한양대학교
HANYANG UNIVERSITY

研究工具Study-1 LDA介绍与预备

目标：我们使用2-3周实现掌握LDA的使用，
以及了解Pycharm（Python工具）的简单使用

한양대학교
HANYANG UNIVERSITY

SSCI (Social Science Citation Index)

SSCI收录的期刊代表了**社会科学领域最具影响力的研究**，阅读这些文献有助于了解学科最新发展方向和前沿成果。提高学术研究质量。

SSCI文献通常经过严格的同行评审，**研究方法更加规范，结论更加可靠**，为我们的研究提供高质量的参考。拓宽研究视野。

英文SSCI文献来自全球各地的研究者，提供了多元化的研究视角和跨文化的比较研究，**有助于拓宽我们的学术视野**。

我英文不好？英文文献对我不友好？要不要放弃英文文献？

PC端操作步骤

1.下载安装有道词典：访问官网下载最新版PC客户端

2.启动截屏翻译：

- 快捷键：按下Ctrl+Alt+D组合键
- 菜单启动：点击主界面的"截图翻译"按钮

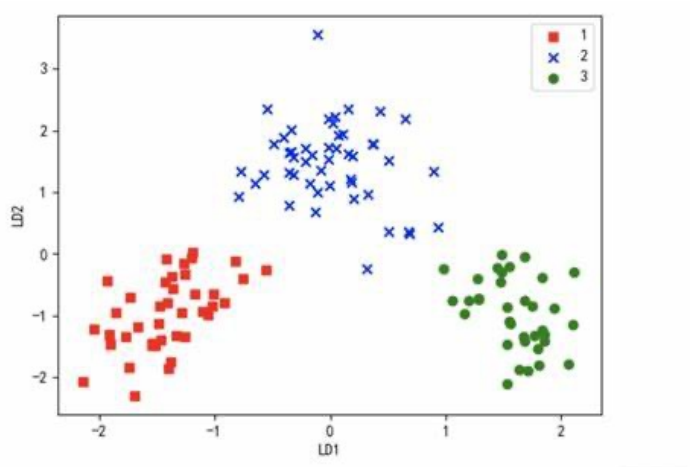
3.选择需要翻译的区域：鼠标拖拽选择文献中需要翻译的部分

4.查看翻译结果：翻译结果会在弹出窗口中显示

什么是LDA?

定义

- **LDA （Latent Dirichlet Allocation）** - 潜在狄利克雷分配
- 由David Blei、Andrew Ng和Michael Jordan于2003年提出
- 一种生成式概率模型，属于**主题模型**的一种
- 一种**无监督机器学习算法**，无需标注数据
- 用于从文本集合中**自动发现潜在主题结构**



LDA的核心思想

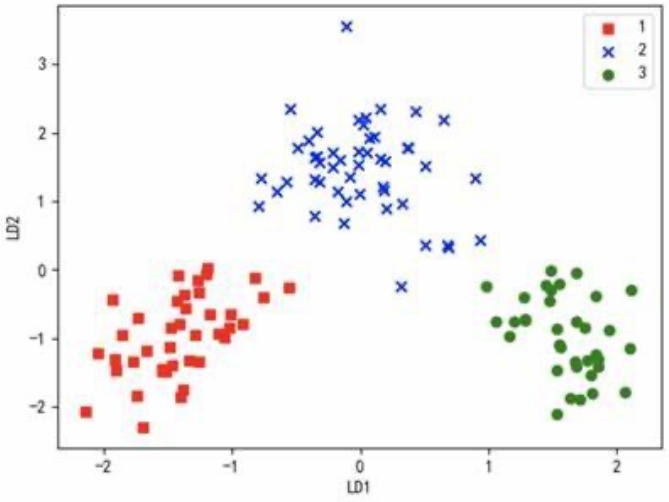
文档是主题的混合

•每个文档可以表示为主题的概率分布

•例如：文档1 = 30%主题A + 50%主题B + 20%主题C

•不同文档对主题的权重不同

•这种表示捕捉了文档的语义结构



主题是词汇的混合分布

•每个主题可以表示为词汇的概率分布

•例如：主题A = 0.15×“经济” + 0.12×“增长” + 0.10×“投资” + ...

•高概率词汇共同定义了主题的语义内容

•同一个词可以在多个主题中以不同概率出现

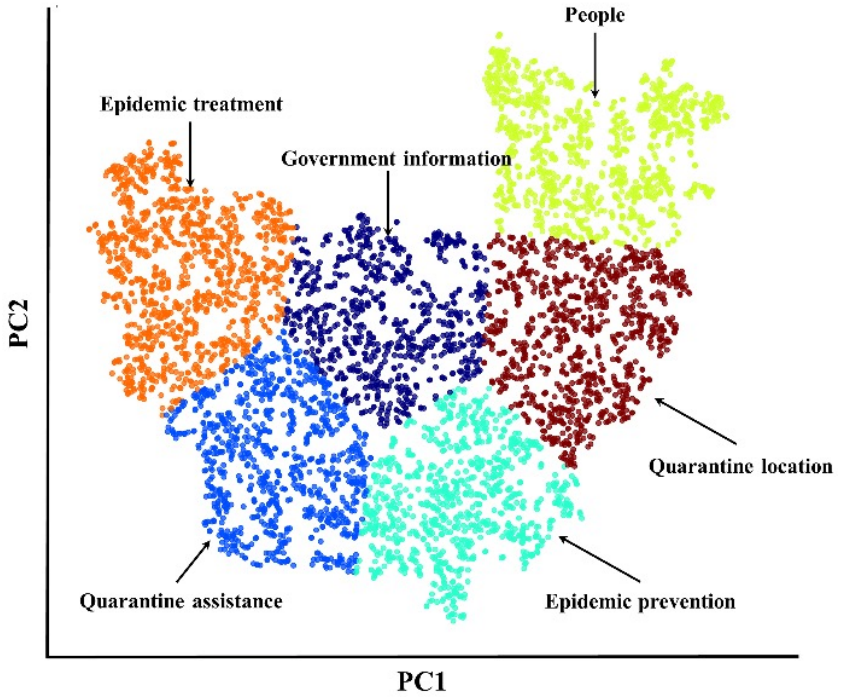
LDA的训练过程

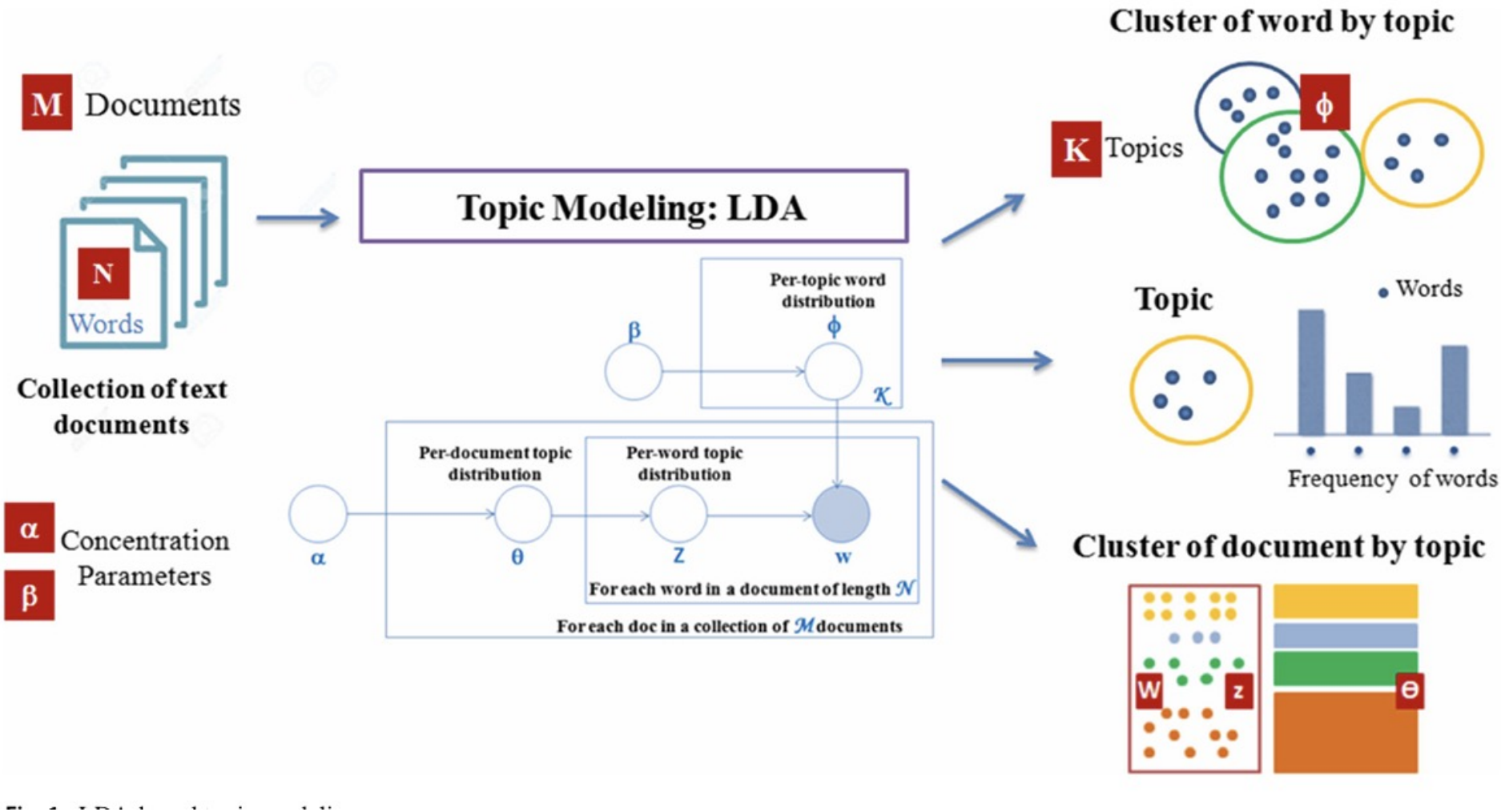
训练目标

- 根据观察到的词语，估计隐含的主题结构
- 具体目标：估计每个文档的主题分布θ和每个主题的词分布φ

常用训练算法

- **变分推断**（Variational Inference）
  - 通过简化的变分分布近似后验分布
  - 计算效率较高





Source: Multilingual topic modeling for tracking COVID-19 trends based on Facebook data analysis

LDA的应用领域：文本挖掘与分析

文档聚类与组织

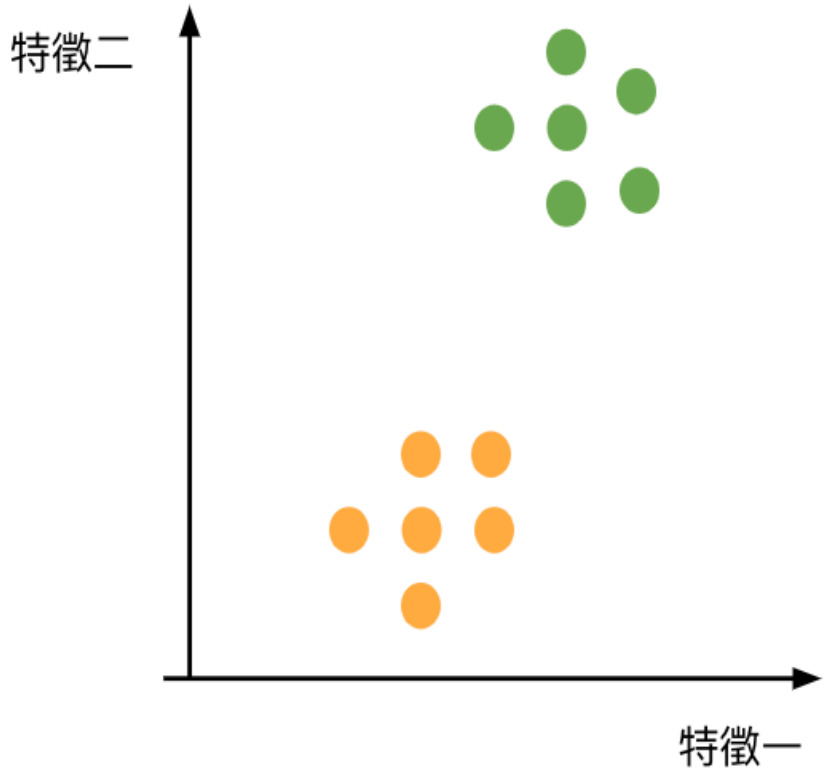
- 自动对大量文档进行主题分组
- 构建分层文档库和知识图谱
- 应用实例：自动组织学术论文库

内容推荐系统

- 基于用户阅读历史的潜在主题兴趣建模
- 推荐相似主题分布的新内容
- 应用实例：新闻推荐、学术文献推荐

趋势发现与追踪

- 分析主题随时间的变化趋势
- 发现新兴主题和消退主题
- 应用实例：科技趋势监测、社交媒体分析



LDA在社会语言学中的应用

词汇语义学研究

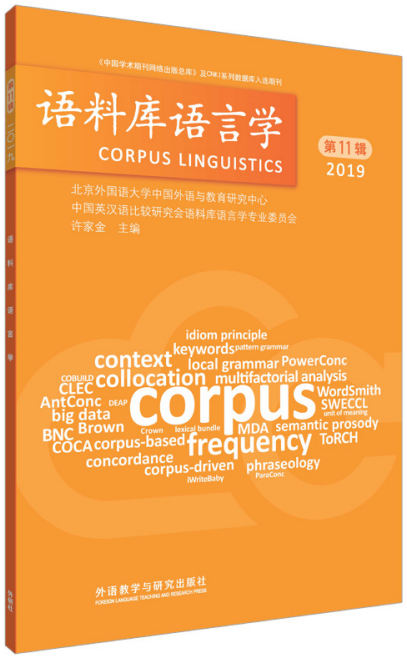
- 分析词汇的主题分布
- 探索词义随语境的变化
- 应用实例：多义词研究

体裁分析

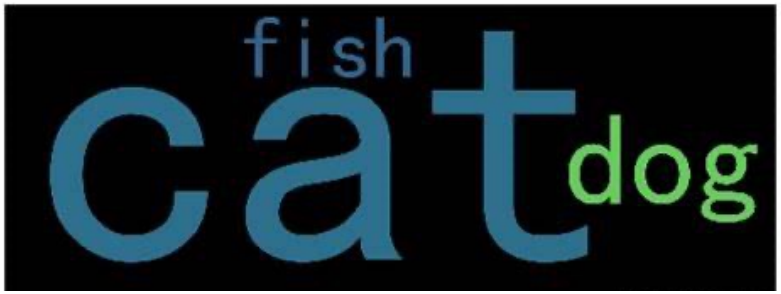
- 识别不同体裁的典型主题分布
- 研究体裁的语言特征
- 应用实例：学术论文、新闻报道、小说比较

历时语言变化

- 分析语言主题随时间的演变
- 探索新词义的产生与旧词义的消失
- 应用实例：历史语料库研究



LDA实践：预处理步骤



文本清洗

•去除特殊字符、数字和标点

•转换为小写形式

•分句和分词处理

停用词移除

•去除高频功能词（the， is， of等）

•去除领域特定的无信息词

•可使用预定义停用词表或自定义列表

词形还原与词干提取 （中文无此项）

•**词形还原**（Lemmatization）：还原词到词典形式（running → run）

•**词干提取**（Stemming）：提取词的词干（connection → connect）

•减少词汇空间，聚合相关词形

LDA主要SSCI期刊采用情况

- **Journal of Sociolinguistics** (IF: 3.2) - 2018年起显著增加LDA应用
- **Language in Society** (IF: 2.8) - 2019-2023年发表18篇LDA研究
- **Discourse & Society** (IF: 2.6) - 批评话语分析结合LDA的创新应用
- **International Journal of Bilingualism** (IF: 2.3) - 双语研究中的主题挖掘

- Applied Linguistics (IF: 4.1) - 语料库语言学与LDA结合
- Journal of Pragmatics (IF: 2.4) - 语用学研究中的LDA应用
- Language Resources and Evaluation (IF: 2.7) - 多语言资源开发与LDA
- Journal of Language and Social Psychology (IF: 2.9) - 社会心理语言学
- Language Learning & Technology (IF: 3.5) - 教育技术与语言学习研究

Topic Modeling and Sentiment Analysis of Twitter Discussions on COVID-19 from Spatial and Temporal Perspectives

Iyad AlAgha*
Faculty of Information Technology, The Islamic University of Gaza, Gaza, Palestine
E-mail: ialagha@iugaza.edu.ps

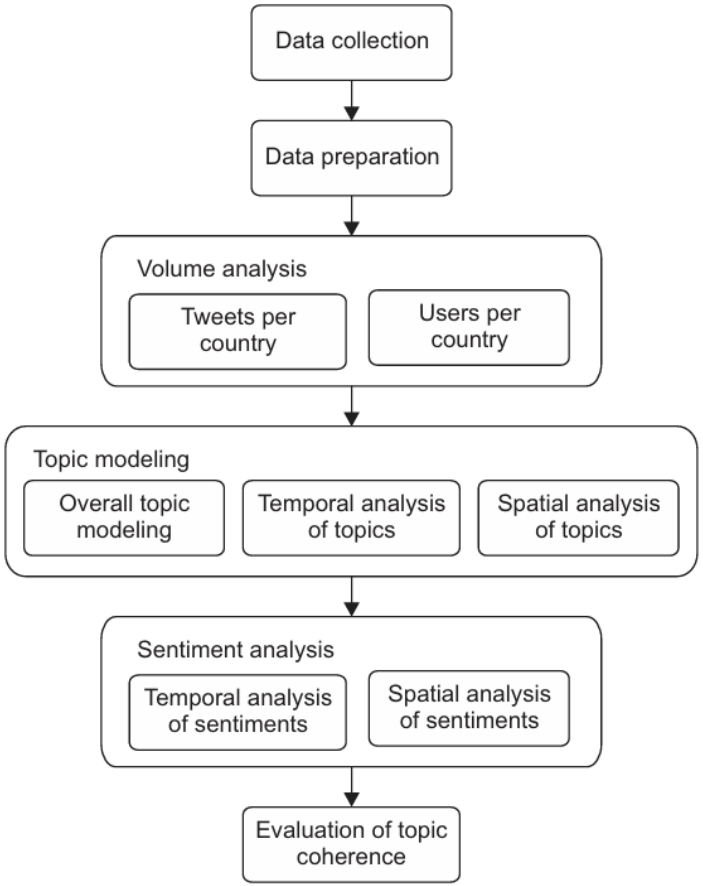


Fig. 1. Experimental methodology.

研究概述 | Research Overview

本研究对Twitter上关于COVID-19的讨论进行了主题建模和情感分析，以评估COVID-19讨论的主题和观点。研究从时间和空间角度分析了疫情期间的推文，揭示了以前文献中未报告的多种模式。

This research conducted topic modeling and sentiment analysis of COVID-19 discussions on Twitter to evaluate topics and opinions. The study analyzed tweets during the pandemic from both temporal and spatial perspectives, revealing patterns not previously reported in the literature.

研究方法 | Research Methodology

研究使用基于作者潜在狄利克雷分配(LDA)方法生成了二十个讨论疫情不同方面的主题。同时进行了主题时间序列分析，探索不同主题讨论如何随时间变化及其潜在原因，并通过比较不同国家在各主题上的推文百分比进行空间分析。

The study used Author-pooled Latent Dirichlet Allocation (LDA) to generate twenty topics discussing different aspects of the pandemic. Time-series analysis explored how discussions on each topic changed over time, while spatial analysis compared the percentage of tweets on each topic among different countries.

数据收集与处理 | Data Collection and Processing

通过Twitter API收集了2020年2月15日至5月31日期间包含关键词"coronavirus"、"COVID"、"COVID-19"、"ncov19"和"ncov2019"的975万条推文。经过清洗和预处理后，最终分析了约606万条英文推文。

Using Twitter API, 9.75 million tweets containing keywords "coronavirus," "COVID," "COVID-19," "ncov19," and "ncov2019" were collected between February 15 and May 31, 2020. After cleaning and preprocessing, approximately 6.06 million English tweets were analyzed.

国家分布 | Country Distribution

美国是讨论COVID-19最活跃的国家，贡献了28.2%的推文总量。排名前五的国家还包括英国(690,693条推文)、加拿大(442,286条)、澳大利亚(387,757条)和新西兰(254,466条)。使用英语为主要语言的国家在语料库中占主导地位。

The USA was the most active country in COVID-19 discussions, contributing 28.2% of total tweets. The top five countries also included the UK (690,693 tweets), Canada (442,286), Australia (387,757), and New Zealand (254,466). English-speaking countries dominated the corpus.

用户分布 | User Distribution

语料库中的推文由超过155万用户发布，平均每用户发布4.9条推文。用户的国家分布与推文分布高度一致，超过四分之一的用户来自美国。这可能是因为在研究期间，美国是受疫情影响最严重的国家。

Tweets in the corpus were posted by over 1.55 million users, with an average of 4.9 tweets per user. The distribution of users across countries was highly consistent with the distribution of tweets, with over a quarter of users from the USA. This may be because the USA was the most affected country during the study period.

主题建模结果 | Topic Modeling Results

研究识别出了20个主要讨论主题，包括：

- 祈祷与祝福：包含"stay_home"、"bless"、"safe"、"god"等关键词
- 封锁政策：讨论居家政策、社交距离、信息传播等
- 病毒起源：关于病毒来源的讨论，包含"china"、"wuhan"、"animal"等关键词
- 症状与防护：涉及咳嗽、发热等症状及口罩、洗手等防护措施
- 医疗保健：讨论医疗系统、医护人员、医院资源等议题

The research identified 20 main discussion topics, including:

- Prayers and blessings: containing keywords like "stay_home," "bless," "safe," "god"
- Lockdown policies: discussing stay-at-home policies, social distancing, information dissemination
- Virus origin: about the source of the virus, containing keywords like "china," "wuhan," "animal"
- Symptoms and protection: involving symptoms like coughing and fever, and protection measures like masks and hand washing
- Healthcare: discussing medical systems, healthcare workers, hospital resources, etc.

主题建模结果(续) | Topic Modeling Results (Continued)

- 政治与经济：包含政治领袖对疫情的回应及疫情对经济的影响
- 受感染国家：关于不同国家疫情状况的比较
- 封锁期间活动：讨论隔离期间的活动，如观看Netflix、游戏、在线活动等
- 阴谋论：围绕病毒起源的各种阴谋论，包括5G技术与病毒的关联、生物武器说等
- 消费行为：讨论疫情期间的消费习惯变化，包括在线购物增加、特定产品需求激增等
- Politics and economy: including political leaders' responses and economic impacts
- Infected countries: comparing outbreak situations in different countries
- Lockdown activities: discussing isolation activities like watching Netflix, gaming, online activities
- Conspiracy theories: various conspiracy theories about virus origins, including 5G technology and bioweapon claims
- Consumer behavior: discussing changes in consumption habits, including increased online shopping and surging demand for specific products

时间维度分析 | Temporal Analysis

研究发现主题讨论随时间变化明显，与疫情发展阶段高度相关：

- 政治与经济主题：3月中旬开始大幅增加，与全国社交距离政策和学校关闭相关；5月随着多数国家度过疫情高峰而下降
- 症状与防护主题：3月初开始上升，4月达到峰值，5月开始下降
- 情感变化：2月15日至4月11日以负面情绪为主，4月中旬后开始转向积极

The study found that topic discussions changed significantly over time, highly correlating with pandemic development stages:

- Politics and economy topics: increased significantly from mid-March, coinciding with national social distancing policies and school closures; declined in May as most countries passed the peak
- Symptoms and protection topics: began to rise in early March, peaked in April, and declined in May
- Sentiment changes: dominated by negative emotions from February 15 to April 11, shifted toward positive after mid-April

空间维度分析 | Spatial Analysis

不同国家在主题关注度和情感表达上存在显著差异:

- 美国：政治主题讨论占比最高(25%)，远高于英国(13%)和加拿大(8%)
- 英国：医疗保健主题讨论比例最高(21%)，高于美国(14%)和加拿大(9%)
- 加拿大：整个研究期间保持积极情绪，平均情感分数为0.2053
- 英国：三国中情绪最负面，平均情感分数为-0.2279
- 美国：平均情感分数为-0.0261，4月底开始转为积极

Different countries showed significant differences in topic focus and emotional expression:

- USA: highest proportion of political topic discussions (25%), much higher than the UK (13%) and Canada (8%)
- UK: highest proportion of healthcare topic discussions (21%), higher than the USA (14%) and Canada (9%)
- Canada: maintained positive sentiment throughout the study period, with an average sentiment score of 0.2053
- UK: most negative emotions among the three countries, with an average sentiment score of -0.2279
- USA: average sentiment score of -0.0261, turning positive from late April

情感分析结果 | Sentiment Analysis Results

情感分析显示公众情绪与疫情发展和政策变化密切相关：

- 2月15日至4月11日：负面情绪占主导，表现为恐惧和焦虑情绪强烈，对医疗服务准备和政府应对计划的激烈辩论
- 4月中旬后：情绪开始向积极方向转变，多国宣布重新开放计划和放宽限制措施产生积极影响
- 5月：随着重返正常生活的讨论增加，积极情绪继续上升

Sentiment analysis showed that public emotions were closely related to pandemic developments and policy changes:

- February 15 to April 11: negative emotions dominated, showing strong fear and anxiety, and intense debate about healthcare service preparation and government response plans
- After mid-April: sentiment began to shift in a positive direction, with reopening plans and easing of restrictions having a positive impact
- May: positive emotions continued to rise as discussions about returning to normal life increased

主题模型评估 | Topic Model Evaluation

研究比较了不同主题建模技术的效果：

- 使用UMass一致性指标评估不同LDA实现方式的效果
- 结果表明主题数量在20-30之间时模型表现最佳
- 基于标签池化的LDA效果最好，但由于只有27.8%的推文含有标签，作者池化的LDA是更实用的选择

The study compared the effectiveness of different topic modeling techniques:

- Using the UMass coherence index to evaluate different LDA implementation methods
- Results showed that the model performed best when the number of topics was between 20-30
- Hashtag-pooled LDA performed best, but since only 27.8% of tweets contained hashtags, author-pooled LDA was a more practical choice

研究意义 | Research Significance

该研究通过分析Twitter上的COVID-19讨论，展示了社交媒体数据如何帮助我们理解：

- 公众对全球性卫生危机的关注焦点和反应
- 不同国家和不同时期的公众情绪变化
- 消费者行为和信息传播模式
- 错误信息的传播途径和影响

这些发现为公共卫生危机沟通和政策制定提供了宝贵参考。

This research, by analyzing COVID-19 discussions on Twitter, demonstrates how social media data can help us understand:

- Public focus and reactions to global health crises
- Changes in public sentiment across different countries and time periods
- Consumer behavior and information dissemination patterns
- Pathways and impacts of misinformation spread

These findings provide valuable references for public health crisis communication and policy making.

结论 | Conclusion

本研究通过主题建模和情感分析揭示了COVID-19疫情期间Twitter讨论的动态变化。研究显示，主题和情绪都受到疫情发展的重大事件影响，不同国家根据其准备情况和受感染程度表现出不同的讨论主题和情绪。这些发现对于理解公众对全球性危机的反应和态度变化具有重要意义。

This study revealed the dynamic changes in Twitter discussions during the COVID-19 pandemic through topic modeling and sentiment analysis. The research shows that both topics and emotions were influenced by significant events in the pandemic's development, with different countries exhibiting different discussion topics and emotions based on their preparedness and infection levels. These findings are significant for understanding public reactions and attitude changes toward global crises.

•Discourse & Society (IF: 2.6)-话语与社会

Sage Journals

Search this journal | Enter search terms...

Advanced search

Access/Profile Cart

Browse by discipline Information for

Discourse & Society

Impact Factor: **2.4** 5-Year Impact Factor: **2.4** / Journal Homepage **Submission Guidelines**

Discourse & Society is a leading international peer-reviewed journal whose major aim is to publish outstanding research at the boundaries of discourse analysis and the social sciences. It focuses on explicit theory formation and analysis of the relationships between the structures of text, talk, language use, verbal interaction or … | View full journal description

This journal is a member of the Committee on Publication Ethics (COPE).

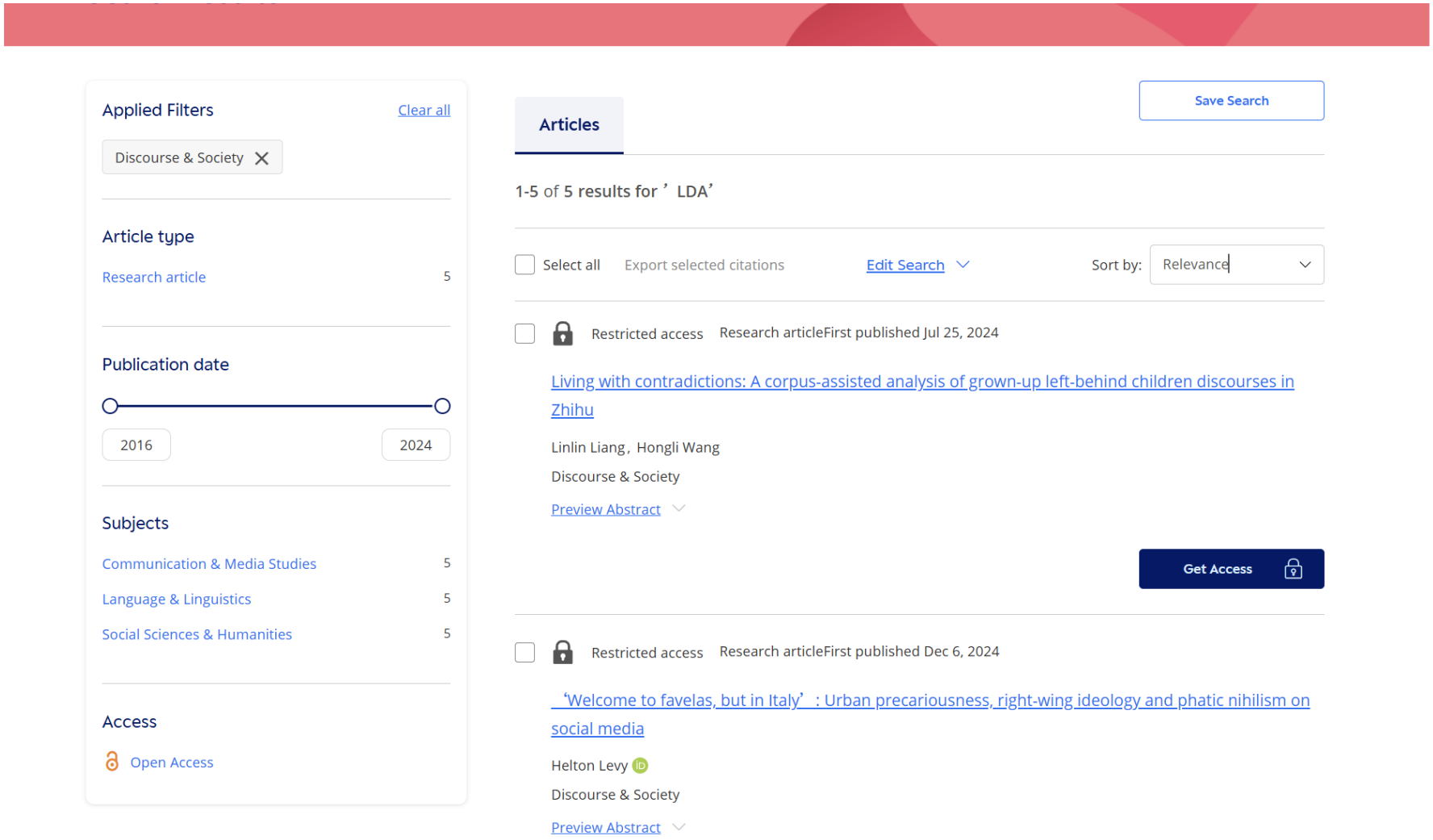
Browse journal

Current issue

OnlineFirst

All issues

Free sample



Living with contradictions: A corpus-assisted analysis of grown-up left-behind children discourses in Zhihu

中国留守儿童成长经历与社会影响：基于知乎语料库辅助话语研究

Living with contradictions: A corpus-assisted analysis of grown-up left-behind children discourses in Zhihu

liang-wang-2024-living-with-contradictions-a-corpus-assisted-analysis-of-grown-up-left-behind-children-discourses-in.pdf

Figure 1. Intertopic distance map (via multidimensional scaling).

研究概述 | Research Overview

本研究采用语料库辅助话语研究(CADS)方法，结合潜在狄利克雷分配(LDA)主题建模技术，对中国成年留守儿童在知乎平台上的个人叙事进行分析。研究揭示了留守儿童成长过程中面临的挑战和机遇，及其对个人发展的长期影响。

This study employs Corpus-Assisted Discourse Studies (CADS) combined with Latent Dirichlet Allocation (LDA) topic modeling to analyze personal narratives of grown-up left-behind children (LBC) on China's Zhihu platform. The research reveals the challenges and opportunities faced by LBC during their growth, and the long-term impacts on their personal development.

研究背景 | Research Background

中国留守儿童现象源于改革开放后的大规模劳动力迁移，由于户籍制度限制和城市高昂的生活成本，农村劳动力被迫将子女留在农村，由祖父母或其他监护人照顾。据2021年中国民政部数据，中国农村地区共有643.6万留守儿童。早期一代留守儿童如今已成年并融入社会，了解他们的成长历程及当前状况具有重要意义。

The phenomenon of left-behind children in China originated from the large-scale labor migration after the Reform and Opening-up. Due to household registration restrictions and high living costs in cities, rural workers were forced to leave their children in rural areas under the care of grandparents or other guardians. According to 2021 data from China's Ministry of Civil Affairs, there were 6.436 million left-behind children in rural China. The early generation of LBC has now reached adulthood and integrated into society, making it significant to understand their growth experiences and current status.

研究方法 | Research Methodology

研究选取知乎平台上"长大后的留守儿童现状如何？"话题下的966篇帖子作为语料库，总计727,619字。研究采用语料库辅助话语研究(CADS)方法，结合LDA主题建模、词汇搭配和语境分析，并运用布朗芬布伦纳生态系统理论(EST)分析影响留守儿童发展的因素。

The research analyzed 966 posts with 727,619 words from the topic "What is the current status of grown-up left-behind children?" on Zhihu. The study utilized Corpus-Assisted Discourse Studies (CADS), combining LDA topic modeling, collocation and concordance analysis, and applied Bronfenbrenner's Ecological Systems Theory (EST) to analyze factors influencing LBC development.

主题建模结果 | Topic Modeling Results

通过LDA主题建模，研究确定了五个主要话题：

1. **人格特质与心理健康**：包含"自卑"、"安全感"、"孤独"、"独立"等关键词
2. **家庭关系**：关注与父母和祖父母的关系，包含"父亲母亲"、"奶奶"、"回家"等关键词
3. **留守童年经历**：描述童年生活，包含"留守"、"孩子"、"村庄"、"欺凌"等关键词
4. **教育问题**：讨论学校教育与家庭教育，包含"高中"、"学校"、"初中"、"成绩"等关键词
5. **现状与未来**：探讨成年后的状况，包含"成长"、"工作"、"婚姻"、"努力"等关键词

Through LDA topic modeling, the research identified five main topics:

1.Personality and Mental Health: Including keywords like "inferiority," "sense of security," "loneliness," "independence"

2.Family Relationships: Focusing on relationships with parents and grandparents, with keywords like "father and mother," "grandmother," "going home"

3.Left-behind Childhood Experiences: Describing childhood life, with keywords like "left-behind," "child," "village," "bullying"

4.Education Issues: Discussing school and family education, with keywords like "high school," "school," "junior high school," "grades"

5.Current Status and Future: Exploring adulthood conditions, with keywords like "growing up," "work," "marriage," "effort"

人格特质与心理健康 | Personality and Mental Health

研究发现成年留守儿童的人格特质呈现以下几对对比性特征：

自尊与自卑：一方面，留守经历培养了强烈的自尊心，成为他们追求卓越的动力；另一方面，缺乏父母陪伴导致深层次的自卑感。

缺乏安全感与独立性：成年留守儿童普遍缺乏安全感，同时表现出极强的独立性，甚至到了疏离的程度。

乐观与抑郁：部分成年留守儿童保持乐观开朗的性格，但更多人表现出孤独感和抑郁倾向。

The research found that grown-up LBC exhibit several contrasting personality traits:

Self-esteem and Inferiority: On one hand, the left-behind experience cultivated strong self-esteem, serving as motivation for excellence; on the other hand, lack of parental companionship led to deep-seated feelings of inferiority.

Lack of Security and Independence: Grown-up LBC generally lack a sense of security while exhibiting strong independence, sometimes to the point of aloofness.

Optimism and Depression: Some grown-up LBC maintain optimistic personalities, but more exhibit feelings of loneliness and tendencies toward depression.

家庭关系 | Family Relationships

父母的长期缺席对成年留守儿童的家庭关系产生深远影响：

缺乏家庭概念：缺少父母陪伴使成年留守儿童对"家庭"的理解和认同感有限。

与祖父母关系紧密：祖父母在很大程度上承担了父母角色，成为农村留守儿童成长中的关键人物。

与父母关系疏远：成年后与父母之间存在情感距离，关系生硬。

亲密关系建立困难：童年缺乏亲情导致成年后在表达和接受爱方面存在困难。

The prolonged absence of parents has profound impacts on the family relationships of grown-up LBC:

Limited Concept of Family: Lack of parental companionship results in limited understanding and identification with the concept of family.

Close Relationships with Grandparents: Grandparents largely assumed parental roles, becoming key figures in the development of rural LBC.

Distant Relationships with Parents: Emotional distance and stiffness in relationships with parents persist into adulthood.

Difficulty in Establishing Intimate Relationships: Childhood lack of affection leads to difficulties in expressing and receiving love as adults.

留守童年经历 | Left-behind Childhood Experiences

研究揭示了留守儿童面临的多重挑战：

物质贫困：留守儿童常生活在相对贫困的环境中，物质条件匮乏。

亲情缺失：父母外出务工导致亲情缺失，情感需求得不到满足。

智能手机依赖：部分留守儿童对手机产生依赖，用以填补情感空白。

同伴欺凌：在学校遭受欺凌，甚至受到邻居嘲笑，对心理造成严重伤害。

The research revealed multiple challenges faced by LBC:

Material Poverty: LBC often lived in relatively impoverished environments with scarce material resources.

Absence of Parental Affection: Parents' migration led to a lack of familial affection, leaving emotional needs unmet.

Smartphone Dependence: Some LBC developed dependence on mobile phones to fill emotional voids.

Peer Bullying: Experiences of bullying at school and ridicule from neighbors caused severe psychological harm.

教育问题 | Education Issues

研究发现留守儿童在教育方面面临双重挑战：

学校教育资源不足：由于区域经济发展不平衡，农村地区教育资源匮乏，导致部分留守儿童初中或高中辍学。

家庭教育缺失：留守儿童的父母或监护人普遍文化水平较低，教育理念落后，如依赖体罚而非引导式教育。这使得留守儿童长大后常表现出脆弱、敏感、缺乏社交技能等特点。

The research found that LBC face dual challenges in education:

Insufficient School Educational Resources: Due to unbalanced regional economic development, rural areas lack educational resources, leading to middle or high school dropouts among some LBC.

Lack of Family Education: Parents or guardians of LBC generally have lower educational levels and outdated educational beliefs, such as relying on corporal punishment rather than guidance-based education. This results in grown-up LBC often exhibiting vulnerability, sensitivity, and lack of social skills.

成年留守儿童的现状 | Current Status of Grown-up LBC

成年留守儿童在工作、婚姻和未来规划方面展现出独特特点：

工作态度积极：成年留守儿童表现出较强的工作动力和韧性精神，更加成熟独立。

婚姻与家庭观念：部分成年留守儿童因负面童年经历而不愿生育；同时，他们强烈希望陪伴自己的孩子，打破留守儿童家庭的循环模式。

积极向上的态度：尽管童年经历不理想，成年留守儿童仍保持积极向上的态度，努力改变自身处境，追求幸福美好生活。

Grown-up LBC exhibit unique characteristics in work, marriage, and future planning:

Positive Work Attitude: Grown-up LBC demonstrate strong work motivation and resilience, with greater maturity and independence.

Marriage and Family Views: Some grown-up LBC are reluctant to have children due to negative childhood experiences; simultaneously, they express a strong desire to be with their own children, breaking the cycle of LBC families.

Positive Outlook: Despite less-than-ideal childhood experiences, grown-up LBC maintain a positive attitude, working to improve their circumstances and pursue happy, fulfilling lives.

生态系统理论分析 | Ecological Systems Theory Analysis

研究运用布朗芬布伦纳生态系统理论分析影响留守儿童发展的多层次因素：

微系统 (Microsystem)：留守儿童直接面临父母陪伴缺失、学校欺凌、邻居嘲笑等挑战，影响其情感健康，但同时培养了自尊、独立和韧性等积极品质。

中系统 (Mesosystem)：家庭与学校的关系不协调，留守儿童缺乏家庭和学校的双重支持，导致许多人辍学，社交行为和人际技能表现出脆弱性和敏感性。

外系统 (Exosystem)：低社会经济家庭地位和媒体对留守儿童的负面塑造间接影响其心理健康，导致幸福感降低。

宏系统 (Macrosystem)：贫困、城乡差距、社会歧视和户籍制度限制等结构性因素限制了留守儿童的教育机会和生活轨迹。

The research applied Bronfenbrenner's Ecological Systems Theory to analyze the multi-level factors influencing LBC development:

Microsystem: LBC directly face challenges such as absence of parental companionship, school bullying, and ridicule from neighbors, affecting their emotional health while fostering positive qualities like self-esteem, independence, and resilience.

Mesosystem: The uncoordinated relationship between family and school leaves LBC lacking support from both systems, leading to dropouts and exhibiting vulnerability and sensitivity in social behavior and interpersonal skills.

Exosystem: Low socioeconomic family status and negative media portrayals indirectly impact psychological well-being, resulting in lower levels of happiness.

Macrosystem: Structural factors such as poverty, urban-rural divide, social discrimination, and household registration restrictions limit educational opportunities and life trajectories for LBC.

研究意义 | Research Significance

理论贡献：本研究首次将CADS与布朗芬布伦纳生态系统理论相结合，拓展了话语分析的理论框架，为研究其他弱势或边缘化社会群体提供了有价值的研究模型。

实践意义：通过分析成年留守儿童在社交媒体上的个人叙事，提供了理解这一群体的新视角，为制定相关政策和社会干预措施提供了基础。研究强调需要同时解决留守儿童面临的直接挑战(如物质贫困、父母陪伴缺失、学校欺凌)和更广泛的系统性因素(如家庭和学校教育不足、社会态度和政策障碍)。

Theoretical Contributions: This study is the first to combine CADS with Bronfenbrenner's EST, expanding the theoretical framework of discourse analysis and providing a valuable research model for studying other vulnerable or marginalized social groups.

Practical Significance: By analyzing personal narratives of grown-up LBC on social media, the study offers a new perspective for understanding this group and provides a foundation for formulating relevant policies and social interventions. The research emphasizes the importance of addressing both direct challenges faced by LBC (such as material poverty, lack of parental companionship, and school bullying) and broader systemic factors (including inadequate family and school education, societal attitudes, and policy barriers).

我们如何实现LDA的使用

How we utilize the LDA in our research?

PyCharm简介 | Introduction to PyCharm

PyCharm是一款专业的Python集成开发环境(IDE)，由JetBrains开发。它具有强大的代码分析、调试工具和版本控制集成功能，是进行LDA主题建模等数据分析工作的理想选择。

PyCharm is a professional Python Integrated Development Environment (IDE) developed by JetBrains. It features powerful code analysis, debugging tools, and version control integration, making it an ideal choice for data analysis tasks such as LDA topic modeling.

PyCharm安装与设置 | PyCharm Installation and Setup

安装之前请先下载Python　https://www.python.org/downloads/

1. **安装PyCharm**：从官方网站下载并安装PyCharm（专业版或社区版）
2. **创建新项目**：选择File > New Project，设置项目名称和位置
3. **配置虚拟环境**：选择"New environment using Virtualenv"创建独立环境
4. **安装必要包**：使用集成终端安装numpy、pandas、scikit-learn、gensim、matplotlib、nltk、jieba和wordcloud
5. **配置项目解释器**：确保项目使用正确的Python解释器和虚拟环境

什么是LDA？ | What is LDA?

LDA（Latent Dirichlet Allocation，潜在狄利克雷分配）是一种无监督机器学习算法，用于发现文本集合中的主题。它基于以下假设：

- 每个文档由多个主题组成
- 每个主题由多个词组成

LDA将文档表示为主题的概率分布，同时将主题表示为词汇的概率分布。

LDA (Latent Dirichlet Allocation) is an unsupervised machine learning algorithm used to discover topics in a collection of texts. It is based on the following assumptions:

- Each document consists of multiple topics
- Each topic consists of multiple words

LDA represents documents as probability distributions over topics, while representing topics as probability distributions over vocabulary.

数据收集方法：

- 网络爬虫（使用Scrapy或Requests）
- API接口（如Twitter API）
- 现有数据集（如Kaggle）
- 手动收集

LDA实现步骤 | LDA Implementation Steps

使用Python实现LDA主题建模的主要步骤：

1. **数据收集与预处理**：收集文本数据并进行初步整理
2. **文本分词与清洗**：对文本进行分词、去除停用词和特殊符号
3. **特征提取**：构建文档-词矩阵或词袋模型
4. **构建LDA模型**：设置合适的参数训练LDA模型
5. **主题可视化与解读**：对模型结果进行可视化和分析

步骤1：数据收集与预处理 | Step 1: Data Collection and Preprocessing

数据来源:

- 网络爬虫收集的社交媒体数据
- API接口获取的结构化数据
- 现有文本数据集
- 手动收集的文本资料

预处理内容:

- 文件格式转换
- 文本编码处理
- 初步数据清洗
- 数据结构组织

步骤2：文本分词与清洗 | Step 2: Text Tokenization and Cleaning

中文文本处理:

- 使用jieba进行中文分词
- 去除URL、表情符号、特殊字符
- 过滤中文停用词
- 保留有意义的词汇

英文文本处理:

- 文本小写化
- 使用NLTK进行分词
- 去除停用词
- 词形还原或词干提取

步骤3：特征提取 | Step 3: Feature Extraction

特征提取方法:

- 词袋模型（Bag-of-Words）
- TF-IDF向量化
- 创建文档-词矩阵
- 构建Gensim语料库

关键参数设置:

- 最小文档频率（min_df）
- 最大文档频率（max_df）
- 词袋大小限制（max_features）
- n-gram范围

步骤4：主题可视化与解读 | Step 4: Topic Visualization and Interpretation

可视化方法:

- 主题词云：直观展示主题关键词
- 一致性得分曲线：评估最佳主题数量
- pyLDAvis交互式可视化：展示主题间关系和关键词

结果解读:

- 分析主题-词分布：识别主题的核心含义
- 分析文档-主题分布：了解文档内容结构
- 应用于文本分类或聚类任务

LDA Application Recommendations:

- **Start with Small-scale Data**: Test and adjust parameters on smaller samples first
- **Combine Quantitative and Qualitative Analysis**: Use both statistical metrics and manual assessment for topic quality
- **Try Different Numbers of Topics**: Systematically test different numbers of topics and compare results
- **Consider Domain Knowledge**: Utilize domain expertise to assist model training and result interpretation
- **Combine with Other Methods**: Use LDA in conjunction with other text analysis methods for more comprehensive understanding