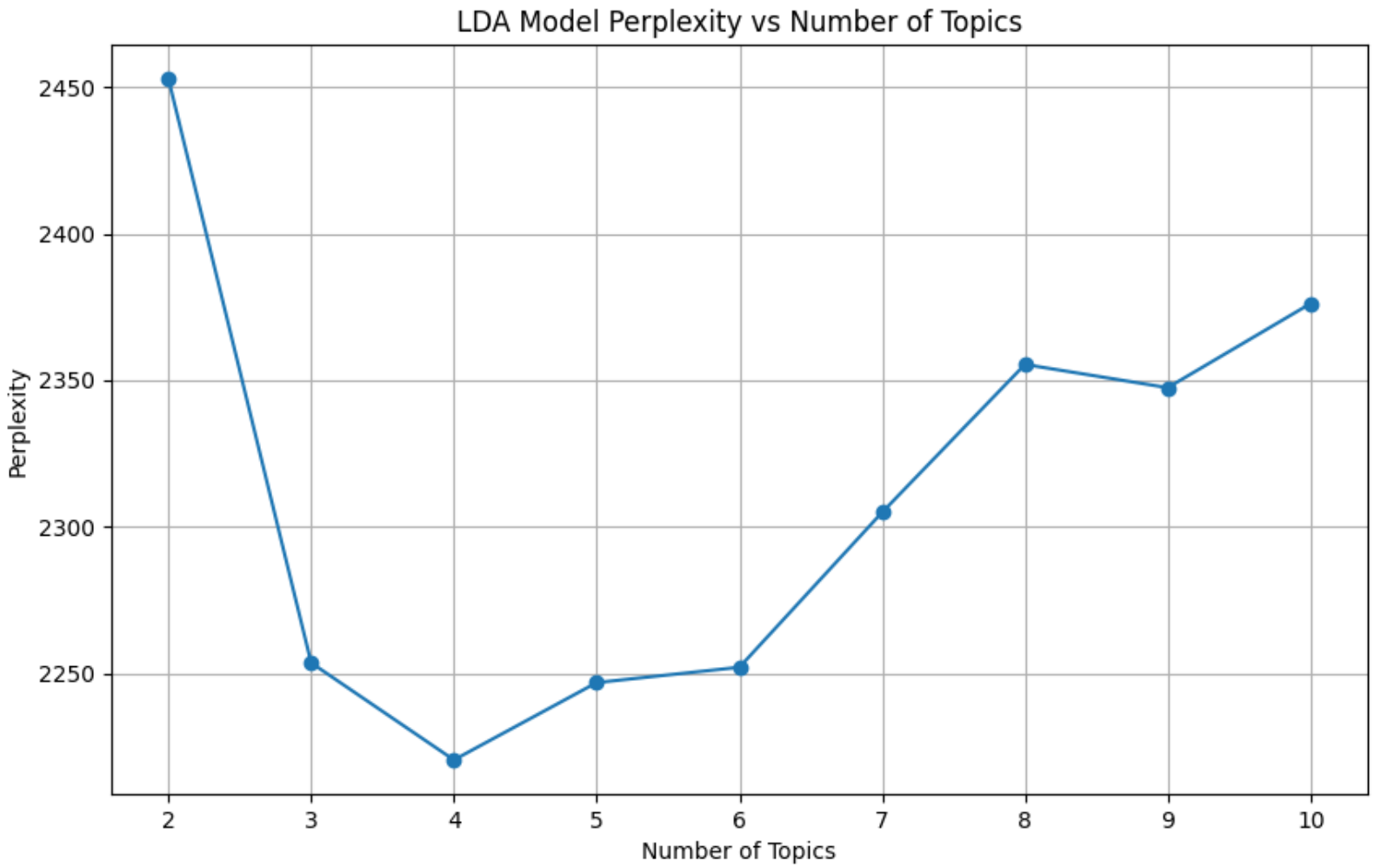
Copy

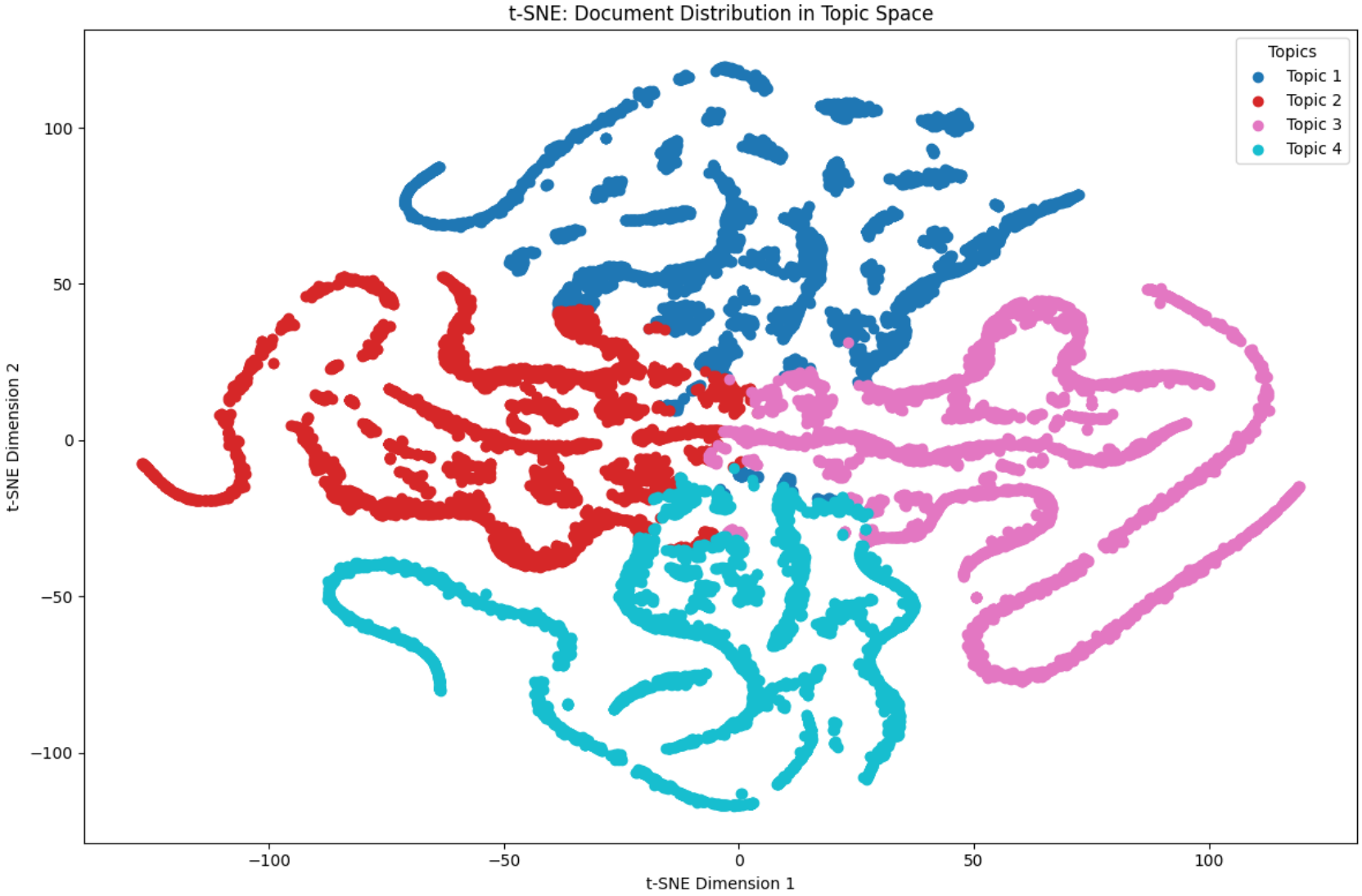
最终LDA模型训练

```python
# 使用t-SNE将文档映射到二维空间进行可视化
tsne = TSNE(n_components=2, random_state=0,
            perplexity=min(30, len(doc_topics) - 1))
tsne_results = tsne.fit_transform(doc_topic_features)

# 为每个文档分配主要主题
doc_main_topic = np.argmax(doc_topics, axis=1)

# 创建散点图可视化不同主题的文档分布
plt.figure(figsize=(12, 8))
for topic_idx in range(n_final_topics):
    indices = np.where(doc_main_topic == topic_idx)[0]
    plt.scatter(
        tsne_results[indices, 0],
        tsne_results[indices, 1],
        label=f'Topic {topic_idx + 1}'
    )
```

主题可视化 - t-SNE

- # 生成的文件:
- # 1. 困惑度曲线: perplexity_curve.png
- # 2. 主题分布t-SNE可视化: topic_distribution_tsne.png
- # 3. 主题相关性热图: topic_correlation.png
- # 4+ 各主题词云: topic_1_wordcloud.png, topic_2_wordcloud.png, ...
- # 5. 详细文本分析报告: topic_distribution.txt


LDA Model Perplexity vs Number of Topics
Perplexity
2450
2400
2350
2300
2250
2
3
4
5
6
7
8
9
10
Number of Topics


t-SNE: Document Distribution in Topic Space
Topics
Topic 1
Topic 2
Topic 3
Topic 4
t-SNE Dimension 2
t-SNE Dimension 1
100
50
0
−50
−100
−100
−50
0
50
100

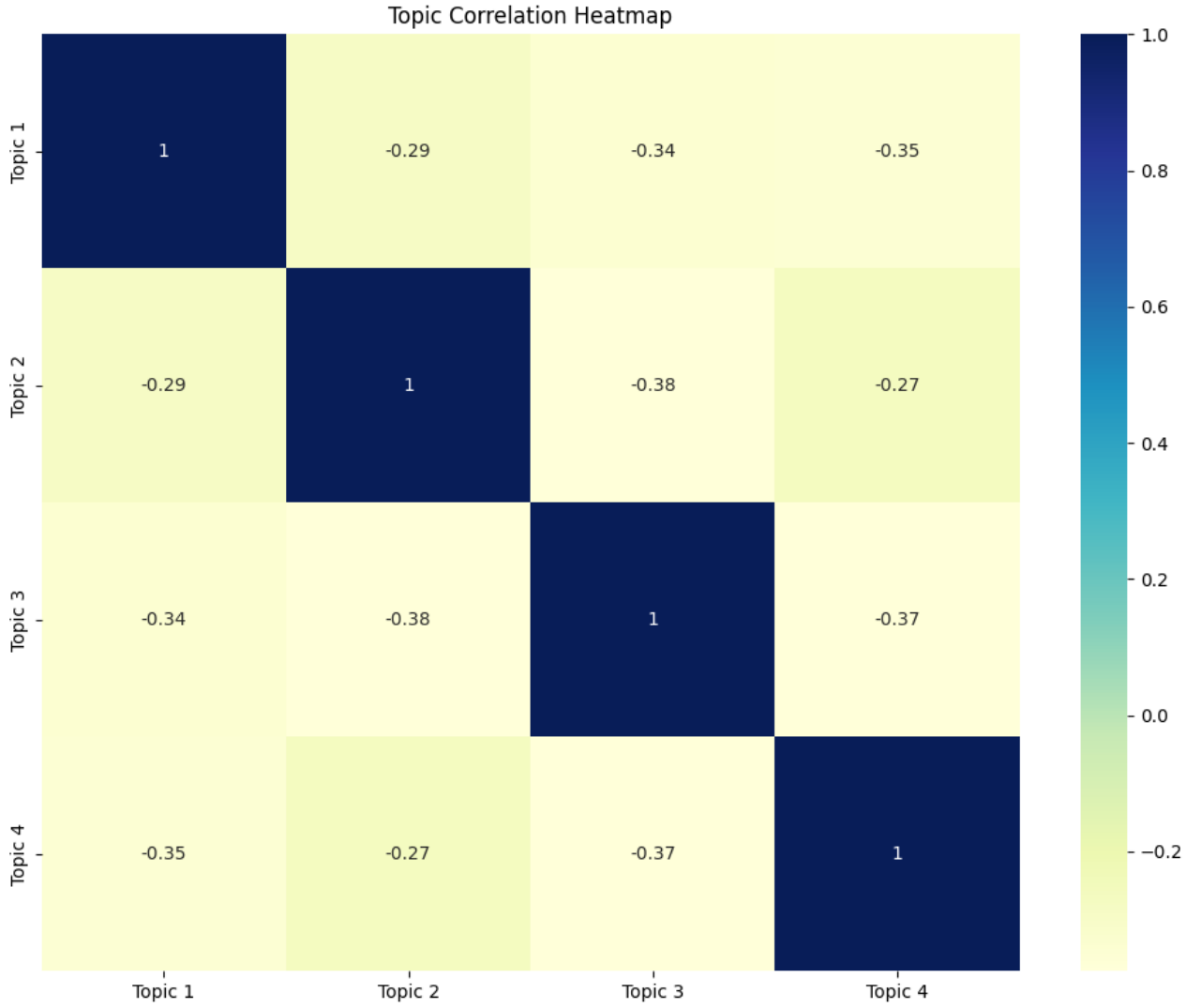

Topic 1
我们
问题
双方
保持
和平
通过
局势
继续
推动
支持
危机
解决
立场
各方
对话
作用
希望
王毅
政治
安全
地区
冲突
俄罗斯
发挥


Topic 2
企业
机构
事件
措施
媒体
日本
报告
人员
政策
政府
没有
实施
包括
安全
核污染
监测
我们
近日
卫视
疫情
调查
表示
计划
公民
香港

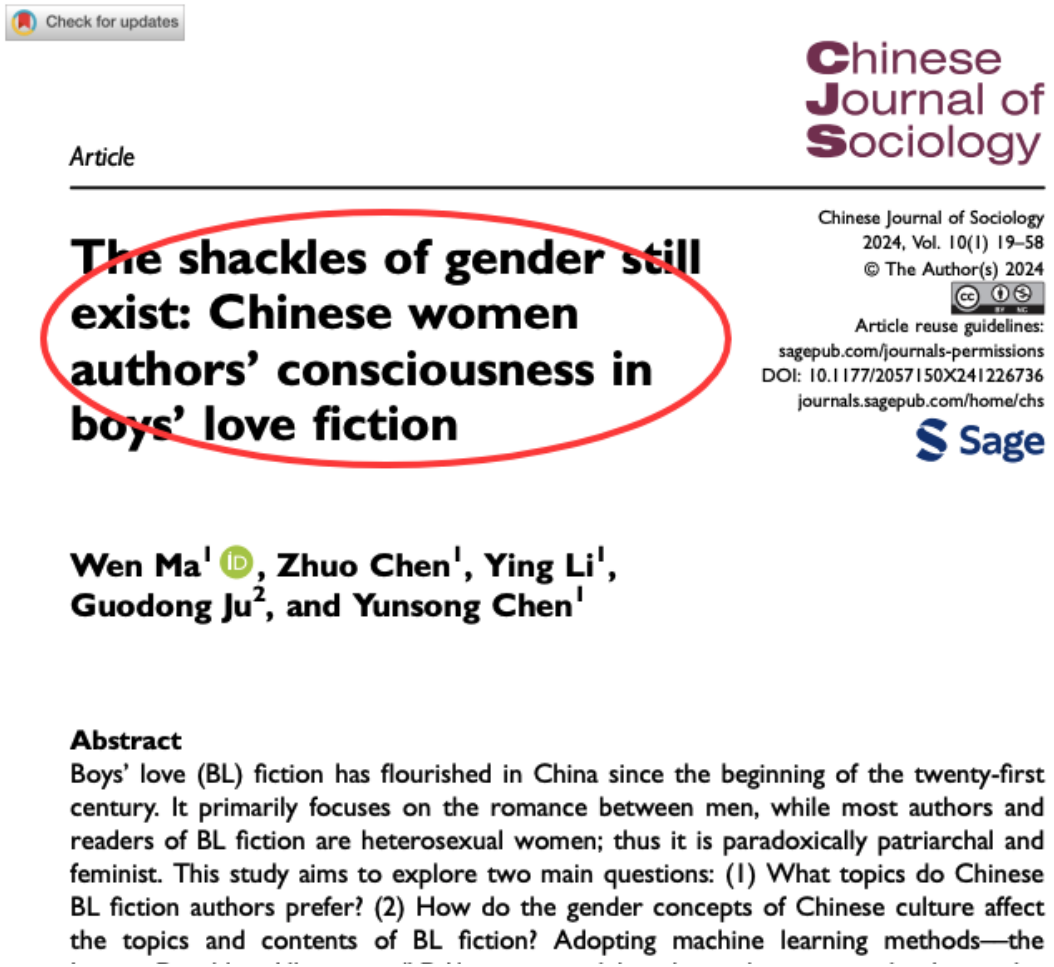

Topic 3
带
推动
战略
持续
稳定
倡议
共同
各国
加强
共建
世界
重要
积极
促进
共同体
一路
我们
领域
合作
命运
一带
发展
建设
次
地区
国际
全面
人民
安全
双方

Topic 4


原则
国际
台湾
坚决
社会
和平
美方
反对
问题
世界
严重
自身
所谓
维护
地区
安全
制裁
停止
主权
美国
我们
政治
台独
稳定
支持
任何
人权
不是
敦促


Topic Correlation Heatmap
Topic 1
Topic 2
Topic 3
Topic 4
1
-0.29
-0.34
-0.35
-0.29
1
-0.38
-0.27
-0.34
-0.38
1
-0.37
-0.35
-0.27
-0.37
1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

阅读分析



Check for updates

Chinese Journal of Sociology

Article

Chinese Journal of Sociology
2024, Vol. 10(1) 19–58

DOI: 10.1177/2057150X241226736
journals.sagepub.com/home/chs

Sage

The shackles of gender still exist: Chinese women authors' consciousness in boys' love fiction

Wen Ma[1], Zhuo Chen[1], Ying Li[1], Guodong Ju[2], and Yunsong Chen[1]

Abstract

Boys' love (BL) fiction has flourished in China since the beginning of the twenty-first century. It primarily focuses on the romance between men, while most authors and readers of BL fiction are heterosexual women; thus it is paradoxically patriarchal and feminist. This study aims to explore two main questions: (1) What topics do Chinese BL fiction authors prefer? (2) How do the gender concepts of Chinese culture affect the topics and contents of BL fiction? Adopting machine learning methods—the

探究中国BL小说

主题选择与性别观念的机器学习研究

研究概述

- ✓ **研究对象**：72,548部中国原创BL小说（2003-2019年）
- ✓ **数据来源**：晋江文学城（最大的中文BL小说平台）
- ✓ **研究方法**：潜在狄利克雷分配(LDA)主题建模和词向量技术
- ✓ **研究目标**：
  - 探究中国BL小说作者偏好的主题
  - 分析中国文化中的性别观念如何影响BL小说的主题和内容

BL小说的背景

- **BL**：Boys' Love，专注于男性之间的浪漫关系
- **创作者与读者**：主要是异性恋女性（"腐女"）
- **悖论性**：同时包含父权制和女权主义元素
- **数据特点**：晋江文学城有超过8万部BL小说，5095万注册用户，女性用户占91%

研究方法

- **LDA主题模型**：将文本聚类为主题
  - 文本集（语料库）→ 文档 → 词语
  - 每个主题是词语的概率分布
  - 每个文档是主题的概率分布
- **词向量技术**：预测与性别相关关键词的具体语境
  - Word2Vec模型捕捉词语间的距离
  - Skip-gram算法计算词向量间的余弦相似度

中国BL小说的典型主题（同步分析）

1.校园：纯真恋爱、师生关系

2.童话：西方背景的浪漫故事

3.中国神话：轮回、前世今生、道教元素

4.古代中国：帝王将相、江湖门派

5.日常幸福：生活中可及的幸福

6.戏剧性爱情：总裁、狗血剧情

7.娱乐圈：明星、老板、婚姻

8.游戏：网游、生存游戏

9.奇幻：未来世界、超自然元素

10.故事标签：功能性主题（如HE、悲剧）

- 在排名前四分之一的作品中，新增了**A/B/O**主题（Alpha/Beta/Omega）

性别相关语境分析：女性（同步分析）

与"女人"最接近的词：

- **同妻**（排名第一）：嫁给同性恋男性的异性恋女性
- **婚姻相关**：待嫁、初夜、嫁了人
- **生育相关**：母爱、母子、传宗接代、生小孩
- **道德规范**：好人家、不孕不育被视为"异端"
- 女性在BL小说中主要与婚姻和生育相关联，成为工具性角色

性别相关语境分析：男性（同步分析）

与"男人"最接近的词：

- **女人**（相互关联）
- **婚姻相关**：婚纱、打光棍

与"男性"最接近的词：

- **生育相关**：生育、生育能力、受孕
- **双性人**：卵生、药物
- **性别流动**：雌雄、不分性别
- 男性被赋予女性的生理特征，尤其是生育能力

性别相关词汇的历时变化（2003-2019）

- **"同妻"**：2010年后才出现
- **"传宗接代"**：频率从2003年到2019年逐渐下降
- **"生育能力"**：2009年首次出现，2013年后迅速增加
- **"双性人"**：与"生育能力"趋势相似，表明两个主题相互关联
- 热门作品中，父权制约束更明显，商业市场仍偏好婚姻和生育相关主题

主要研究发现

1.中国女性BL作者偏好超现实主题

1. 历史和未来背景
2. "爱情神话"展现异性恋特征

2.中国传统文化强烈影响BL小说

1. 创造独特美学风格
2. 宗族观念促使作者关注婚姻和生育

3.中国BL小说揭示作者性别观念的悖论性

1. 通过男性角色感受力量
2. 将生育能力转移给男性
3. 回避讨论自身性别，仍受父权制束缚

研究意义

1.理论贡献：通过女性主义视角研究BL小说，展示中国传统文化和伦理如何影响当代文学创作

2.方法论贡献：机器学习方法实现全面、稳健的文本分析，避免忽视小众作品

3.社会意义：揭示中国女性通过文学创作展现的矛盾性别观念：

1. 既渴望反抗父权制
2. 又无法完全摆脱宗族观念的影响

读者访谈补充发现

- 15位22-32岁异性恋女性读者
- **确认发现**：偏好中国背景的超现实作品
- **确认发现**：承认女性角色在BL小说中的边缘化
- **有差异的发现**：
  - 12位受访者表达对男性生育和A/B/O主题的反感
  - 大多数受访者认为阅读BL小说只是消遣
  - 一位作者兼读者认为写作比阅读更能唤醒女性主义意识

结论

BL小说的悖论性：

- 女性作者创作男男爱情，但与同性恋现实生活脱节
- 表达女性欲望，但缺乏女性角色
- 唤醒某种女性主义意识，但仍保留父权制婚姻观和生育观

后现代叙事中：

- 现实与虚构的界限模糊
- BL小说试图消除女性存在，但最奇幻的情节也无法逃脱作者的生活经历
- 即使在女性创作的男男爱情中，厌女情绪仍然存在