分类号_____ 密级 _____
U D C _____ 编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

（专业学位）

论文题目 基于 LDA 模型的中国古典诗词在不同历史时期的主题发现

研 究 生 姓 名：李伦珑

指导教师姓名、职称：王永瑜 教授

学 科、专 业 名 称：统计学 应用统计硕士

研 究 方 向：大数据分析

提 交 日 期：2021 年 6 月 6 日

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：李佺珏 签字日期：2021.6.6

导师签名：[signature] 签字日期：2021.6.6

导师(校外)签名：[signature] 签字日期：2021.6.6

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入 CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：李佺珏 签字日期：2021.6.6

导师签名：[signature] 签字日期：2021.6.6

导师(校外)签名：[signature] 签字日期：2021.6.6

Theme discovery of Chinese classical poetry in different historical periods based on LDA model

Candidate : Li Lunlong

Supervisor: Wang Yongyu Professor

摘要

古典诗词是中国文学史上重要的文化形式之一，对于记录当时文人的所思所想，以及反映诗词所处时代政治、经济、文化与社会背景具有重要的意义。然而，中国历史各朝各代留存下来的古典诗词数不胜数，如果要从宏观的角度来分析这些诗词，研究者要翻阅、整理大量的书籍，以至于在诗词收集整理阶段就需要投入更多的时间与精力，而且后期的统计与分析阶段，也会因为研究者的主观偏好，带来结论的不客观性。诗词是历史的沉淀，同时影响深远，也在塑造当时的历史。古典诗词与当时历史时期的写作背景有密切的关系。丰富的古典诗词、悠久的历史是中华民族的长期以来不断汲取文化营养的书籍。

主题是文学作品或者社会活动等所要表现的中心思想，分析诗词主题是研究诗词创作背景的重要渠道，这有助于勾勒社会样貌。从时间序列角度出发，通过对古典诗词不同历史时期的主题进行研究，比较不同历史时期的古典诗词主题差异，呈现主题内容演化。无论是盛世时代、还是乱世时代，反映在古典诗词作品中的情感与内容都可以帮助我们更好得理解一个健康且进步的社会形态，并最终引领现实社会，推动社会向前发展。

LDA（Latent Dirichlet Allocation）模型是一种经典主题概率生成模型。LDA 模型通常也被叫做三层贝叶斯模型，即包含词项、主题和文档三层结构。作为一种无监督机器学习方法，LDA 模型擅长识别规模较大的文档数据集中潜在的主题。本文的古典诗词数据按照各个历史时期进行划分，共计 738881 条数据，结合 LDA 模型进行文本挖掘，共获得 46 个有效主题，分布在不同的历史时期。主要使用 Python 语言完成对数据的预处理和 LDA 模型的建模和实现，并把具有相同对象的主题划分为不同的主题类，文章以“国家意识”主题类为例对主题内容演化进行可视化分析，可以看出，“国家意识”的转变经历了三个阶段，一个是以魏晋、南北朝、隋为代表的圣皇崇拜阶段，一个是以唐、宋为代表的国家主体观阶段，一个是以明、清、近代为代表的人民主体观阶段，“国家意识”概念逐渐下沉，同时也变得越来越广；通过计算主题-文档概率矩阵，计算不同主题下的主题强度，利用主题强度的大小排序，结合诗词文本内容，以分析不同历史时期的古典诗词的时代背景，可以看出王朝的更迭、社会的动荡、人口的迁徙、科

举的鼎盛、文化的繁荣、经济的复苏、军事的强盛与革命的乐观等都反映在了各个时代文人的精神面貌当中。

本文研究认为，各个历史时期古典诗词的主题大致存在一定的上下波动，无论是国家意识类，人生羁旅类，离愁类，还是爱情、友情、乡情类等，但由于历史阶段社会背景的不同，各个主题存在一定的差异，并且强度不一，社会演化趋势显著。通过古典诗词的文本分析，从宏观角度丰富了中国古典诗词的计算化研究。同时，借古思今，为后来研究者能够从政治、经济、文化、教育、军事等在社会中所处的重要地位出发，为当今社会坚定“四个自信”意识，推进“五位一体”总体布局，建设一个更好的社会提供参考和启迪。

关键词：LDA 模型 古典诗词 不同历史时期 主题发现

Abstract

Classical poetry is one of the important cultural forms in the history of Chinese literature. It is of great significance to record what the literati thought at that time, and to reflect the political, economic, cultural and social background of the times. However, there are so many classical poems in Chinese history. If we want to analyze these poems from a macro perspective, researchers have to read and sort out a large number of books, so that they need to invest more time and energy in the poetry collection and sorting stage. Moreover, in the later stage of statistics and analysis, because of the subjective preference of researchers, The conclusion is not objective. Poetry is the precipitation of history, at the same time, it has far-reaching influence, and it is also shaping the history at that time. Classical poetry is closely related to the writing background of that historical period. Rich classical poetry and a long history are the books of the Chinese nation that have been absorbing cultural nutrition for a long time.

Theme is the central idea of literary works or social activities. Analyzing the theme of poetry is an important channel to study the background of poetry creation, which helps to outline the social appearance. From the perspective of time series, this paper studies the theme of classical poetry in different historical periods, compares the

theme differences of classical poetry in different historical periods, and presents the evolution of theme content. Whether it is the prosperous age or the turbulent age, the emotions and contents reflected in the classical poetry can help us better understand a healthy and progressive social form, and ultimately lead the real society and promote social development.

LDA (latent Dirichlet allocation) model is a classical topic probability generation model. LDA model is also known as three-tier Bayesian model, which includes three-tier structure of words, topics and documents. As an unsupervised machine learning method, LDA model is good at identifying potential topics in large-scale document datasets. In this paper, the classical poetry data is divided according to each historical period, a total of 738881 data, combined with LDA model for text mining, a total of 46 effective topics, distributed in different historical periods. This paper mainly uses Python language to complete the data preprocessing and LDA model modeling and implementation, and divides the theme with the same object into different theme classes. Taking the theme class of "national consciousness" as an example, this paper makes a visual analysis of the evolution of the theme content. It can be seen that the transformation of "national consciousness" has gone through three stages: one is the Wei and Jin Dynasties, the northern and Southern Dynasties, the Southern Dynasties, the northern and Southern Dynasties

During the period of emperor worship represented by Sui Dynasty, the concept of "national consciousness" gradually sank and became more and more popular; By calculating the theme document probability matrix, the theme intensity of different themes is calculated, and the order of theme intensity is used to analyze the background of classical poetry in different historical periods. We can see the change of dynasties, social unrest, migration of people, prosperity of imperial examination, prosperity of culture, recovery of economy, and the change of culture The prosperity of the military and the optimism of the revolution are reflected in the mental outlook of the literati in various times.

This paper argues that the themes of classical poetry in different historical periods fluctuate up and down, whether it's national consciousness, life fetters, loneliness, or love, friendship, nostalgia, etc. but due to the different social backgrounds in different historical stages, there are certain differences in various themes, and the intensity is different, and the social evolution trend is significant. Through the text analysis of classical poetry, it enriches the computational research of Chinese classical poetry from a macro perspective. At the same time, thinking about the present through the past can provide reference and Enlightenment for later researchers to start from the important position of politics, economy, culture, education and military in the society, to strengthen the "four self-confidence" consciousness, to promote the "five

in one" overall layout, and to build a better society.

Keywords: LDA model; Classical poetry; Different historical periods; Theme discovery

目录

1 绪论

1.1 研究背景

党的十八大以来，习近平总书记围绕中国传统文化发表了系列讲话，以表达自己对传统文化与传统思想价值观的认可与敬仰。中华文化源远流长，具有五千年的悠久历史，古典诗词无论是内容、文字还是韵律，都凝聚了中华文化的重要精髓。对自身文化价值的坚定信心有助于一个民族认定自己、认定长远，并为之积极践行。在中国共产党成立 95 周年大会上，习近平总书记提出“四个自信”，其中“文化自信”是中国自信的本质，我们民族的灵魂。今年恰是中国共产党成立 100 周年，百年恰是风华正茂，今天的中国人一定会坚定中国自信，构建富有中国特色的价值观体系。

中国素有“诗国”之称，其灿烂的诗词文化，对整个民族及其每一个个体成员都产生潜移默化的深远影响。当读到“蒹葭苍苍，白露为霜。所谓伊人，在水一方。”，会让人感受到诗中的主人公对意中人深深的企慕但求而不得的伤感；当读到“九天阊阖开宫殿，万国衣冠拜冕旒。”，让人感受到盛唐时期万国来朝时的雄威庄严；当读到“羁怀病思不禁秋，又报西风大火流。”，又让人感受到百姓对宋朝国力羸弱的忧愁与悲愤；当读到“千里冰封，万里雪飘。望长城内外，惟余莽莽；大河上下，顿失滔滔。山舞银蛇，原驰蜡象，欲与天公试比高。”，毛主席身上涌现出的革命乐观主义会让人豪情万丈。受不同朝代的政治、思潮、文化的影响，古典诗词在不同历史时期会呈现不同的主题和内容演变。诗词是历史的沉淀，同时影响深远，也在塑造当时的历史。中国典藏的古典诗词作品浩如烟海，如果需要经过研究者阅读大量素材之后再进行整理、记录、翻阅、分析这一系列过程，最后对诗词做出判断，不仅会耗费过量的时间精力，在得到结论时也容易受到个人主观想法的影响。而且过高的专业要求也会阻碍很多普通的诗词爱好者们对研究古典诗词的热情。不断发展创新的计算机科学技术和统计理论，使得人民越来越关注利用文本挖掘技术对古典诗词进行研究分析，也越来越希望通过这种方法把我国优秀的传统文化在当代更好的发扬光大。通过分析海量的古典诗词，从而回到写作诗词的那个时代，去探讨研究那个时代的社会背景。

近些年来，随着机器学习与大数据分析技术的迅猛发展，使用统计理论与机器学习方法对古典诗词进行的探索与研究方兴未艾。本文使用自然语言处理技术（NLP-Natural Language Processing）中的 LDA 主题模型（Latent Dirichlet Allocation），以挖掘诗词中的主题特征，以帮助我们能够宏观地把握中国古典诗词在不同历史时期的历史背景与内容演化规律。

1.2 研究意义

基于以上的研究背景，本文的研究意义在于以下几点：

（1）从宏观层面提高文学爱好者或者研究者对古典诗词主题的认知效率。从先秦到唐宋，再到当代，古典诗词有几十万首，篇幅浩大。要对古典诗词主题有整体的认知，不仅要读古典诗词，还要对历史背景有一定的了解。使用计算机的方法对古典诗词进行分词切块，再进行主题聚合，利用文本可视化可以将古典诗词在不同历史时期的主题及其演化更为直观与宏观的呈现，进而窥见古诗词当时的历史状况。

（2）探索基于自然语言处理技术的古典诗词研究方法。中国历朝历代的古典诗词，跨度大，篇幅广，相比于以往依靠人工进行阅读与分析的方法，利用自然语言处理技术可以节省古典诗词研究者大量的时间与精力。在前人研究的基础上，利用 LDA 主题模型对古典诗词进行无监督学习，无论是对自然语言处理技术的应用研究还是对古典诗词文学的方法研究都是一种有益的探索。

（3）扩大古典诗词在当代研究的范围。当前对古典诗词及其主题的研究还只是停留在文字总结描述的形式上或者某一个单一诗词文本的分析上，缺乏直观、宏观、有时间趋势的呈现；还有就是对古典诗词的风格进行判析，且主要是二分类的监督学习方法。本文利用大数据可视化技术对古典诗词主题结论进行可视化呈现，受众可以从分析中把握主题内容及主题演化过程，为在当代新媒体语境下扩大古典诗词的研究领域探索新的方法。

（4）为新时代坚定文化自信，探索古典文化精髓做出贡献。古典诗词是中华民族的文化表达的最初呈现形态。直抒胸臆有之，隐逸内敛有之，但都成就了中华文化昂扬向上的自信：“不要人夸颜色好，只留清气满乾坤。”。古典诗词与中华文化自信交相辉映，为坚定道路自信、理论自信、制度自信奠定了重要基

础。

（5）为描绘社会提供了一个发现视角。无论是古典诗词还是当代文章、文学作品、民众评论等，都是该时代的现实反映。国家自信，则人民自信，国家富强，则人民富强。这些对现实社会的情感流露都会在百姓的文学作品与日常评论中得到体现。同时，文字具有传播效应，百姓的情感表达也在反作用于现实社会，一个健康且进步的社会就会被塑造出来。这也为从国家层面推进了“五位一体”，促进了国家全面发展。

1.3 研究内容与技术路线

1.3.1 研究内容

古典诗词是我国文学史上的重要文化瑰宝，其极高的艺术价值与情感体验，渊源不断地为我们整个中华民族提供丰富的精神食粮。而不同历史时期的古典诗词，主题内容差异很大，这离不开当时创作的历史背景。政治、思潮、文化都对诗词内容产生巨大的影响。

中国古典诗词跨度大，上至先秦时期。范围广，著作者众多，名人志士有之，无名之士亦有之。此外，从宏观视角研究古典诗词需要阅读的著作卷帙浩繁，对于普通受众来说，也难以通过大量的阅读对古典诗词的主题内容与历史样貌的演化进行全面掌握。因此，本文对中国各个历史时期的古典诗词进行收集，通过数据清洗、分词切块、主题聚类等大数据分析方法，利用大数据可视化技术对不同历史时期的古典诗词的主题内容进行呈现，尝试为研究古典文学提供一种思路与方法。

LDA（Latent Dirichlet Allocation）模型是一种经典主题概率生成模型。LDA 模型通常也被叫做三层贝叶斯模型，即包含词项、主题和文档三层结构。作为一种无监督机器学习方法，LDA 模型擅长识别规模较大的文档数据集中潜在的主题。自先秦到近代的古典诗词有几十万首，本文结合 LDA 模型对古典诗词数据集进行文本挖掘，共获得若干个有效主题。主要使用 Python 语言完成对数据的预处理和 LDA 模型的建模和实现，得出古典诗词共 46 个主题，分布在各个历史时期阶段中，按照主题类划分，可以分为“国家意识”主题类，“人生羁旅”

主题类，“离愁”主题类等，并利用可视化工具对主题演化进行深入挖掘；对各个历史时期的主题强度进行排序，并在此基础上，结合当时背景，进行诗词与历史相结合的分析。从古典诗词角度探析中国各个历史时期的社会、政治与文化背景。

按照章节划分如下：

第一章为文章的绪论部分，主要阐述古典诗词计算化研究的时代背景、研究意义、研究内容和本文的技术框架，以及本文研究创新之处。

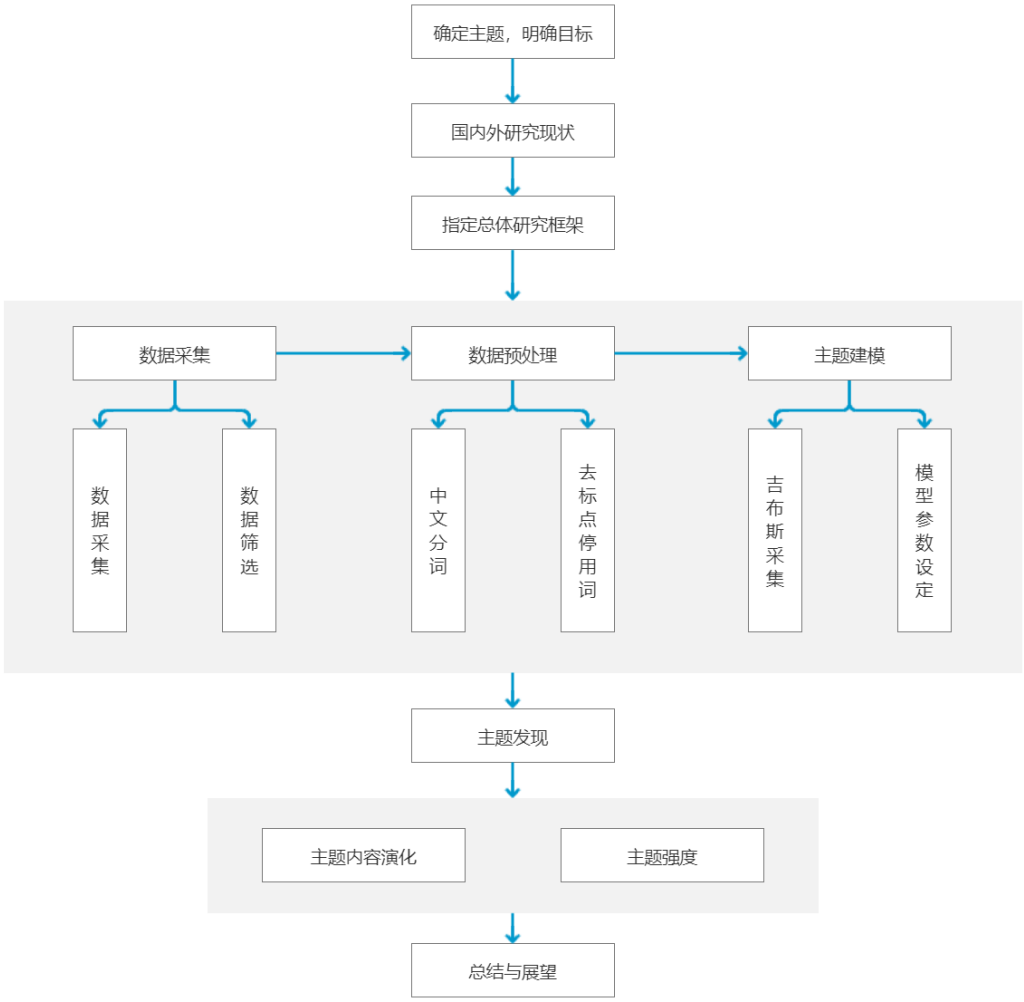
第二章为文献综述与主题模型理论以及 LDA 模型。文献综述即阐述国内外的研究现状，分为 LDA 主题模型的研究现状，以及古典诗词计算化的研究现状；同时对主题模型理论与 LDA 模型进行总结和介绍，如贝叶斯定理、吉布斯抽样、文本聚类以及 LDA 模型的使用等。

第三章为古典诗词在不同历史时期的主题模型实现。详细介绍了主题模型的实现过程，主要包括数据采集、数据清洗、分词、模型参数设置、主题输出等等，本章节详细介绍了 LDA 主题模型实现的所有技术，并对得到的 46 个主题模型进行初步的梳理，为下一章主题模型的深入分析做好准备。

第四章为古典诗词的内容演化与主题强度差异分析。主要对 LDA 模型计算出来的主题按照对象进行主题类的划分，并且以“国家意识”主题类为例，对主题内容进行可视化呈现与分析，以及利用主题-文档概率矩阵计算每个主题的文档支持度，从而得出各个主题的强度，通过强度大小分析古典诗词的写作背景与社会历史。

第五章为全文的总结与展望部分，通过对全文的总结梳理古典诗词与历史背景的联系，以及提出自己的建议，为后期的古典诗词计算化研究提供参考，并希冀通过对宏观视角下古典诗词的研究，为建设健康且进步的社会目标提供启迪。

1.3.2 技术路线



1.4 本文创新点

本文的创新之处有以下几个方面：

（1）研究方法的创新：现有的研究古典诗词文本挖掘的文献，分类类别多为有监督学习，即事先已打分类标签，且多为二分类，例如豪放派婉约派分类等。本文依据的是无监督学习方法 LDA，在事先不打标签的基础上，进行主题类别分析。

（2）研究视角的创新：现有的研究古典诗词的文本挖掘的文献，多为某本文集（例如《全唐诗》、《全宋词》）进行分析，缺少从时间序列角度进行宏观的研究。

2 文献综述和相关理论

2.1 文献综述

2.1.1 文本挖掘国内外研究综述

文本数据是我们工作生活中最常见的数据类型之一，随着大数据时代的来临，文本数据日益增长，数据中包含的信息越来越得到重视。近些年大数据研究领域的重点之一就是如何利用文本挖掘理论与技术从文本数据“大海”中提取出有价值的信息。

文本挖掘在发展初期，聚焦于统计思想，例如 Luhn H P.[1]采用词频统计来提取文本摘要。Maron 和 Kuhns[2]在二十世纪 60 年代提出基于朴素贝叶斯理论的关键词分类技术，这是学术界第一次讨论文本分类技术，在此之后，越来越多的学者开始对文本分类技术进行讨论与研究。

Salton G.（1975）[3]提出了向量空间模型（VSM），在文本集中，一篇文档由若干词组成，根据文档中的不同词对组成和相应权重可以将一篇文档转为一组向量，而词就表示该组向量中的分量，从而可以将所要研究的文本集构成一个向量空间。然而，这种技术又高度依赖专业技术人员的支持，需要事先为每个类别定义大量的推理模板。

Hearst（1996）[4]证明了“聚类假设”，假设认为在进行聚类后，相关性高的文档距离更近，相关性低的文档距离远，从而在进行相关研究中，可以缩小搜索范围。

二十世纪 80 年代~90 年代，文本挖掘技术在研究应用中逐渐得以凸显。Craig Silverstein 等（1998）[5]利用文本挖掘技术研究客户购买商品的关联规则，从而确定市场商品篮子的范围。

David M. Blei 等（2003）[6]提出了 LDA 模型的基本理论，LDA 模型是一个包含了文档-主题-词语的三层贝叶斯概率模型，其中文档选中某个主题、主题选中某个词语都服从一定的概率。

Bolton 等（2004）[7]等通过实验调查，比较分析是否在线反馈会对市场交易

的效率产生影响。结果发现反馈机制在促进交易效率大幅提高的同时，也构建了整个社会的信用评价体系。

Antweiler 和 Frank（2004）[8]首先采用支持向量机算法对华尔街日报流行专栏的内容进行情感定量，研究认为，媒体悲观预测股票价格会对股票交易产生较大影响。

Dongshan Xing 和 Mark Girolami（2007）[9]基于 LDA 概率分布可以对电信诈骗者的活动进行粗略分析，以提升诈骗电话识别的准确率。

Boiy 和 Moens（2009）[10]基于机器学习方法对三类语言相关文本进行情感分析，得出不同语言文本的分类准确率。

我国互联网发展得相对较晚，以至于上世纪国内几乎很少有学者在文本挖掘与机器学习领域有相关研究成果。另外，中文字词与英文字词不同，英文每个单词之间，就有空格进行切分，而中文字词切分涉及到语义等复杂情况。这就导致了直到最近二十年才逐渐出现相关领域的文献。我国目前在文本挖掘的研究主要集中在方法改进与应用研究方面，并且成果相当丰富。

王继成等（2000）[11]在研究中首次探讨了网络挖掘的相关理论，重点分析了包括文本的特征表示、文本分类与文本聚类等网络文本挖掘的方法。

李凡等（2001）[12]在进行文本分类研究中，通过比较分析现有的基于不同评估函数的特征筛选方法与适用范围，提出了一种新的评估函数从而替代 TFIDF 法中 IDF 函数的新算法。

黄晓斌等（2009）[13]提出网络舆情信息挖掘分析模型，并以博客一热搜事件的内容进行文本挖掘，为舆情背景下，公共事件的正确处理提供建议。

张彦（2011）[14]通过优化核函数参数与惩罚因子的方法对支持向量机进行改进，最后得出改进后的支持向量机泛化能力强、分类准确率高等优点。

杨丹（2018）[15]针对传统 Kmeans 算法在选择聚类中心时易受异常值影响等特点，提出了新的函数来选择聚类中心。准确率、召回率和 F 度量值是评估聚类算法改进效果的重要方法。通过实验表面，改进之后的聚类算法，准确率提高，聚类效果更好，适合于文本聚类方法的研究。

LDA 分析是文本挖掘的重要方法之一，国内外许多学者运用 LDA 模型进行科学文献主题发现。李湘东等（2014）[16]在研究科技期刊主题演化的过程中纳

入时间因素，采用困惑度这一指标确定最优主题数目，并从主题具体内容与主题强度两个方面对主题演化进行研究。

关鹏（2016）等[17]对常见的科学文献文本语料库进行分类：关键词、摘要、关键词+摘要，从文本挖掘中的语料库角度出发，研究不同语料库背景下进行 LDA 主题抽取的效果，并得出不同数据源背景下的主题抽取的广度与细颗粒度存在明显差异的结论。

谭春辉（2020）等[18]收集 1998 年-2018 年 CNKI 及 Web of Science 收录的数据挖掘领域核心期刊论文，通过 LDA 主题模型抽取研究主题，并基于主题生命周期识别热点主题，结合时间片构建主题的演化路径，从数据挖掘研究的理论维度和应用维度来对比分析国内外数据挖掘领域热点主题演化的区别与联系。

2.1.2 古典诗词计算化研究综述

从上世纪九十年代开始，我国陆续有学者开始利用计算机相关技术和统计理论对古典诗词进行辅助研究。包括文本数据提取、词频统计、语料库建立、诗词风格辨析、诗词模仿创作等方面的研究，并且成果颇丰。

周思源（1992）[19]借助统计方法，对《红楼梦》中黛钗二人的诗词进行分析，进而探讨曹雪芹通过诗词来塑造人物性格的主要表现形式。

周昌乐（2003）[20]在所著《心脑计算举要》中，第一次提出“计算诗学”的重要思想。旨在借助计算机技术与统计理论将诗词计算化，从而达到宏观分析诗词的目的。“计算诗学”这一概念提出来之后，很多相关学者尝试利用自然语言处理技术对中国的古典诗词进行研究，包括格律、风格、情感等方面，并且成果颇丰。

易勇（2005）[21]通过建立唐诗、宋词和对联语料库，基于机器学习中的朴素贝叶斯等分类方法，提出了利用模型来分辨古典诗词的豪放和婉约风格；利用机器学习方法对语料库的学习，第一次提出不限字数的对联语应对生成的统计模型。

游维（2007）[22]以汉语古典诗词为研究对象，基于遗传算法模型，建立了宋词生成系统，并给出了系统框架、主要生成流程与宋词生成实例进行说明。

苏劲松（2007）[23]通过多重松弛迭代计算方法，研究了宋词词语中的情感

标准问题，为之后的词句情感意义研究提供了参考。

吴春龙等（2008）[24]在建立和完善宋词语料库之后，提出了一种宋词风格表示模型，通过实验表明，该模型可以很好地应用于诗词风格分类之中。

赖兴邦（2008）[25]利用宋词格律与节奏方面具有严格限制的特征，展开对宋词格律方面的应用研究，并针对宋词特殊体裁，采用词聚类对宋词文本进行抽样。

钱鹏等（2015）[26] 基于全唐诗语料库，采用主题模型对分词之后的唐诗文本进行数据建模，在此基础上进行主题演变、诗人群体风格网络和探索性分析。

申资卓等（2019）[27]以《全唐诗》、《全宋词》中的有关“八音”的诗句、词句作为研究对象，使用 LDA 和 NMF 的主题挖掘方法，从整体到局部、从宏观到微观，多视角研究了唐诗宋词中的中国古典乐器。

张馨怡（2020）[28]基于 TextCNN 对古典诗词中的爱国情怀进行研究，该方法本质上是监督学习的二分类方法，并且经过实现打好便签以便后续计算机训练。

2.1.3 文献述评

综合上述对研究情况的整理，文本挖掘和文本分类都已经发展的越来越成熟，而目前国内古典诗词文本挖掘的研究，发现相关研究论文非常少，且有一定的不足。

1.以有监督学习研究为主。现有的古诗词文本挖掘大多是对研究古诗词的分类，比如对宋词的豪放派与婉约派的分类，从无监督学习的视角对古诗词进行研究较少；

2.缺乏从时间序列角度对古诗词进行宏观角度的研究。近年来，更多的是研究侧重于对某个特定朝代的古典诗词进行文本挖掘分析，从时间序列的角度对古典诗词进行主题的宏观分析较少。

2.2 主题模型理论

2.2.1 主题和主题模型

（1）主题的概念

想要理解主题模型的相关知识，首先要弄懂主题的概念。主题是指在文艺作品中或社会活动中所体现出的中心思想，通常也被称为主要内容。此外，在一些描绘性艺术作品中，主题还会涉及到艺术家的自身经验与价值判断。具体表现为一系列相关的词语来表示一段文本的主要内容。而主题模型就是对文本中隐含主题的一种建模方法[29]。

在传统的信息检索领域里，通过统计文章中每个词语出现概率的大小来判断词语成为主题词的可能性，认为词频越高的词语成为主题词的可能性越大。这种仅仅通过统计词频的方式而忽略词语语义的作用很不明显，对于表示主题的准确性是远远不够的。

（2）主题模型

用主题来表示文档内容实际是对文档进行降维处理的一种思想，它先将文档主题与词汇进行关联，然后把文档按照主题进行分类。主题模型是一种以无监督学习方式对文档中的隐含语义结构进行聚类的统计模型。自然语言处理中的主题挖掘与语义分析问题是该模型的主要应用领域。主题模型起源于隐形语义索引。与传统的文献计量学相比，主题模型能够深度挖掘语义之间的关系，例如对新闻、文学作品、科技文献等进行研究，对之进行主题发现、监督、跟踪与预测。

主题模型有多种类型[30]，常见的有：一元模型（Unigram model）、一元混合模型（Mixture of Unigrams model）、潜在语义分析模型（Latent Sematic Analysis，LSA）[31]、概率潜在语义分析模型[32]（Probabilistic Latent Semantic Analysis，PLSA）以及潜在迪利克雷分布模型（Latent Dirichlet Allocation，LDA）。LDA 模型是主题模型中最常见的模型，由 Blei 提出，理论基础是在 PLSA 模型的基础上进行贝叶斯化。

一元模型的模型图如下图所示。

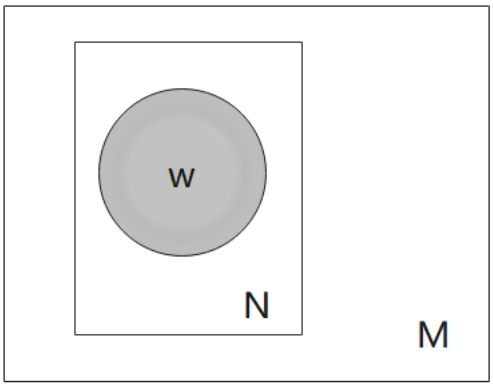


图 2.1 一元模型

对于文档$\mathrm{w}=(w_1,w_2,\ldots,w_N)$，用$\mathrm{p}(w_i)$表示词$w_i$的先验概率，生成文档的概率为：

$$\mathrm{p}(\mathrm{w})=\prod_{i=1}^{N}p(w_i) \qquad (2.1)$$

其中，w_i表示离散的词汇，$\mathrm{p}(w_i)$表示获取一篇文档词w_i的概率分布。本质上，这个模型就是对每个单词的词频进行统计。

一元混合模型的概率模型图如下图所示。

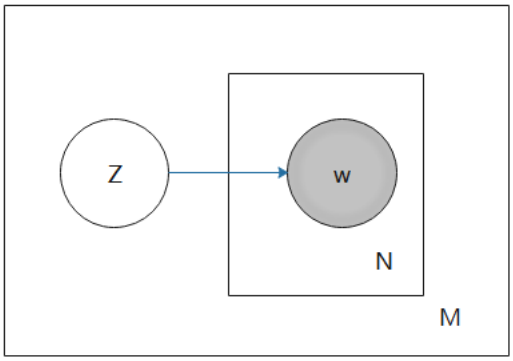


图 2.2 一元混合模型

一篇文档只由一个主题生成。给定某个文档先选择一个主题z，再根据该主题生成文档，该主题中的所有词都来自一个主题。假设主题有$z_1,z_2,\ldots,z_k$，生成语料w的概率为：

$$\mathrm{p}(\mathrm{w})=\mathrm{p}(z_1)\prod_{n=1}^{N}p(w_n|z_1)+\cdots+\mathrm{p}(z_k)\prod_{n=1}^{N}p(w_n|z_k)$$

$$= \sum_z p(z) \prod_{n=1}^{N} p(w_n|z) \quad (2.2)$$

其中，w表示可观测变量，z表示未知的隐变量，p(w)表示获取语料w的概率分布。

潜在语义分析模型（LSA）将“词-文档”由高维稀疏向量空间映射到低维向量空间，该空间也被称为“潜在语义”空间，由 Scott Deerwester 等人（1988）提出。在低维的潜在语义空间，词项之间会呈现一定的语义信息，通过该信息可以进行高层次的分析。

LSA 通过一种奇异值分解技术（Singular Value Decomposition，SVD），将词项与文档都投射到低维的空间，用一个确定性长度为 T 的向量来表示每个词项与每篇文档，不同的维度表示不同的潜在语义类别，而每个值表示词项与文档跟这个潜在语义之间的关联程度。如果词项与文档投射的恰好是等同的潜在语义空间，则词汇-文档，词汇-词汇，文档-文档之间的分析就容易进行，词汇层面的信息也容易获取。降维过程如下图所示：

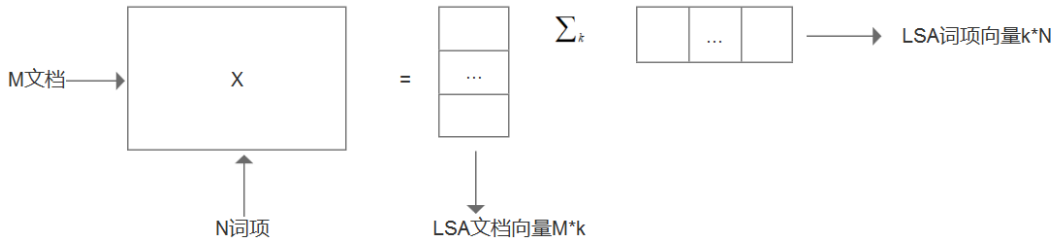


图 2.3 潜在语义分析模型

然而，可以看出，缺点是显而易见的。LSA 的缺陷在于，在进行信息提取过程中，词项在一篇文档中的顺序等语法信息被忽视了，该模型认为在文本的语义表达过程中，语法等结构处于次要的位置。因此，本质上而言，它仍然是一种采用词语向量的线性加总来表示文本向量，从而表示文本含义的方法。然而，由于语法结构等信息的缺失，在一篇文档中，LSA 模型并没有很好的表示出不同词语之间的更深层次的语义关联信息。

概率潜在语义分析（PLSA）是在潜在语义分析模型基础上，引入了概率主题，进而演化而来。由 Hoffmann 等人（1999）提出。通过引入一个隐变量“主题”的方式，就能够对文档进行降维与语义分析，为解决 LSA 模型对一词多义

无从下手的问题，可以采用 EM 算法对参数实行估计。

概率潜在语义分析模型图如下图所示。

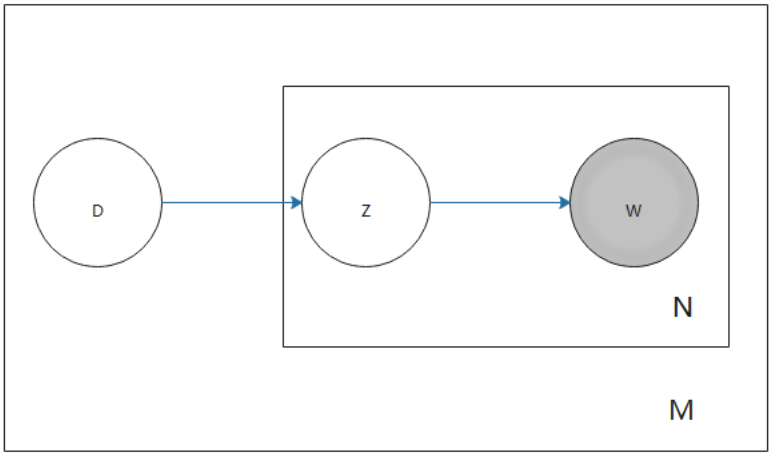


图 2.4 概率潜在语义分析模型

根据该概率模型图，可以得出：

$$\mathrm{p}(\mathrm{d}, w_n) = \mathrm{p}(\mathrm{d})\mathrm{p}(\mathrm{w}|\mathrm{d}) = \mathrm{p}(\mathrm{d}) \sum_z p(w, z|d)$$

$$= \mathrm{p}(\mathrm{d}) \sum_z p(w|z)p(z|d) \tag{2.3}$$

对比于 LSA 模型，通过引入概率主题，PLSA 模型的概率假设更贴切文本特性。但该模型还有一些不足，PLSA 并不是一个完整的生成模型，因为该模型在处理新的文档时必须要事先确定文本，然后才可以进行随机抽样。另外，训练数据的增加带来的一个结果是训练参数也随之线性增加，复杂度会增高，EM 推断算法计算量也会增加。针对这些存在的问题，为更好地表示文本的生成过程，就有研究者提出了 LDA 主题模型。

LDA 主要包含词项、主题和文档三层结构。LDA 模型的基本思想可以这样理解：在一篇文章之中，每个词都服从一定的概率分布，而词语之间存在关联性，不同的词语可以聚类为不同的主题，用出现概率最大的一个或者多个主题表示该篇文章的意思。其中，文档到主题、主题到词都服从多项式分布。

2.2.2 LDA 模型

LDA 主题模型是一种概率生成模型，它通过假设文档-主题分布、主题-词汇分布的参数都服从狄利克雷分布，以此来构造一个三层贝叶斯网络结构。生成模型指那些能够反映给定输入与模型输出之间生成关系的模型，因此 LDA 模型能够通过给定的先验参数，随机产生一些有意义的观测数据。如果要构成一篇文章，其文本信息的每个词语出现的概率为：

$$P(词项|文档) = \sum 主题P(词项|主题) * P(主题|文档)$$

公式用矩阵表示如下图所示：

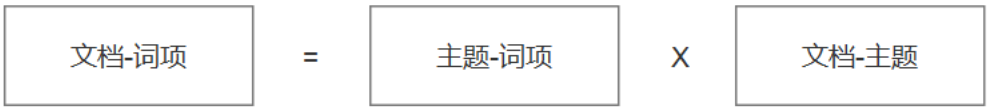


图 2.5 文档-主题-词汇矩阵表达

“文档-词项”矩阵表示的是每个词项在某篇文档中的概率分布；

“主题-词项”矩阵表示的是每个词项代表某个主题的概率分布；

“文档-主题”矩阵表示的是每个主题出现在某篇文档中的概率分布。

LDA 模型的基本思想是以词袋模型为假设前提，所谓词袋模型，就如同把文档中所有词项都打乱放进一个密封的袋子里，认为词项出现的位置都是可以交换的，完全不考虑每个词项出现的先后顺序。词袋假设的意义就在于可以将文本信息转化为方便建模的数字信息。下图展现的是 LDA 模型的图模型。

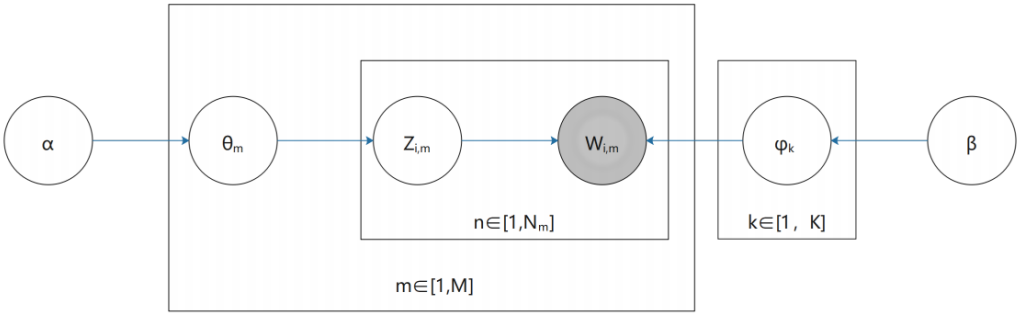


图 2.6 LDA 图模型

表 2.1 变量解释

符号	解释
M	文档的数量
K	主题的个数
V	词袋的长度
N_m	第 m 篇文章中单词的总数
α	是**每篇文章的主题分布**的先验分布 Dirichlet 分布的参数（也被称为超参数）通常是手动设定的
θ_m	θ是一个 M*K 的矩阵，θ_m表示第 m 篇文章的主题分布，$\theta_m \sim Dir(\alpha)$是我们要求的参数
$Z_{i,m}$	第 m 篇文章第 i 个词被赋予的主题，隐变量
$W_{i,m}$	第 m 篇文章第 i 个词，这个是可以被我们观测到的
β	是**每个主题的词分布**的先验分布 Dirichlet 分布的参数（也被称为超参数）通常是手动设定的
φ_k	φ是一个 K*V 的矩阵，φ_k表示第 k 个主题的词分布，$\varphi_k \sim Dir(\beta)$是我们要求的参数

Rickjin（2013）将 LDA 模型的生成过程比喻为投掷多面骰子的游戏，并给出以下两个等价的 LDA 生成过程。其中文档-主题分布、主题-词汇分布对应不同的骰子，这些骰子具有不同的先验。语料库中的每个词汇都是按照“文档-主题-词汇”的顺序进行生成。这是因为 LDA 模型假设一篇文档可以包含多个主题，而且这些主题都能由文档中的词汇表示，语料库中词汇的生成方法具体如下：

i. 对语料库中的每个文档，根据Dirichlet(α)分布随机抽取一个文档-主题骰子θ_m；
ii. 根据Dirichlet(β)分布进行随机抽样得到 K 个主题-词汇骰子φ；
iii. 对该文档中的每个词汇：

a) 先根据Multinomial(θ_m)随机抽样一个主题$Z_{i,m}$；

b) 再根据Multinomial($\varphi_{Z_{i,m}}$)随机抽样一个词汇$W_{i,m}$。

对于同样的模型假设，第二个生成过程将“文档-主题-词汇”的生成过程进行了割裂。即在该过程中，模型会先生成文档中所有可能包含的主题，在根据这些主题分布生成相应的词汇。

i. 根据Dirichlet(α)分布随机抽取 D 个文档-主题骰子θ；

ii. 根据Dirichlet(β)分布进行随机抽样得到 K 个主题-词汇骰子φ；

iii. 对语料库中的每个词汇：

a) 先根据Multinomial(θ_m)随机抽样一个主题$Z_{i,m}$；

b) 再根据Multinomial($\varphi_{Z_{i,m}}$)随机抽样一个词汇$W_{i,m}$。

上述生成过程反映了该模型的基本思想：待分析文档集合中蕴含若干个独立的隐含主题，每个文档都是部分主题的混合；每个主题都能由构成文档集合的词汇表示。其中每个主题都是文档集合上相同或相关信息的精简表示。因此 LDA 模型能够将由高维稀疏的词集合表示的文档，映射为低维表示。

LDA 主题模型通过各个词汇在每篇文档中共现次数的概率统计挖掘自身潜在的语义结构，其词袋假设不仅使得文档当中存在的不确定性和噪声干扰得到有效解决，还同时具备降维能力。更重要的是，LDA 主题模型属于贝叶斯网络模型，扩展能力较好，对于各种元数据、结构化信息等皆可被引入到模型中，不同主题之间的碰撞也能产生恰当且新颖的主题结构。

根据 LDA 原理可得，模型生成过程中的已知参数和未知参数的联合分布为：

$$\mathrm{p}(w_m, z_m, \theta_m, \varphi|\alpha, \beta) = \prod_{n=1}^{N_m} p(w_{m,n}|\varphi_{z_{m.n}}) p(z_{m,n}|\theta_m) p(\theta_m|\alpha) p(\varphi|\beta) \quad (2.4)$$

一般采用 Gibbs Sampling 与 EM 算法等方法进行参数估计。

2.2.3 LDA 模型理论基础

（1）贝叶斯定理

贝叶斯定理也被称为贝叶斯法则，它是在概率统计领域中，基于所观察到的现象对概率分布的先验概率进行修正的一种方法，当所研究的样本量大到接近总体样本量时，该事件发生的概率将接近于总体中事件发生的概率值。

贝叶斯公式

$$\mathrm{P(AB)} = \mathrm{P(A)} * \mathrm{P}(B/A) = \mathrm{P(B)} * \mathrm{P}(A/B)$$

$$\mathrm{P}(A/B) = \mathrm{P}(\mathrm{B/A}) * \frac{\mathrm{P(A)}}{\mathrm{P}(B)} \quad (2.5)$$

注释：

P(A)是 A 的先验概率或边缘概率，称作“先验”是因为它不考虑 B 因素；

$\mathrm{P}(A/B)$是已知 B 发生后 A 的条件概率，又称作 A 的后验概率；

$\mathrm{P}(B/A)$是已知 A 发生后 B 的条件概率，又称作 B 的后验概率，称作似然度；

P(B)是 B 的先验概率或边缘概率，也被称作标准化常量；

P(B/A)/P(B)称作标准似然度

贝叶斯法则又可表述为：

后验概率 = （似然度 * 先验概率）/标准化常量 = 标准似然度 * 先验概率

（2）迪利克雷分布

迪利克雷分布是概率数学知识中一种重要的分布形式，与多项式分布存在共轭关系，又叫贝塔多项式分布。定义：

i. 在[0，1]间均匀分布中，抽取n个随机数字，即$X_1, X_2, \dots, X_n \sim Uniform(0,1)$;

ii. 将以上n个随机数依次排序，得到的顺序统计量为：$X_1, X_2, \dots, X_n$.

Dirichlet 函数是 Beta 函数的高维推广，则其一般形式的概率密度函数为：

$$\mathrm{f}(X_1, X_2, \dots, X_k; \alpha_1, \alpha_2, \dots, \alpha_k) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha^i - 1} \quad (2.6)$$

其中

$$B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha^i)}{\Gamma\left(\sum_{i=1}^{k} \alpha^i\right)} \qquad \alpha = (\alpha_1, \dots, \alpha_k) \quad (2.7)$$

Dirichlet 分布和多项式分布存在共轭的关系，用贝叶斯思想表达：

$$\mathrm{Dir}(\vec{p}|\vec{\alpha}) + \mathrm{Multcount}(\vec{\mathrm{m}}) = \mathrm{Dir}(\vec{p}|\vec{\alpha} + \vec{m}) \quad (2.8)$$

迪利克雷分布在自然语言处理等方面运用的非常广泛，一方面可以给模型带来较好的概率分布解释，另一方面，也可以减少参数推断算法的计算量，减少了模型参数估计时的复杂度。

（3）吉布斯采样

参数推理是 LDA 模型的核心内容，想要对参数进行直接求解非常困难，需要借助文本中的词汇作为可观测变量，所以主题的产生就是对 LDA 模型参数求

解的过程。常见的参数推理算法有很多种，主要包括 EM 算法、变分(Variational inference)算法、Gibbs 算法等。Gibbs 算法是马尔科夫链最简单的实现形式之一，它是一种马尔科夫链蒙特卡洛理论中用来获取一系列近似等于多维概率分布观察样本的算法。具有计算效率较高且实现难度较低的优点，在科学计算中被广泛使用。Markov 方法相当于一种迭代方法，其中心思想就是从复杂的概率分布中抽取一定的样本值作为可观察对象。然后生成输出一套从参数后验分布中抽取样本的规则。

后验概率为$\mathrm{p}(\mathrm{z}_i = \mathrm{k}|\mathrm{z}_{-i}, \mathrm{w}_i)$，计算公式如下：

$$\mathrm{p}(\mathrm{z}_i = \mathrm{k}|\mathrm{z}_{-i}, \mathrm{w}) = \frac{n_{-i,d}^{(k)} + \alpha_k}{\sum_{s=1}^{K}(n_{-i,d}^{(s)} + \alpha_s)} \cdot \frac{n_{-i,k}^{(t)} + \beta_t}{\sum_{f=1}^{V}(n_{-i,k}^{(f)} + \beta_f)} \quad (2.9)$$

在这里，$\mathrm{z}_i = \mathrm{k}$表示文本中任意一个词语 w 可能属于主题 k，$\mathrm{z}_{-i}$代表去掉下标为i的词对应的主题后的主题分布，$n_{-i,d}^{(k)}$代表在第 d 篇文档中，第 k 个主题的词的个数，$n_{-i,k}^{(t)}$代表第 k 个主题中，第 t 个词的个数。

Gibbs 抽样算法过程如下：

i. 训练流程：

a) 选择合适的主题数 K，选择合适的超参数α，β；

b) 对应语料库中的每一篇文档的每一个词，随机赋予它一个主题编号z；

c) 重新扫描语料库，对每一个词，利用 Gibbs 采样公式更新它的主题编号，并更新语料中该词的编号；

d) 重复第 3 步的基于坐标轴轮换的 Gibbs 采样，直到 Gibbs 采样收敛为止；

e) 统计语料库中的各个文档各个词的主题，得到文档主题分布θ_d，统计语料库中各个主题词的分布，得到 LDA 的主题与词的分布φ_k。

ii. 预测流程：

a) 对应当前文档的每一个词，随机的赋予一个主题编号z；

b) 重新扫描当前的文档，对于每一个词，利用 Gibbs 采样公式更新它的主题编号；

c) 重复第 2 步的基于坐标轴轮换的 Gibbs 采样，直到 Gibbs 采样收敛；

d) 统计文档中各个词的主题，得到该文档的主题分布。

$$\varphi_W = \frac{n_i^{(W)} + \beta}{n_i^{(\cdot)} + W_\beta} \quad (2.10)$$

$$\theta_{z=i} = \frac{n_i^{(d)} + \alpha}{n^{(d)} + T_\alpha} \quad (2.11)$$

其中W_β是服从Dir(β)分布的所有词总数，T_α是服从Dir(β)分布的所有主题总数。

（4）EM 算法

EM 算法通过采用极大似然估计的方法对参数的值进行近似值估计，不断地迭代，每次迭代产生的结果作为下次迭代的初始值，直到获取到某一个理想的结果。

E-步骤：

根据参数初始值或上一次迭代的模型参数来计算出隐性变量的后验概率，其实就是隐性变量的期望。作为隐藏变量的现估计值：

$$Q_i(z^{(i)}) := p(z^{(i)}|x^{(i)};\theta) \quad (2.12)$$

M-步骤：

将似然函数最大化以获得新的参数值：

$$\theta := \text{argmax} \sum_i \sum_{z^{(i)}} Q_i \left(z^{(i)} log \frac{p(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})} \right) \quad (2.13)$$

这个过程不断迭代，最终可以求得使似然函数L(θ)最大化的参数θ了。

2.2.4 LDA 模型最优主题数确定方法

利用什么方法来确定最优的主题数目决定着后面的主题挖掘效果。本文采用的方法为困惑度指标 [33]。困惑度一般用来评估 LDA 主题模型的优劣程度，能够对模型的性能进行量化评价，困惑度越低表示该主题数目下的模型泛化能力越强，计算公式如下：

$$\text{Perplexity(D)} = \exp\{-\frac{\sum_{m=1}^{M} log_D P(w_m)}{\sum_{m=1}^{M} N_m}\} \quad (2.14)$$

$$P(w_m) = \sum_d \prod_{n=1}^{T} \sum_{j=1}^{T} p(w_j|z_j = j) p(z_j = j|w_m) p(d) \quad (2.15)$$

其中，D 表示语料库中的测试集，M 表示文档数量，N_m表示在文档m中拥

有的单词数量，$P(w_m)$表示w_m产生的概率，困惑度反映了模型的预测能力。

对于一篇文档 d，我们的模型对文档 d 属于哪个主题的不确定性程度就是困惑度，在其他条件固定的情况下，主题数越多，困惑度就越小，但是与之带来的问题是过拟合。所以恰当大小的困惑度是确定主题数的重要参考标准。

3 古典诗词在不同历史时期的主题模型实现

3.1 数据来源

本文数据来源于 GitHub 公开数据集[①]，共收集了 738881 首古典诗词，其中历史时期分布为从先秦到近代（注：本文历史时期划分没有细化，例如没有呈现中国历史上的西夏、金等朝代）。该数据基本包含了在文献中有记载的，且较为完整的古典诗词。具体数据汇总见表 3.1。

表 3.1 古典诗词在不同历史时期分布情况

历史时期	时间跨度	诗词数	作者数
先秦	旧石器时期～前 221 年	570	8
秦	前 221 年～前 207 年	2	2
汉	前 202—220 年	363	83
魏晋	220 年—420 年	3020	251
南北朝	420 年～589 年	4586	434
隋	581 年-618 年	1170	84
唐	618 年-907 年	49195	2736
辽	907 年—1125 年	22	7
宋	960 年~公元 1279 年	287114	9446
元	1271 年—1368 年	37375	1209
明	1368 年—1683 年	236957	4439
清	1636 年-1912 年	90088	8872
近代	1840 年-1949 年	28419	790
总和	/	738881	28361

3.2 数据预处理

数据预处理主要包括清洗掉原始数据集中不满足条件的数据（如一些显示有

[①] Dataset webpage.https://github.com/Werneror/Poetry

误的字的数据）、分词和制作停用词表词典等。本文首先用 Excel 对获得的原始古典诗词经过人工数据清洗。数据清洗的目的是为了使数据满足模型计算所需格式的要求，以结构化的形式表示出来，然后进行分词后，制作停用词表和词典。最后得到满足 LDA 模型计算的数据集。以下是数据预处理的主要步骤，其中分词与制作停用词表和词典是一个循环往复的过程。

图 3.1 预处理流程图

分词是进行主题模型分析必不可少的环节，而且精准地分词对正确理解古典诗词有着极为重要的作用。中文分词方法大致可以分为统计分词法、理解分词法和词典分词法 3 类。目前国内可以用于分词工具有很多，例如 Jieba、SnowNLP、Hanlp、PKUSeg、LTP、THULAC、BaiduLac 与阿里云 NLP 等，笔者对比了 Python 的几种常用分词工具对王昌龄诗作“烽火城西百尺楼，黄昏独上海风秋。更吹羌笛关山月，无那金闺万里愁。”（《从军行七首 其一》）进行分词，以便查看其效果[34]。

表 3.2 几种常用分词工具对古典诗词的分词效果

分词工具	分词结果
Jieba 分词	|烽火|城西|百尺|楼|，|黄昏|独|上海|风秋|。 |更|吹|羌笛|关山月|，|无|那金|闺|万里|愁|。
SnowNLP 分词	|烽火|城西|百尺|楼|，|黄昏|独|上海|风|秋|。 |更|吹|羌笛|关|山月|，|无|那|金|闺|万|里愁|。
PKUSeg 分词	|烽火|城西|百尺楼|，|黄昏|独|上海|风秋|。 |更|吹|羌笛|关山月|，|无|那|金|闺|万里|愁|。
THULAC 分词	|烽火城|西|百|尺|楼|，|黄昏|独|上海|风秋|。 |更|吹|羌笛|关山月|，|无|那|金闺万|里|愁|。
BaiduLac 分词	|烽火|城西|百尺|楼|，|黄昏|独|上海|风秋|。 |更|吹|羌笛|关山月|，|无|那|金闺|万里|愁|。

分词的结果不同可能会造成句子的歧义，例如“黄昏独上海风秋”这一句，

以上的分词工具都把“上海”单独作为一个词语，但原意为“独上|海风|秋”，独自登上戍楼台，任凭从沙海吹来的秋风撩起自己的战袍；此外，“关山月”是羌笛的曲子名，SnowNLP 将之分成“关|山月”;“无那金闺万里愁”,“无那”是“无奈”的意思，“金闺”特指自己的妻子，以上分词的效果显然不佳。

究其原因，在于当前常用的中文自然语言处理工具包大多是基于现代汉语语料训练的分词模型，对于切分古典诗词效果有限，例如，上段对于“独上|海风|秋”的切分，“上海”是现代汉语语料中的常用地名词，但原文诗句意思并非如此。

Github 有一个叫“Jiayan”的文言文 NLP 处理工具包[35]，能够提供文言文的分词、断句、标点和词性标注功能。Jiayan，取“甲骨文言”之意，是一款专注于古汉语处理的 NLP 工具包。使用 Jiayan 对“烽火城西百尺楼，黄昏独上海风秋。更吹羌笛关山月，无那金闺万里愁。”进行分词，可以看到如下结果：

|烽火|城西|百尺楼|，|黄昏|独|上|海风|秋|。

|更|吹羌笛|关山|月|，|无|那|金闺|万里|愁|。

我们发现，即使有一定程度上的信息损失，但是分词效果要明显好于其他分词工具的结果。

3.3 LDA 模型实现

为便于主题聚类，笔者将诗歌数量低于 1000 的历史时期进行剔除，最终对以下历史时期的古典诗词做分析。

表 3.3 诗词主题聚类的选定历史时期

历史时期	诗词数
魏晋	3020
南北朝	4586
隋	1170
唐	49195
宋	287114
元	37375
明	236957
清	90088
近代	28419

应用 LDA 模型对古典诗词进行建模之前，需要估算其最优主题数目，因而

在实验中选取不同的值来运行 Gibbs 抽样算法，最终确定最优主题数目。对 LDA 模型进行相同次数的迭代，除了主题数 K 的选择作为变量以外，其他的参数都是相同的。利用“甲言”分词工具，输出的格式[36]如下所示：

表 3.4 分词输出格式

[M]
D_1{[W_{11}],[W_{21}],[W_{31}],[W_{41}],[W_{51}],……[$W_{I_1 1}$]}
D_2{[W_{12}],[W_{22}],[W_{32}],[W_{42}],[W_{55}],……[$W_{I_2 2}$]}
D_3{[W_{13}],[W_{23}],[W_{33}],[W_{43}],[W_{53}],……[$W_{I_3 3}$]}
D_4{[W_{14}],[W_{24}],[W_{34}],[W_{44}],[W_{54}],……[$W_{I_4 4}$]}
D_5{[W_{14}],[W_{24}],[W_{34}],[W_{44}],[W_{55}],……[$W_{I_5 5}$]}
……
D_m{[W_{1m}],[W_{2m}],[W_{3m}],[W_{4m}],[W_{5m}],……[$W_{I_m m}$]}

其中，M 为语料总体。D_i为文本语料集的一个文档记录， W_{ij}为D_i中的第 j 个词。首先对“魏晋”时期的古典诗词进行分词，因此 M=3020，分词结果如下所示：

表 3.5 魏晋时期古典诗词的分词输出格式

M=3020
D_1{[吴王],[剑客],[百姓],[疮瘢],[楚王],……[饿死]}
D_2{[兰若],[生],[春],[阳],[涉冬],……[狂痴]}
D_3{[兰草],[自然],[香],[生],[镰],……[束薪]}
D_4{[孟冬],[寒气],[至],[北风],[惨栗],……[愁多]}
D_5{[悠悠],[四顾],[茫茫],[东风],[百草],……[盛衰]}
……
D_{3020}{[柳条],[恒],[着地],[弱柳],[荫],……[倚松]}

对古典诗词分词之后，进行频率统计，会发现有很多例如“兮、长、将、所、

在、哉、斯、乃、亦…”的停用词出现，目前的常用中文停用词库，多是基于现代语料库训练出的结果，例如川大、哈工大以及百度停用词库，还尚未有古典诗词专门的停用词库直接可以利用。针对古典诗词中出现的一些无意义的停用词，目前研究中主要采用两种方法，一种方法是根据词频统计分析，人为的将这些分词之后出现的停用词加到停用词库中来，还有一种是利用 sklearn 里面的 CountVectorize 函数过滤掉在文档中出现超过一定频率与低于一定频数的术语，分别用 max_df 与 min_df 两个参数表示，实验过程中，文章采用的是 max_df=0.5，min_df=10。

确定文本聚类的最佳主题数目对于后续分析尤为重要，通过困惑度对最佳主题数进行确定是一个可行的方法。根据手肘法，当曲线位于最低的拐点处困惑度最小的位置就是最佳主题的位置。绘制魏晋时期古典诗词的困惑度与主题数的折线图，我们发现当K^*取 2 的时候，主题颗粒度过大，主题过于集中；当K^*越往后取的时候，颗粒度过小，主题过于分散，分析的实际意义都有所丧失。经过反复测试，笔者取第二个肘折点的位置，即$K^*=5$。

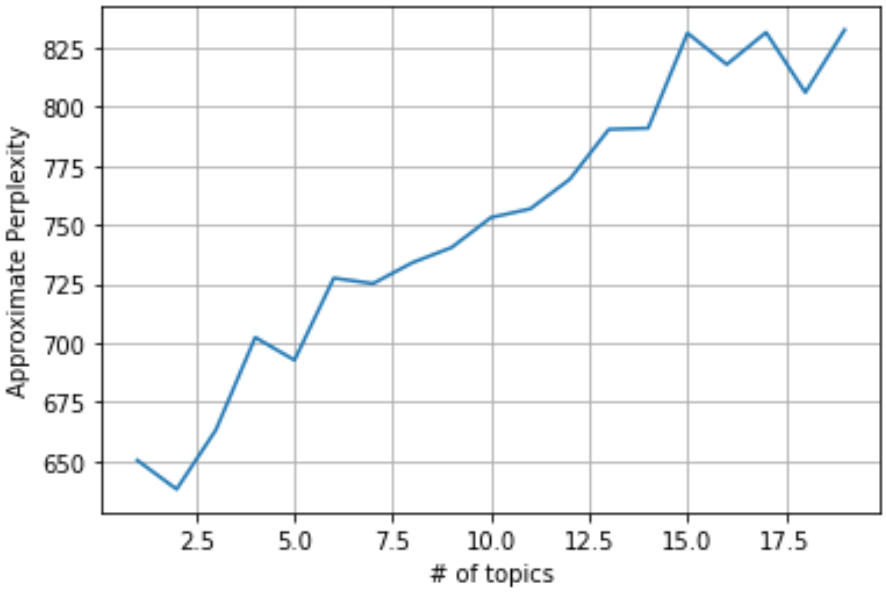


图 3.2 魏晋时期古典诗词的困惑度与最佳主题数折线图

在设置最佳主题个数之后，开始迭代实验。设置最大迭代次数 max_iter=100，保存记录的步长 evaluate_every=10，样本训练影响消除参数 learning_offset=150.，文档-主题先验概率$\alpha = {}^{1}/_{K^*}$， 主题-词先验概率$\beta = {}^{1}/_{K^*}$，得到魏晋时期的运行

结果如下表所示。

表 3.6 魏晋时期古典诗词的主题输出

主题	关键词（部分）
Topic 0 相思恋人	明月 佳人 叹息 与君 蟋蟀 秋风 岁暮 悲风 今日 远望 灼灼 憔悴 漫漫 烈烈 凄凄 思君 西山 牵牛 终日 彷徨 常恐 织女 鸡鸣 依依 踟蹰 忽如 春华 萧索 罗裳 悲鸣 泣涕 遥遥 泪下 弦歌 北风 褰裳 露沾 被服 秋霜 贱妾 金石 朱颜 秋夜 独无 幽人 暧暧 远行 我心 零落 泠泠
Topic 1 翩翩少年	千里 翩翩 松柏 四海 我心 少年 出门 徘徊 青云 九州 羽翼 万里 乘云 蓬莱 五岳 遥望 飞鸟 神仙 悠悠 平生 黄鹄 贫贱 眇眇 白马 玄云 扬州 富贵 西南 西北 华盖 千年 登高 嵯峨 东南 亲友 陛下 荆棘 一朝 来归 须臾 参天 驱车 忧思 归来 父母 随风 赤松 欢乐 游子 遨游
Topic 2 逍遥君子	君子 逍遥 慷慨 天下 俯仰 千载 古人 自然 天地 扶桑 日月 远游 松乔 清流 朝阳 穷达 风尘 世间 良辰 感物 清歌 绸缪 弱冠 闲居 凤凰 昆仑 窈窕 凯风 众鸟 无穷 自古 羲和 天道 各异 丹霞 峨峨 长生 明德 夫子 六龙 弹琴 洪波 王子 风流 达人 紫霞 相忘 随风 沧海 登城
Topic 3 离别惆怅	窈窕 鸳鸯 同心 殷勤 交颈 芙蓉 徘徊 佳人 日月 相思 流水 登高 好音 高山 形影 辛苦 行人 泛舟 公子 折杨柳 行役 戢翼 别离 清风 故人 缠绵 我心 惆怅 踟蹰 郁郁 故乡 山川 饥寒 万里 长叹 苦心 淹留 骨肉 河水 衣裳 一人 借问 望舒 夜光 春秋 冬夏 感物 命驾 寤寐 百忧
Topic 4 圣皇神明	穆穆 四海 万国 圣皇 我皇 赫赫 天地 四方 天下 六合 文武 神明 巍巍 天子 宇宙 受命 无疆 万邦 盛德 圣德 礼乐 神武 神祇 八风 峨峨 天命 八音 煌煌 享祀 率土 万物 龙飞 应天 皇祖 克昌 鹰扬 多士 天人 文皇 顺天 祖考 万世 福禄 圣明 昊天 皇极 大业 开元 永世 邦家

按照同一步骤，再对南北朝、隋、唐、宋、元、明、清、近代时期分别进行分析，可以得到最终的 LDA 主题聚类结果如表 3.8 所示。

表 3.7 中国古典诗词在不同历史时期的主题分布

朝代	Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
魏晋	相思恋人	翩翩少年	逍遥君子	离别惆怅	圣皇神明	
南北朝	闺中佳人	君臣之道	与君别离	羁旅生涯	千里思君	
隋	道法自然	太上神明	春风杨柳	万国皇祖		
唐	家国无事	读书圣朝	岁月悠悠	千古风流	浪迹天涯	人间风尘
宋	国事愁绝	人间富贵	春光十里	世间幽居	读取功名	
元	回首当年	人间世事	君王朝廷	神仙逍遥	借景排忧	
明	山中幽栖	纵横江山	世间问道	他乡惆怅	与君离别	
清	将军封侯	他乡归来	我心伤悲	茫茫浮生	富贵功名	往事回首
近代	苍茫人间	江山忧患	壮士报国	神州英雄	相思天涯	

4 古典诗词的内容演化与主题强度差异

基于 LDA 的分析结果，可以看出，古典诗词在不同的历史阶段有不同的主题分布。在历史阶段变更过程中，主题内容具有一定的演化趋势，同时，各历史时期的各个主题强度存在差异，本部分分别从主题内容演化角度和主题强度差异[37-42]角度，对不同历史阶段的古典诗词进行分析。

4.1 主题内容演化

古典诗词主题在不同历史阶段呈现一定的差异，但可以看出，具有相同对象的主题会形成同一主题类。例如“国家意识”主题类，“人生羁旅”主题类，“离愁”主题类等。本文根据分析出来的主题，以“国家意识”主题类为例，对不同历史时期的主题内容演化进行分析。

主题演化可视化可以选取桑基图来实现。桑基图（Sankey Diagram）又称为桑基能量分流图，起源于 1898 年的“蒸汽机的能源效率图”。[43]在桑基图中，元素块代表对象，其中的连线表示对象产生能量的流动方向以及联系，利用该特性可以直观表现主题内容随时间推移产生的变化。本文通过同一主题类下不同主题的共词构建联系，从而发现主题内容的具体演化。

在古代，“国家”概念更多的是和“皇帝”相关联，所以在古典诗词中，会出现圣皇、皇极、上帝等词语，在魏晋时期与隋朝时期到达顶峰。根据主题分析结果，本文把以下主题归为“国家意识”主题类，见表 4.1。

表 4.1 古典诗词的“国家意识”主题类

历史时期	“国家意识”主题类		
魏晋	圣皇神明		
南北朝	君臣之道		
隋	太上神明	万国皇祖	
唐	家国无事		
宋	国事愁绝		
元	君王朝廷		
明	纵横江山		
清	将军封侯		
近代	神州英雄	壮士报国	江山忧患

为便于可视化，本文分阶段绘制“国家意识”主题类演化桑基图。图中，大的元素块是主题，小的元素块是词。

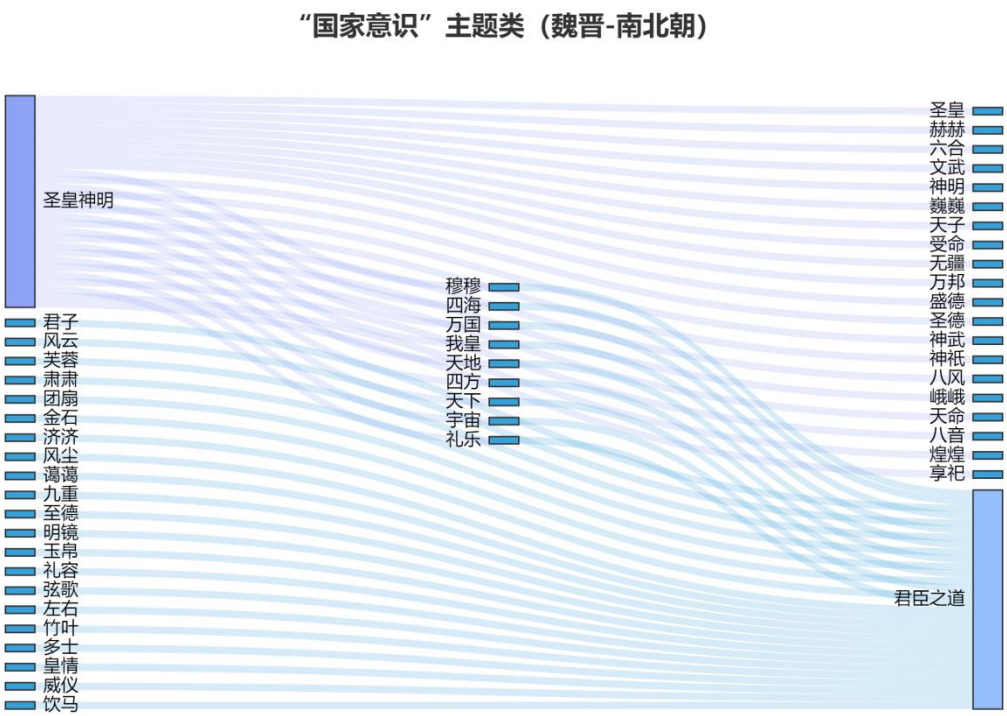


图 4.1 魏晋-南北朝时期的“国家意识”主题类演化

在魏晋时期-南北朝时期，主题由“圣皇神明”演化为“君臣之道”。在魏晋时期的古典诗词中，君王通常和神明具有极强的联系，所以我们看到该该主题下的词更多的是表现君王的盛德、祭祀的庄重与神明的威严。伴随着主题的演化，南北朝时期在继承了魏晋时期“圣皇”主题的基础上，并没有延续“神明”主题，而是逐渐转换为“君与臣”之间的关系中来，我们看到该主题下，有君子、济济、多士与皇情等词。这种转换更多的反映了当时军阀割裂背景下，君王对有才之士的渴望。

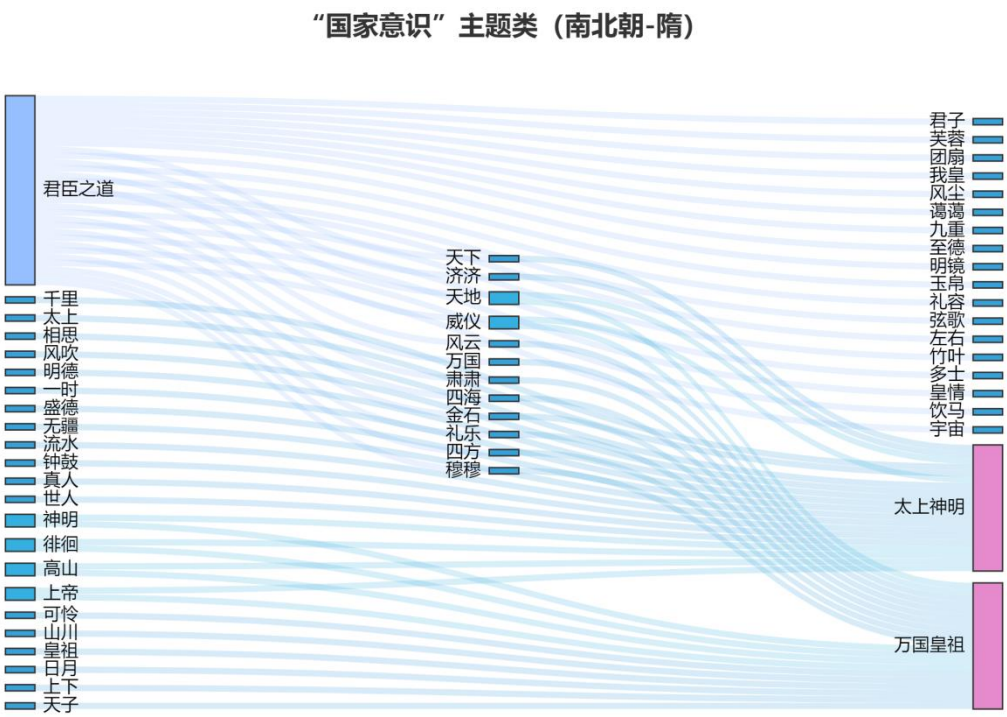


图 4.2 南北朝-隋时期的“国家意识”主题类演化

在南北朝时期-隋时期，主题由“君臣之道”演化为“太上神明”与“万国皇祖”。隋朝的统一结束了古代中国自东晋到南北朝以来的战乱不断，王侯割据的局面，使中华大地重新统一于中央集权之下。这个时候，普天之下，莫非王土，于是天下对君王的尊崇重新到达了一个高峰。而“君与臣”之间的关系就弱化了。

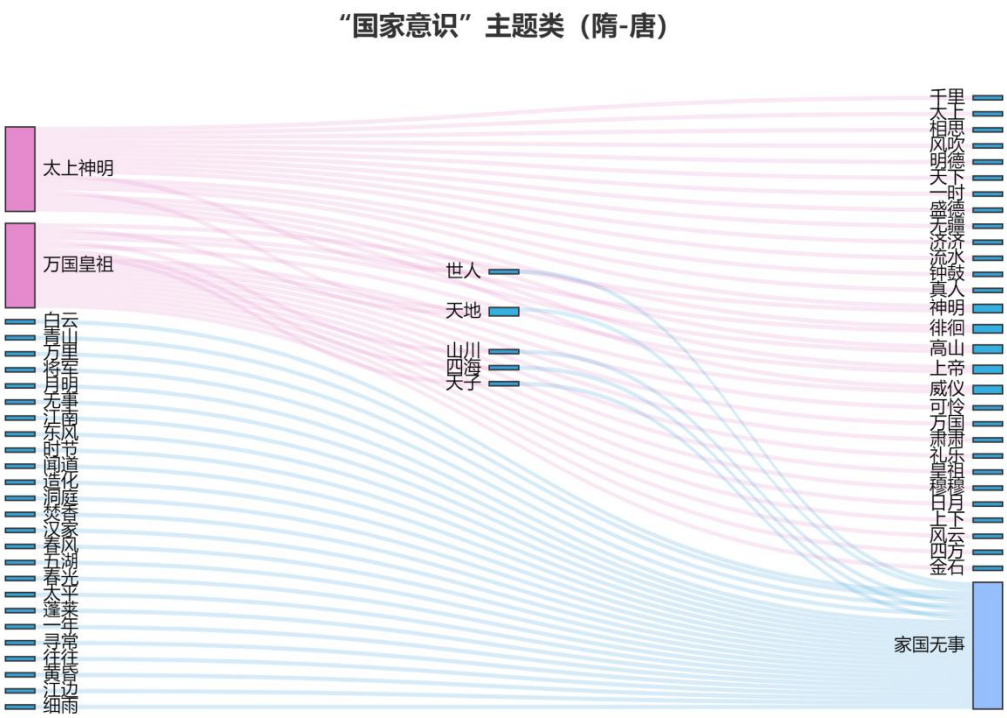


图 4.3 隋-唐时期的“国家意识”主题类演化

到了唐朝时期，主题演化为“家国无事”。唐朝时期，国力强盛，诗歌在这时期到达了顶峰，更多的自然意象在这一时期也出现在诗作中，而这在魏晋时期与南北朝时期是很少出现的。同时将军守边、天下太平的时代环境造就了这一时期“家国无事”的主题。

图 4.4 唐-宋时期的“国家意识”主题类演化

到了宋朝时期，主题演化为“国事愁绝”。宋朝自建立以来，一直与周边游牧民族等政权形成对立局面[44]，这种现象在之前是很少见到的。“国家意识”主题色彩多了好几分忧患、愁苦。呜呼、愁愁、怅然等直抒胸臆的情感表达词增多；旧时、故国、忆昔等思恋过往的词也经常出现；老夫、白发、孤城等内心的无奈更是溢于言表。

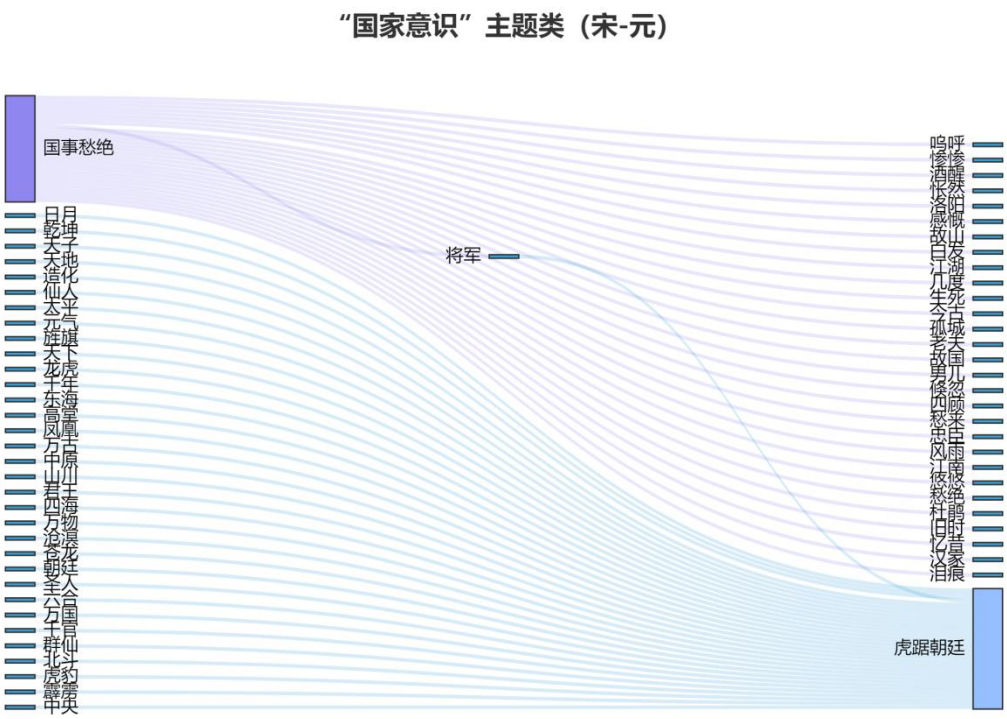


图 4.5 宋-元时期的“国家意识”主题类演化

到了元朝，主题演化为“君王朝廷”。元朝是中国历史上首度由游牧民族建立的大一统帝国，其所建立的疆域空间深刻影响了后代的版图意识[45]。此外，元代诗词（包括元曲）豪放、清丽，意象清晰，同时该时期关于天子威严的作品是自隋朝以来的又一个高峰。

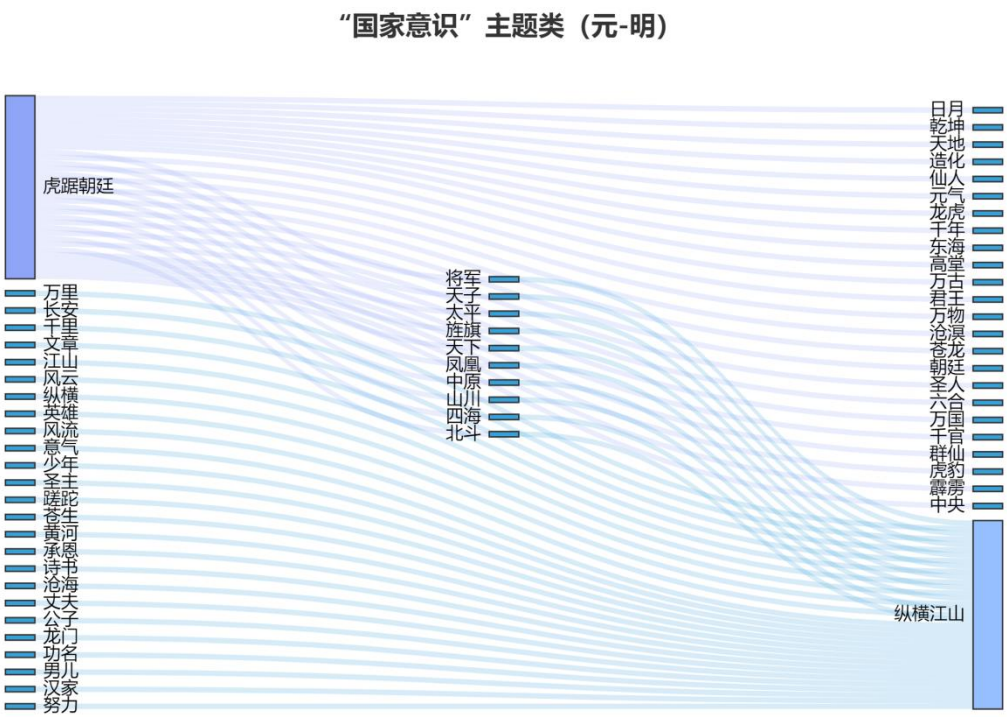


图 4.6 元-明时期的“国家意识”主题类演化

到了明朝，主题演化为“纵横江山”。一些元朝时期特有的诗词作品元素不再出现，取而代之的是关于个人功名与天下苍生的作品元素。在明朝的古典诗词作品中，没有唐朝“家国无事”主题中经常出现的自然意象；也没有宋朝“国事愁绝”主题中悲怨的感情色彩。国家意识在这一时期更多地会与个人的功名紧紧联系在一起，例如作品中出现的龙门、男儿、努力等词，这与明朝重视教育，将科举发扬光大，勉励天下苍生努力考取功名的历史背景具有一定的关系。

图 4.7 明-清时期的“国家意识”主题类演化

清朝与明朝的“国家意识”主题具有鲜明的对比，清朝非常重视开疆拓土，因此，在清朝的古典诗词作品中，“国家意识”主题含有很多的征战、封侯等元素。清朝前期 121 年的开疆拓土，奠定了当时中华大地的全盛疆域，但同时，在外征战的过程中，也流露出别离、辛苦、鸿雁与客心等离愁情感。

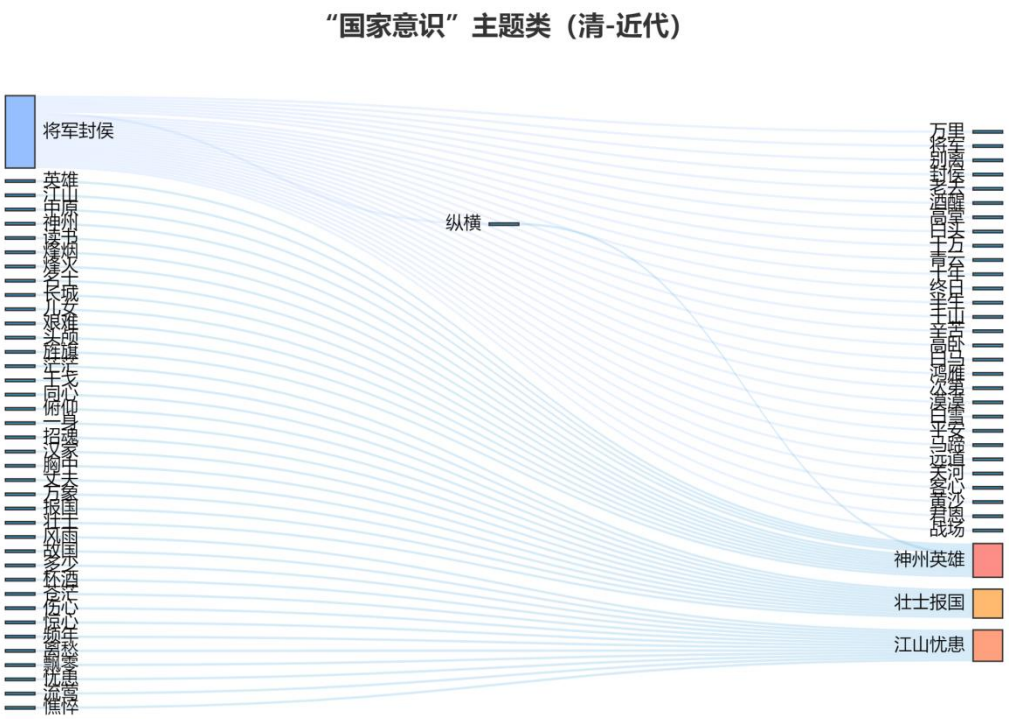


图 4.8 清-近代时期的“国家意识”主题类演化

中国近代是从 1840 年 6 月鸦片战争爆发到 1949 年中华人民共和国成立的历史时期，历经清朝晚期、中华民国临时政府时期、北洋军阀时期和国民政府时期。这段时期，国家的羸弱导致了外敌不断入侵，是一段充满灾难、落后挨打、内忧外患的屈辱史，同时我们也看到既有“读书报国心仍在，种树成林志不灰。”，也有“壮夫报国轻一死，惜哉不见收华中。”，无数的中华儿女抵御外敌、以身报国，这段历史时期，我们牺牲了无数的烈士，付出了惨烈的代价，中国古典诗词中体现出来的对中华大地的忧患意识到达了历史的顶峰。同时，我们崇尚英雄，无数的英雄赞歌这一时期得以涌现，“喜看稻菽千重浪，遍地英雄下夕烟。”，“襄樊大将人中龙，转危为安真英雄。”。

通过以上古典诗词的主题演化分析，可以看出，“国家意识”的转变经历了三个阶段，一个是以魏晋、南北朝、隋为代表的圣皇崇拜阶段；一个是以唐、宋为代表的国家主体观阶段；一个是以明、清、近代为代表的人民主体观阶段。尤其到了近代时期，普通群众走向历史的视野，经历了长时间的斗争、救亡图存，越来越多的人意识到人民才是国家的主体。从朝廷到百姓，从顶层到万众，“国家意识”概念的外延变得越来越广。

4.2 主题强度差异

本文对主题强度的衡量采用的是支持度指标。支持度指标（Support Index，SI）主要揭示主题的强度维度特征，体现主题的受关注度。通过 LDA 模型得出的文档-主题θ_m矩阵，可以得到主题在文档中的概率分布，进而可以得到一个主题涉及的支持文档合集。支持文档合集指的是以较大概率涉及某个主题的文档的合集，本文中支持文档判定概率的阈值确定为 10%，即若某主题在某个文档中的组成概率大于等于 10%，则将该文档视为该主题的一个支持文档。在主题模型中一篇文档代表一首古典诗词，因此主题支持度是同一历史时期下主题支持文档占总古典诗词的比重[46-49]。用公式 4.1 所示。

$$\mathrm{SI}_{all}^{sup_k} = \frac{Sum_{period}(support_k)}{Sum_{period}(all)} \quad (4.1)$$

式中$period$表示某一历史时期，$support_k$表示主题k下的支持文档，all 表示该历史时期下的所有文档。

笔者首先利用 LDA 模型计算出每个文档的每个主题出现的概率，见表 4.1。然后利用 Excel 筛选出文档-主题概率值大于等于 10%的古典诗词，标注该诗词为主题k，见表 4.2。最后统计该主题下的古典诗词的数量，最后计算该主题下的诗词数量占该历史时期古典诗词总数的比重。需要注意的是，同一个古典诗词可能隶属于多个主题。

表 4.1 魏晋时期古典诗词的文档-主题θ_m矩阵

	0	1	2	3	4
0	0.100002	0.100002	0.100001	0.100003	0.599992
1	0.051513	0.050359	0.05033	0.796194	0.051605
2	0.100333	0.10058	0.598069	0.100436	0.100581
3	0.898313	0.025776	0.025264	0.025574	0.025074
4	0.257085	0.655944	0.029452	0.028662	0.028857
5	0.202905	0.758494	0.012943	0.012874	0.012784
…	…	…	…	…	…
3019	0.103101	0.100002	0.59569	0.101204	0.100002

表 4.2 魏晋时期古典诗词对应的主题标签

	内容	主题
0	吴王好剑客，百姓多疮瘢。…	相思恋人 翩翩少年 逍遥君子 离别惆怅 圣皇神明
1	兰若生春阳，涉冬犹盛滋。…	离别惆怅
2	兰草自然香，生于大道旁。…	相思恋人 翩翩少年 逍遥君子 离别惆怅 圣皇神明
3	孟冬寒气至，北风何惨栗。…	相思恋人
4	回车驾言迈，悠悠涉长道。…	相思恋人 翩翩少年
5	驱车上东门，遥望郭北墓。…	相思恋人 翩翩少年
…	……	……
3019	柳条恒着地，弱柳荫修衢。…	相思恋人 翩翩少年 逍遥君子 离别惆怅 圣皇神明

最后，通过支持度指标计算公式，可以得出魏晋时期每个主题下的文档支持度，分别为：$\mathrm{SI}_{all}^{sup_0} = 0.619536$，$\mathrm{SI}_{all}^{sup_1} = 0.642384$，$\mathrm{SI}_{all}^{sup_2} = 0.687417$，$\mathrm{SI}_{all}^{sup_3} = 0.596689$，$\mathrm{SI}_{all}^{sup_4} = 0.580132$。因此在魏晋时期，主题强度依次排序为：逍遥君子>翩翩少年>相似恋人>离别惆怅>圣皇神明。

魏晋诗词作品中“隐逸”、“逍遥”风格突出，这在一定程度上继承了老庄思想。因此，“逍遥君子”、“翩翩少年”等主题强度在当时历史时期占据了主导地位；同时，很多诗词作者在宫廷斗争中郁郁不得志，远离朝廷，逃避到自然山水之间，“圣皇神明”主题强度最弱。

魏晋时期涌现出很多著名诗词作家，比如曹丕、陶渊明、竹林七贤（嵇康、阮籍、山涛、向秀、刘伶、王戎、阮咸）等。魏晋，是人觉醒的时代，该时代的文人墨客多致力于人格的独立与主体的自由，他们常常思维敏捷，或淡泊朴素，或逍遥自在，又或超然物外。形成该时代这种特殊现象的原因，一方面是老庄思想所带来的影响，另一方面是因为当时社会上下的动荡。如嵇康“齐物养生，与道逍遥。”、曹丕“遨游快心意，保己终百年。”、陶渊明“平生不止酒，止酒情无喜。”。从中无不感受到当时时代逍遥自在的独特特征[50-52]。

通过计算不同历史时期的文档支持度指数，可以得到汇总表 4.3。

表 4.3 不同主题所对应的支持度指数

时期	$SI_{all}^{sup_0}$	$SI_{all}^{sup_1}$	$SI_{all}^{sup_2}$	$SI_{all}^{sup_3}$	$SI_{all}^{sup_4}$	$SI_{all}^{sup_5}$
魏晋	0.619536	0.642384	0.687417	0.596689	0.580132	
南北朝	0.556912	0.511339	0.558875	0.493894	0.55604	
隋	0.923077	0.922222	0.910256	0.909402		
唐	0.560443	0.555707	0.522573	0.555748	0.552475	0.524383
宋	0.647043	0.638314	0.612255	0.640547	0.65808	
元	0.610462	0.422181	0.413565	0.397351	0.770649	
明	0.619467	0.596408	0.640715	0.597518	0.681562	
清	0.464623	0.467232	0.640152	0.559153	0.564481	0.48406
近代	0.5604	0.559168	0.534783	0.552518	0.657975	

表 4.4 各历史时期古典诗词的主题强度顺序

时期	主题强度
魏晋	逍遥君子 > 翩翩少年 > 相似恋人 > 离别惆怅 > 圣皇神明
南北朝	与君别离 > 闺中佳人 > 千里思君 > 君臣之道 > 羁旅生涯
隋	道法自然 > 太上神明 > 春风杨柳 > 万国皇祖
唐	家国无事 > 千古风流 > 读书圣朝 > 浪迹天涯 > 人间风尘 > 岁月悠悠
宋	读取功名 > 国事愁绝 > 世间幽居 > 人间富贵 > 春光十里
元	借景排忧 > 回首当年 > 人间世事 > 君王朝廷 > 神仙逍遥
明	与君离别 > 世间问道 > 山中幽栖 > 他乡惆怅 > 纵横江山
清	我心伤悲 > 富贵功名 > 茫茫浮生 > 往事回首 > 他乡归来 > 将军封侯
近代	相思天涯 > 苍茫人间 > 江山忧患 > 神州英雄 > 壮士报国

由表 4.4 可知，在南北朝时期，“别离”、“相思”主题占据当时时代诗词的主流，例如江淹的“远与君别者，乃至雁门关。”，范云的“东风柳线长，送郎上河梁。”，陈叔宝“佳人在北燕，相望渭桥边。团团落日树，耿耿曙河天。”。南北朝时期是中国历史上一个处于分裂、动荡的特殊时期，王朝的更迭、南北朝的对峙以及与之带来的饥荒、瘟疫、大规模人口的迁徙，这些社会现象都反映在了

该时代的古典诗词当中，都反映在了该时代文人的精神面貌当中。在心灵敏感的文人眼里，对社会动荡所感到的世间无常，对自身渺小所感到的无能为力，所有感受不仅体现在游子思妇的离愁情绪，更体现在对自己对与亲人的离别、对社会的关注、对个体生命的忧患。然而，诗词作者并没有把眼光停留在伤感之中，而是投向了更广阔的社会视野，对连年战争、社会残破、民生疾苦等进行了描写。如谢灵运诗“家本秦川。贵公子孙。遭乱流寓。自伤情多。幽厉昔崩乱。桓灵今板荡。伊洛既燎烟。函崤没无象。整装辞秦川。”庾信诗“萧条亭障远，凄惨风尘多。关门临白狄，城影入黄河。秋风别苏武，寒水送荆轲。谁言气盖世，晨起帐中歌。”，这都曲折地反映出战乱给人民造成的巨大灾难，反映出诗人对人民苦难的深切感慨与同情[53]。

隋统一天下后，道教文风盛行。南北朝大规模的人口迁徙带来了文化的相互交流，道教文化是其中重要的一部分。北方道教文化中的古朴豪放、南方道教文化中的教理教义，相互交流借鉴。隋代道教南北河流体现在了文人对仙圣合一的追求当中，例如隋诗“学仙行为急，奉戒制情心。虚夷正气居，仙圣处相寻。”与“吾故及弱龄，弃世以学道。”等，都正面反映了隋代“修仙问道”文化盛行的时代背景。另外，该时代背景与隋炀帝与隋文帝对道教的推崇与政治定位有精密的联系[54-56]。

唐朝是我国封建社会的全胜时期，也是中国历史上最开放、最包容、最鼎盛的时代。贞观之治期间的知人善用，重视科举，以农为本，休养生息，稳固边疆带来了清明政治、经济复苏和文化繁荣。开元盛世期间的革新吏治，精修律法，广建书院，检田括户，孳息兵马，畅通丝路，使得该时代成为历朝历代前所未有的极盛之世。正所谓“我皇膺运太平年，四海朝宗会百川。自古几多明圣主，不如今帝胜尧天。”。唐代古典诗词中“家国无事/天下太平”的主题占据了当时的文学主流。同时，作为与政治和文化息息相关的科举制度，在唐朝得到了社会的高度重视，唐朝在官吏任用上，已经完全摒弃了过去的门第观念，几乎全部依赖于开科举士。朝中官员大多是进士出身。有鉴于此，形成了一种良性循环，士子以科举为毕生追求，中举做官后，又格外重视科举选士，“读书圣朝”就形成了当时历史时期古典诗词的时代氛围。有了这时代背景，古典诗词中所表现出来的浪漫洒脱的人格魅力就被释放出来，比如李白有诗“浪迹天涯去，南荒必动情。”，

自在有诗“经行宴坐闲无事，乐道逍遥三不归。”。家国、社会、个人的理想主义色彩在这一时代绽放光芒[57-58]。

北宋时期受“杯酒释兵权”思想的影响，当权朝廷重视文人士大夫，广布恩泽，免除兵役，文人的地位得到空前的提高。因此，通过科举选拔文官的制度开始成为宋朝皇帝推行文官政治、维护统治的有效手段，科举考试也由此获得了极大的发展空间，得到了进一步的完善。因此，该时期的历史基调呈现为“重文轻武”。在这样的历史基调下，加上还有其他很多因素的多重叠加作用，“读取功名”就成为了当时包括很多平民百姓在内的广大书生的重要出路。“可笑此公何太惑，读书写字到三更。”，“瑶姬来自状元家，真是姚黄第一花。”。但显而易见，这种“重文轻武”的历史氛围所显现出来的社会弊端也是突出的。习武之人地位低下，将士们建功立业、保卫社稷的雄心也难以得到支撑，“虚外守内”的消极防御政策占据了主流。“澶渊竟要盟，老莱气安吐。”，“羁怀病思不禁秋，又报西风大火流。”，“孤臣惭舜辅，愁绝望苍梧。”，举国上下，为国而忧，为国而愁。而这也给宋代诗词烙下很多“愁”的印子，“少年不知愁滋味，爱上层楼，为赋新词强说愁。”，国愁、人愁相互叠加，杨花不再是杨花，细雨不再是细雨，流水也不再是流水……，在漫长的中华诗词历史过程中，宋代是一个独特、浓墨重彩的时代[59-62]。

在元代，很多诗词作家通过“借景”来“排忧”。白云、西风、东风、明月、流水、秋风、夕阳、杨柳、落花、芳草、斜阳、鸿雁等词是元代诗词中的常见白描意象，例如马致远有诗“枯藤老树昏鸦，小桥流水人家，古道西风瘦马，夕阳西下，断肠人在天涯。”其语言相当干净简练，字字精心，句句入心，遣词断句，匠心独运，犹如一幅自然淳朴，干净简约，意境开阔的中国“工笔画”。品析之，宜怀着秋思之情，在无尽的秋意之中，感受作者在秋日时光里对人生的忧愁。元代，是中国历史上少有的少数民族统治的时代，在蒙古贵族的统治下，科举取士时行时辍，很多文人士大夫往往仕进无门，社会地位一落千丈。这种时代的无望暗淡，造成元代文人只能眼睁睁看着岁月蹉跎的空虚与感伤，这一切的一切，最终凝聚成一份断肠的心绪，这种断肠，恰恰是元代文人在当时社会环境中心灵的一曲悲歌。“秋深故国梦，应与逝川东。”，“李侯作画述者钱，想见温公当国年”，回思故往，空悲今日。“浮生碌碌。算由天由命，也由人福。”，“万里功名，半生

湖海。十五年间颜改。”，叹人生世事，只有“孤城孤客孤舟。”。文人以各种方式调侃科举、讽刺时政也是他们作品中常出现的主题，“龙虎相交，倒把黄河卷。”，“明时进用多英杰，迂腐深惭守一经。”。到了元代后期，文人开始崇尚隐逸。“樵夫觉来山月底，钓叟来寻觅。你把柴斧抛，我把鱼船弃，寻取个稳便处闲坐地。”，“不恋功名，不求富贵，不惹闲非。”,“独余洒然脱颖，任运止逍遥，自在无拘。”[63-64]。

在明代，“与君离别”主题愈发突出，有友人离别、游子离别、佳人离别、故乡离别等等。高启有诗云“重臣分陕去台端，宾从威仪尽汉官。四塞河山归版籍，百年父老见衣冠。函关月落听鸡度，华岳云开立马看。知尔西行定回首，如今江左是长安。”,写的就是送好朋友沈左司郎跟随汪御史中丞上任陕西参政时的心情。四塞的山川已纳入大明朝的版图,告诫朋友此次西行要能够常常回望故乡，如今金陵已经成了大明的都城。从高启诗句中，可以轻易看出作者的欣喜之情，以及百姓对国家安定和谐的期盼，新王朝之下官员远赴任职，这离别之情中表现出的是新时期的朝气。“杨柳毿毿弄晓晴，柳边持酒送君行。”，元代时期文人所表现出来的黯然、萧瑟与孤独，到了明代，一扫而尽。到了明代的中期，整个朝廷上下既有皇帝的荒诞不经、又有朝臣的阿谀奉承，更有宦官的独断专权。在文学思想领域，八股取士制度变得极为严格，宋明理学变成了思想的桎梏，高压政策、暗杀活动频仍。因此这一时期，文人们开始转而求佛、取道，以安抚对国家、对社会、对自己前途命运的不安，以至明代的道观众多，狂禅之风盛行。道学在这时期到达了高潮。“云际扪萝倚蔚蓝，圣屏北面憩禅庵。”,“宇宙何茫茫，起灭同一尘。”。阳明心学更是将这种参禅悟道推到极致。“一卧禅房隔岁心，五峰烟月听猿吟。”,“忆昔与君约，玩《易》探玄微。”。，王阳明的思想主要以“心外无理”、“知行合一”、“致良知”等为主，是典型的主观唯心主义思想。在这样一种思想背景之下，很多文人开始崇尚格物致知，热衷隐逸，“境幽人迹少，林暗鸟声惊。隐约棋中趣，从容物外颜。”，但却始终关心世道，关注国家兴衰。“遐想随时倦，幽栖与世违。”[65-66]。

到了清代,“伤悲”主题又开始占据时代主流。清代的很多诗风偏清丽婉约、缠绵悱恻，包含的情感丰富，爱情、友情、亲情、乡情、家国之情等等，相互杂糅，悲到极致。清代的诗词具有强烈的现实主义。清代初期，中国出现资本主义

生产关系的萌芽，商品经济日趋活跃，使得贵族地主阶级物质生活享受的眼界扩大，他们加紧了对劳动人民的剥削与掠夺，这在一定程度上激化了农民与地主阶级的矛盾。清朝中期，文字狱盛行，统治者会主观性地从文人作品中摘取字段，构成了无数冤狱，文人心中自我压抑。清朝末期，多遭列强入侵，北洋水师全军覆没，维新派开始主张君主立宪又遭封建势力毁灭，以八国联军为代表的帝国主义企图瓜分中国，主权和国家领土面临严重丧失的重大危机。中华民族到了最危险的时刻。“皇华与大汉，第供异族谑。”，“芒芒问禹迹，何时版图廓？”，伤悲中带着悲愤，悲愤中带着绝望[67-68]。

清朝末期的衰败无能导致了中国近代的探索。国家沦为半殖民地半封建社会，世间颠沛流离，多少人被迫离乡，近至香港澳门，远至东洋西洋。此去一别，不知何时再相见，只有相思天涯，“班生此去意何云，破碎神州日已曛。”，“子夜新声碧玉环，可怜肠断念家山。”。叹苍茫人间，“孤愤满腔何处诉。”。当悲愤到了极致，就会反抗。我们看到无数的有志之士，冲上街头，反抗侵略，捍卫国土完整的决心坚定不移，他们为国家主权呐喊着，他们对帝国主义怒吼着，这些爱国青年，用那一腔腔热血来唤醒了四万万中国人民的救亡图存意识，这是一个中国人觉醒的时代；我们看到，南昌城头第一枪，揭开了中国共产党独立领导武装斗争和创建革命军队的序幕。离愁不再是离愁，悲愤不再是悲愤，“抗战今开第五年，男儿志在复幽燕。”，化悲愤为力量，这一时期，产生了无数的时代英雄，救国家于危难，“穷人自有英雄胆，塌下青天双手擎。”，革命变得乐观，无数的人民走上历史舞台，“萧瑟秋风今又是，换了人间。”[69-70]。

5 总结与展望

5.1 总结

本文基于 LDA 模型对中国不同历史时期的古典诗词进行文本挖掘，通过困惑度计算，确定每个历史时期的最优主题数，并最终获取了若干有效主题。并根据古典诗词内容，根据不同历史时期描述的对象，把主题按类划分，分为“国家意识”主题类，“人生羁旅”主题类，“离愁主题类”等，并重点分析了“国家意识”主题类的内容演化过程，利用桑基演化图对该演化过程进行了可视化。同时，

采用文档支持度指标计算了各个历史时期的不同主题的强度，通过主题强度大小排序，以探究、分析当时历史时期的历史背景。古典诗词是历来文人豪士表达自我情感的重要形式，借古思今，本文利用 LDA 主题模型，从古典诗词出发提供了一个看待历史背景的角度，也为描绘社会提供了一个发现视角。

主要工作如下：

（1）通过文献调研主题模型的研究现状，发现，LDA 模型在处理舆情、研究热点、政策演化等方面，应用广泛，同时也发现了 LDA 应用场景的不足，可以进一步的扩展。笔者同时调研，目前古典诗词计算化领域的研究现状，发现，目前古典诗词计算化研究方面主要在两块，一块是机器学习分类问题，与之带来的诗词的风格分析，比如宋词婉约派/豪放派；一块是分析单个诗词文本的数据挖掘，例如分析《全唐诗》、《全宋词》的主题内容，从宏观历史角度分析古典诗词的文献较少；

（2）对有记载的、较为完整的、被研究较多的古典诗词进行收集，并且按照不同历史时期划分，利用 Python 对各文本进行分词，并且调用 sklearn 包执行 LDA 模型代码，利用 Excel 与在线工具桑基图（https://www.zxgj.cn/g/sankey）对概率进行计算和内容演化可视化。

（3）总结现有主题模型的种类，并对 LDA 模型的基本原理和建模过程进行研究，结合各个历史时期的古典诗词文档，每个历史时期各获得 4-6 个主题，共 46 个主题，并按照文本内容划分若干主题类。

（4）利用 Python 计算每个历史时期的各个主题-文档概率，引入支持度指标计算各个主题的强度；将历史时期变量加入到主题演化分析中，以“国家意识”主题类为例，绘制了 9 个历史时期的 8 个演化分析图。结合文本内容和时代背景进一步分析主题强度与主题演化的内在逻辑。

本文研究认为，“国家意识”的转变经历了三个阶段，一个是以魏晋、南北朝、隋为代表的圣皇崇拜阶段；一个是以唐、宋为代表的国家主体观阶段；一个是以明、清、近代为代表的人民主体观阶段。尤其到了近代时期，普通群众走向历史的视野，经历了长时间的斗争、救亡图存，越来越多的人意识到人民才是国家的主体。从朝廷到百姓，从顶层到万众，“国家意识”概念的外延变得越来越广。

同时，各个历史时期古典诗词的主题大致存在一定的上下波动，无论是国家意识类，人生羁旅类，离愁类，还是爱情、友情、乡情类等，但由于历史阶段社会背景的不同，各个主题存在一定的差异，并且强度不一。魏晋时期的古典诗词强调隐逸、逍遥，这与魏晋时期的政局变幻、崇尚道教等时代特征密切相关；南北朝时期的古典诗词强调“别离”、“相思”，这与南北朝时期的战乱不断、大规模的人口迁徙密相关；隋代古典诗词中的“道法自然”离不开南北道教河流，当朝皇上出于政治的需要，对道教积极扶持的社会背景；唐代古典诗词中的的“家国无事”更是离不开大唐王朝的政治清明，经济复苏，文化繁荣的治世局面；宋代古典诗词中的“读取功名”源自于宋太祖推行的文官政治，“重文轻武”的历史基调；元代古典诗词中的“借景排忧”是广大文人墨客仕进无门的一曲悲歌；明代古典诗词中的“与君别离”中溢出的却是百姓对国家安定和谐的期盼与喜悦；清代古典诗词中的“我心伤悲”弥漫着清朝初期社会阶级矛盾、中期文字狱、末期列强入侵的时代压抑；近代的古典诗词中的“相思天涯”是国家沦为半殖民地半封建社会背景下，世间颠沛流离，百姓前途未卜，对亲人、友人与爱人的无限担忧。诗词是历史的沉淀，同时影响深远，也在塑造当时的历史。

5.2 展望

本文通过 LDA 模型对各个历史时期的古典诗词进行文本挖掘，最终获得各个历史时期古典诗词的主题，结合诗词文本对当时历史时期的社会背景进行深入讨论，但在研究过程中还存在以下问题，需要在后续研究中进行改进：

（1）LDA 模型本身存在不足，经过十几年的发展已演化为各种改进版本。本文缺乏对所有主题模型在古典诗词主题分析应用方面的量化比较；本文计算的主题强度比较仅限于某一单个历史时期，缺乏多个历史时期主题强度的演化过程，以及不同历史时期主题间的相关性度量，从而为“主题类”的划分提供一个计算依据；主题词在历史时期变化过程中也存在着演变、融合与消亡的过程，这一点笔者没有提供多余笔墨进行说明。

（2）去停用词和词典制作方法可能还不够完善。本文去停用词采用的是文本向量化计算过程中去高频词与低频词的方法，并且频率阈值设定的不同，对结果也有一些影响。笔者会在后续尝试多种方法，以改善其结果。

（3）古典诗词属于文学与历史学的范畴，笔者作为应用统计专业的学生对文献计量和文本挖掘颇有学习，但在文学与历史学的相关知识尚显不足，所以在对古典诗词与历史背景关系进行解读的过程中难免会有疏漏，挂一漏万。

（4）文字记录在一定程度上是该时代的现实反映，笔者研究的文本数据仅限于古典诗词，反映的是中国不同历史时期的社会。但是无论是古典诗词还是当代文章、文本作品、民众评论等，都可以反映一个社会样貌。一个向上、崇尚进步的社会一定会体现在人民积极与自信的文字之中，道路自信、理论自信、制度自信与文化自信是当代中国人民自信的深刻内涵。所以在今后的文本数据挖掘之中，笔者会尝试不同的数据源以反映一个更加细致的社会样貌。同时，为从国家层面上推进“五位一体”总体布局，促进社会进步提供了重要启迪意义。

参考文献

[1] Luhn H P. Auto-encoding of documents for information retrieval systems[M]. IBM Research Center, 1958.

[2] Maron, M.E. and Kuhns, J.L. (1960) On Relevance, Probabilistic Indexing and Information Retrieval. Journal of the ACM (JACM), 7, 216-244.

[3]Salton G. A Vector space model for automatic indexing[J]. Communications of the ACM, 1975, 18(11): 613-620.

[4]Hearst M A. Reexamining the cluster hypothesis: scatter/gather on retrieval results[C]// Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR-1996), New York, NY, USA. ACM Press, 1996.

[5]Craig Silverstein,Sergey Brin,Rajeev Motwani. Beyond Market Baskets: Generalizing Association Rules to Dependence Rules[J]. Data Mining and Knowledge Discovery,1998,2(1).

[6] Blei D M, Jordan M I. Modeling annotated data[J]. ACM SIGIR FORUM, 2003(Special):p.127-134.

[7] Bolton G E , Katok E , Ockenfels A . How Effective Are Electronic Reputation Mechanisms? An Experimental Investigation[J]. Management Science, 2004, 50(11):1587-1602.

[8] Antweiler W , Frank M Z . Does Talk Matter? Evidence From a Broad Cross Section of Stocks. University of British Columbia Working Paper, 2004.

[9]Dongshan Xing,Mark Girolami. Employing Latent Dirichlet Allocation for fraud detection in telecommunications[J]. Pattern Recognition Letters,2007,28(13).

[10]Boiy, Erik, Moens, et al. A machine learning approach to sentiment analysis in multilingual Web texts[J]. Information Retrieval, 2009.

[11]王继成,潘金贵,张福炎.Web 文本挖掘技术研究[J].计算机研究与发展,2000(05):513-520.

[12]李凡,鲁明羽,陆玉昌.关于文本特征抽取新方法的研究[J].清华大学学报(自然科学版),2001(07):98-101.

[13]黄晓斌,赵超.文本挖掘在网络舆情信息分析中的应用[J].情报科学,2009,27(01):94-99.

[14]张彦. web 中文文本的数据挖掘技术研究[D].山东大学, 2011.

[15]杨丹,朱世玲,卞正宇. 基于改进的 K-means 算法在文本挖掘中的应用[J].计算机技术与

发展, 2019, 29(04): 68-71.

[16] 李湘东, 张娇, 袁满. 基于 LDA 模型的科技期刊主题演化研究[J]. 情报杂志,2014,33(07):115-121.

[17]关鹏, 王曰芬, 傅柱. 不同语料下基于 LDA 主题模型的科学文献主题抽取效果分析[J]. 图书情报工作, 2016,60(02):112-121.

[18]谭春辉,熊梦媛.基于 LDA 模型的国内外数据挖掘研究热点主题演化对比分析[J]. 情报科学. 2021(1) :1-12.

[19]周思源.统计批评:黛钗诗词差异论[J].红楼梦学刊,1992(04):181-195.

[20]周昌乐.心脑计算机举要[M].北京清华大学出版社, 2003.

[21]易勇.计算机辅助诗词创作中的风格辨析及联语应对研究[D].重庆大学,2005.

[22]游维.基于遗传算法的宋词自动生成研究[D].厦门大学,2007.

[23]苏劲松.全宋词语料库建设及其风格与情感分析的计算方法研究[D].厦门大学,2007.

[24]吴春龙,周昌乐.基于频繁关键字共现的诗词风格分类模型研究[J].厦门大学学报, 2008(01): 41-44.

[25]赖兴邦.宋词格律分析的计算方法及其应用研究[D]. 厦门大学, 2008.

[26] 钱鹏, 黄萱菁. 中国古诗统计建模与宏观分析[J]. 江西师范大学学报(自然版),2015,000(002):117-123.

[27] 申资卓, 杨莹, 邵艳秋. 基于主题模型的古典乐器诗词文本挖掘[J]. 中文信息学报,2019,33(03):79-86.

[28]张馨怡.基于 TextCNN 的古典诗词爱国情怀研究[D].上海师范大学,2020.

[29]徐戈,王厚峰.自然语言处理中主题模型的发展[J].计算机学报,2011,34(08):1423-1436.

[30]勒孚刚. 基于 LDA 模型的专利文本分类及演化研究[D].江西理工大学,2017.

[31]Scott Deerwester,Susan T. Dumais,George W. Furnas,Thomas K. Landauer,Richard Harshman. Indexing by latent semantic analysis[J]. Journal of the American Society for Information Science,1990,41(6).

[32]Thomas Hofmann. Unsupervised Learning by Probabilistic Latent Semantic Analysis[J]. Machine Learning,2001,42(1-2).

[33] 赵凯, 王鸿源.LDA 最优主题数选取方法研究:以 CNKI 文献为例[J]. 统计与决策,2020,36(16):175-179.

[34]石凤贵.中文文本分词及其可视化技术研究[J].现代计算机,2020(12):131-138+148.

[35]文本挖掘实录：用文本挖掘剖析 54 万首诗歌

[EB\OL]. http://www.woshipm.com/data-analysis/4160560.html

[36]李斌斌. 基于 LDA 模型的我国文化政策主题演化研究（1979-2017）[D].上海大学,2019.

[37]刘自强,王效岳,白如江.多维度视角下学科主题演化可视化分析方法研究——以我国图书情报领域大数据研究为例[J].中国图书馆学报,2016,42(06):67-84.

[38]唐果媛,张薇.基于共词分析法的学科主题演化研究进展与分析[J].图书情报工作,2015,59(05):128-136.

[39]王平.基于层次概率主题模型的科技文献主题发现及演化[J].图书情报工作,2014,58(22):70-77.

[40]祝娜,王芳.基于主题关联的知识演化路径识别研究——以 3D 打印领域为例[J].图书情报工作,2016,60(05):101-109.

[41]文琦,郑殿元,施琳娜.1949—2019 年中国乡村振兴主题演化过程与研究展望[J].地理科学进展,2019,38(09):1272-1281.

[42]刘敏娟,张学福,颜蕴.基于核心词、突变词与新生词的学科主题演化方法研究[J].情报杂志,2016,35(12):175-180.

[43]姜婷婷,肖卫东,张翀,葛斌.基于桑基图的时间序列文本可视化方法[J].计算机应用研究,2016,33(09):2683-2687+2692.

[44]袁从秀,李恩泉.认识多元一体趋势　强化中华民族认同——“辽、西夏与北宋的并立”一课的教学分析与设计[J].历史教学(上半月刊),2019(11):64-70.

[45]陈彩云.元朝疆域观演变与多民族国家的空间认知[J].民族研究,2021(01):120-132+142.

[46]林丽丽,马秀峰.基于 LDA 模型的国内图书情报学研究主题发现及演化分析[J].情报科学,2019,37(12):87-92.

[47]李海林,邬先利.基于时间序列聚类的主题发现与演化分析研究[J].情报学报,2019,38(10):1041-1050.

[48] 朱晓霞,宋嘉欣,孟建芳.基于动态主题—情感演化模型的网络舆情信息分析[J].情报科学,2019,37(07):72-78.

[49]桂小庆,张俊,张晓民,于鹏飞.时态主题模型方法及应用研究综述[J].计算机科学,2017,44(02):46-55.

[50]王惠.从魏晋山水诗文看士人山水审美意识的觉醒[J].名作欣赏,2019(26):137-138.

[51]刘红宁. 论魏晋名士的自然人格[D].青岛大学,2007.

[52]刘俊,化金荣,肖瑞阳.浅谈魏晋时期诗歌风格的转变[J].才智,2014(17):79.

[53]郑宏萍.论魏晋南北朝送别诗的审美内涵[J].文教资料,2009(27):8-9.

[54]杨会萍.隋朝二帝与道教[J].江西金融职工大学学报,2009,22(S1):218-219.

[55]蒋振华,邓超.隋代道教文学创作倾向的仙圣合一和神仙意象化[J].中国文学研究,2011(02):44-47.

[56]史话. 隋炀帝“大一统”思想的形成与实践[D].中国社会科学院研究生院,2017.

[57] 祝陶然. 国际视野下的国家形象塑造与传播——盛唐为例[J]. 黑龙江史志, 2014(03):196.

[58]王述尧.唐太宗与盛唐气象[J].绥化师专学报,2004(02):71-72.

[59] 魏爱玲. 人生自是有情痴,此恨不关风与月——宋词愁情之审美意象探析[J]. 安徽文学(下半月), 2015, 000(012):58-59.

[60] 钮敏, 李仁霞. 试论宋代"重文轻武"的社会风气[J]. 兰台世界, 2015, 000(033):55-57.

[61]周安邦.家国情怀溢韵坛——曾巩诗歌的忧国怀民意识[J].东华理工大学学报(社会科学版),2019,38(03):223-226.

[62]徐雨婷. 南宋爱国词研究[D].陕西理工大学,2019.

[63] 张道元.《天净沙 秋思》的白描艺术赏析[J]. 云南教育:中学教师, 2017, 000(009):14-15.

[64]杜肇昆.元散曲风格流派二分界定辩疑——兼论俗俏、旷达、清丽三分界定表达[J].中国韵文学刊,2015,29(04):86-98.

[65]吴倩.二十世纪明代诗词研究索引(二)[A]. .中国诗歌研究动态（第一辑）[C].:中国诗歌研究中心,2004:32.

[66]吴倩.二十世纪明代诗词研究索引(二)[J].中国诗歌研究动态,2004(00):162-193.

[67]王光华.清代卜奎流人诗词的苦难意蕴[J].牡丹江教育学院学报,2015(05):4+126.

[68]段天顺.清代的诗人和诗词[J].海内与海外,2006(09):56-58.

[69]翁晓宇.中国近代古诗词艺术歌曲的创作与发展[J].当代音乐,2020(07):77-78.

[70]周甲辰.试看天下谁能敌——毛泽东诗词和古典诗词风格比较分析[J].黑龙江农垦师专学报,2000(04):41-44.

附件

表 1 魏晋时期古典诗词的主题输出

主题	关键词（部分）
Topic 0 相思恋人	明月 佳人 叹息 与君 蟋蟀 秋风 岁暮 悲风 今日 远望 灼灼 憔悴 漫漫 烈烈 凄凄 思君 西山 牵牛 终日 彷徨 常恐 织女 鸡鸣 依依 踟蹰 忽如 春华 萧索 罗裳 悲鸣 泣涕 遥遥 泪下 弦歌 北风 褰裳 露沾 被服 秋霜 贱妾 金石 朱颜 秋夜 独无 幽人 暧暧 远行 我心 零落 泠泠
Topic 1 翩翩少年	千里 翩翩 松柏 四海 我心 少年 出门 徘徊 青云 九州 羽翼 万里 乘云 蓬莱 五岳 遥望 飞鸟 神仙 悠悠 平生 黄鹄 贫贱 眇眇 白马 玄云 扬州 富贵 西南 西北 华盖 千年 登高 嵯峨 东南 亲友 陛下 荆棘 一朝 来归 须臾 参天 驱车 忧思 归来 父母 随风 赤松 欢乐 游子 遨游
Topic 2 逍遥君子	君子 逍遥 慷慨 天下 俯仰 千载 古人 自然 天地 扶桑 日月 远游 松乔 清流 朝阳 穷达 风尘 世间 良辰 感物 清歌 绸缪 弱冠 闲居 凤凰 昆仑 窈窕 凯风 众鸟 无穷 自古 羲和 天道 各异 丹霞 峨峨 长生 明德 夫子 六龙 弹琴 洪波 王子 风流 达人 紫霞 相忘 随风 沧海 登城
Topic 3 离别惆怅	窈窕 鸳鸯 同心 殷勤 交颈 芙蓉 徘徊 佳人 日月 相思 流水 登高 好音 高山 形影 辛苦 行人 泛舟 公子 折杨柳 行役 戢翼 别离 清风 故人 缠绵 我心 惆怅 踟蹰 郁郁 故乡 山川 饥寒 万里 长叹 苦心 淹留 骨肉 河水 衣裳 一人 借问 望舒 夜光 春秋 冬夏 感物 命驾 寤寐 百忧
Topic 4 圣皇神明	穆穆 四海 万国 圣皇 我皇 赫赫 天地 四方 天下 六合 文武 神明 巍巍 天子 宇宙 受命 无疆 万邦 盛德 圣德 礼乐 神武 神祇 八风 峨峨 天命 八音 煌煌 享祀 率土 万物 龙飞 应天 皇祖 克昌 鹰扬 多士 天人 文皇 顺天 祖考 万世 福禄 圣明 昊天 皇极 大业 开元 永世 邦家

表 2 南北朝时期古典诗词的主题输出

主题	关键词（部分）
Topic 0 闺中佳人	可怜 千金 日暮 洛阳 徘徊 佳人 零落 芳树 夜月 闺中 光景 徙倚 梅花 别离 薄暮 河桥 几时 寄语 霜雪 未央 惆怅 叹息 寸心 歌吹 独无 凄凄 良人 寂寂 相看 宛转 长叹 飘飘 洞房 叶落 美人 鸟声 倦游 蕙草 踯躅 绮罗 相忆 朝云 芳菲 风飘 窗前 青楼 池上 华烛 夜长 今朝
Topic 1 君臣之道	君子 天下 风云 万国 芙蓉 肃肃 天地 团扇 四海 金石 济济 我皇 风尘 蔼蔼 九重 至德 礼乐 明镜 玉帛 礼容 弦歌 左右 四方 竹叶 多士 穆穆 皇情 威仪 饮马 宇宙 金羁 千载 长城 古人 百川 仁义 紫微 大梁 君臣 明德 受命 元首 万寿 长卿 龙门 桑榆 太平 功名 八风 建章
Topic 2 与君别离	徘徊 春风 江南 千里 淹留 佳人 洞庭 罗衣 秋月 无人 秋风 萧条 珠帘 琴瑟 采莲 美人 淮南 含情 桂枝 岁暮 置酒 佳期 白雪 离别 西园 飞盖 沾衣 空自 寂寞 伫立 朝日 鸡鸣 连翩 迢递 巧笑 长袖 靡靡 凄凄 昔时 今日 别离 憔悴 日夜 合欢 徒自 桃花 绮窗 采菱 红尘 沧海
Topic 3 羁旅生涯	黄金 千里 霜露 阳台 花落 春日 金鞍 游子 白马 岁月 倾城 千载 风急 使君 纵横 陈王 摇落 怀抱 但愿 宿昔 寒风 羁旅 长安 绸缪 王孙 乔木 宝剑 忘忧 来归 天道 西归 公子 昔闻 游鱼 楚王 含笑 陇西 寂寥 严霜 富贵 玉盘 高枝 风雨 鸟飞 飘飖 客心 大道 攀折 罗裙 征马
Topic 4 千里思君	相思 千里 万里 思君 流水 长安 将军 一朝 西北 关山 陇头 杨柳 相望 黄河 故乡 高楼 鸳鸯 洛阳 窈窕 相见 春草 同心 离宫 归来 踟蹰 胡笳 飞燕 凤凰 留连 花开 平原 未归 落叶 妾心 别离 逢迎 织素 日暮 佳人 风波 对酒 日晚 烽火 夫婿 流星 今夜 玉阶 舞衣 一人 辛苦

表 3 隋朝时期古典诗词的主题输出

主题	关键词（部分）
Topic 0 道法自然	万里 天地 大道 长安 芙蓉 自然 风云 辽东 天子 白日 升天 日月 浮云 无极 流水 学道 长生 二仪 千里 太上
Topic 1 太上神明	千里 太上 相思 风吹 明德 天下 一时 神明 盛德 天地 无疆 济济 流水 徘徊 高山 钟鼓 真人 世人 上帝 威仪
Topic 2 春风杨柳	明月 杨柳 悠悠 阳春 萧萧 洛阳 流水 春风 辽东 真人 十五 四方 九天 自然 金石 钟鼓 上下 芙蓉 二仪 大道
Topic 3 万国皇祖	可怜 万国 肃肃 礼乐 山川 皇祖 徘徊 穆穆 四海 上帝 日月 威仪 上下 风云 四方 天地 神明 金石 高山 天子

表 4 唐朝时期古典诗词的主题输出

主题	关键词（部分）
Topic 0 家国无事	白云 青山 万里 天子 将军 月明 无事 江南 东风 时节 闻道 造化 洞庭 焚香 汉家 春风 五湖 春光 太平 四海 山川 蓬莱 天地 一年 寻常 世人 往往 黄昏 江边 细雨 花枝 莓苔 风雨 马蹄 罗衣 湖上 梦里 白首 门前 秋色 柳色 世事 去年 尘埃 他年 儿童 鬼神 今朝 衣冠 骑马 童子 杨柳 平生 长江 秋水 飞去 永日 窈窕 南山 旌旗
Topic 1 读书圣朝	十年 芳草 白日 纷纷 人间 风尘 故乡 悠悠 零落 日月 夜深 憔悴 清风 萧萧 今年 三月 四十 终日 草堂 一身 几回 世间 潺湲 朝朝 吟诗 风景 功成 自古 洞庭 山下 烟霞 一枝 羽翼 圣主 绮罗 独立 清明 桃花 回首 迟迟 读书 苍生 百里 归来 寂寂 苦辛 飞鸟 四时 少年 流年 玉树 云间 文章 金陵 幽人 江山 圣朝 心期 洞房 美人
Topic 2 岁月悠悠	明月 年年 悠悠 今日 故人 花开 早晚 岁月 迢迢 人间 秋草 日日 野人 朱门 风月 佳期 莲花 虚空 三年 昔年 出门 次第 自然 九天 清净 今朝 草木 上天 水中 明日 清秋 未曾 思归 十二 玉堂 忘机 风云 夜月 往事 平生

	封侯 归来 处处 重重 白雪 未成 应须 老人 薄暮 风飘 九霄 迢递 老夫 浮云 天地 木落 几年 玄珠 尘中 雨中
Topic 3 千古风流	春风 寂寞 可怜 当时 少年 芙蓉 洛阳 鸳鸯 风流 君子 与君 惆怅 千载 古人 风吹 佳人 万古 长安 颜色 秋月 衣裳 江水 管弦 天地 春深 新诗 江海 芳菲 才子 松柏 登临 千金 歌舞 白马 年少 婵娟 兄弟 黄金 江南 南国 劝君 多情 富贵 平生 古来 蛟龙 随风 歌声 含情 月色 烟雨 意气 江头 相思 春日 三千 宇宙 宾客 第一 怜君
Topic 4 浪迹天涯	万里 行人 东西 裴回 日月 别离 归去 一朝 山水 明朝 孤舟 终日 扁舟 苍苍 断肠 流水 寒山 春水 别后 南北 文章 暮雨 千古 万事 东山 山色 山川 四邻 天涯 飘飖 圣人 潇湘 杯酒 出门 故园 平生 古木 游子 万国 旧游 风雨 我心 世间 夫子 东风 五陵 苍梧 猿啼 泉声 浮云 浮生 朝夕 回首 来往 昔时 关山 仙家 南山 一笑 知音
Topic 5 人间风尘	秋风 白发 夕阳 三十 春色 茫茫 相逢 长安 逍遥 相思 公子 肠断 归路 青春 怅望 君王 王孙 故人 人间 落日 今朝 游人 走马 青云 笙歌 殷勤 镜中 多病 残月 叹息 欢娱 使君 登楼 昔日 伤心 玉楼 红粉 画堂 天地 楼台 须知 明年 月照 远客 行乐 扬州 青楼 南山 阳春 容易 尚书 银河 泪痕 凤凰 杨花 回首 谢公 一夜 梦魂 红尘

表 5 宋朝时期古典诗词的主题输出

主题	关键词（部分）
Topic 0 国事愁绝	呜呼 惨惨 酒醒 怅然 将军 洛阳 感慨 故山 白发 江湖 几度 生死 今古 孤城 老夫 故国 男儿 倏忽 四顾 愁来 忠臣 风雨 江南 悠悠 愁绝 杜鹃 旧时 忆昔 汉家 泪痕 秋风 回首 扁舟 夕阳 萧萧 可怜 故人 长安 消息 终日 芳草 天涯 寂寂 底事 空山 心事 参差 归路 沧海 儿女 叹息 悠然 茫茫 斜阳 缥缈 漠漠 梦魂 花落 秋来 徘徊

	无奈 潇湘 细雨 婆娑 落叶 鸥鹭 归心 老去 独自 思量
	伤心 孤云 肠断 冥冥 凄凉 经年 孤舟 沧溟 客愁 闭户
Topic 1 人间富贵	人间 富贵 天上 天下 千年 风流 先生 君子 神仙 三千
	天子 山水 梦中 堂堂 虚空 君王 精神 春风 和气 世间
	君家 声名 世人 纵横 日月 工夫 自古 风云 山河 尧夫
	佳句 闻道 画图 风味 功名 南山 经纶 千岁 秋月 老人
	赵州 松柏 暮云 文章 相随 雨馀 苍生 星斗 雨露 凭栏
	从容 扰扰 堂上 艰难 超然 我心 南北 百年 佳处 青天
	千秋 光明 翰墨 万年 九重 玉树 鸡犬 犹自 四面 白鸥
	笔端 寒暑 香火 勋业 三世 弟兄 竹林 桑麻 龙蛇 百岁
Topic 2 春光十里	春风 白云 东风 青山 明月 桃李 春色 年年 花开 幽人
	深处 重来 芙蓉 西湖 春光 江头 园林 夜来 黄昏 多情
	拄杖 海棠 昨夜 一曲 故乡 佳人 笑语 回首 几度 雨过
	一年 居士 燕子 乘兴 月下 翠微 雨后 长生 芳菲 锦绣
	鸳鸯 江东 碧云 春归 乔木 溪水 百花 老翁 岁岁 牡丹
	青春 青衫 少年 野水 笙歌 垂杨 花枝 众生 乐事 江梅
	精神 行乐 碧玉 天气 梨花 踪迹 南枝 颜色 月明 北窗
	荷花 啼鸟 荆棘 暗香 春意 游子 琵琶 随风 绿阴 歌舞
Topic 3 世间幽居	无人 相逢 寂寞 风流 尘埃 山林 登临 使君 太平 故人
	黄花 将军 相思 杨柳 生涯 一笑 落花 无事 渊明 春来
	林下 山色 清明 逍遥 三径 梅花 仙人 客来 徘徊 岁寒
	朝廷 清香 洞庭 杯酒 生死 老夫 春秋 夜雨 往事 苍苔
	柴门 重阳 身世 逢人 岁晚 秋声 烟霞 丘壑 断肠 流水
	高卧 松竹 昔人 霜雪 泉石 庐山 清夜 田园 道路 岁月
	登高 梦里 故园 菊花 姓名 天涯 茅屋 江湖 茅檐 新诗
	白雪 野人 西湖 东坡 北斗 崎岖 东篱 陈迹 寒梅 幽居
Topic 4 读取功名	平生 乾坤 功名 江山 千载 四海 古人 文章 胸中 岁月
	三年 读书 江湖 先生 古今 世事 万古 明月 青云 十年
	天下 万物 丹青 白首 夫子 少年 文字 吟诗 中原 峥嵘

	诗书 万象 宇宙 丈夫 万顷 寂寥 事业 黄金 一笑 中兴 世间 五湖 人心 风波 江海 圣贤 书生 邂逅 一朝 俯仰 衣冠 英雄 金石 知音 谈笑 凛凛 一念 老去 万卷 梦回 次第 千金 老僧 意气 清净 圣人 门户 梧桐 暮年 人情 九州 万人 赋诗 骚人 中秋 著书 庭前 公卿 造物 醉乡

表 6 元朝时期古典诗词的主题输出

主题	关键词（部分）
Topic 0 回首当年	今日 相逢 归来 故人 江湖 风雨 扁舟 百年 平生 千载 玉堂 白发 悠悠 英雄 少年 秋风 人间 今年 当年 文章 千古 岁月 回首 白头 今朝 当时 去年 昨日 功名 寂寂 黄花 燕子 王孙 公子 万事 十载 茫茫 古来 新诗 他年 九州 春深 邂逅 杯酒 四海 春秋 沧海 忆昔 我心 知己
Topic 1 人间世事	富贵 君子 长生 百年 功名 纷纷 诗书 黄金 辛苦 潇洒 平生 读书 人间 诸公 锦绣 呜呼 礼乐 浮生 世事 白发 客来 青云 老夫 蹉跎 寂寥 俯仰 世间 后来 艰难 丈夫 天地 长安 人情 谈笑 人心 四海 从容 书生 清谈 丹心 坐看 岁寒 晚来 道路 夫子 男儿 御史 圣主 宫中 随缘
Topic 2 君王朝廷	蓬莱 日月 将军 乾坤 自然 天子 天地 黄金 造化 仙人 太平 元气 旌旗 天下 龙虎 千年 东海 高堂 凤凰 万古 中原 山川 君王 四海 明珠 万物 四方 沧溟 苍龙 朝廷 圣人 光明 六合 万国 千官 群仙 北斗 千里 万年 浮云 和气 凌云 虎豹 光辉 苍生 霹雳 中央 大道 城南 人间
Topic 3 神仙逍遥	清风 逍遥 神仙 天地 古今 冥冥 物外 太古 云中 虚空 归去 清净 瑶台 洞天 道人 星辰 更无 飘飘 丹砂 功成 尘世 孤云 世人 性命 修行 上天 仙家 鬼神 真人 骑马 龙门 牛羊 松风 瑶池 琼花 紫芝 道士 瑶草 清虚 桃源 知音 野鹤 爽气 青鸾 吹笙 紫云 仙翁 蟠桃 玉京 行路难
Topic 4	白云 春风 青山 风吹 江南 西风 东风 明月 流水 秋风

借景排忧	落日 夕阳 回首 杨柳 相思 萧萧 无人 落花 西湖 芳草 悠悠 花落 寂寞 春色 浮云 空山 夜深 日暮 秋色 徘徊 画图 秋水 人家 白发 溪上 山色 故园 黄昏 参差 杏花 草堂 别离 诗人 昔年 漠漠 红尘 佳人 斜阳 鸿雁 惆怅

表 7 明朝时期古典诗词的主题输出

主题	关键词（部分）
Topic 0 山中幽栖	白云 烟霞 青山 明月 清风 扁舟 山中 主人 登临 蓬莱 流水 缥缈 悠然 苍茫 白石 桃花 秋色 云霄 沧洲 深处 琅玕 幽人 青天 苍苍 名山 风景 云中 寂寥 山色 夕阳 长啸 清秋 秋月 山人 洞庭 神仙 松风 秋风 冰雪 桃源 石上 远近 参差 山水 蒹葭 崔嵬 白鸥 幽栖 山僧 江流 云霞 白鹤 云间 瑶草 潇湘 阶前 回首 无人 仙家 重重
Topic 1 纵横江山	万里 长安 千里 当年 文章 将军 江山 风云 中原 四海 天子 千年 太平 天下 九重 纵横 英雄 翩翩 风流 玉树 意气 少年 青云 关山 圣主 临风 从容 蹉跎 山河 苍生 风尘 黄河 旌旗 凤凰 霄汉 兄弟 承恩 明珠 玉堂 诗书 沧海 古今 乾坤 山川 司马 丈夫 紫气 公子 春秋 北斗 感慨 东山 麒麟 龙门 礼乐 沧溟 功名 男儿 汉家 努力
Topic 2 世间问道	仙人 世事 今古 万事 归来 天地 乾坤 日月 平生 百年 悠悠 君子 浮云 千载 古人 茫茫 无穷 世间 吾道 万壑 须臾 五色 宇宙 冥冥 逍遥 山中 俯仰 世人 我心 江海 造化 一笑 乘兴 把酒 金石 叹息 山水 万物 道人 武陵 青天 长生 鸡鸣 登台 丘壑 圣人 寄语 西山 吁嗟 胸中 偶然 神仙 丹心 风尘 日夕 百岁 翩翩 星斗 佳气 万象
Topic 3 他乡惆怅	人间 梅花 明月 西风 君王 惆怅 题诗 草堂 夜深 风流 白头 斜阳 柴门 梦里 黄花 细雨 终日 伤心 夜月 银河 孤舟 歌舞 莲花 他年 多情 漠漠 今夜 一曲 还家 白马 黄昏 月下 萧萧 风霜 游人 风前 随风 苍苔 红颜 他乡

	摇落 灯前 白雪 江南 山阴 离别 鹦鹉 蛾眉 燕山 琵琶 客路 西湖 风雨 夜夜 东篱 桃李 相忆 寂寞 谷口 春草
Topic 4 与君离别	故人 千里 青山 东风 江上 天涯 江湖 芳草 相思 万里 回首 杨柳 桃花 萧萧 落日 美人 落花 流水 悠悠 故园 秋水 归去 相对 车马 夕阳 相见 使君 与君 明朝 读书 风尘 萧条 别离 载酒 迢迢 依依 鸳鸯 日暮 浮生 故乡 秋色 同心 知己 佳人 门前 行人 鸿雁 看花 相看 送君 春光 怅望 红尘 远游 知音 心事 生涯 焚香 游子 花落

表 8 清朝时期古典诗词的主题输出

主题	关键词（部分）
Topic 0 将军封侯	万里 将军 风雨 东南 别离 梧桐 世事 封侯 老去 尊前 酒醒 高堂 三百 千万 白头 十万 青云 十年 终日 半生 千山 辛苦 纵横 高卧 白马 鸿雁 次第 漠漠 白雪 未成 万象 玉楼 逢人 悲秋 平安 阴阴 马蹄 远道 关河 从今 翡翠 客心 黄沙 君恩 客来 倏忽 夜夜 闲情 战场 夫人
Topic 1 他乡归来	归来 苍茫 夕阳 十载 斜阳 杏花 西风 夕照 满地 他年 秋色 杨花 落叶 梦回 黄叶 苍苔 萧萧 窗纱 花间 暮雨 重阳 阶前 溪水 消魂 独自 楼阁 罗衣 空庭 桃花 恹恹 红叶 炉烟 无人 清溪 离离 忘机 寂寞 京华 荒凉 崎岖 他乡 西窗 禾黍 卷帘 残雪 薄暮 吾家 蛱蝶 对酒 苔痕
Topic 2 我心伤悲	东风 相思 斜阳 西风 芳草 黄昏 落花 鸳鸯 伤心 凄凉 风吹 憔悴 燕子 阑干 夕阳 依旧 回首 多情 风雨 无端 匆匆 惆怅 销魂 秋水 清明 断肠 梦里 寂寞 飘零 昨夜 肠断 寂寂 海棠 庭院 无情 往事 风流 依依 花落 心事 思量 旧时 梦中 相逢 萧瑟 无人 故园 一曲 梦醒 鹦鹉
Topic 3 茫茫浮生	白云 芙蓉 风尘 读书 登临 三十 江湖 茫茫 烟波 扁舟 平生 茅屋 君子 回首 离别 故人 天际 与君 云气 江上 浮云 空山 蝴蝶 知音 长啸 俯仰 夕阳 城郭 沧海 归去

	浮生 孤城 幽人 日暮 萧然 晚风 烟云 逍遥 五十 烽烟 岁寒 生死 西山 红尘 泠泠 浩荡 道路 看山 凄凄 秋气
Topic 4 富贵功名	人间 文章 千秋 英雄 万里 先生 平生 百年 天下 黄金 中原 千载 功名 少年 富贵 一朝 一笑 五更 天子 长安 衣冠 江山 第一 纵横 古人 十年 春风 意气 慷慨 儿女 风云 风流 松柏 太平 男儿 流光 摩挲 文字 兴亡 冥冥 呜呼 七十 知己 诗人 丈夫 山河 鱼龙 九州 须眉 三千
Topic 5 往事回首	秋风 故人 十年 相逢 回首 缥缈 残月 蹉跎 长安 明月 身世 楼头 佳人 经年 艰难 旧事 往来 悲歌 今日 春梦 掩映 弹指 诗书 堪怜 一年 惆怅 吟诗 廿年 使君 屈指 落日 酒杯 东坡 搔首 娟娟 光阴 风霜 匆匆 飘零 远游 辛苦 古木 重来 双鬓 相识 杜陵 风露 因缘 游子 归梦

表 9 近代时期古典诗词的主题输出

主题	关键词（部分）
Topic 0 苍茫人间	故人 寂寞 百年 当年 人间 花开 风雨 底事 生涯 苍茫 浮云 风光 功名 红尘 未能 梦回 花落 胭脂 岁岁 江山 多少 潇湘 依旧 他乡 世事 重阳 天涯 知音 沈吟 风雪 富贵 神仙 文章 秋风 事业 钟声 他年 萧萧 深深 载酒 黄花 落叶 无声 归去 人情 刘郎 踪迹 几回 无端 空自
Topic 1 江山忧患	夕阳 万里 长安 风雨 风尘 回首 将军 人间 江湖 男儿 故国 歌舞 多少 灯前 杯酒 茫茫 婆娑 苍茫 太平 平生 一笑 伤心 心事 惊心 独自 南北 频年 落叶 霓裳 离愁 飘零 明镜 天际 忧患 流莺 憔悴 江南 身世 斜晖 烟波 翠微 故园 音书 寸心 沉沉 中原 诗酒 凄凉 苍苍 红楼
Topic 2 壮士报国	平生 天地 天下 沧海 少年 乾坤 十年 归来 四海 精神 茫茫 干戈 终古 世间 同心 俯仰 一身 横流 日夜 苍生 肝胆 招魂 汉家 千载 胸中 万古 诗书 堂堂 山川 匆匆 万里 徘徊 风雷 江海 烟云 秋月 白首 儿孙 丈夫 万象

	报国 壮士 兴亡 风云 浩浩 意气 穷途 人心 未可 至今
Topic 3 神州英雄	风流 英雄 江山 中原 千里 神州 慷慨 山河 风波 弹指 江南 纵横 读书 文章 悲歌 烽烟 风雨 浩劫 风云 烽火 龙蛇 江湖 江东 辛苦 莽莽 海外 名士 乾坤 垂老 兴亡 佳节 登临 长城 儿女 蹉跎 半壁 艰难 百花 胜地 人物 头颅 中兴 书生 旌旗 高歌 老去 钟山 波涛 云气 江流
Topic 4 相思天涯	相思 天涯 斜阳 西风 芳草 黄昏 飘零 梦里 惆怅 当年 杨柳 断肠 落花 鸳鸯 燕子 回首 多情 憔悴 匆匆 往事 梦中 肠断 思量 凄凉 阑干 寻常 相逢 清明 依依 灯火 清风 销魂 一夜 归来 芙蓉 庭院 美人 伤心 旧时 无情 流光 黄叶 看花 人间 情怀 珍重 心事 零落 如今 今宵

后记

时光飞逝，三年前的初春，来参加兰州财经大学的研究生复试，这是我第一次来到兰州。回蚌埠之前，在兰州的大学同学带我来到中山大桥，站在大桥上，我发现自西而来的黄河并不浑浊，滚滚向东，流向远方。

兰州很纯粹，好似一碗牛肉面，“一清二白三红四绿五黄”，所有面馆的做法几乎是统一的；兰州又很独特，不同的面馆风味又千差万别。

在兰州，我遇到我的导师王永瑜教授与师母马新惠老师。王老师知识渊博、治学严谨、品德高尚、诲人不倦。在论文的写作过程中，从选题到最后定稿，老师都会细致入微地给出自己的分析。师母品德高尚、和蔼可亲，在读三年里，无微不至地关心着我们的生活。

在兰州，我遇到了很多给予我关心与帮助的统计学院的老师们，黄恒君老师、韩君老师、刘明老师、郭精军老师、韩海波老师、杨盛菁老师、高海燕老师、庞智强老师、傅德印老师、田茂再老师与张崇歧老师等，无论是一封邮件的回复、电话里细心的建议还是课堂上不一样的见解，这些都给了我很大的收获。

在兰州，我遇到了我的师兄师姐、师弟师妹与同届同学们，感谢王蕊师姐很多细致的回复，感谢许程程与杨亨莉在学习上的互帮互助，师门建立的深厚友谊无疑是我今后不可多得的精神财富；感谢我的室友与同班同学们，亲密合作，共同进取；我也感谢“Social Listening 与文本挖掘”公众号作者高长宽老师，写作过程中，关于分词工具的使用给了很多建议；此外还有 CSDN 和 GitHub 等平台，带着问题学习，在上面可以学到很多东西。

在最后，我感谢我的父母，感谢父母的养育之恩与做人做事的教诲。家庭永远是最温馨的港湾。

李伦珑

2021 年 5 月 10 日