西南科技大学

Southwest University of Science and Technology

本科毕业设计（论文）

题目名称：诗酒数据可视分析系统设计与实现

学院名称	计算机科学与技术
专业名称	软件工程
学生姓名	姚永坤
学号	5120188039
指导教师	彭莉娟 讲师 王桂娟 助教

二〇二二年六月

西南科技大学
本科毕业设计（论文）学术诚信声明

本人郑重声明：所呈交的毕业设计（论文），是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

作者签名：

日期：　　年　月　日

西南科技大学
本科毕业设计（论文）版权使用授权书

本毕业设计（论文）作者同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权西南科技大学可以将本毕业设计（论文）的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本毕业设计（论文）。

保密□，在__年解密后适用本授权书。

本论文属于

不保密□。

（请在以上方框内打“√”）

作者签名：　　指导教师签名：

日期：　年 月 日　　日期：　年 月 日

诗酒数据可视分析系统设计与实现

摘要：为了解决“诗酒文化”普及程度不广、探索分析诗歌情感主题方法单一以及分析结果呈现方式不直观、难理解等问题，本文提出一套基于深度学习和数据可视化技术分析方法，并依据该方法设计实现了诗酒数据可视分析系统。本系统采用诗歌文本情感分析法和诗歌主题提取法探索从先秦至现当代的酒诗词数据中所蕴含的情感和主题，同时对算法模型运行结果采用数据可视化的方法设计图表，最后将可视化布局和交互相结合，在时间、空间和主题情感等多个维度实现可视分析系统。案例分析得出的结论表明：本文系统能够完成探索酒诗词情感随朝代和年份发展的变化趋势、结合历史背景分析不同地理位置对酒诗词情感的影响、分析诗人社交网络和诗歌情感等研究任务。本系统可以为诗酒文化的研究提供一种新的思路，同时可以帮助普通民众、甚至外国友人了解和学习诗酒文化，从而为诗酒文化的普及提供一种新的途径。

关键词：诗酒文化；深度学习；诗歌情感和主题分析提取法；数据可视化

Design and implementation of a visual analysis system for Poetry and Liquor data

Abstract: This thesis proposes a set of analysis methods based on deep learning and data visualisation techniques, and designs and implements a visual analysis system for poetry and liquor data based on this method to address the problems of the lack of popularity of "poetry and liquor culture," the single method of exploring and analysing the emotional themes of poetry, and the lack of intuition and understanding of the presentation of analysis results. The system employs poetry text sentiment analysis and poetry theme extraction to investigate the emotions and themes present in liquor poetry data from the pre-Qin dynasty to the present day, while also employing data visualization methods to create charts and graphs for the algorithm model's output, and finally combining visual layout and intersection to realize the visual analysis system in multiple dimensions such as time, space, and thematic emotions. The conclusions from the case studies show that the system outlined in this thesis is capable of completing research tasks such as examining changing trends in wine poetry emotions across dynasties and years, analyzing the impact of different geographical locations on liquor poetry emotions in historical contexts, and analyzing poets' social networks and poetry emotions. This system has the potential to create a new way of thinking about poetry and liquor culture, as well as a new means of popularizing poetry and liquor culture by assisting the general public, including international friends, in understanding and learning about poetry and liquor culture.

Key words: Poetry and liquor culture; Deep learning; Extraction methods for sentiment and theme analysis of poetry; Data visualization

目 录

第 1 章 绪论

1.1 选题背景

中国文化博大精深。我国是诗的国度，也是酒的故乡。酒，除了被当作食物，它往往还寄托着饮者的多情善感。中国的酒文化具有鲜明的民族特色，是国家文化大数据不可或缺的重要组成部分，我国酿酒和饮酒历史悠久，经过数千年的发展和积淀，形成了独特的风格[1][2]。自古以来，酒贯穿社会社会生活的方方面面，既是国家和社会生活重要事件的见证者，也是人民情感和社交的寄托。诗，因为其言语精炼、汇聚情感、富有内涵等特点深受古代文人学者的喜爱。而当诗人饮酒作诗，那么诗与酒就在机缘巧合下结下了不解之缘[1][2]。酒和诗两者相辅相成，酒时常在不同的场景中出现同时表达着别样的情感：送别友人往往会流露出不舍之情，宴请宾客往往会表达出开心和喜悦，而独酌则会体现出孤寂和相思……。酒所代表的酒文化和诗代表的诗文化相互融合形成了“诗酒文化”。我们通过对“诗酒文化”研究，可以分析诗歌中蕴含的情感，折射出当时诗人自身的情况和大的历史背景，同时对于我们研究我国古代历史提供捷径。

以往对于“诗酒文化”的研究，往往需要这一方面的专家学者进行研读，然后结合大的历史背景和自身所学才能够将研究任务完成。由于诗歌中含有丰富的感情，利用文本情感分析和机器学习相关方法对诗歌中蕴含的情感进行提取分析，这样可以更加高效、科学[3]。同时，对于得到的结果通常情况下也只能用文字进行表述，不是很直观、易懂。数据可视化[4]运用图形化的方式，结合多视图的交互将不同的结果呈现给用户，使其更加直观便于理解。由于诗酒数据具有时间和空间等多维特征，所以我们将通过文本情感分析得到的结果与这两个维度进行结合，这样就可以分析酒诗词情感随朝代、君主和年份发展变化趋势，以及不同地理位置酒诗词中蕴含的情感风格。

本课题通过设计并实现诗酒数据可视分析系统，对诗酒文本类数据进行情感分析，同时结合时间和空间双维度对其进行展示并提供交互用于分析情感随时间和空间位置发生的变化，进而分析我国古代历史大小事件对于诗人情感产生的影响等。

1.2 国内外研究现状

近几年，国家内外学者对于中华传统文化研究的越来越多，而诗歌，因为其短小精悍、探骊得珠等的特点让其光芒四射，拥有很高的研究价值。特别是对于“诗酒文

化”的研究颇多，他们从诗歌学、美学、文化学角度做出学理阐释与新的论说、探寻“精神文化”——诗与“物质文化”酒之间的关系、感悟古人情怀以及观照历史变迁等。

1.2.1 国内研究现状

王玉成等[5] 首先对于唐朝“诗酒文化”的概念进行了解释，然后总结得出其三个特征，针对不同特征又分开进行了阐释，最后针对其形成原因从主观和客观两个方面进行了总结：主观因素因为饮酒作诗表达情感已经成为了重要的传统，客观原因主要是当时具备经济和时间这两个条件；吴亚东等[6]针对中国白酒文化在可视分析领域进行了探索与研究，他们将可视化技术应用到中国白酒文化领域当中，论证了其应用场景、流程等方面，然后针对不同类型的白酒文化数据也提出了不同的创作思路以及不同类别的可视化设计方案并罗列多样的可视化图表进行展示论证，最后又从 3 个角度分析了中国白酒文化可视化研究现状以及未来发展；张玮等[6]在对宋词的研究中运用了可视化的手段，提出了一种可视分析方法并将其应用在宋词的研究上，然后分层次对这一个可视分析系统进行详细的介绍，针对不同类别的数据也设计了不同类型的可视化图表，具有较强的创新性，这种处理方法为宋词的研究开辟了新视角，同时为该领域的未来提供了新的解决方案；欧阳剑[8]通过分析现有对于古籍文本的研究中存在的缺陷，提出了将数字人文应用在中国古籍当中以古籍文本为基础，通过大数据实时分析技术来解决数据量较大的问题，同时提取并阐述古籍中人物和历史事件的关系，展示各时期古籍的空间分布和同步变化，展示语言在时间和空间上的发展和变化，最后也提出并实现了一个系统用于验证效果；封颖超杰等[9]提出了一种生成算法用在中国古典诗歌上，同时设计实现了一种评分方法来评定诗歌的等级，在以上所述之上创建了一个个可视化交互创作系统“为你写诗”用来呈现中国古典诗歌，为用户提供创作意见然后用作改正，最后通过论证获得了不错的有效性和实用性；李斌等[10]为了解决传统文学研究中速度慢等问题，提出了将机器分析与人工校对相结合的方式对古籍进行处理和分，然后以《左传》为例进行词切分等一系列操作，将可视化手段应用在这上面，得到了非常好的效果，同时为古典文本的内容标注等提供了新的研究路径;王妮满等[11]为人文科学中的层次分析创建了基于 WebGIS 的分析系统，通过对诗歌、文言文等史料进行提取分析，以玄奘、欧阳修等历史名人的人生轨迹进行空间可视化与分析得出了多个重要结论，为多维度文本数据可视分析提供了新角度，同加强了不同学科间的互

通交流。

1.2.2 国外研究现状

Luis Meneses 等[12]通过回顾以前与诗歌可视化有关的研究总结出其中的短板，然后描述他们所开发的工具的原理并探讨其特点，创建这一套可视化工具的目的是为了：帮助学者综合通过视觉强调叙事的结构、诗歌的组织、语言元素和使用的隐喻来分析诗歌，最后他们通过综合标定的方式完成了目标；Mittmann A 等[13]设计一种诗歌的多层次可视化方案，在不同层次中不同的元素被代表，它们的属性以不同的方式被编码。允许用户从单个诗歌的语音分析开始，跳转到基于网格的诗歌整体可视化，跳转到多首诗的视图，其中音节属性是彩色编码的，最后，在基于网格的可视化中，一次获得数百首诗的大图视图；O Musaoglu 等[14] 他提出了一个可扩展和可适应的可视化系统，创造了新的方法来表示每首诗的结构、节奏和情感基调。还包括一种跟踪整个语料库的特征随时间变化的连续性（或不连续性）的方法。该系统的创建有一个用于存储元信息的数据库和一个用于建立和连接的可视化网站。

1.2.3 总结

综上，从国内研究现状可以看出：国内的专家学者对于古代文学的研究一直都有涉足，近些年也开始将数字科学与文学研究进行一些融合，利用机器学习中的算法、数据可视化的手段来满足整个研究工作，但是从整体上看在这一方面的研究成果还是较少，另外针对“诗酒文化”的研究也比较少；从国内外研究现状可以看出：由于国外的诗歌体裁和形式相较于我们国家的传统诗歌还是有较大的区别，这也就导致了专家学长在开展研究时其研究的侧重点就会有较大的偏差。总的来说，国内外的研究中对于诗歌情感的分析有所涉足，将可视化手段应用在诗歌上也有少部分成果，但是将机器学习算法中文本情感分析和主题提取的方法以及数据可视化的手段相结合同时应用在“诗酒”领域更是少之又少了，所以对该选题的研究必要性很强。

1.3 系统目标

对于本系统主要的目标是设计并实现诗酒数据可视分析系统。利用文本情感分析的方法对诗酒数据中的主题和情感进行提取和分析，然后利用可视化的手段从时间、空间和主题情感等多个维度创建可视化系统。充分将酒诗词情感随朝代、君主和年份

变化的趋势以及随地理位置呈现不同风格这一特点展示出来。能够让大众更加直观的了解“诗酒文化”，同时也能为研究人员提供有效的分析和探索工具。

1.4 本章小结

本章从“诗酒文化”的历史背景开始，对其研究价值和意义进行阐述。紧接着通过调研国内外对于“诗酒文化”的研究可以得知对于“诗酒文化”这一领域缺少创新型的研究手段，如数据可视化与机器学习算法相结合同时应用在诗酒文化研究中，进而提出整个系统的研究目标：设计并实现诗酒数据可视分析系统。

第 2 章 诗酒数据可视分析系统需求分析

2.1 可行性分析

诗酒数据可视分析系统的可行性分析，大致包括以下三个方面：技术可行性、经济可行性、开发可行性，本小节将从以上三个方面分节展开论述。

2.1.1 技术可行性

对于整个系统的开发所要利用的技术主要分为三部分：数据爬取、算法、可视化实际实现（如图 2-1）。

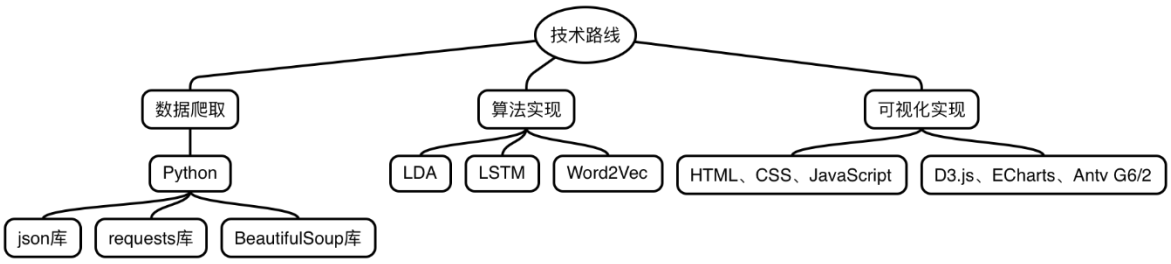

图 2-1 诗酒数据可视分析系统技术路线

数据爬取部分主要利用 Python 语言，使用 request 库和 BeautifulSoup 库来完数据的爬取，使用 json 库实现数据的存储。Requests 是一个简短而直接的第三方库，用于处理 HTTP 请求，它经常被用来抓取和评估服务器的响应数据。BeautifulSoup 是一个第三方软件包，可以解析和处理 HTML 和 XML[15]。

算法实现部分主要分为两大部分，文本情感分析算法：LSTM（Long Short-Term Memory）+Word2Vec；主题提取算法：LDA（Latent Dirichlet Allocation）。算法部分均使用 Python 实现。

可视化实现主要分为两大部分：前端界面绘制和可视化图表绘制。其中前端界面的绘制主要使用前端三件套外加 Vue 框架，可视化图表的绘制则使用第三方 JS 可视化图库。以上提到的所有技术均为学习或实践尝试过，不会存在临时学习等影响整个开发进度因素。

2.1.2 经济可行性

诗酒数据可视分析系统为前后端分离的 Web 应用，其开发仅需要一台电脑，系统相关代码的编写工具均为开源免费，满足开发的标准，不存在影响开发的因素。

2.1.3 开发可行性

通过第一章中对与国内外研究现状的调查研究可以发现，现存的对于“诗酒文化”领域的研究较少，同时将多样化技术应用在该领域的成果较少，所以本系统的开发具有较强的开发目的和意义，满足开发的可行性。

2.2 功能需求

诗酒数据可视分析系统整体的功能需求满足传统应用可视化系统的基础功能，另外由于本系统基础数据为文本类型数据，所以应当在对诗词以及其情感和主题等文本类数据进行可视化展示的同时，还需要设置视图内交互和视图间联动来让用户充分观察和了解整体情况。

2.2.1 业务流程

业务流程图如图 2-2。本系统业务流程从用户访问使用诗酒数据可视分析系统开始，访问存在成功访问和失败访问，系统访问可能会因为网络、用户电脑等外界因素导致访问系统失败，此时需要系统做出相应的报错提醒，如果浏览器亦或者其他应用能够给与用户一些报错提醒，那么系统可以直接使用该报错提醒；当用户成功访问系统后，系统前端会向后端发送数据请求，而此时后端将会从数据库中请求数据，由于请求到的数据或多或少与可视化图表的需求数据存在一定的偏差，所以此时还需要在后端进行一部分的数据处理操作，在这些操作执行完毕后才会将数据返回至前端，前端首先加载整个页面的框架布局，在收到后端数据后才会进行可视化图表的渲染，至此页面加载完成，此时支持用户进行交互操作，而当用户进行交互操作时，每执行一次都会将相对应视图发生更新，如果交互前后显示数据相同则不会重新渲染图表，直至用户关闭系统。

2.2.2 功能列表及其说明

整个系统的功能模块进行分类大致可以分为五类，详细内容见表 2-1。

表 2-1 功能列表

功能模块	功能	备注
地图模块	诗酒数据空间分布	
时间轴模块	诗酒数据时间分布	

功能模块	功能	备注
诗人间关系模块	展示关系数据	
诗词文本模块	呈现诗词文本数据	
诗词主题模块	展示朝代主题	

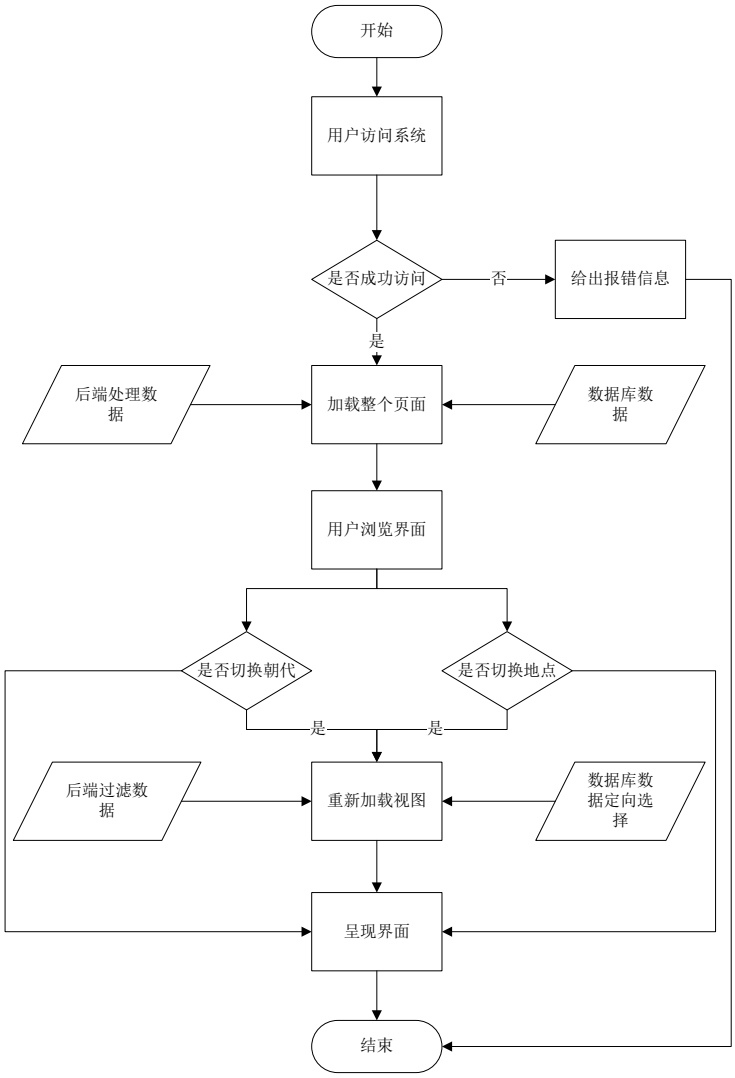

图 2-2 诗酒数据可视分析系统业务流程图

2.3 数据需求

本系统基于诗酒数据进行算法的执行和可视化分析与设计，本小节将从四个方面展开陈述本系统的数据需求。

2.3.1 数据来源

本系统的数据来源大致分为三部分（见图 2-3）：现有数据库中的文本数据、诗人关系数据库和其他数据。现有数据库和关系数据库为已存在的数据只需要对其进行处理即可作为基础数据来使用。其他数据是需要通过爬虫技术重新获取的诗歌补充数据。

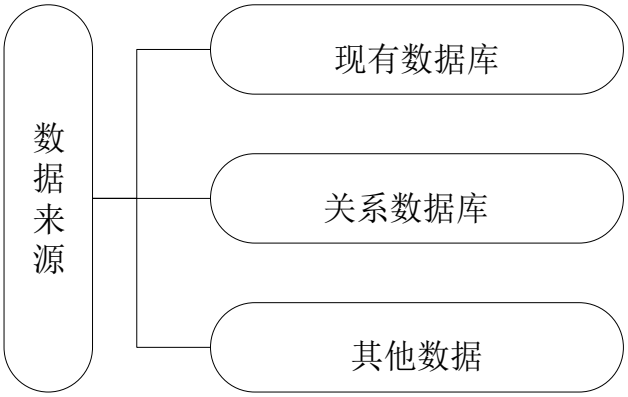

图 2-3 诗酒数据可视分析系统数据来源

2.3.2 数据规模

原始数据由三部分构成：诗歌基础数据、诗歌情感数据、诗人关系数据这三部分数据。数据按诗人进行整合后总数大约在在 10 万条左右。其中针对诗歌基础数据里面单条数据应当包括以下内容：诗歌名称、朝代、作者、诗歌内容等，如图 2-4 所示。另外两部分数据诗歌情感数据需要与诗歌基础数据进行整合，诗人关系数据作为单一数据集分开进行存储。

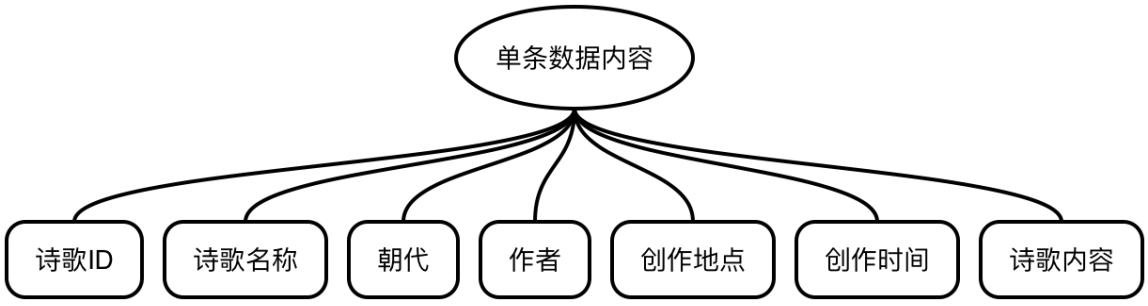

图 2-4 单条数据内容

2.3.3 存储需求

由于数据总量预估在 10 万条左右，所以对于原始数据需使用数据库进行存储。处

理分析后的数据可按照数据类型不同进行划分，分别使用 csv 或者 json 文件进行单一存储。这种储存方式一方面，方便新获取数据的存储，另一方面也方便可视化系统对于数据进行操作与处理，避免多次调用数据库造成性能损耗。

2.3.4 预处理需求

在对数据进行整理汇总之前需要对数据进行预处理。预处理需要对原数据库中的诗歌数据进行过滤，同时需要提取关系数据库中的诗人关系数据。预处理需要使用 Python 进行数据的处理和转存，另外还需要安装 SQLiteSpy 软件来对数据库进行查看和操作。

2.4 性能需求

本系统的性能需求大致分为四个性能指标用于性能测试，详情见表 2-2。

表 2-2 性能需求表

性能指标	要求
系统首屏加载时间	首屏加载时间控制在 8 秒内
系统加载性能（LCP）	页面的 LCP 应保持在 2.5 秒或更少
系统交互性（FID）	交互响应时间控制在 100 毫秒内
系统视觉稳定性（CLS）	页面的 CLS 应保持在 0.1 或更少

2.5 本章小结

本章从四个方面分析了系统的需求。可行性分析主要从技术可行性、经济可行性和开发可行性三个方面对诗酒可视分析系统开发的可行性进行了论证；紧接着对功能需求进行展开论证，使用流程图和功能表全方位的阐释功能需求，为后续开发设计提供参考；因为一个可视化系统最重要的部分就是数据，所以对数据进行需求分析这部分是不可或缺的，同时在数据需求这一小节多角度来对数据需求进行解释；最后一小节对系统的性能也提出了相对应的性能指标以及要求便于系统测试时使用。

第 3 章 诗酒数据可视分析系统算法实现

3.1 算法概述

本系统主要用到三个算法：Word2Vec、LSTM、LDA。Word2Vec 和 LSTM 算法主要是用于分析诗歌中表达的情感，而 LDA 算法主要时用于提取诗歌主题词。

3.1.1 Word2Vec

Word2vec 是一种单词嵌入技术，使用实数集合来表示一个字符串，这是一种用于自然语言处理（NLP）的技术，使用深度学习从指定文件中提取数据，Word2Vec 使用两层神经网，以统计或矢量形式表示输入文本，Word2vec 的用途超出了句子分析；它可以在任何有明确模式的应用中得到利用[16]。

3.1.2 LSTM

LSTM 是一种特殊的循环神经网络——长短期记忆网络，能够学习长期依赖关系。它主要通过三个门控逻辑：遗忘、输入、输出实现。它是为了解决长序列训练过程中的梯度消失和梯度爆炸问题而提出的[17][18]。它常用于基于语言模型去预测下一个单词、词性标注等。

3.1.3 LDA

LDA 是一种对语料库进行建模的无监督生成性概率方法，是最常用的主题建模方法。LDA 假设每个文档都可以被表示为潜伏话题的概率分布，并且所有文档中的话题分布都有一个共同的 Dirichlet 先验。LDA 模型中的每个潜伏话题也被表示为单词的概率分布，话题的单词分布也共享一个共同的 Dirichlet 先验[20]。

3.2 诗歌情感计算

诗歌是由一句句诗组合而成的。每一句诗往往蕴含着 1～2 种情感，而将诗歌中每一句诗的情感统计汇总起来，然后按权重进行计算就可以大概率推测出整首诗的情感，这就是诗歌情感计算的整个过程[19]，详情见图 3-1。

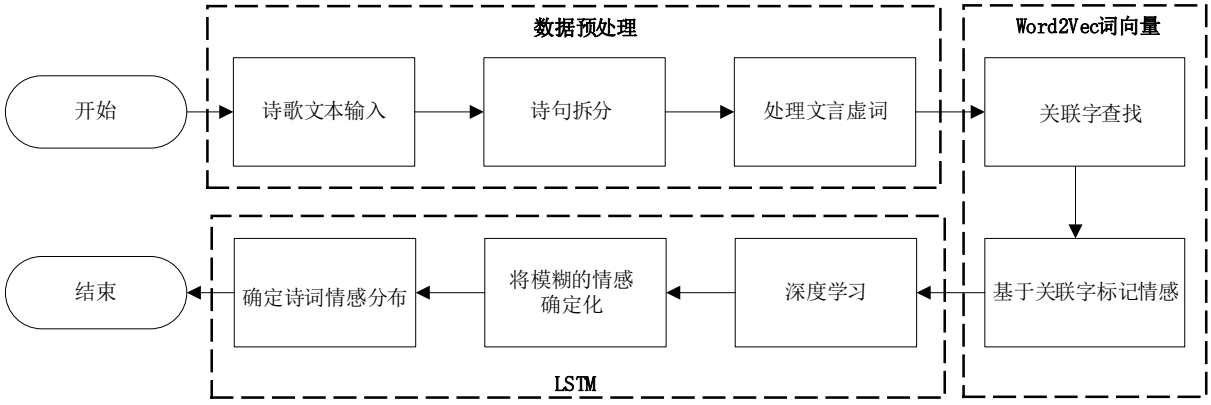

图 3-1 诗歌情感计算算法执行步骤

3.2.1 算法执行流程

本节将对诗酒数据情感分析法具体执行流程进行解释，如图 3-2。首先将提前准备好的实验语料库加入到词向量模型当中，通过该模型将五种情感的关联词数据集。接下来需要将该数据进行合并得到执行算法的训练集，然后在执行算法的步骤中同时引入 LSTM 模型、训练集和测试集。执行完毕算法后得到的结果中包含整首诗歌的情感以及诗歌中每句诗的情感，如果诗句与其五种感情的关联度不高则表述为无情感。

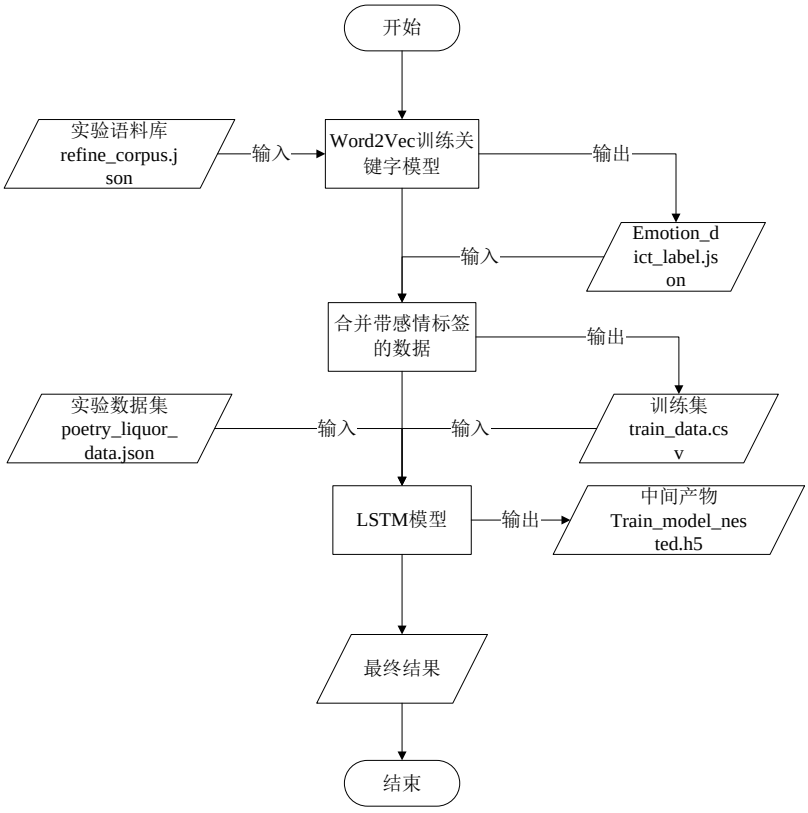

图 3-2 诗歌情感计算流程图

3.2.2 诗句情感分类

1、基于 Word2Vec 的关联字查找

首先，对语料库中所有的诗词做预处理操作，预处理目的是：将原数据进行预处理转化为模型的输入数据。如图 3-3 所示，原数据为诗酒文本数据，数据量在 10 万条左右，单条数据内容包括诗歌名称、朝代、作者、创作地点、创作时间、诗歌内容六个维度。由于本算法仅需要使用诗歌内容这一维度的数据，所以需要对原数据进行提取，在提取数据的同时需要对诗歌中的标点符号、特殊字符、文言虚词等无关内容进行过滤删除。然后将每一首诗歌的每一句话作为一个词向量，从语料库所有的句子中抽取长度大于 1 的向量作为 Word2Vec 模型的训练数据。最终得到的训练数据，数据大小为 75 万条左右，该数据将用作 Word2Vec 关联字查找算法模型的输入。

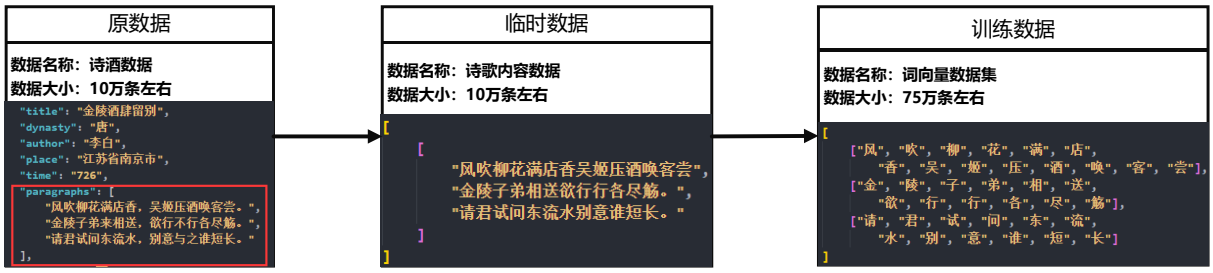

图 3-3 Word2Vec 关联字查找算法预处理

其次，创建 Word2Vec 关联字查找模型。这里需要将训练数据引入，紧接着需要对模型参数进行设置（如表 3-1 所示）。由于本语料库（训练数据）较大，所以在设置词向量维度时需要将值设定为 300，这样可以避免因为值太小导致词映射产生冲突进而影响最终结果，同样如果设置的值过大也会导致内存消耗变大使得计算速度变慢；然后设置基准词频为 5，这样设置可以忽略只出现一两次的单词。

表 3-1 Word2Vec 模型参数列表

参数	说明	设置数值	备注
sentence	语料库（训练数据）	database	
sg	训练算法	0	0（默认）：CBOW 算法（训练速度较快） 1：skip-gram 算法（训练速度慢，对罕见字有效）
vector_size	词向量的维度	300	默认为 100
min_count	最低词频数	5	默认为 5
workers	线程数	1	默认当前运行机器的处理器核数

然后，设置诗词情感标签。将酒诗词情感标签分为五类：喜、乐、怒、哀、思。由于古文中“怒”字常表述为发怒、气势强盛与“怒”这种情绪在意思上有一定的出入，所以为了提高模型的准确性，在本算法中将“怒”用“恚”来代替。接着，执行 Word2Vec 关联字查找模型。将情感标签逐个放入 Word2Vec 模型中会得到关联程度不同的汉字集合，关联值在 0 到 1 之间分布；每个情感标签得到的关联字集合按照关联程度由高到低选出 100 个关联字作为情感标签的关联字集合（如图 3-4），最终将结果存储在 emotion_dict_label.json 文件中。

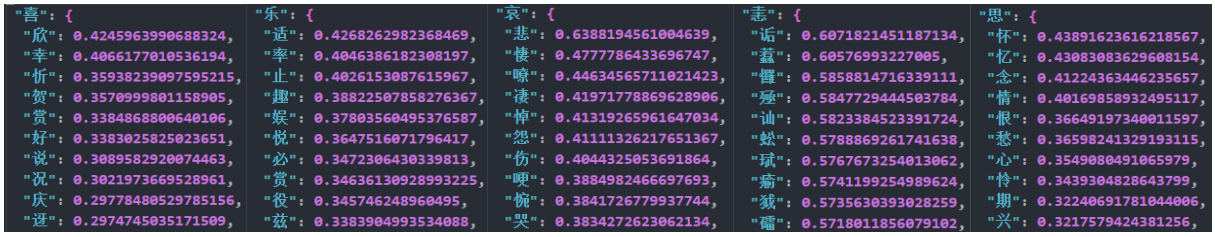

图 3-4 五种情感标签的关联字集合

2、基于关联字标记诗歌情感

在得到 emotion_dict_label.json 后，需要依据关联字来标记诗句和诗歌的情感。首先，我们将关联字集合转化为如图 3-5 所示的数据格式，以便于后续操作。接下来遍历五类情感标签的所有关联字，如果不同情感标签间存在重复的关联字，那么只保留关联得分最大的一项，其余重复项均删除。

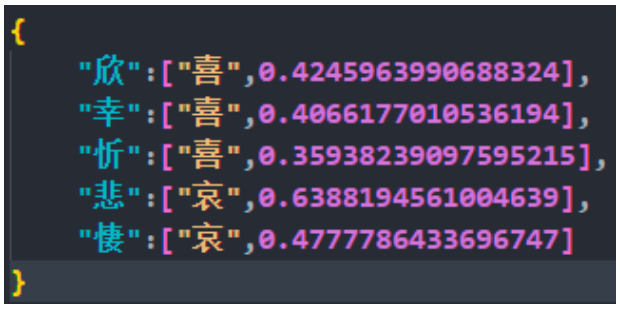

图 3-5 关联字集合转换后的数据格式

然后计算每句诗对应五类情感标签的关联得分，这里我们假设 x_i 为一个汉字，n 是诗句的长度，那么使用公式 3-1 就可以表示一句诗；设置 e 为感情类别标签编号，$value\left(x_i\right)_e$ 是汉字 x_i 对于 e 类标签的关联度，那么诗句 x 的在 e 类标签的感情关联得分可以使用式 3-2 来表示；诗句最大的标签关联度可以使用式 3-3 来表示。在具体实现上：首先，判断诗句中是否存在关联字，如果存在，那么就将该关键字对应情感标签

的关联得分记作这句诗在该情感标签上的关联得分，最终通过累加的方式求得整句诗对应五类情感标签的关联得分（关联得分初始值均为 0）。接着通过比较整句诗对应五类情感标签关联得分的大小来确定这句诗的最大标签关联度。

$$\mathrm{x} = \left\{\mathrm{x_i}\right\}_{\mathrm{i=1}}^{\mathrm{n}} \quad \text{（式 3-1）}$$

$$x_E = \max\left\{score_{x,e}\right\}_{e=1}^{5} \quad \text{（式 3-2）}$$

$$score_{x,e} = \sum_{i=1}^{n} value\left(x_i\right)_e \quad \text{（式 3-3）}$$

接着需要根据五类标签对应的关联字字典来确定每一种情感标签的阈值T_e，该阈值将用于判断诗句对应的情感。在实验中，五类情感标签设定的阈值分别为$T_{喜} = 1.2$，$T_{乐} = 1.6$，$T_{哀} = 1.1$，$T_{思} = 1.5$，$T_{怒} = 1.2$，在不同的实验环境下阈值可以根据实际的情况再调整。诗句x的最终感情类别E分为三大类情况，见式 3-4，第一种情况x_E大于等于人为设定的阈值T_e，则诗句的情感标签可以归为e，第二种情况是x_E的关联度不强，此刻得出的情感标签准确定较低，属于未能识别出情感的诗句，为 unknown 状态。第三种情况是$x_E = 0$，即该诗句和所有的情感关联词无交集，此刻诗句的情感极不明显，本算法将这类诗句设置成无情感诗句。

$$E = \begin{cases} label - emotion & if\ x_E \geq T_e \\ unknown & if\ T_e > x_E > 0 \\ non - emotion & if\ x_E = 0 \end{cases} \quad \text{（式 3-4）}$$

3.2.3 确定诗句情感

在执行完毕以上模型后得到了大约有 1500 条诗句可以确定具体的情感，大部分的诗句的情感式无法确定的。接下来可以通过现有的情感标签，结合深度学习框架提炼文本的特征和语义信息，做出诗歌情绪的判断。

第一步，为 LSTM 模型准备数据集（如图 3-6）。先将已经得到的情感的诗句添加至数据集中，其次通过随机取样的方式从未识别到情感的诗句中随机提取 3 万条数据并将其添加至数据集，接着在数据集中添加一部分已经确定情感的诗句，这部分诗句

属于附加数据。综合以上数据，整个数据集大小在 3 万条左右，然后将这部分数据存储在 train_data.csv 文件中作为 LSTM 算法模型的其中一个输入，另一个输入是：诗酒文本数据。

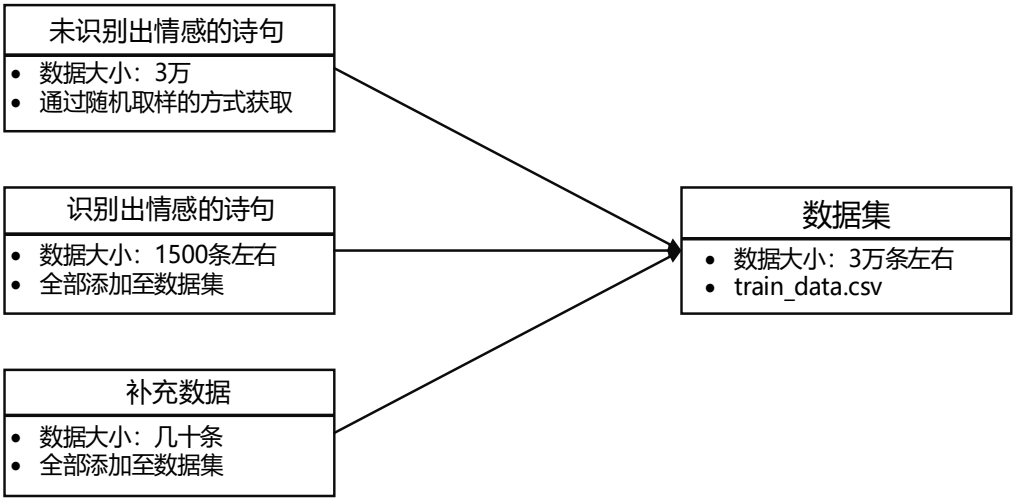

图 3-6 LSTM 模型数据集转换

第二步，执行 LSTM 模型预处理操作（如图 3-7）。执行这一步的目的是：为了将数据序列化以满足后续 LSTM 模型的训练。这里需要创建情感标签字典和词汇字典来将情感标签和诗句中每一个字进行数值化，同时创建输出字典用于将数值化数据转化为文本类型数据（如图 3-8 所示）。由于诗句的长度（字数）不相同，所以需要执行序列补充操作，先找出最长的序列，再将不满足最大序列长度的序列用 0 进行补充。接下来需要将数据集划分训练集和测试集，两者比例为 7:3，即训练集大小为 22156，测试集大小为 9496。

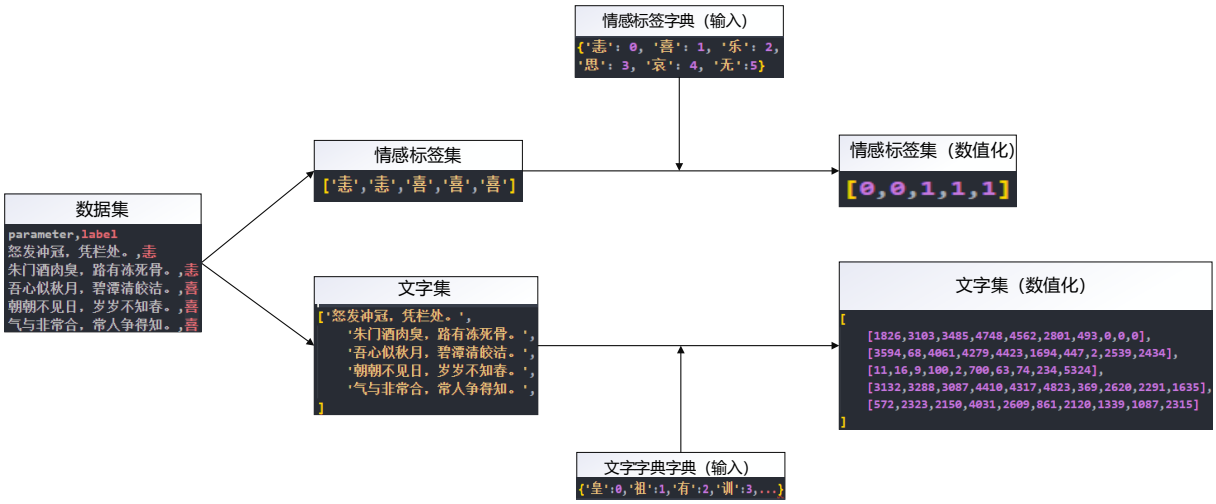

图 3-7 LSTM 模型数据预处理步骤

```
input_label_dictionary = {'恚': 0, '喜': 1, '乐': 2, '思': 3, '哀': 4, '无':5}
input_word_dictionary = {'皇':0,'祖':1,'有':2,'训':3,...}

output_label_dictionary = {0: '恚', 1: '喜', 2: '乐', 3: '思', 4: '哀', 5: '无'}
output_word_dictionary = {0:'皇',1:'祖',2:'有',3:'训',...}
```

图 3-8 情感标签和词汇输入字典（上）和输出字典（下）

第三步，创建并执行 LSTM 深度学习算法模型。本算法使用三层序列模型：Embedding + NLSTM + Softmax 来构建整体的框架，以上三层分别对应嵌入层、NLSTM 层和卷积层、全连接层。另外在卷积层与全连接层之间增加池化层、Dropout 层和 Flatten 层来对数据进行降维、防止过拟合等操作。 七层神经网络结构的详细内容见表 3-2。

表 3-2 七层神经网络结构表

神经网络结构	描述
嵌入层	将每一个汉字用一个唯一的数字表示，每一种情感用单编码向量表示为唯一的向量。一句诗被编码为一个等长的数字向量，诗篇则为多个数字向量组成的向量矩阵。将诗词编码成的向量作为嵌入层的输入
NLSTM 层	Nested LSTMs(NLSTM)，是一种具有多级记忆的新型 RNN 结构。它通过嵌套增加 LSTMs 的深度。相比于传统的 LSTM，Nested LSTM 有更高的自由度，能处理更长时间记忆规模的内部记忆。
卷积层	采用 Conv1D 卷积层。该层创建了一个卷积核，该卷积核以单个空间维上的层输入进行卷积，以生成输出张量。该层的激活函数采用整流线性单元。
池化层	采用 MaxPooling1D 层。该层将对象划分成若干区域，每个区域取最大值。池化层会不断减小数据空间，是降采样的实现。
Dropout 层	用于防止模型的过拟合。
Flatten 层	将输入展平，但不损失数据。
全连接层	该层的每一个神经元与前一层所有的神经元连接。该层用于整合具有类别区分性的局部信息，采用 softmax 逻辑回归进行情感分类作为最后的输出。

在算法模型的实现上，本算法采用 Keras 框架中 Sequential 模型来构建以上七层神经网络。首先第一层为嵌入层，在这里我们将文本传递给嵌入层，因为有数以万计的字词，所以我们需要比单编码向量（One-Hot Encoded Vectors）更有效的表示来输入数据。这里将使用上一小节提到的 Word2vec 字向量模型，用预先训练的词嵌入（Word

Embedding）来引入的外部语义信息，结合设定的情绪字典，做迁移学习（Transfer Learning）。为了取得更好的效果，本模型采用最新的 NestedLSTM+Conv1D 的深度学习模型来做情绪判断，它能较好的提炼文本里的特征和语序信息，记住更长的语义依赖关系，做出较为精确的情绪判断。在卷积层与全连接层之间首先创建池化层用于降采样成比缩小特征图的长和宽，本算法模型的池化层采用 MaxPooling1D—1D 输入的最大池化层。同时如果模型的参数太多，而训练样本又太少，那么训练出来的模型很容易产生过拟合的现象，所以为了避免上述问题在本模型还设置了 Dropout 层，它可以比较有效的缓解过拟合的发生，在一定程度上达到正则化的效果。接下来添加 Flatten 层，该层主要是为了将在不损失数据的前提下将数据扁平化，用作全连接层之前的过度使用。全连接层采用 softmax 逻辑回归进行情感分类作为最后的输出。在创建完毕模型后，需要设置相对应参数，然后训练模型，具体参数列表见表 3-3。

表 3-3 LSTM 模型参数列表

参数	说明	设置数值	备注
n_units	模型参数数量	200	-
batch_size	训练一次网络所用的样本数	32	数值大小影响模型的优化程度和速度
epochs	迭代次数	6	-
verbose	0：不显示日志 1：显示进度条	1	默认为 1

第四步，评估模型准确率。这里将测试集输入训练后的模型中的得到预测出的情感标签，然后利用分类准确率得到预测出的情感标签与真实的情感标签正确分类的比例，最终准确率为 94.86%，详情见表 3-4。

表 3-4 LSTM 模型准确率评估表

项	值
待评估模型	LSTM 模型
测试集大小	9496
评估方法	accuracy_score（分类准确率）
评估结果	准确率：94.86%

第五步，利用模型获取酒诗词情感。具体步骤如图 3-9 所示，引入诗酒文本数据，针对每一首诗歌，将诗歌中每一句诗输入到 LSTM 模型当中得到这句诗对应的情感标

签，当获取到整首诗歌每一句诗的情感后，对情感标签进行分类统计，占比较高的情感标签即为整首诗歌表述出的情感，至此整个算法模型执行完毕。

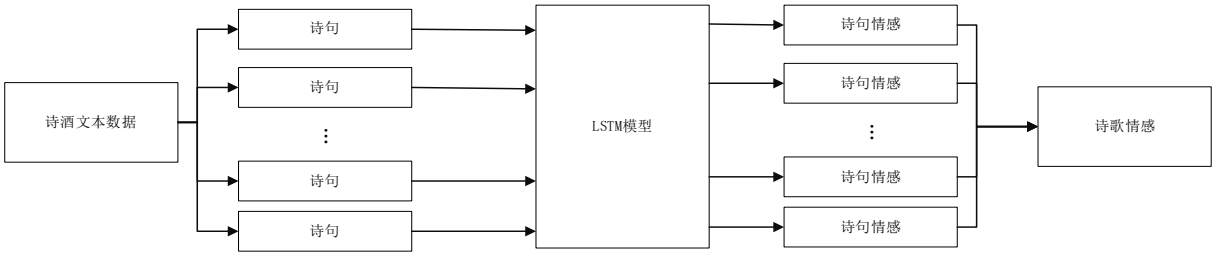

图 3-9 诗酒文本数据情感获取模型

3.3 诗歌主题词提取

3.3.1 隐含狄利克雷分布（LDA）

诗歌主题词提取使用了“隐含狄利克雷分布（LDA）”算法[20]。它是一种流行的主题建模技术，用于从给定的语料库中提取主题。隐含一词传达了一些存在但尚未开发的东西。LDA 生成用于形成主题的单词的概率，并最终将主题分类为文档。诗歌主题词提取整个过程执行两项任务：从语料库（文本文档集合）中找到主题，同时将这些主题分配给同一语料库中存在的文档。图 3-10 总结了 LDA 过程。

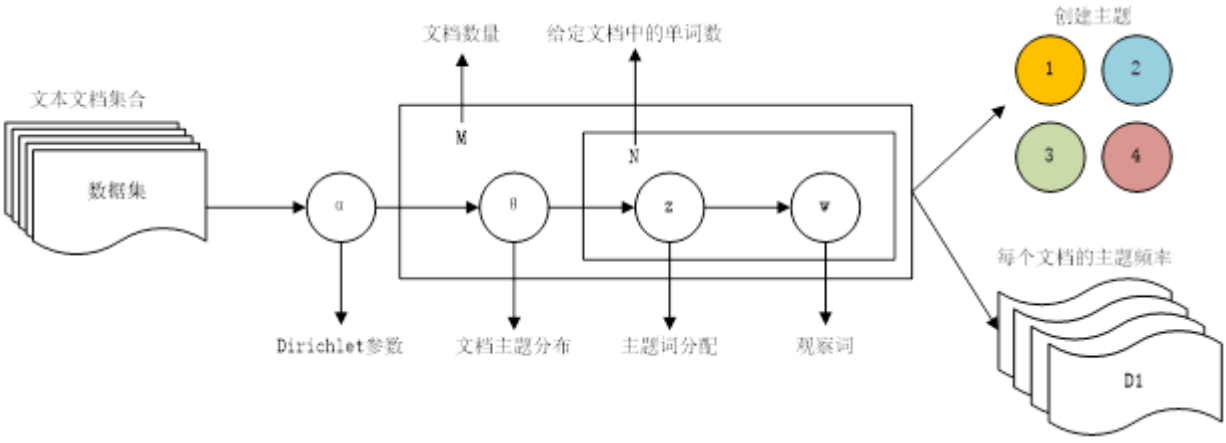

图 3-10 LDA 模型

3.3.2 LDA 向量空间

整个 LDA 空间及其数据集如下图 3-11 所示。黄色框是指语料库中的所有文档（M）。红色框是文档中的单词数，由 N 表示。在红色框中可以有很多字，其中一个词是 w，它位于蓝色圆圈中。根据 LDA，每个词都与潜在主题存在关联，此处使用 z

表示，将 z 分配给这些文档中的主题词给出了语料库中存在的主题词分布由 θ 表示。LDA 模型有两个控制分布的参数：α（控制每个文档的主题分布）和 β（控制每个主题词分布）。

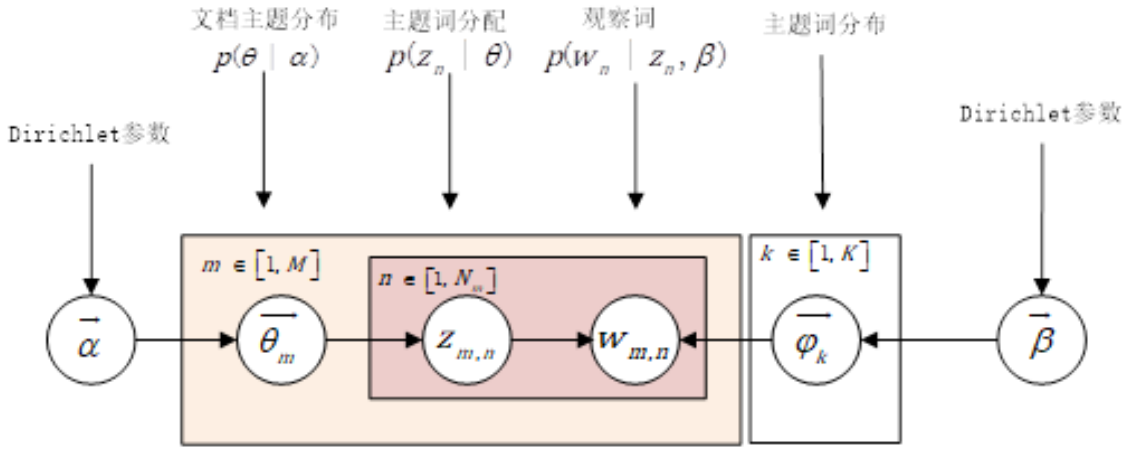

图 3-11 LDA 空间及其数据集

3.3.3 LDA 工作流程

LDA 的工作流程如下图 3-12 所示。需要明确一点 LDA 提出了两个关键假设：文档是主题的混合体和主题是标记（或单词）的混合并且，这些主题使用概率分布生成单词。在统计语言中，文档被称为主题的概率密度（或分布），而主题是单词的概率密度（或分布）。任何作为文档集合的语料库都可以表示为文档词，也称为 DTM。

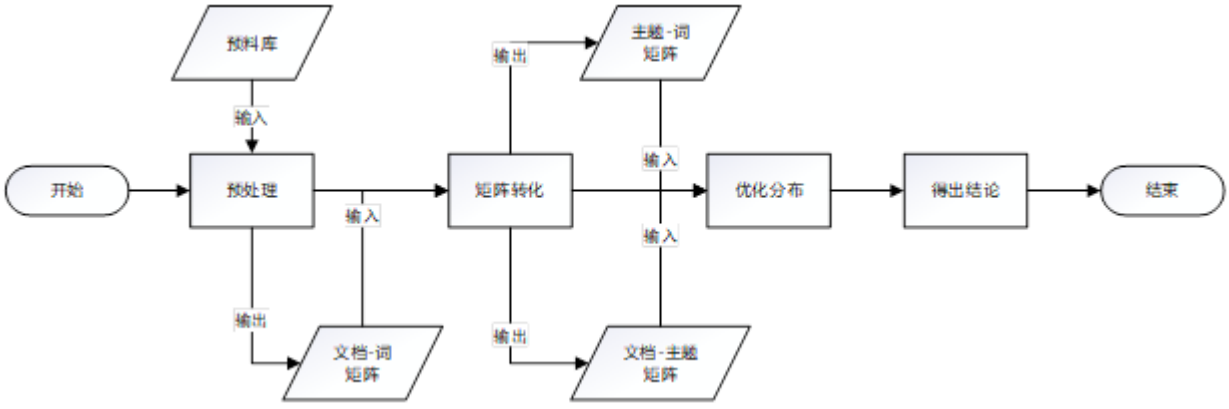

图 3-12 LDA 工作流程

首先，LDA 将以上两个重要假设应用于给定的预料库，执行预处理操作，其中预处理操作包括三部分：文本数据的清理、处理并将其标记为单词。对文档进行预处理后就可以得到文档-词矩阵，这里我们假设输入的文档为 D_i，文档数量为 I，单词由 W_s

表示，单词数量为S，那么文档D_i可以用式 3-5 来表示，文档-词矩阵就可以使用式 3-6 表示。其中每一行是一个文档，每一列是 tokens 或 words。

$$D_i = (W_1, W_2, W_3, \cdots, W_{s-1}, W_s) \quad \text{（式 3-5）}$$

$$\begin{pmatrix} W_{1,1} & W_{1,2} & \cdots & W_{1,s-1} & W_{1,s} \\ W_{2,1} & W_{2,2} & \cdots & W_{2,s-1} & W_{2,s} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ W_{i-1,1} & W_{i-1,2} & \cdots & W_{i-1,s-1} & W_{i-1,s} \\ W_{i,1} & W_{i,2} & \cdots & W_{\mathrm{i},s-1} & W_{i,s} \end{pmatrix}_{i\times s} \quad \text{（式 3-6）}$$

接下来 LDA 将此文档-词矩阵执行矩阵优化转换为另外两个矩阵：文档-主题矩阵（式 3-7）和主题-词矩阵（式 3-8），其中主题词有 k 个。这两个矩阵中：文档-主题矩阵已经包含了文档可以包含的可能主题；主题-词矩阵则具有这些主题可以包含的词。

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,k} & a_{1,k+1} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,\mathrm{k}} & a_{2,k+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{i-1,1} & a_{i-1,2} & \cdots & a_{i-1,k} & a_{i-1,k+1} \\ a_{i,1} & a_{i,2} & \cdots & a_{i,k} & a_{i,k+1} \end{pmatrix}_{i\times(k+1)} \quad \text{（式 3-7）}$$

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,s-1} & a_{1,s} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,s-1} & a_{2,s} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{k,1} & a_{k,2} & \cdots & a_{k,s-1} & a_{k,s} \\ a_{k+1,1} & a_{k+1,2} & \cdots & a_{k+1,s-1} & a_{k+1,s} \end{pmatrix}_{(k+1)\times s} \quad \text{（式 3-8）}$$

得到以上两个矩阵后将会执行优化分布这一步。LDA 的最终目标是找到文档-主题矩阵和主题-词矩阵的最优表示，从而找到最优化的文档-主题分布和主题-词分布。由于 LDA 假设文档是主题的混合体，而主题是词的混合体，因此 LDA 从文档级别回溯（迭代）以确定哪些主题会生成这些文档以及哪些词会生成这些主题。在第一次迭代

的过程中，它会将主题随机分配给文档中的每个单词，主题由字母 k 表示。因此，在我们的语料库中，文档中的单词将与一些随机主题相关联，这就会得到式 3-9 中的结果。这将输出作为文档，其中包含主题（式 3-10）和由单词组成的主题（式 3-11）。同样的 LDA 会给出其他主题的单词组合。

$$D_i = \left(W_1\left(K_5\right), W_2\left(K_1\right), W_3\left(K_k\right), \cdots, W_{s-1}\left(K_2\right), W_s\left(K_4\right)\right) \quad \text{（式 3-9）}$$

$$D_i = K_6 + K_{k-1} + K_3 + \cdots + K_1 + K_{10} \quad \text{（式 3-10）}$$

$$K_k = W_3 + W_s + W_7 + \cdots + W_{18} + W_5 \quad \text{（式 3-11）}$$

此时，LDA 做了另一个假设，即除了当前单词之外，所有已分配的主题都是正确的。因此，基于那些已经正确的主题词分配， LDA 将迭代每一个文档 D_i 和每个单词 W_s，通过计算两个概率：P_1（文档 D_i 中当前分配给主题 K_k 的单词的比例，式 3-12）和 P_2（来自的单词 W_s 的所有文档中分配给主题 K_k 的比例，式 3-13）来纠正和调整当前单词的主题分配。现在，使用这些概率 P_1 和 P_2，LDA 估计一个新的概率，它是 $\mathrm{P}_1 \times \mathrm{P}_2$ 的乘积，通过这个乘积概率，LDA 识别出新的主题，也就是与当前单词最相关的主题。通过 $\mathrm{P}_1 \times \mathrm{P}_2$ 的乘积概率将文档 D_i 的单词 W_s 重新分配给新主题 K_k。现在，为选择新主题 K_k 再次执行 LDA 同时进行大量迭代，直到获得稳定状态。LDA 在提供文档-词矩阵和主题-词矩阵的最佳表示时收敛。这样就完成了 LDA 的整个工作流程。

$$\mathrm{P}_1 = \frac{K_k}{D_i} \quad \text{（式 3-12）}$$

$$\mathrm{P}_2 = \frac{W_s}{K_k} \quad \text{（式 3-13）}$$

3.3.4 具体实现

本算法主要用于提取诗歌主题词，汇总统计后得到不同朝代的主题词。

第一步，数据预处理。首先明确一点，本算法的原数据为诗酒文本数据，所以需要先把诗歌内容提取出来，其次将提取出来的内容进行分词处理，即将一句诗拆分为多个汉字或者词语。在对分词方法的选择上，中文文本分词常用 Jieba 分词库，但由于

本算法针对的数据内容为诗歌这一类型的古汉语，Jieba 分词库主要以现代汉语为核心预料库，对古汉语的处理效果并不是很理想，所以本算法采用甲言分词。通过观察分词结果（图 3-13）可以发现其中存在大量的标点符号、文言虚词等，在此需要剔除这类数据，这里本算法采用哈工大停词表、百度停用词表、四川大学机器智能实验室停用词库四份停用词表进行了合并去重后的得到的中文停词表。由于本算法主要是用于提取各朝代诗歌的主题词，所以当执行完以上步骤后就需要按朝代将数据进行汇总来便于后续算法输入执行操作。

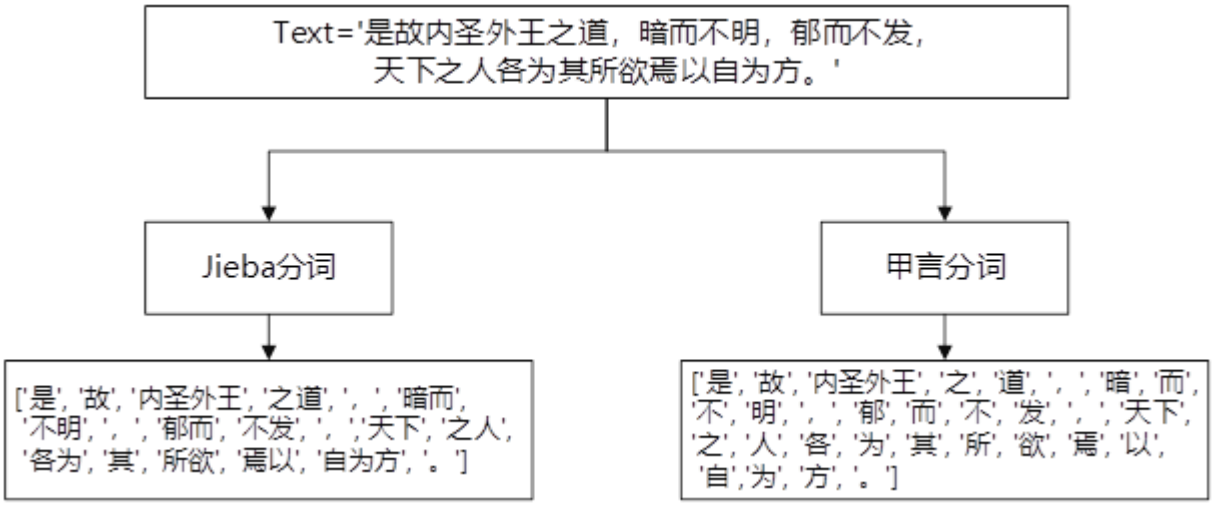

图 3-13 jieba 分词与甲言分词结果比较

第二步，构建词典并向量化语料。本算法使用 gensim 来做自然语言处理。首先使用 gensim.corpora 模块来对上一步处理后的数据进行字典的构建，同时被赋索引。其次，将文档列表转换成 DT 矩阵，用于表示第几个单词出现了几次（图 3-14）。

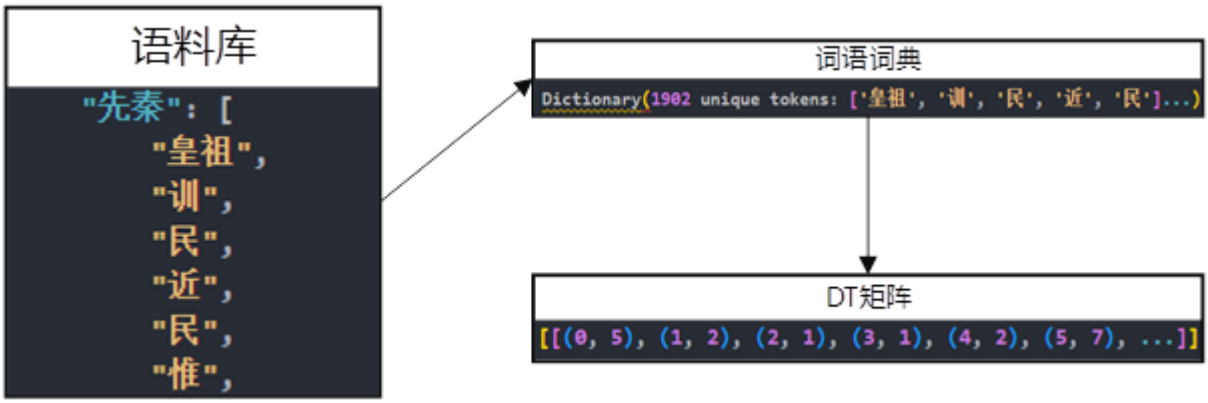

图 3-14 构建词典并向量化语料

第三步，构建 LDA 模型算法。本算法使用 gensim 所提供的 ldamodel 模块来构建

模型。因为针对语料集设置最优的主题数可以使得整个模型准确性更好，所以在这一步中最重要的就是需要确定主题数。本算法使用主题一致性（coherence）来选择最优主题数，一致性越高说明模型越好，主题数也为最优值。首先进行实验将主题数确定为 1-20 范围，其次执行模型，然后获得不同主题数目下 coherence 的大小，最后绘制主题数-coherence 折线图。如图 3-15 为唐朝诗歌主题数-coherence 的折线图，可以观察到在主题数为 8 的时候主题一致性最高，所以此时主题数应设为 8。同时由于不同朝代的诗作数量风格大都不相同，所以需要针对不同的朝代分别确定对应的最优主题数。各朝代对应最优主题数见表 3-5。

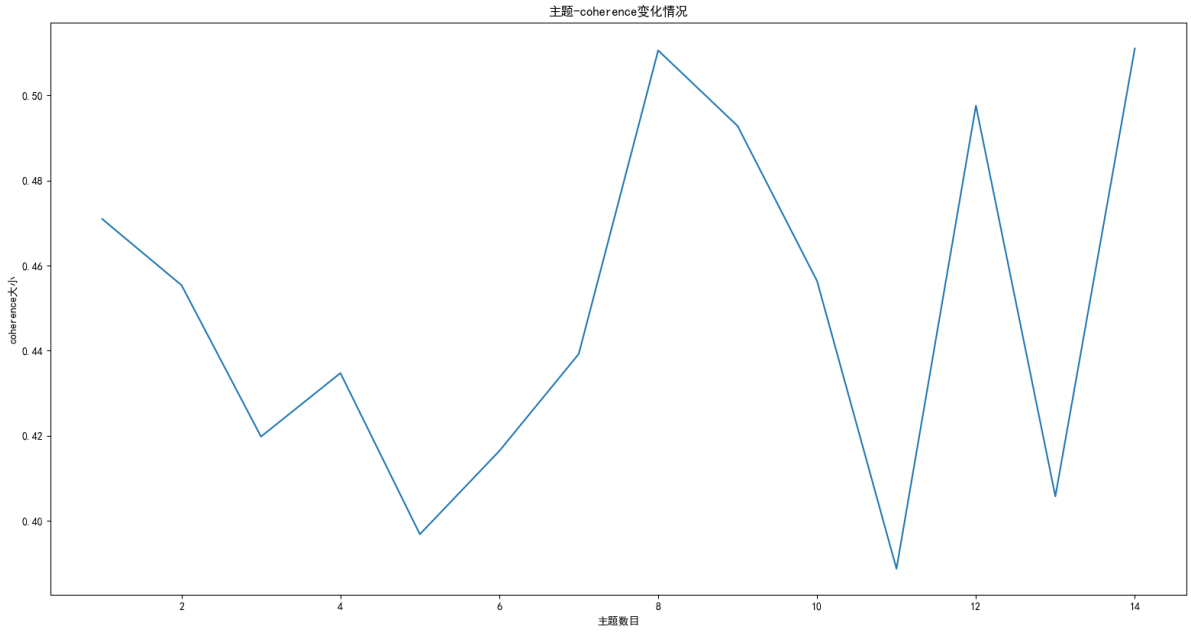

图 3-15 唐朝诗歌主题数-coherence 折线图

表 3-5 各朝代最优主题数

朝代	最优主题数	朝代	最优主题数
先秦	8	元	4
秦	6	金	11
汉	1	明	18
三国	1	清	2
魏晋南北朝	3	民国	11
隋	7	当代	1
唐	8	近现代	1

五代十国	3	现代	17
宋	12	现当代	15

在得到以上结果后，开始构建算法模型。gensim.models.ldamodel 模块中已经将 LDA 模型封装好了，这里只需调用传参即可，参数列表见表 3-6。

表 3-6 LDA 模型参数

参数	说明	设置数值	备注
corpus	词袋	corpus	DT 矩阵
id2word	词语词典	dictionary	创建语料的词语词典
num_topics	主题个数	-	参照表 3-5
passes	模型训练次数	30	-
random_state	随机种子	1	-

第四步，运行模型并处理数据。由于模型最终运行结果为 Python 元组类型的数据，所以需要将最终数据也按照朝代进行划分存储。

3.4 本章小结

本章主要分为三小节对本系统所使用到的算法进行解释分析。首先对使用到的三个个算法进行概要性描述，接下来两节分别对于两个算法展开描述。诗歌情感计算小节首先介绍诗句情感分类，利用 Word2Vec 算法将诗句中的字、词与情感形成映射，然后得到诗句与关联字间的关联得分将诗句进行情感分类；紧接着介绍了利用深度学习的方法来对未识别出情感的诗句进行训练来得到对应的情感标签；最后对现有的情感标签进行扩充。诗词主题词提取算法小节则是从介绍原理开始介绍算法整个工作流程来描述整个算法。

第 4 章 诗酒数据可视分析系统设计

本系统共分为数据层、数据分析与处理、设计层以及交互层四层（如图 4-1），从数据的获取、数据处理、可视化的设计以及交互的设计上进行划分。数据分析与处理主要包含对原始数据的处理、分析和提取，然后对处理后的数据选择合适的可视化编码方式。可视化设计层主要针对不同类别的数据设计不同的可视化图表。由于本课题系统主要以文本类数据为主，所以需要依靠选择、亮度以及多视图协同作为基础交互来达到研究目标。另外，需要考虑到视图的联动性，所以要安排合理的视图布局。本章将按小节来详细介绍可视化数据层、数据分析与处理、可视化设计层、可视化交互层。

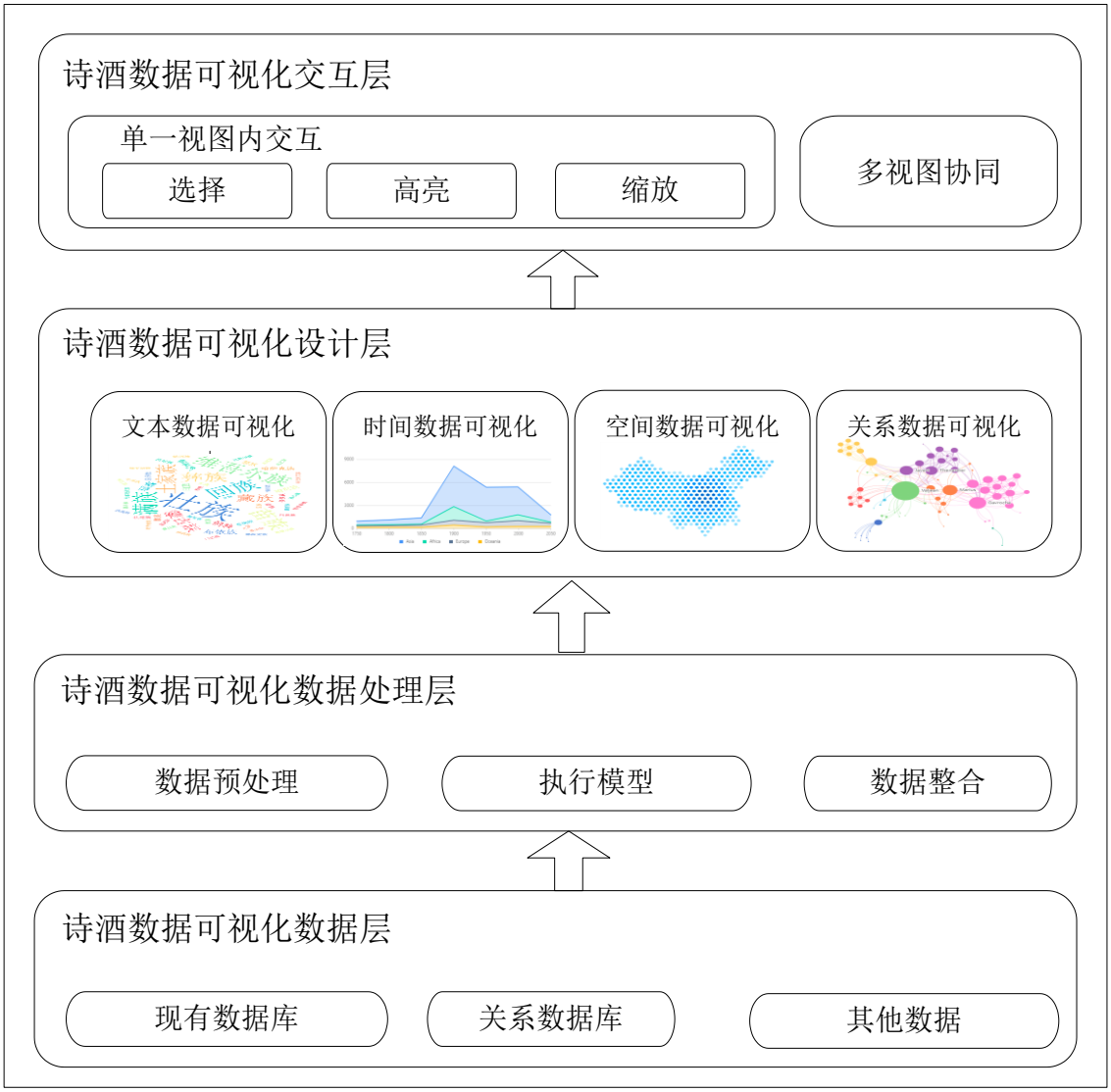

图 4-1 诗酒数据可视分析系统架构图

4.1 可视化数据层

整个系统的数据层包含三部分原始数据：现有数据库中的文本数据、诗人关系数据库和其他数据。现有数据库中的文本数据主要是李白、杜甫等唐代诗人的诗歌数据，数据维度较为完整，但是数据量较少。诗人关系数据库为“中国历代人物传记资料库”（China biographical database project , CBDB），是一个关系型数据库，其中统计搜集了从先秦至近现代大多数文人的生平、社会关系等数据[21]。对于该数据库，可用于提取诗人间的关系数据。其他数据是需要利用爬虫技术来进行数据的获取，获取诗歌的网站为：唐宋文学编年地图[22]。通过对于含“酒”的诗歌进行定向爬取来获得各个朝代的诗酒数据。

4.2 数据分析与处理

数据分析与处理过程需要将基础数据进行数据预处理，其次将预处理后的数据放入算法模型当中执行得到算法模型执行后的数据，最后对以上步骤处理后的数据进行整合以满足后续过程的需要。

4.2.1 数据预处理

由于基础数据分别是几部分数据共同组成的，所以在进行分析与处理的时候难免会遇到问题，比如：命名不统一、数据格式不统一等。所以需要对基础数据进行数据的初步处理。其中预处理步骤见图 4-2。

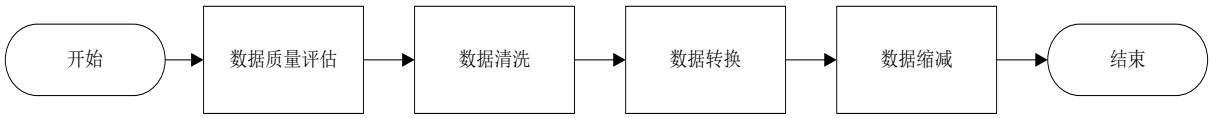

图 4-2 数据预处理步骤

首先需要对各部分数据进行一个评估，评估内容大致分为表 4-1 中几类。通过这一步能够大体了解整个数据集，确定数据与本课题的相关性和一致性。下一个阶段是数据清洗，这一步需要在代码编辑器 PyCharm 中编写 Python 代码来完成整个过程，其他详细内容见表 4-2，针对本数据除了需要进行一些基础处理之外还需要对朝代进行统一编码，比如：北宋、南宋统一归为宋朝，初唐、盛唐、唐末等统一归为唐朝等。第三步需要进行数据转换，具体转换规则见图 4-3，这里强调一点 Python 版本和 Pycharm 版本没有明确要求。最后一步需要进行数据缩减（数据分包），按不同标准进行数据分包缩小单个文件的大小。

表 4-1 数据预处理评估表

评估内容	详情
数据类型是否匹配	从许多不同的来源收集数据时，可能会以不同的格式呈现
是否存在混合数据值	不同的来源对特征使用不同的描述符，例如：唐或唐朝 。
是否存在异常值	—
是否存在缺失值	—

表 4-2 数据清洗

内容	备注
编程语言	Python
IDE 工具	PyCharm
目标	通过数据清洗集中更正、修复或删除不正确或不相关的数据
步骤	检测与处理重复值 检测与处理数据的缺失值 检测与处理数据的异常值 数据类别统一

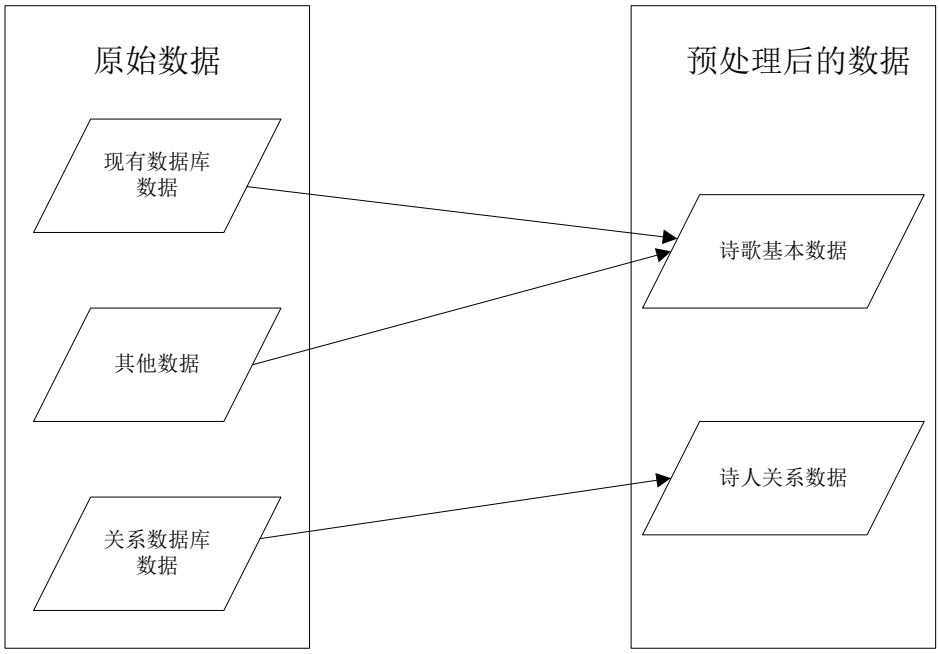

图 4-3 数据转换

4.2.2 执行模型

在完成数据预处理后需要将预处理后的数据分别执行第二章中所描述的诗词情感

分析算法和诗歌主题词提取算法模型。当然在模型的创建和执行过程中还需要针对不同的模型在进行更为细致的数据处理过程，详细内容见第 3 章 3.2.2、3.2.3 和 3.3.4 小节。通过执行模型就可以得到最后一部分数据：诗歌主题和情感数据。

4.2.3 数据整合

经过上述两步对于数据的处理，在原有数据的基础之上又新增了部分数据，所以需要将所有数据进行统一整合。数据整合后整个系统的数据层得到更新，其中包含三部分数据：诗歌基本信息数据、诗人关系数据、诗歌主题和情感数据，如图 4-4 所示。其中诗歌基本数据包括诗歌 ID、诗歌名称、朝代、作者等诗歌背景信息以及诗歌内容；诗人关系数据是通过对“中国历代人物传记资料库”（CDBD）中数据进行筛选过滤后得到的关系数据；而诗歌主题和情感数据包含由诗词情感分析算法得到的诗歌情感信息以及由诗歌主题词提取算法得到的诗歌主题信息，其中诗歌主题信息依据朝代进行划分统计，而诗歌情感则是针对每首诗歌进行统计。为了方便后续的操作和调用，数据均采用 csv 文件或 json 格式文件进行存储。

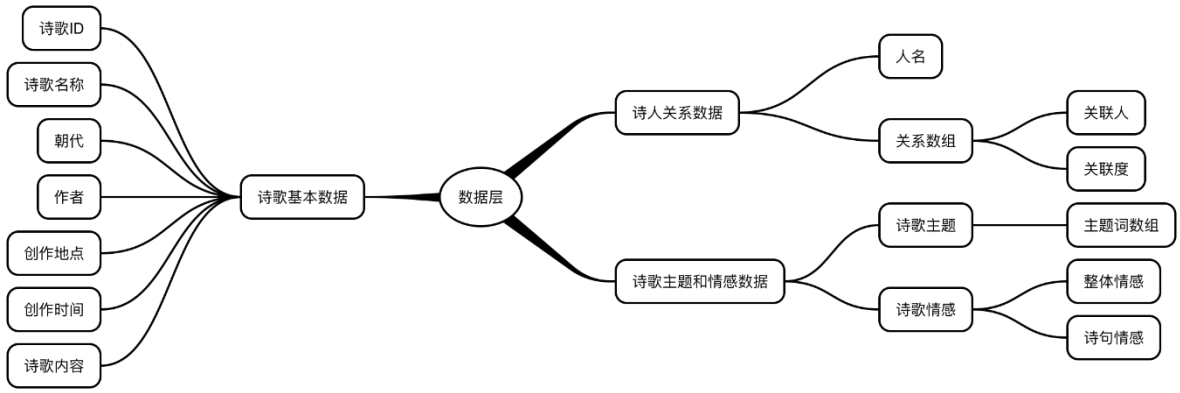

图 4-4 诗酒数据可视分析系统数据规划

4.3 可视化设计层

可视化设计层主要针对不同类别的数据设计不同的可视化图表，同时话需要针对整个可视化系统的布局进行设计。

4.3.1 可视化编码

通过 4.2 小节对于数据的处理，我们按照数据类型可以将数据分为 5 大类（见表 4-3）：文本型数据、时序型数据、空间型数据、关系型数据、其他类型数据。其中，文

本类型数据主要是诗歌内容文本、诗歌的主题以及诗歌中每一句的情感和整首诗的情感，这一部分数据可以使用颜色编码来表示情感类别或者使用词云的形式来呈现出不同情感类别以及占比；由于诗歌可以从时间维度按照不同朝代进行划分，具有较强的时序性特点，针对这一类数据可以采用带时间轴的河流图或者面积图来多方位展示时间和情感数据；同样诗歌从空间维度上看其创作地点也大都不同，这也就形成了诗歌创作地点在空间上的分布，此时就可以使用地图来对这一类数据进行可视化呈现；关系型数据在本系统中特指的是诗人间的关系数据，同一时期的诗人间或多或少或会存在着交际，通常情况下可以采用力导向图来对数据呈现；其他类型数据可以使用柱状图、条形图等来做为统计分析，使用饼图来显示情感占比等。

表 4-3 可视化编码分类表

数据类型	数据项	可视化编码
文本型数据	诗歌内容、诗歌主题词、诗歌情感	颜色编码、词云
时序型数据	按朝代划分的诗歌数据	带时间轴的面积图、河流图
空间型数据	按创作地点划分的诗歌数据	地图
关系型数据	诗人关系数据	力导向图
其他类型数据	统计类数据	柱状图——统计分析 饼图——情感占比

4.3.2 可视化布局设计

通过上一小节对于不同类别的数据进行可视化编码设计可以得到本系统可视化图表的数目大约在 7 个左右。所以为了符合视觉设计，本系统将使用左右双主图布局（见图 4-5）。在图 4-5 中 1 部分用于展示柱状图，表示按地点对诗歌数量进行统计排名；2 部分设计为诗歌列表，呈现不同朝代或者地点的诗歌列表；3 部分用于呈现诗歌文本信息，其中包括诗歌名称、作者、朝代、内容等；4，5 为两个主视图用于展示地图和力导向图，地图用于呈现诗歌情感在空间上的分布，力导向图用于展示诗人的交际关系；6 部分用于呈现词云，呈现不同朝代诗歌创作的主题词数；7 部分为带时间轴的面积图，主要用于呈现诗歌情感在时间上的分布。

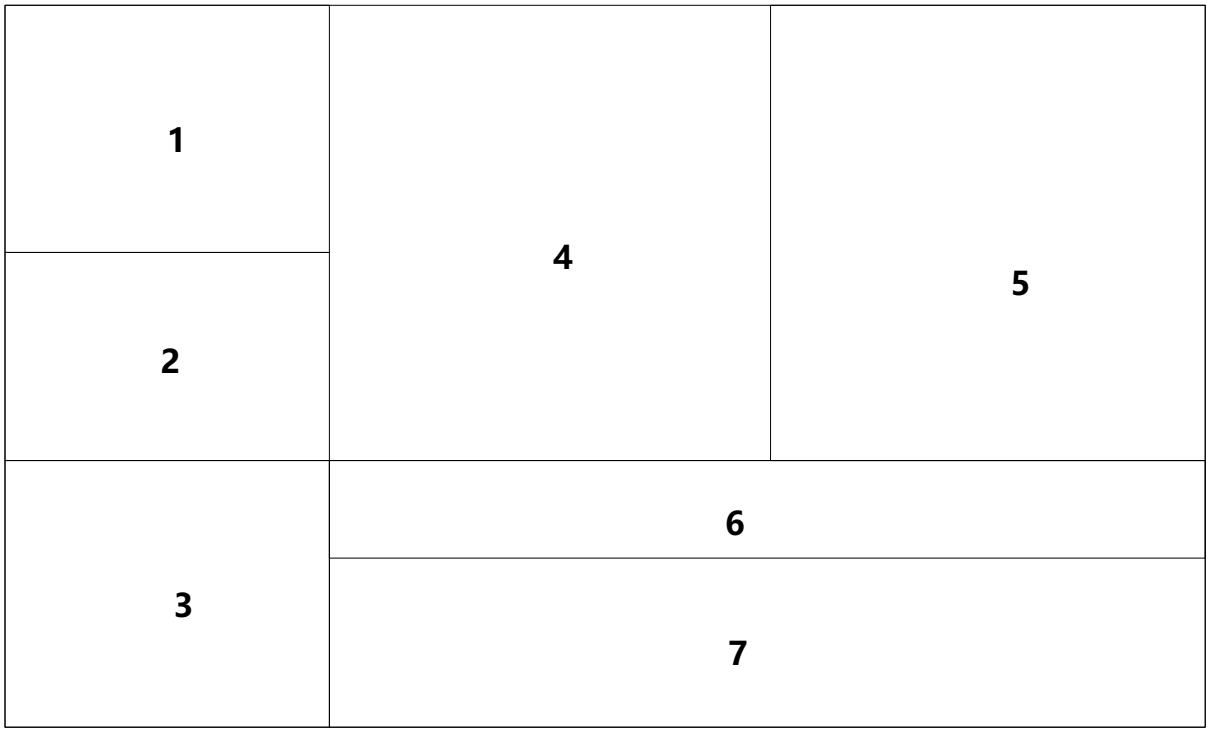

图 4-5 诗酒数据可视分析系统可视化布局设计

4.4 可视化交互层

本系统主要以文本类数据为主，在将数据以不同类别的可视化图表呈现后还需要利用交互的手段来使得整个系统更加的生动，拥有更强的可扩展性。系统的交互设计主要分为单一视图内的交互以及多视图间的联动交互。

4.4.1 单一视图内交互

单一视图内的交互主要分为鼠标的点击事件、滚轴缩放事件、选定后图表的高亮显示、悬停凸显等交互操作，详情见表 4-4。设置这些一方面是为了可视化图表的可展示性更强，另一方面也为了让图表的目的性更强。

表 4-4 单一视图内交互与图表对照表

交互操作	涉及到的可视化图表
鼠标点击事件	地图、诗人关系力导向图、主题词云、面积图
鼠标滚轴缩放	地图、诗人关系力导向图
高亮显示	诗人关系力导向图、主题词云、面积图
鼠标悬停	统计柱状图、主题词云、面积图

4.4.2 多视图联动

另外、在本系统中多视图间的交互联动主要体现在两方面：时间维度和空间维度，通过时间轴选择朝代，通过地图选择地点两者均可用于整个系统视图间的联动，依据选择更新展示不同朝代或者不同地点的数据，进而满足用于对于诗酒数据的探索需求。

4.5 本章小结

本章主要从四个方面来对诗酒数据可视分析系统设计进行解释说明。整个系统共分为四层：数据层、数据处理与分析层、设计层、交互层。数据层作为系统最底层由三部分数据内容共同组成，同时该层也为后续可视化图表设计做铺垫；数据分析与处理包括两部分：执行算法和数据处理，首先需要数据层的数据进行一次初步的整合得到运行算法前的算法基础数据，然后将这些数据分别执行对应的算法得到诗酒情感数据和诗酒主题数据，最后针对算法运行完成后的数据分别使用相对应的处理方法为后续可视化图表创建基础数据；可视化设计按照系统需求和数据特点设计不同的可视化图表，同时对多样的可视化图表进行系统的布局设置；最后为了满足用户的交互需求等设计了单一图表内部的交互和多视图间的联动。

第 5 章 诗酒数据可视分析系统详细设计与实现

基于第二章和第四章对于整个系统的需求分析和可视化设计，本章将对于诗酒数据可视分析系统进行详细介绍。系统主界面见图 5-1，主要包括以下几个部分：诗词情感空间分布（图 5-1 A、D）、诗词情感的时间分布（图 5-1 G）、诗人关系可视分析（图 5-1 E）、酒诗词主题展示（图 5-1 F）、诗歌内容展示（图 5-1 B、C）。

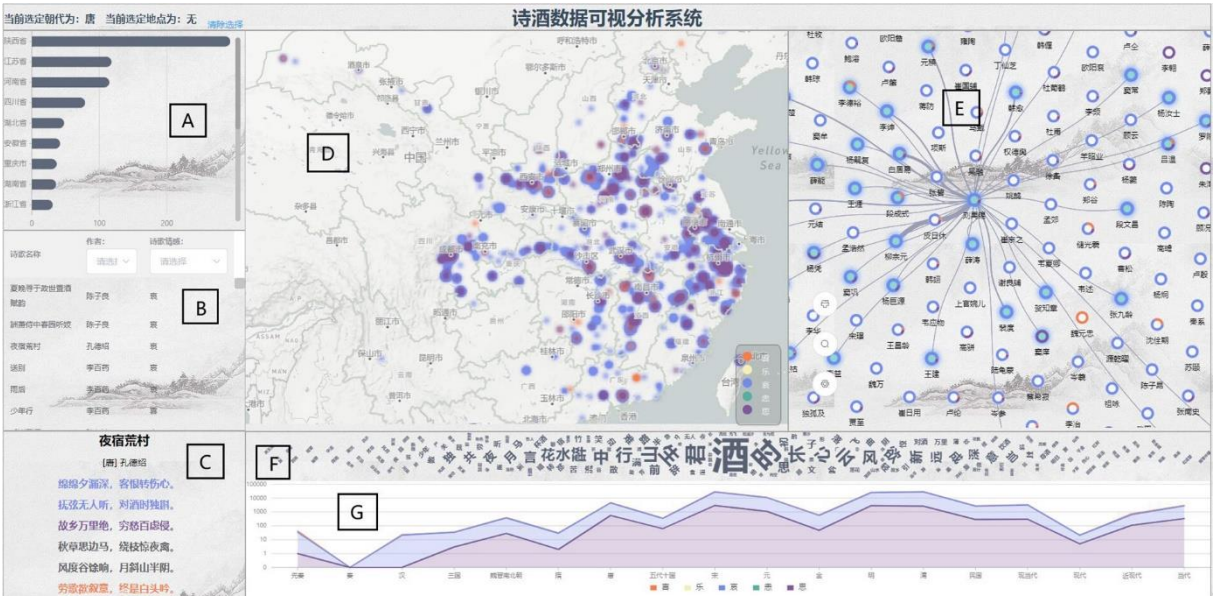

图 5-1 诗酒数据可视分析系统主界面

5.1 酒诗词情感空间分布

由于酒诗词的创作存在创作地点这一维度特点，所以我们将酒诗词情感对应酒诗词的创作地点以热力图的方式将其呈现在地图上（图 5-1 D 区域），同时使用柱状图对地点诗作数进行统计展示（图 5-1 A 区域）。

图 5-1 中 D 区域地图用于呈现酒诗词情感在空间上的分布，地图采用的是中华人民共和国地图。针对不同类别的酒诗词情感使用不同的颜色编码进行区分，然后利用热力图进行呈现。同时可以通过点击右下角图例来切换不同情感的热力图层（如图 5-2）。另外，本系统使用柱状图来对诗作数量在空间分布进行排名和统计（如图 5-1 A），默认排序方式为倒序。该视图还可以通过切换不同朝代来实现数据的更新，同时支持点击左侧省份名称实现地点选择（如图 5-3）。

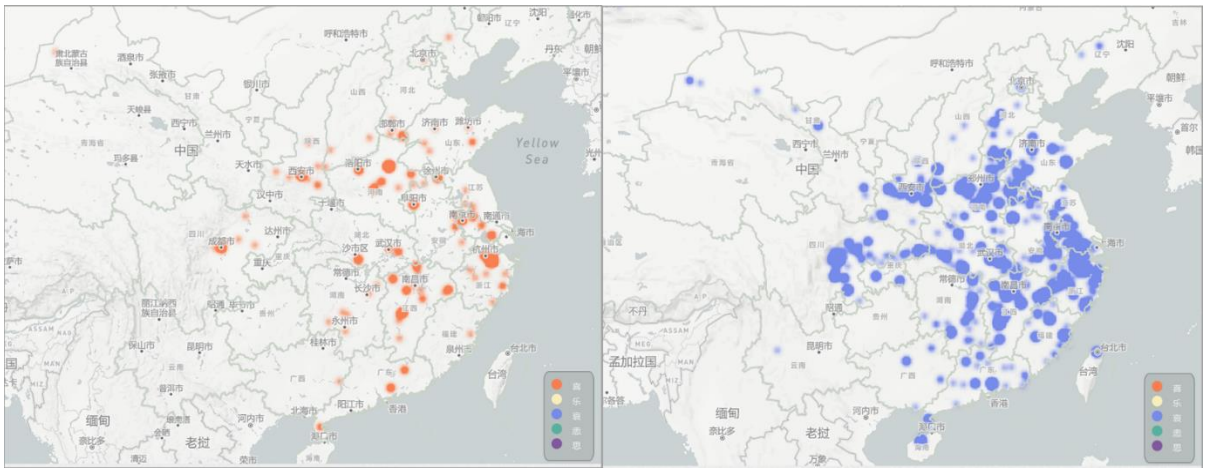

图 5-2 热力图展示酒诗歌情感为“喜”（左）“哀”（右）

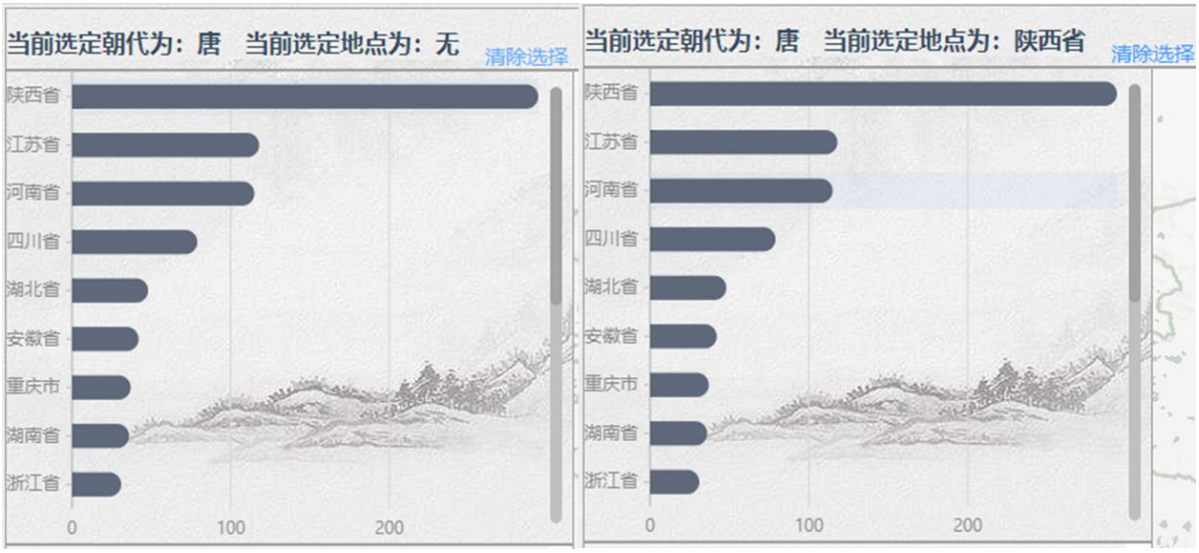

图 5-3 唐朝未选择地点（左）和陕西省（右）诗作数量在空间分布的排名柱状图

5.2 酒诗词内容展示

如图 5-1 中 B、C 两部分均用于呈现诗歌具体内容。B 部分为诗词列表，列表总共分为三栏：诗歌名称、作者、诗歌情感，同时针对作者和诗歌情感两栏还设置了下拉列表用于指定作者和诗歌情感如图 5-4 所示，另外还可以通过点击选择诗作与 C 部分形成交互。C 部分用于呈现诗词具体内容其中包括：诗（词）名、朝代、作者、内容等，其中针对每一句诗歌都将其表达出来的情感通过颜色编码反映在诗句上，如图 5-5 所示，通过颜色编码能够更加直观的将整首诗的情感呈现出来。

图 5-4 诗词列表作者筛选（左）诗歌情感筛选（右）

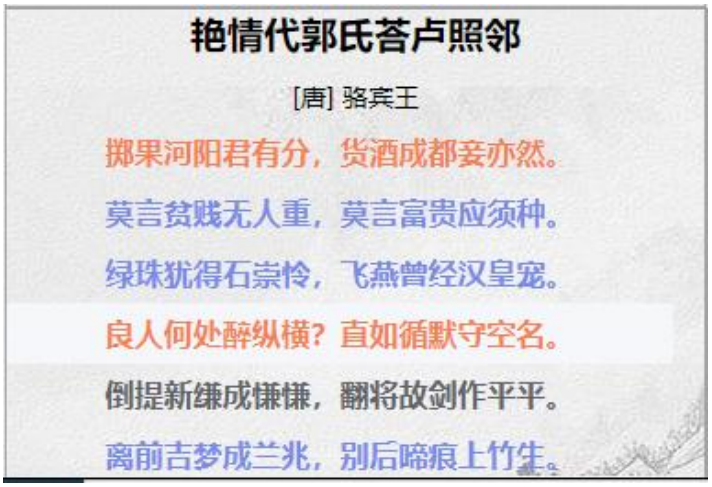

图 5-5 诗词内容颜色编码呈现

5.3 诗人关系可视分析

如图 5-1 中 E 部分所示。首先观察节点，图中每一个节点均表示一个诗人，对于每一个节点都使用环图来表示（图 5-6），环图中的展现的数据是：按照酒诗词情感对该诗人创作的所有诗词进行分类统计，其中酒诗词情感的颜色编码与全局的颜色编码统一。然后我们来观察边，如果两个诗人间有联系，那么两节点之间就会使用一条三次贝塞尔曲线进行连接；反之如果该诗人与其他任何诗人都没有联系，那么就是一个独立的节点，没有任何连线。同时，诗人间联系的密切程度使用曲线的粗细来进行表示，曲线越粗就代表两者关系越密切。由于视图窗口大小限制不能够展示所有节点，所以整个视图窗口支持鼠标拖动平移事件来满足其他节点观察需求。另外针对每一个

节点都添加了鼠标点击效果，该点击效果会高亮与其相连的所有节点和边，同时隐藏其他无关的边（图 5-7）。

图 5-6 诗人关系力导向图节点详情

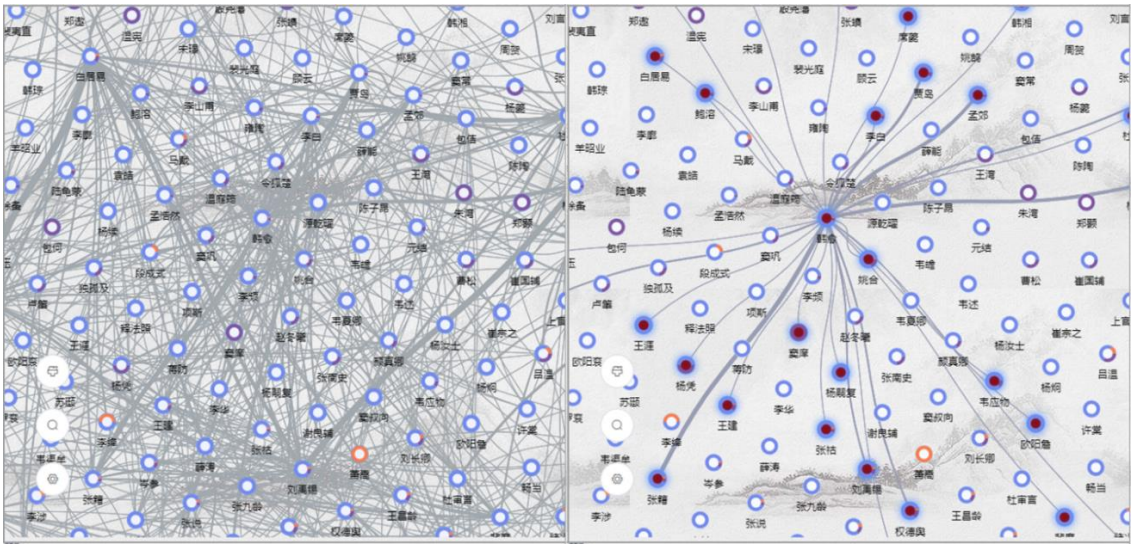

图 5-7 诗人关系力导向图（左）节点点击效果（右）

由于诗人节点数量较多，为了便于观察和分析该视图还添加了鱼眼放大镜功能。该功能可以通过右下角三个按钮进行控制（图 5-8），包括以下几个功能：清除现存的放大镜、开关按钮、鱼眼放大镜的控制模式（鼠标移动、鼠标拖动、点击）。当且仅当鱼眼放大镜功能开启后以下三个按钮才会起作用。

图 5-8 鱼眼放大镜控制按钮

当开启鱼眼放大镜功能后，力导向图中所有文字信息将会消失，当放大镜移动至

对应位置的时候才会进行显示（图 5-9）。

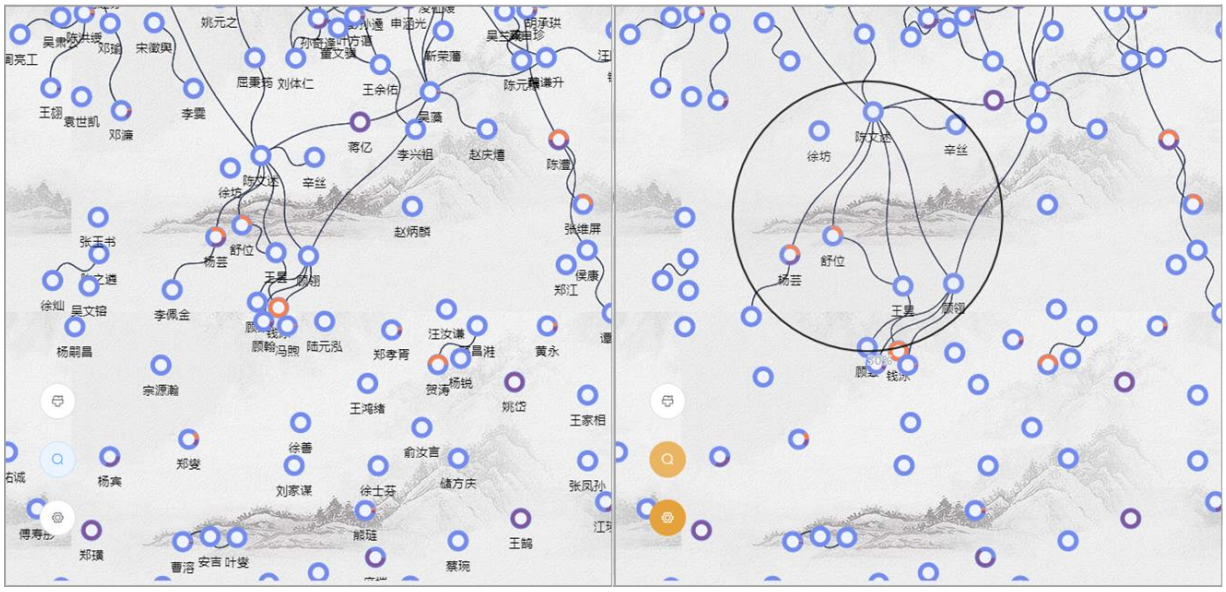

图 5-9 未开启（左）和开启（右）鱼眼放大镜

5.4 酒诗词主题展示

如图 5-1 中 F 部分，使用词云来对各个朝代诗歌的主题词进行展示，如果该主题词出现的次数越多那么其在词云中呈现的字体效果就越明显。另外可以通过切换不同的朝代来达到数据的更新操作。

5.5 酒诗词情感时间分布

如图 5-1 中 G 部分是一个带有时间轴的百分比面积图。分别用于呈现从先秦至当代 18 个时间段的所有酒诗词数据不同情感类别随时间变化的趋势。同时可以通过点击下方图例来单独观察某一种或几种情感的变化趋势（图 5-10）。另外，可以通过点击时间轴对应的朝代来做到全局的交互。

图 5-10 “喜”（上）情感和“哀”（下）情感的变化趋势

第 6 章 案例分析与系统评估

本章节将展示两个实际使用情况：诗酒数据时空态势分析、李白生平及其诗歌情感分析进行案例分析，同时将分析结果作为系统评估的结果。

6.1 案例一：宋朝诗酒数据情感和主题分析

某古文学研究专家想要通过本系统来分析宋朝时期酒诗词中情感的时间和空间分布的特点。首先专家可以通过点击下方时间轴，选择朝代为宋朝，此时系统中各模块更新为宋朝酒诗词数据。通过观察诗作数量在空间分布的排名柱状图（图 6-1）可以发现在浙江省创作的诗歌数量居多其次是河南省和江西省。结合历史我们不难发现宋朝都城定都在了两个地方：北宋汴梁和南宋临安，也就是今河南省开封市和浙江省杭州市，那江西省的排名为什么也那么靠前呢？这就不得不提到宋朝经济繁荣带来的教育事业的需求与发展。在当时由于江西粮食生产位居全国第一，使得当地人民不再因为经济而去担心温饱等问题，此时人民就会花更多的精力来满足仕途和精神文化方面的需求。

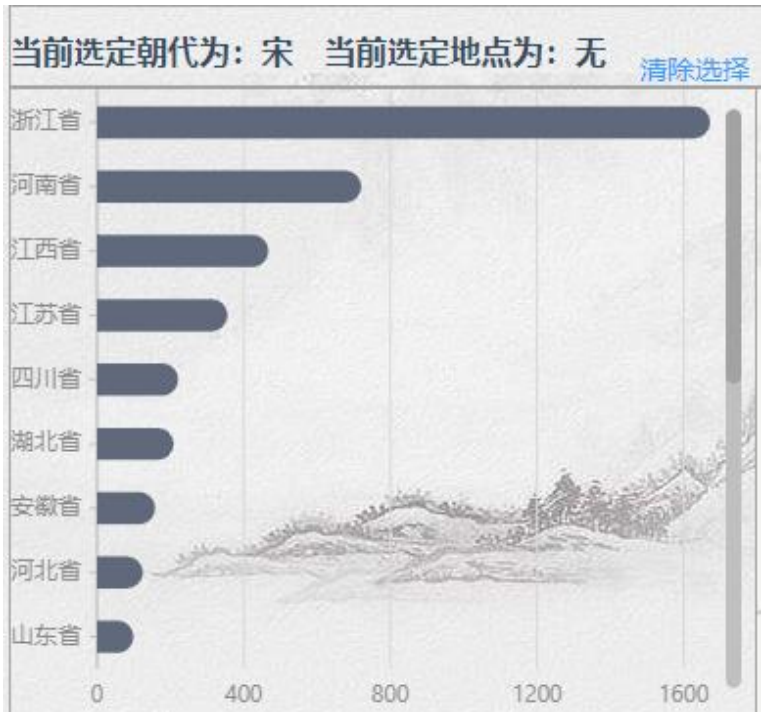

图 6-1 宋朝诗作数量在空间分布的排名柱状图

紧接着我们通过观察地图同样可以对上述分析进行论证，可以通过切换热力图层来分析一下具体的情感分布情况如图 6-2 所示。切换热力图层为喜（图 6-2-左）和哀

（图 6-2-右）可以得出两点结论：第一诗词创作和情感在空间上成聚集性分布；第二情感中哀伤大于喜乐；由于大的历史背景，虽然宋朝相较于五代十国时期的战事混乱有了一定的统一和发展，但是整个领土并没有收复完整，北面仍遭受着外敌的侵略和骚扰，同时在政治方面一直推崇革新，这就导致了宋朝时期的政党之争，直接性的反映在人民的生活当中。

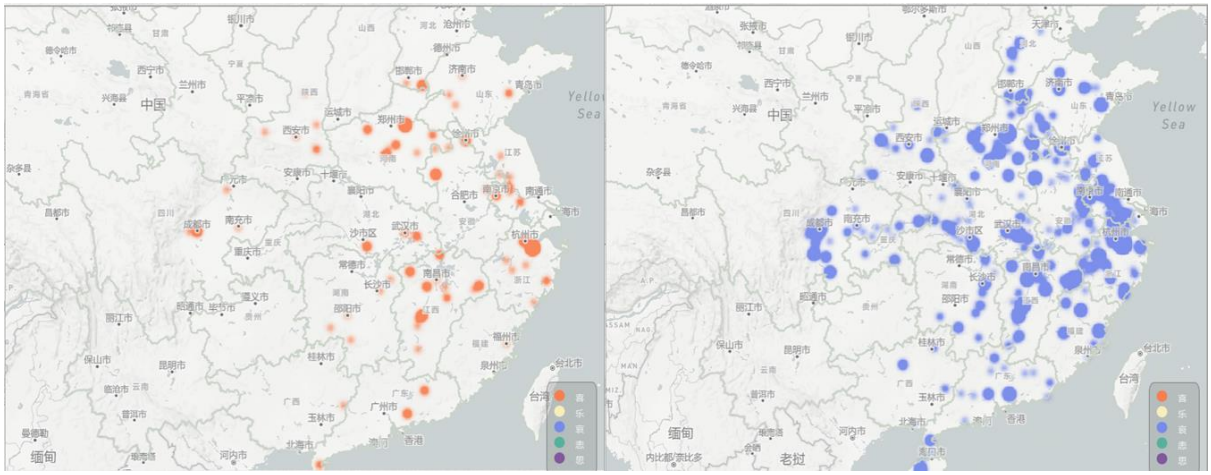

图 6-2 宋朝酒诗词喜（左）和哀（右）情感空间分布

最后我们观察分析主题词云和面积图（图 6-3）。从主题词词云中我们可以发现：“酒”作为主题词占比较高这是毋庸置疑的，除开“酒”字以外，“时”、“生”、“知”、“春”等字出现频率也比较高，通常情况下这些字会出现在以忧国伤时、感时伤逝等为主题思想情感的诗词中。所以从这里不免能看出宋朝战事等对诗人生活状态的影响。同时我们可以直观的从面积图中可以看出“哀”这种情感占了绝大部分，而喜则很少。

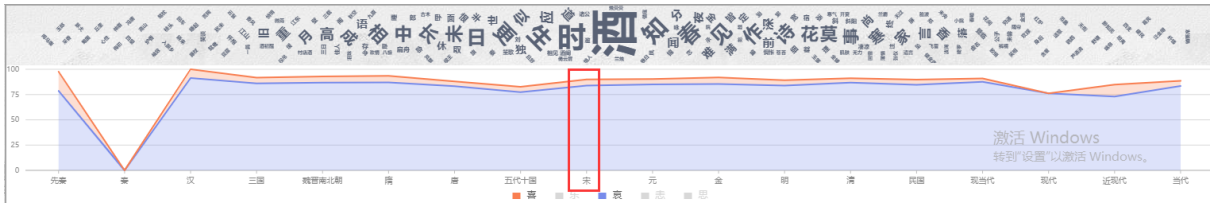

图 6-3 宋朝诗酒数据主题词词云（上）面积图（下）

6.2 案例二：李白社交网络及其诗歌情感分析

首先选择朝代为唐朝，此时所有视图发生更新。接下来在诗人关系力导向图图中找到诗人李白并点击视图发生更新（图 6-4-左），与李白有联系的诗人均被高亮显示。我们通过观察可以发现李白与杜甫、贾至、魏万之间交际较为频繁。开启视图放大镜功能，通过观察李白节点可以发现李白饮酒做诗往往表达出自己的哀伤之情，包括月

下独酌的孤寂忧愁、借酒消愁等。

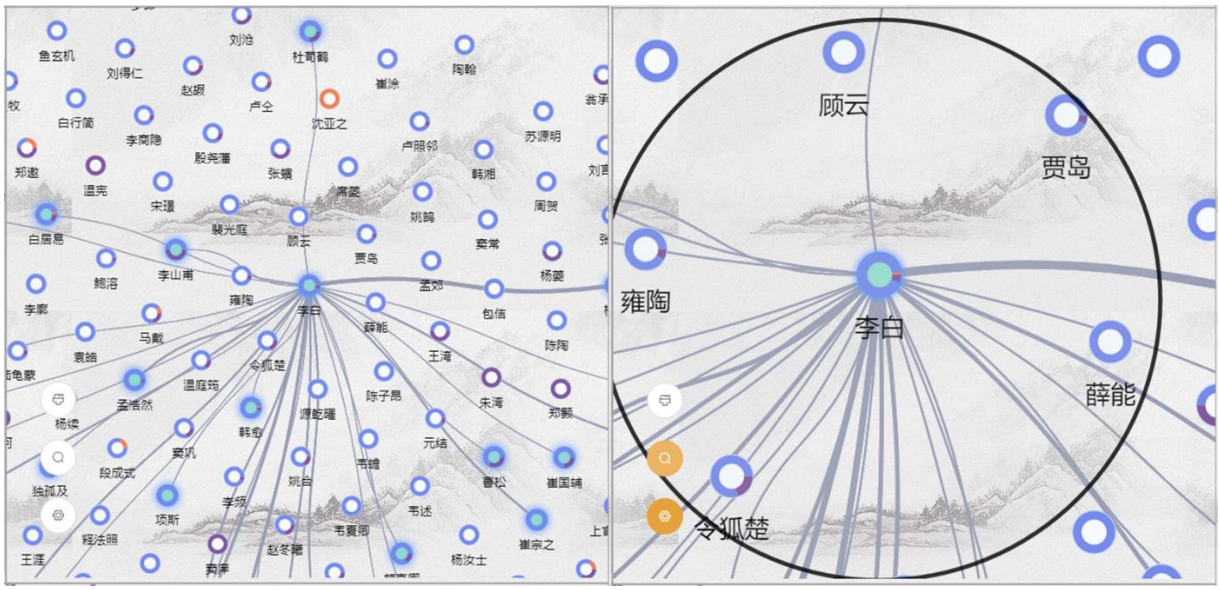

图 6-4 李白人际关系网络（左）和酒诗词情感占比（右）

针对李白酒诗词中情感的分析，我们将《鼓吹曲辞·将进酒》这首诗拿出来举例分析。首先在诗词列表中作者一栏下拉菜单中选择李白，诗词列表展示所有李白的酒诗词数据，然后找到《鼓吹曲辞·将进酒》这首诗，从诗词列表中可以直观看出这首诗的情感为“哀”（图 6-5）。紧接着我们来具体分析这首诗，点击诗词列表可以得到诗作的具体内容（图 6-6）。诗歌中每一句都使用颜色编码来呈现诗句蕴含的情感，通过颜色编码可以更加直观的了解整首诗的情感变化情况。

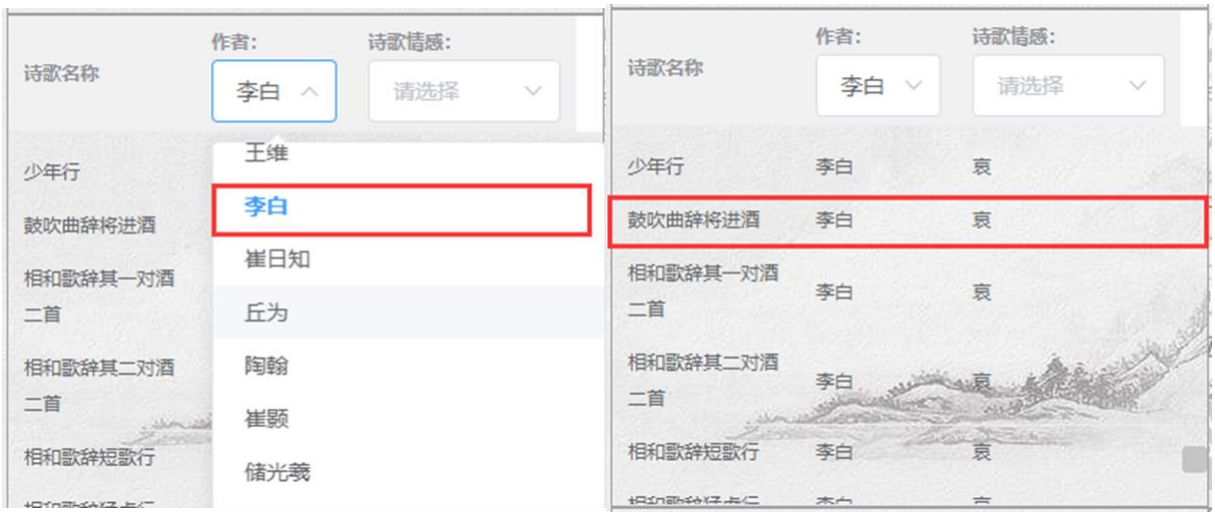

图 6-5 李白酒诗词列表

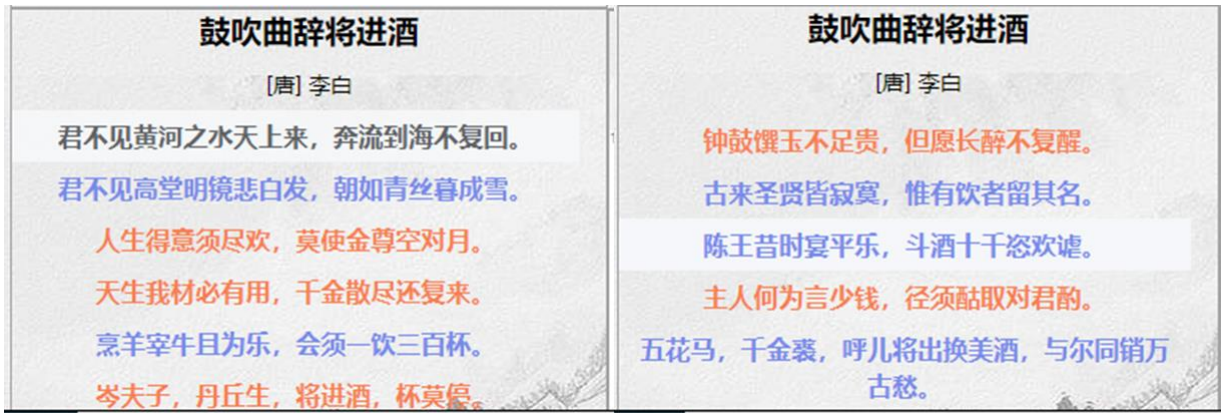

图 6-6 《鼓吹曲辞·将进酒》具体内容

第 7 章 诗酒数据可视分析系统测试

系统测试总共分为两部分：系统功能测试和系统性能测试，本章将从以上两个方面展开。

7.1 系统功能测试

本系统的功能测试将分为前端功能测试和后端功能测试。

7.1.1 后端功能测试

本系统后端主要根据前端请求发送数据。按照数据请求的不同分为六类，其对应的测试内容为六类：诗作数统计数据发送功能测试（表 7-1）、主题词云数据发送功能测试（表 7-2）、诗歌内容数据发送功能测试（表 7-3）、诗人关系数据发送功能测试（表 7-4）、诗词地点数据发送功能测试（表 7-5）、诗词时间数据发送功能测试（表 7-6）。

表 7-1 诗作数统计数据发送功能测试

用例编码	T-001			
功能描述	发送诗作数统计数据			
用例目的	测试诗作数统计数据发送功能			
用例前提	前端诗作数统计柱状图发送数据请求			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-001-001	前端诗作数统计柱状图发送数据请求	发送诗作数统计数据	发送诗作数统计数据	正确

表 7-2 主题词云数据发送功能测试

用例编码	T-002			
功能描述	发送主题词云数据			
用例目的	测试主题词云数据发送功能			
用例前提	前端主题词云向后端发送数据请求			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-002-001	前端主题词云发送数据请求	发送主题词云数据	发送主题词云数据	正确

表 7-3 诗歌内容数据发送功能测试

用例编码	T-003

功能描述	发送诗歌内容数据			
用例目的	测试诗歌内容数据发送功能			
用例前提	前端诗歌内容模块发送数据请求			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-003-001	前端诗歌内容模块发送数据请求	发送诗歌内容数据	发送诗歌内容数据	正确

表 7-4 诗人关系数据发送功能测试

用例编码	T-004			
功能描述	发送诗人关系数据			
用例目的	测试诗人关系数据发送功能			
用例前提	前端诗人关系力导向图发送数据请求			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-004-001	前端诗人关系力导向图发送数据请求	发送诗人关系数据	发送诗人关系数据	正确

表 7-5 诗词地点数据发送功能测试

用例编码	T-005			
功能描述	发送诗词地点数据			
用例目的	测试诗词地点数据发送功能			
用例前提	前端地图模块发送数据请求			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-005-001	前端地图模块发送数据请求	发送诗词地点数据	发送诗词地点数据	正确

表 7-6 诗作数统计数据发送功能测试

用例编码	T-006			
功能描述	发送诗词时间数据			
用例目的	测试诗词时间数据发送功能			
用例前提	前端带时间轴的面积图模块发送数据请求			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-006-001	前端带时间轴的面积图模块发送数据请求	发送诗词时间数据	发送诗词时间数据	正确

7.1.2 前端功能测试

前端功能测试按照图表模块进行。首先测试诗作数统计柱状图查看功能（表 7-7）、诗作数统计柱状图滚动功能（表 7-8）、诗作数统计柱状图坐标轴点击功能（表 7-9）。

表 7-7 诗作数统计柱状图查看功能测试

用例编码	T-007			
功能描述	查看诗作数统计柱状图			
用例目的	测试诗作数统计柱状图的查看功能			
用例前提	收到后端发送诗作数统计数据			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-007-001	查看诗作数统计柱状图	显示按创作地点诗作数降序排列的柱状图	显示按创作地点诗作数降序排列的柱状图	正确

表 7-8 诗作数统计柱状图滚动功能测试

用例编码	T-008			
功能描述	滚动诗作数统计柱状图			
用例目的	测试诗作数统计柱状图滚动功能			
用例前提	查看诗作数统计柱状图			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-008-001	向下滚动诗作数统计柱状图	显示未展示出来的图表项	显示未展示出来的图表项	正确
T-008-002	向上滚动诗作数统计柱状图	回到图表初始位置	回到图表初始位置	正确

表 7-9 诗作数统计柱状图坐标轴点击功能测试

用例编码	T-009			
功能描述	点击诗作数统计柱状图坐标轴			
用例目的	测试点击诗作数统计柱状图坐标轴功能			
用例前提	查看诗作数统计柱状图			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-009-001	点击坐标轴上的浙江省	系统地点选择变为浙江省	系统地点选择变为浙江省	正确
T-009-002	点击坐标轴上的河南省	系统地点选择变为河南省	系统地点选择变为河南省	正确
T-009-003	点击坐标轴上的陕西省	系统地点选择变为陕西省	系统地点选择变为陕西省	正确

其次，对诗歌列表查看功能进行测试（表 7-10）、诗歌列表诗人筛选功能测试（表 7-11）、诗歌列表诗歌情感筛选功能测试（表 7-12）、诗歌列表诗人、情感同时筛选功能测试（表 7-13）。

表 7-10 诗歌列表查看功能测试

用例编码	T-010
功能描述	查看诗歌列表

用例目的	测试诗歌列表的查看功能			
用例前提	后端发送诗歌内容数据			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-010-001	查看诗歌列表	显示当前朝代所有诗歌的列表	显示当前朝代所有诗歌的列表	正确

表 7-11 诗歌列表诗人筛选功能测试

用例编码	T-011			
功能描述	通过筛选点击诗人姓名，呈现该诗人的诗歌列表			
用例目的	测试诗歌列表诗人筛选功能			
用例前提	查看诗歌列表			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-011-001	筛选诗人：李白	呈现李白的诗歌列表	呈现李白的诗歌列表	正确
T-011-002	筛选诗人：苏轼	呈现苏轼的诗歌列表	呈现苏轼的诗歌列表	正确
T-011-003	筛选诗人：陶潜	呈现陶潜的诗歌列表	呈现陶潜的诗歌列表	正确

表 7-12 诗歌列表诗歌情感筛选功能测试

用例编码	T-012			
功能描述	通过筛选点击诗歌情感，呈现该诗歌情感的诗歌列表			
用例目的	测试诗歌列表诗歌情感筛选功能			
用例前提	查看诗歌列表			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-012-001	筛选诗歌情感喜	呈现诗歌情感为喜诗歌列表	呈现诗歌情感为喜的诗歌列表	正确
T-012-002	筛选诗歌情感哀	呈现诗歌情感为哀诗歌列表	呈现诗歌情感为哀的诗歌列表	正确
T-012-003	筛选诗歌情感思	呈现诗歌情感为思诗歌列表	呈现诗歌情感为思的诗歌列表	正确

表 7-13 诗歌列表诗人、情感同时筛选功能测试

用例编码	T-013			
功能描述	通过同时筛选诗人和诗歌情感，呈现该筛选条件下的诗歌列表			
用例目的	测试诗歌列表诗人、情感同时筛选功能			
用例前提	查看诗歌列表			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-013-001	筛选诗人：李白 筛选诗歌情感：喜	呈现李白的诗歌中诗歌情感为喜的诗歌列表	呈现李白的诗歌中诗歌情感为喜的诗歌列表	正确
T-013-002	筛选诗人：李白	呈现李白的诗歌中诗歌情	呈现李白的诗歌中诗歌情	正确

T-013-003	筛选诗歌情感：思 筛选诗人：苏轼 筛选诗歌情感：喜	感为思的诗歌列表 呈现苏轼的诗歌中诗歌情 感为喜的诗歌列表	感为思的诗歌列表 呈现苏轼的诗歌中诗歌情 感为喜的诗歌列表	正确

再次，对诗歌内容模块查看功能进行测试（表 7-14）。

表 7-14 诗歌内容模块查看功能测试

用例编码	T-014			
功能描述	查看诗歌内容模块			
用例目的	测试诗歌内容模块的查看功能			
用例前提	后端发送诗歌内容数据			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-014-001	查看诗歌内容模块	显示诗歌具体内容和诗句情感对应的颜色编码	显示诗歌具体内容和诗句情感对应的颜色编码	正确

然后，对地图模块进行测试。地图查看功能测试（表 7-15）、地图热力图层切换功能测试（表 7-16）。

表 7-15 地图查看功能测试

用例编码	T-015			
功能描述	查看地图			
用例目的	测试地图的查看功能			
用例前提	后端发送诗歌地点数据			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-015-001	查看地图	显示地图轮廓和热力图层	显示地图轮廓和热力图层	正确

表 7-16 地图热力图层切换功能测试

用例编码	T-016			
功能描述	点击图例，切换成对应的热力图层			
用例目的	测试地图热力图层切换功能			
用例前提	查看地图功能			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-016-001	点击图例中“喜”	显示“喜”热力图层	显示“喜”热力图层	正确
T-016-002	点击图例中“哀”	显示“哀”热力图层	显示“哀”热力图层	正确
T-016-003	点击图例中“思”	显示“思”热力图层	显示“思”热力图层	正确

接下来，测试诗人关系力导向图查看功能（表 7-17）、测试诗人关系力导向图拖拽

功能（表 7-18）、测试诗人关系力导向图缩放功能（表 7-19）、测试诗人关系力导向图节点点击功能（表 7-20）、测试诗人关系力导向图放大镜功能（表 7-21）。

表 7-17 诗人关系力导向图查看功能测试

用例编码	T-017			
功能描述	查看诗人关系力导向图			
用例目的	测试诗人关系力导向图的查看功能			
用例前提	后端发送诗人关系数据			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-017-001	查看诗人关系力导向图	显示诗人关系力导向图	显示诗人关系力导向图	正确

表 7-18 诗人关系力导向图拖拽功能测试

用例编码	T-018			
功能描述	拖拽诗人关系力导向图			
用例目的	测试诗人关系力导向图拖拽功能			
用例前提	查看诗人关系力导向图			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-018-001	向左拖拽力导向图	显示左侧未展示出的节点边	显示左侧未展示出的节点边	正确
T-018-002	向右拖拽力导向图	显示右侧未展示出的节点边	显示右侧未展示出的节点边	正确
T-018-003	向上拖拽力导向图	显示上侧未展示出的节点边	显示上侧未展示出的节点边	正确
T-018-004	向下拖拽力导向图	显示下侧未展示出的节点边	显示下侧未展示出的节点边	正确

表 7-19 诗人关系力导向图缩放功能测试

用例编码	T-019			
功能描述	缩放诗人关系力导向图			
用例目的	测试诗人关系力导向图缩放功能			
用例前提	查看诗人关系力导向图			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-019-001	缩小诗人关系导向图	诗人关系力导向图缩小	诗人关系力导向图缩小	正确
T-019-002	放大诗人关系导向图	诗人关系力导向图放大	诗人关系力导向图放大	正确

表 7-20 诗人关系力导向图节点点击功能测试

用例编码	T-020
功能描述	点击诗人关系力导向图中节点，该节点高亮，显示其连接的边，同时隐藏其他边
用例目的	测试诗人关系力导向图节点点击功能

用例前提	查看诗人关系力导向图			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-020-001	点击诗人“李白”节点	高亮该节点，显示其连接的边，隐藏无关的边，高亮与其关联的节点	高亮该节点，显示其连接的边，隐藏无关的边，高亮与其关联的节点	正确
T-020-002	点击诗人“杜甫”节点	高亮该节点，显示其连接的边，隐藏无关的边，高亮与其关联的节点	高亮该节点，显示其连接的边，隐藏无关的边，高亮与其关联的节点	正确
T-020-003	点击诗人“苏轼”节点	高亮该节点，显示其连接的边，隐藏无关的边，高亮与其关联的节点	高亮该节点，显示其连接的边，隐藏无关的边，高亮与其关联的节点	正确

表 7-21 诗人关系力导向图放大镜功能测试

用例编码	T-021			
功能描述	点击诗人关系力导向图放大镜按钮开启（关闭）放大镜			
用例目的	测试诗人关系力导向图放大镜功能			
用例前提	查看诗人关系力导向图			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-021-001	点击放大镜按钮	显示放大镜	显示放大镜	正确
T-021-002	再次点击放大镜按钮	关闭放大镜	关闭放大镜	正确
T-021-003	点击切换放大镜模式	放大镜模式由 drag 切换为 mousemove	放大镜模式由 drag 切换为 mousemove	正确

然后，对主题词云模块进行测试。主题词云查看功能测试（表 7-22）、主题词云悬浮高亮功能测试（表 7-23）。

表 7-22 主题词云查看功能测试

用例编码	T-022			
功能描述	查看主题词云			
用例目的	测试主题词云的查看功能			
用例前提	后端发送主题词云数据			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-022-001	查看主题词云	显示主题词云	显示主题词云	正确

表 7-23 主题词云悬浮高亮功能测试

用例编码	T-023			
功能描述	鼠标悬停至主题词上，该主题词高亮显示			
用例目的	测试主题词云悬浮高亮功能			
用例前提	查看主题词云			
子用例编号	输入	期望输出	实际输出	状态
T-023-001	鼠标悬停至“酒”字	“酒”字高亮显示	“酒”字高亮显示	正确
T-023-002	鼠标悬停至“春”字	“春”字高亮显示	“春”字高亮显示	正确
T-023-003	鼠标悬停至“美”字	“美”字高亮显示	“美”字高亮显示	正确

接下来，对面积图进行功能测试。面积图查看功能测试（表 7-24）、面积图图例点击功能（表 7-25）。

表 7-24 面积图查看功能测试

用例编码	T-024			
功能描述	查看面积图			
用例目的	测试面积图的查看功能			
用例前提	后端发送诗歌时间数据			
子用例编号	输入	期望输出	实际输出	状态
T-024-001	查看面积图	显示面积图	显示面积图	正确

表 7-25 面积图图例点击功能测试

用例编码	T-025			
功能描述	点击面积图图例切换视图，呈现不同诗歌情感下的视图			
用例目的	测试面积图图例点击功能			
用例前提	查看面积图			
子用例编号	输入	期望输出	实际输出	状态
T-025-001	鼠标点击图例中“哀”	隐藏诗歌情感为“哀”的数据	隐藏诗歌情感为“哀”的数据	正确
T-025-002	再次点击图例中“哀”	显示诗歌情感为“哀”的数据	显示诗歌情感为“哀”的数据	正确
T-025-003	鼠标点击图例中“思”	隐藏诗歌情感为“思”的数据	隐藏诗歌情感为“思”的数据	正确

最后，对全局视图联动进行测试。时间维度联动功能测试（表 7-26）、空间维度联动功能测试（表 7-27）。

表 7-26 时间维度联动功能测试

用例编码	T-026			
功能描述	点击面积图坐标轴切换不同朝代，联动试图发生变化			
用例目的	测试时间维度联动功能			
用例前提	系统加载完毕			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-027-001	点击面积图坐标轴中“唐”	诗作统计柱状图、地图、诗人关系力导向图、主题词云、诗歌列表发生更新	诗作统计柱状图、地图、诗人关系力导向图、主题词云、诗歌列表发生更新	正确
T-027-002	点击面积图坐标轴中“宋”	诗作统计柱状图、地图、诗人关系力导向图、主题词云、诗歌列表发生更新	诗作统计柱状图、地图、诗人关系力导向图、主题词云、诗歌列表发生更新	正确
T-027-003	点击面积图坐标轴中“清”	诗作统计柱状图、地图、诗人关系力导向图、主题词云、诗歌列表发生更新	诗作统计柱状图、地图、诗人关系力导向图、主题词云、诗歌列表发生更新	正确

表 7-27 空间维度联动功能测试

用例编码	T-027			
功能描述	点击柱状图坐标轴切换不同朝代，联动试图发生变化			
用例目的	测试时间维度联动功能			
用例前提	系统加载完毕			
子用例编号	**输入**	**期望输出**	**实际输出**	**状态**
T-027-001	点击柱状图坐标轴中“陕西省”	诗歌列表更新	诗歌列表更新	正确
T-027-002	点击面积图坐标轴中“河南省”	诗歌列表更新	诗歌列表更新	正确
T-027-003	点击面积图坐标轴中“浙江省”	诗歌列表更新	诗歌列表更新	正确

7.2 系统性能测试

可视化系统的性能主要分为两部分：系统首屏加载时间（FCLT，First screen load time）和核心 Web 指标。

系统首屏加载时间指的是浏览器从输入 URL 地址到首屏内容加载渲染完成的时间，

其计算公式由式 7-1 表示，系统首屏加载测试表见表 7-28。

$$time = \left(performance.timing.domComplete - performance.timing.navigationStart \right) / 1000 \quad （式 7-1）$$

表 7-28 诗酒数据可视分析系统首屏加载测试表

测试样例	测试结果(秒/s)	是否通过
T-FCLT-1	4.001	通过
T-FCLT-2	6.897	通过
T-FCLT-3	5.440	通过
T-FCLT-4	5.361	通过
T-FCLT-5	2.293	通过
T-FCLT-6	3.96	通过

核心 web 指标终于用户体验的三个方面：加载性能（LCP，Largest Contentful Paint）、交互性（FID，First Input Delay）和视觉稳定性（CLS，Cumulative Layout Shift）。核心 web 指标的测试使用的是 Web Vitals Chrome 测试工具。

加载性能最大内容绘制，测量加载性能。为了提供良好的用户体验，LCP 应在页面首次开始加载后的 2.5 秒内发生（表 7-29）。

表 7-29 诗酒数据可视分析系统加载性能 LCP 测试表

测试样例	测试结果（秒/s）	是否通过
T-LCP-1	1.845	通过
T-LCP-2	1.731	通过
T-LCP-3	0.922	通过
T-LCP-4	1.632	通过
T-LCP-5	1.924	通过
T-LCP-6	1.932	通过

交互性首次输入延迟，测量交互性，为了提供良好的用户体验，页面的 FID 应为 100 毫秒或更短，由于本系统涉及到大体积数据的加载和图表渲染，所以针对这种情况允许更长的响应时间（表 7-30）。

表 7-30 诗酒数据可视分析系统交互性 FID 测试表

测试样例	测试结果（毫秒/ms）	是否通过
T-FID-1	14.000	通过

T-FID-2	17.500	通过
T-FID-3	26.700	通过
T-FID-4	30.400	通过
T-FID-5	18.400	通过
T-FID-6	25.300	通过

视觉稳定性累积布局偏移，测量视觉稳定性，为了提供良好的用户体验，页面的 CLS 应保持在 0.1 或更少（表 7-31）。

表 7-31 诗酒数据可视分析系统视觉稳定性 CLS 测试表

测试样例	测试结果	是否通过
T-CLS-1	0.006	通过
T-CLS-2	0.001	通过
T-CLS-3	0.003	通过
T-CLS-4	0.001	通过
T-CLS-5	0.007	通过
T-CLS-6	0.004	通过

结论

本文研究设计并实现了诗酒数据可视分析系统，通过使用文本情感分析方法和诗词主题情感提取算法对诗酒数据进行处理和分析，同时从时间空间两个维度对算法分析得到的结果和诗酒数据本身进行可视化编码，最后通过对可视化图表进行布局排版、设置丰富的用户交互来实现整个可视分析系统。

本系统可以满足不同类型人群使用。对于文学研究学者来说本系统是一个便捷的研究工具，可以在传统的研究基础之上结合使用本系统来使得研究结果更加准确；对于普通大众，本系统可以作为一个知识普及、了解文化知识的渠道；而对于外国友人来说，通过本系统可以直观了解诗词情感的变化，不需要使用太多的时间在理解诗词的意思上，反过来可以通过表达出来的感情再来推测诗句想要表达的意思。以上几点体现了本系统相较于传统方法更加便捷、直观、易懂等特点。

由于数据获取途径有限，本系统中使用到的诗酒数据其数量并没有做到全方位多层次的覆盖；其次，一首诗所表述的内容和情感具有主观性，不同人看来是有所不同的，所以系统中针对诗歌情感的表述仅为算法运行出的较为准确的结论，仅作为参考。

相信未来对于诗酒文化的研究会一直向前，将人工智能与机器学习的方法应用在对于传统文化的研究上也会有很大的作用，可以为专家学者的研究工作减轻负担。其次，使用可视化方法用于呈现数据也可以为诗酒数据的分析工作提供便捷。最后，通过可视分析系统的方式来对诗酒文化进行宣传也是一个不错的方式。

致谢

首先，我要感谢我的研究导师彭莉娟、王桂娟。两位老师在学术研究上对我很严格，同时她们也时常给我鼓励，在情绪不好时也会帮助我调节情绪，如果没有她们对我的严格要求和关系以及对整个过程每一步的专注参与，我的毕业设计（论文）也没有那么顺利的完成。然后我还很感谢实验室的学长（姐）和学弟（妹）们，在完成整个毕业设计的过程中，当我遇到问题时或者向他们需求一些意见的时候，他（她）们也给我提了很多的建议帮助我解决问题同时也让我的思路更加开阔了。如果没有指导老师和同学们的帮助我的毕业设计也不会这么顺利的完成。

转眼间大学四年也只剩下为数不多的几天了，再回首我认为我是一个非常幸运的人。在大一学年下半学期很幸运能通过实验室的考核加入虚拟现实与可视化实验室。在实验室里参与科研项目，同时作为队长带领实验室学弟、学妹们参加了各种挑战赛，并取得了很好的荣誉。从第一次站在实验室前面进行学术汇报，到跨实验室汇报，再到国家级比赛上台答辩，自己的成长也是一步一个脚印踏踏实实的走出来的，在这里还要再次感谢实验室王桂娟老师和陈华荣老师，没有老师们的悉心指导我也不可能取得这么长足的进步，同时也要感谢学校学院能够为我们提供这样一个学习和发展的平台。

最后，还是要感谢父母和自己。感谢父母将我培养成人，您们辛苦了；感谢一下自己，从大一到大四时间说长也不长说短也不短，虽然自己大学四年并没有太多精彩的瞬间，但至少来说现在回顾这四年自己做的每一个决定，走的每一步可以说是问心无愧的，辛苦了，来日方长，接下来继续加油！

参考文献

[1] 葛景春. 诗酒风流-试论酒与酒文化精神对唐诗的影响[J]. 河北大学学报：哲学社会科学版, 2002, 27(2):6.

[2] 杨利. 酒文化及酒的精神文化价值探微[J]. 邵阳学院学报：社会科学版, 2005, 4(2):2.

[3] 杨立公, 朱俭, 汤世平. 文本情感分析综述[J]. 计算机应用, 2013, 33(6):1574-1607.

[4] 陈为, 沈则潜, 陶煜波. 数据可视化[M]. 电子工业出版社, 2013.

[5] 王玉成, 邢慧斌. 唐代诗酒文化特征及形成原因初探[J]. 河北师范大学学报：哲学社会科学版, 2009, 32(3):4.

[6] 陈超,吴亚东,付朝帅,童兴,李攀,褚琦凯,王雪楠.中国白酒文化可视化研究[J].大数据,2021,7(2):78-98

[7] 张玮, 谭思危, 刘凯,等. 宋词研究的新视角:文本关联与时空可视分析[J]. 计算机辅助设计与图形学学报, 2019, 31(10):11.

[8] 欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报,2016(2):66-80. DOI:10.13530/j.cnki.jlis.160011.

[9] 封颖超杰,周姿含,张玮,等. "为你写诗":面向中国古典诗歌的可视化交互创作系统[J]. 计算机辅助设计与图形学学报,2021,33(9):1318-1325. DOI:10.3724/SP.J.1089.2021.18980.

[10] 李斌, 王璐, 陈小荷,等. 数字人文视域下的古文献文本标注与可视化研究——以《左传》知识库为例[J]. 大学图书馆学报, 2020, 38(5):10.

[11] 王妮满, 秦昆, 罗俊,等. 历史名人轨迹的空间可视化与分析[J]. 地球信息科学学报, 2020, 22(5):11.

[12] Meneses L, Furuta R. Visualizing poetry: Tools for critical analysis[J]. paj: The Journal of the Initiative for Digital Humanities, Media, and Culture, 2015, 3.

[13] Mittmann A, von Wangenheim A, dos Santos A L. A multi-level visualization scheme for poetry[C]//2016 20th International Conference Information Visualisation (IV). IEEE, 2016: 312-317.

[14] Musaoglu O, Dag Ö, Küçükakça Ö, et al. A Generic Tool for Visualizing Patterns in Poetry[C]//DH. 2017.

[15] 白雪丽. 浅析基于 Python 爬虫技术的特性及应用[J]. 山西科技,2018,33(2):53-55. DOI:10.3969/j.issn.1004-6429.2018.02.015.

[16] 陶娅芝. 基于 word2vec 和词性的用户评论情感分析研究[J]. 数字化用户,2019,025(003):267269

[17] 薛晓宇,龙杰,方义成. 基于 LSTM 算法的无线网络流量预测研究[J]. 长江信息通信,2021,34(10):4-

6. DOI:10.3969/j.issn.1673-1131.2021.10.002.

[18] Yu Y, Si X, Hu C, et al. A review of recurrent neural networks: LSTM cells and network architectures[J]. Neural computation, 2019, 31(7): 1235-1270.

[19] 赵刚, 徐赞. 基于机器学习的商品评论情感分析模型研究[J]. 信息安全研究, 2017, 3(2):5.

[20] Jelodar H, Wang Y, Yuan C, et al. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey[J]. Multimedia Tools and Applications, 2019, 78(11): 15169-15211.

[21] The President and Fellows of Harvard College. China Biographical Database Project (CBDB)[DB/OL]. [2022-05-31]. https://projects.iq.harvard.edu/cbdb.

[22] 搜韵网．唐宋文学编年地图[CP/OL]．[2022-05-31]．https://sou-yun.cn/poetlifemap.aspx.