**1**

**人工智能学院**

**实践环节报告**

**课程名称：** 大数据工程实践一

**课题名称：** 全唐诗数据分析系统的设计与实现

**专 业**： 数据科学与大数据技术

**班 级：** 数科221

**学 号：** 202219274127

**学生姓名：** 李文彬

**指导教师：** 张茜

**2024年 12月22日**

**中原工学院人工智能学院**

**大数据工程实践一设计任务书**

|  |  |  |
| --- | --- | --- |
| **姓 名** | 李文彬 | 数据科学与大数据技术 专业 221 班 |
| **题 目** | 全唐诗数据分析系统的设计与实现 | |
| **设**  **计**  **任**  **务** | 全唐诗数据分析系统的目的是通过大数据技术对全唐诗进行系统化、规模化的分析，从海量文本中挖掘出隐藏的语言规律、情感倾向和主题分布，为唐诗研究提供更加客观、全面的数据支持。  主要功能包括：   1. 诗词展示，诗人信息展示，搜索功能 2. 统计分析，相关字查找，情感预测   开发工具：python，vscode | |
| **时**  **间**  **进**  **度** | 第1周（2024-12-23~2024-12-27）：设计系统框架，数据获取与清洗  第2周（2024-12-30~2025-1-3）：数据挖掘与可视化，实现系统功能，添加机器学习模型  第3周（2025-1-6~2025-1-10）：完善系统功能，优化代码逻辑，编写项目报告 | |
| **原 主**  **始 要**  **资 参**  **料 考**  **与 文**  **献** | [1] 周志华. 机器学习 [M]. 北京: 清华大学出版社, 2016.  [2] 王小艳. 基于Python编程的《全唐诗》数据分析[J]. 新课程导学,2022(16):50-52.  [3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111-3119).  [4] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. | |

**指导教师签字： 2024年 12月 22日**

# 摘要

本课程设计

目录页

# 第一章课题介绍

## 1.1背景

近年来，随着全球化进程的加速和信息技术的迅猛发展，人们的生活方式和思维方式发生了深刻的变化。然而，正是在这种快速变迁的背景下，越来越多的人开始重新审视和珍视中华传统文化的价值。弘扬中华传统文化的现象级综艺节目如《中国汉字听写大会》、《中国成语大会》、《中国谜语大会》、《中国诗词大会》等应运而生，这些节目不仅在社会上引发了广泛关注，更在潜移默化中唤起了人们对传统文化的认同感和自豪感。

这些节目的成功，背后有着深刻的社会心理和文化背景。首先，随着现代社会的快节奏发展，人们的生活逐渐被碎片化的信息所包围，网络语言、流行语汇层出不穷，虽然这些新词汇丰富了语言的表达形式，但也使得人们逐渐感受到一种“语言荒漠”的窘境。在这种背景下，人们对传统文化中那些精致、典雅的文字产生了强烈的向往和膜拜心理。诗词作为中华文化的瑰宝，承载着古人的智慧、情感和审美，成为了人们心灵寄托的重要载体。

作为一名大数据相关专业的学生，我在学习过程中接触到了许多数据挖掘、自然语言处理（NLP）和机器学习的技术与方法。这些技术在现代文本分析中已经得到了广泛应用，但在古汉语文本处理领域，尤其是像唐诗这样的古典文学作品，仍然存在许多值得探索的问题。例如，古汉语与现代汉语在词汇、语法和表达方式上存在显著差异，如何有效地对唐诗进行分词、情感分析和主题建模，是一个具有挑战性的课题。

## 1.2课题意义

全唐诗作为中国古代文学的瑰宝，承载了唐代社会的历史、文化、思想和情感，是中华优秀传统文化的重要组成部分。随着大数据技术的快速发展，将现代数据分析方法应用于古典文学研究，不仅能够为唐诗研究提供新的视角和工具，还能为传统文化的传承与创新开辟新的路径。本课题的意义主要体现在以下几个方面：

深化对唐诗语言与文化的理解 。唐诗以其精炼的语言、深邃的意境和丰富的情感著称，但其语言风格、表达方式和文化内涵往往需要深入解读才能被现代人理解。通过词频分析、主题建模和情感分析等技术，可以系统地揭示唐诗中的高频词汇、常见意象和情感表达规律，帮助人们更直观地理解唐诗的语言特点和文化价值。例如，通过分析不同时期诗人的作品，可以揭示唐代社会变迁对文学创作的影响；通过情感分析，可以探讨诗人在不同人生阶段的情感变化。

提升数据分析能力与实践经验。对于我们而言，本课题不仅是一次对古典文学的探索，更是一次对大数据技术的实践与应用。在课题实施过程中，需要综合运用数据爬取、文本预处理、统计分析、机器学习和数据可视化等技术，这将极大地提升数据分析能力和解决实际问题的能力。同时，课题中涉及的跨学科思维和创新方法，也将为未来的学术研究或职业发展奠定坚实的基础。

弘扬中华优秀传统文化。在全球化背景下，中华优秀传统文化的传承与弘扬显得尤为重要。通过对全唐诗的深度分析，可以挖掘其中蕴含的智慧、情感和价值观，并将其与现代社会的需求相结合，为当代人提供精神滋养和文化自信。这种基于数据分析的文化研究，不仅能够增强人们对传统文化的认同感，还能为中华文化在国际上的传播提供新的视角和工具。

## 1.3任务重述

### 1.3.1数据挖掘

从全唐诗的数据库或文本中提取有价值的信息，包括以下内容：

诗人信息：如诗人的姓名、生平、所属时期、社会地位等。

诗歌内容：如诗歌的标题、正文、创作时间、体裁等。

主题与风格：通过文本分析技术，挖掘诗歌的主题（如山水、边塞、咏史等）和风格（如豪放、婉约、写实等）。

用词特征：统计高频词汇、常见意象（如“月”、“酒”、“花”等）以及特定词汇的使用频率。

数据挖掘的目标是为后续分析提供高质量、结构化的数据基础，同时解决古汉语分词、文本清洗等技术难题。

### 1.3.2数据分析

对挖掘出的数据进行深入的统计分析，探索唐诗中的规律与特点，具体包括：

主题分析：研究不同时期（如初唐、盛唐、中唐、晚唐）诗歌的流行主题，分析主题随时间的变化趋势。

诗人风格分析：对比不同诗人的写作风格，分析其用词偏好、情感倾向和创作特点。

情感分析：通过情感词典或机器学习模型，分析诗歌中的情感倾向（如喜悦、悲伤、愤怒等），并研究情感表达的变化规律。

数据分析的目标是揭示唐诗中的语言规律、文化内涵和历史价值，为古典文学研究提供数据支持。

### 1.3.3可视化

利用数据可视化技术，将分析结果以直观、易懂的形式呈现，使用图表展示，如柱状图、折线图、饼图等，用于展示词频统计、主题分布、情感趋势等分析结果。

可视化的目标是让分析结果更加生动、易于理解，同时为公众提供一种欣赏唐诗的新方式，助力传统文化的传播与普及。

# 第二章数据收集与预处理

## 2.1源数据获取及处理

在项目初期，我们从老师提供的参考项目中获取了《全唐诗.txt》作为初始数据集。然而，通过初步浏览和分析，我们发现该数据集存在诸多问题，例如格式不统一、乱码现象严重等。这些问题直接影响了数据的可用性和后续分析的准确性。因此，我们决定采用多线程网络爬虫技术重新获取全唐诗数据，以确保数据的完整性和质量。

在互联网上，许多网站提供了全唐诗的部分诗集，但这些数据往往分散且格式各异，难以满足我们的需求。为此，我们参考了王小艳前辈的爬虫代码，并对其进行了多线程优化，以提升数据抓取的效率。最终，我们从“国学导航”网站成功获取了完整的全唐诗数据，并将其存储为《全唐诗2.txt》。

为了便于后续的计算和读取，我们对获取到的数据进行了初步处理。通过正则表达式匹配，我们将数据转换为统一的格式（如图2-1所示）。然而，在处理过程中，我们发现数据中仍存在一些不符合格式要求的问题，主要包括：题目中含有特殊字符；部分诗歌缺失作者信息；诗文内容中包含特殊字符。

经过进一步分析，我们了解到这些特殊字符主要是由于繁体字转换为简体字时，部分字符无法正常匹配所致。为了解决这一问题，我们尝试使用爬虫技术重新获取正确的诗文内容并进行替换。然而，由于替换效果不理想，最终我们决定采用过滤特殊字符的方案，以确保数据的整洁性和可用性。

图2-2展示了经过处理后的数据样例。通过上述步骤，我们成功获取并清洗了全唐诗数据，为后续的数据分析和挖掘奠定了坚实的基础。

## 2.2诗人信息获取及处理

在我们已有的全唐诗数据中，仅包含诗歌内容，缺乏诗人的个人信息。为了丰富数据集并支持后续的分析，我们需要获取诗人的详细信息，包括姓名、生卒年份、所属时期等。为此，我们基于第一步处理得到的诗人列表，利用爬虫技术从互联网上爬取诗人信息。

在初步尝试中，我们选择了搜狗百科作为数据源，因其提供了较为工整的诗人信息。然而，在实际爬取过程中，搜狗百科的反爬虫机制（如人机验证）对数据获取造成了较大阻碍。为了不影响项目进度，并确保课程目标侧重于数据分析而非反爬虫技术的攻克，我们更换了数据源，最终从网站（http://hkxy.xcz.im）成功获取了所需的诗人信息。

获取到诗人信息后，我们对其进行了数据清洗，以确保数据的完整性和一致性。在检查过程中，我们发现诗人的生卒信息存在以下三类问题：仅包含出生年份或死亡年份：对于这类情况，我们根据已知的出生或死亡年份，推断诗人所处的时期（如初唐、盛唐、中唐、晚唐）。无生卒信息：对于缺乏生卒年份的诗人，我们将其所处时期统一标记为“唐”，以表示其属于唐代，但具体时期未知。

通过上述处理，我们得到了格式统一、信息完整的诗人数据集。图2-3展示了清洗后的诗人信息样例，包括诗人姓名、生卒年份以及所属时期。这一步骤为后续的诗人风格分析、时期对比等研究提供了重要的数据支持。

## 2.3诗词主题及体裁的获取

在现有的数据集中，我们已经拥有了较为完整的诗词内容和诗人信息，但缺乏诗词的体裁和主题信息。为了丰富数据集并支持后续的主题分析和体裁研究，我们需要从诗词内容中提取这些信息。以下是我们在这一过程中遇到的挑战及解决方案。

### 2.3.1体裁获取

在获取诗词体裁的过程中，我们首先考虑了几种可能的方案。

定义体裁类别：这种方法实现简单且处理速度快，但存在明显的局限性。由于体裁的定义完全依赖于人工判断，忽略了诗词的具体内容，导致处理结果与实际体裁可能存在较大偏差，难以满足数据分析的精度要求。

从现有网站爬取体裁信息：这种方法看似合理，但在实际操作中，我们发现能够获取到的有效数据非常有限。许多网站并未提供完整的体裁信息，或者数据格式不统一，难以直接使用。

为了解决上述问题，我们小组经过深入讨论和研究，最终提出了一种折中

的方案：调用国内先进大模型（DeepSeek）的API接口。具体步骤如下：将每首诗的正文内容作为输入，发送至DeepSeek API；通过API返回的结果，提取每首诗的体裁信息（如五言绝句、七言律诗等）。由于数据量庞大（全唐诗共四万余首），我们在实际调用中采用了多线程技术，显著提升了处理效率。

### 2.3.2主题获取

对于诗词主题的获取，我们面临类似的挑战。现有的网站和资源中，很难找到符合我们需求的主题信息。因此，我们决定沿用体裁获取的方案，继续调用DeepSeek API，通过输入诗词内容，获取每首诗的主题信息（如山水、边塞、咏史、送别等）。

通过上述方法，我们成功获取了全唐诗的体裁和主题信息，为后续的诗词分类、主题分析和风格研究提供了重要的数据支持。这一过程不仅解决了数据获取的难题，也为我们积累了调用大模型API和多线程优化技术的宝贵经验。

# 第三章探索性分析

## 3.1数据分布分析

为深入探究全唐诗数据集中诗人所处的历史时期分布特征，本研究对数据集中的"period"列进行了系统的频次统计分析。

从全唐诗数据集中提取关键字段"period"，该字段详细记录了每位诗人所属的历史分期。根据唐诗发展的历史脉络，主要划分为以下时期：初唐（618-712年）、盛唐（713-765年）、中唐（766-835年）、晚唐（836-907年）等。这些分期不仅反映了唐代政治、经济、文化的变迁，也体现了诗歌风格的演变轨迹。

采用Python编程语言中的Pandas库进行数据处理，具体使用value\_counts()方法对"period"列进行频次统计。该方法能够快速、准确地计算每个历史时期出现的频次，并自动按降序排列，为后续分析提供基础数据支撑。核心代码如下：

period\_stats = self.poets\_df['period'].value\_counts()

## 3.2诗歌体裁分布分析

3.2.1 数据提取

从全唐诗数据集中提取关键字段"type"，该字段详细记录了每首诗歌的体裁分类。唐代诗歌体裁丰富多样，主要包括：五言绝句、七言律诗、乐府、五言律诗、七言绝句、古体诗。

采用Python编程语言中的Pandas库进行数据处理，具体步骤如下：使用value\_counts()方法对"type"列进行频次统计，计算每种体裁出现的绝对数量。基于统计结果，进一步计算各类体裁的相对占比，以揭示其分布特征。核心代码如下：

type\_stats = self.poetry\_data['type'].value\_counts()

total\_poems = len(self.poetry\_data)

type\_percentages = type\_stats / total\_poems

## 3.3诗歌长度分布分析

为深入探究全唐诗数据集中诗歌的篇幅特征，本研究对数据集中的"content"字段进行了系统的字数统计与区间分布分析。具体研究过程如下：从全唐诗数据集中提取"content"字段，该字段完整记录了每首诗歌的文本内容。为确保统计结果的准确性，在分析前进行了以下数据清洗工作：去除标点符号、统一繁简字体、处理异常值（如缺失值、重复值）。

诗歌字数计算：采用Python编程语言中的Pandas库进行数据处理，使用str.len()方法精确计算每首诗歌的字数。为更清晰地展现诗歌长度的分布特征，将诗歌字数划分为9个区间，并使用pd.cut()方法进行区间统计。区间划分充分考虑了唐代诗歌的体裁特征和篇幅特点核心代码如下：

poetry\_lengths = self.poetry\_data['content'].str.len()

bins = [0, 20, 40, 60, 80, 100, 150, 200, 300, float('inf')]

labels = ['1-20字', '21-40字', '41-60字', '61-80字', '81-100字',

'101-150字', '151-200字', '201-300字', '300字以上']

length\_dist = pd.cut(poetry\_lengths, bins=bins, labels=labels).value\_counts().sort\_index()

## 3.4字频统计

在字频统计分析环节，本研究采用系统性方法对《全唐诗》进行量化研究。为确保统计结果的全面性和准确性，我们将诗歌正文与题目内容均纳入统计范围，数据源为《全唐诗2.txt》。具体研究步骤如下：

首先，进行停用词处理。基于语言学理论和诗歌文本特征，本研究构建了一个包含常见虚词（如"之"、"乎"、"者"等）的停用词列表。通过过滤这些功能性词汇，可以有效降低数据噪声，突出诗歌文本中的核心实词，从而提高后续分析的准确性。

其次，实施中文字符提取。在完成停用词处理后，采用正则表达式方法从文本中精确提取中文字符，确保统计对象均为有效汉字。这一步骤不仅保证了数据的纯净度，也为后续的量化分析奠定了可靠基础。

最后，进行字频统计与数据保存。运用Python编程语言中的Collections模块，对提取的中文字符进行频率统计，并将统计结果以结构化数据形式保存。统计结果如图X所示（请补充具体图表编号），直观展示了《全唐诗》中汉字的分布特征和使用规律。

值得注意的是，本研究采用的字频统计方法具有普适性，可推广应用于其他文本数据分析。例如，通过对legal.csv文件中的诗人作品进行类似分析，可以统计每位诗人的常用字频，进而深入挖掘诗人的用词偏好、创作特点和风格演变。这种量化分析方法为古典文学研究提供了新的视角和方法论支持。

## 3.5四季词频

基于char\_frequency.csv文件中的字频统计数据，本研究对"春"、"夏/暑"、"秋"、"冬"四个季节主题词在《全唐诗》中的出现频次进行了系统性分析。统计结果显示，四季主题词的使用呈现显著差异："春"字以13132次高居榜首，"秋"字以9768次位列第二，而"夏"和"冬"的出现频次分别为1439次和908次，较前两者低一个数量级。

## 3.6色彩分析

采用定量分析方法，对《全唐诗》中的色彩词汇进行系统性考察。首先，基于色彩学原理和汉语词汇特征，建立了四个主要色系的字符字典：红色系（红、丹、朱、赤、绛等）、黑色系（暗、玄、乌、冥、墨等）、绿色系（绿、碧、翠、苍等）及白色系（白、素、皎、皓等）。通过处理char\_frequency.csv文件，对各色系词汇进行提取和统计。

## 3.7不同时期主题的变化

采用数据融合方法，对唐诗主题的历时演变进行定量分析。数据源整合，以poets为关键字段，将poets\_info.csv中的诗人时期信息与legal.csv中的诗词信息进行关联合并，构建包含诗词文本、作者及时期信息的完整数据集。时期划分，依据传统唐诗分期理论，将数据划分为初唐、盛唐、中唐、晚唐四个时期，确保分期标准与文学史研究范式一致。

数据统计，对合并后的DataFrame进行分组聚合操作，计算各时期不同主题的数量分布和比例关系。趋势分析，运用时间序列分析方法，考察特定主题在不同时期的演变轨迹。

# 第四章数据挖掘

## 4.1Word2Vec模型

Word2Vec 是一种广泛使用的词嵌入模型，能够将词语映射到低维向量空间，从而捕捉词语之间的语义关系。在自然语言处理任务中，文本数据需要被转换为数值形式以供机器学习模型使用。尽管存在多种文本表示方法，如 One-Hot 编码、矩阵分解和 NNLM（神经网络语言模型），但这些方法在本项目中并不适用。相比之下，Word2Vec 因其简单高效且能够有效捕捉词语之间的语义关系，成为本项目的理想选择。由于本数据分析系统需要实现相关词提取和情感预测功能，因此本文重点介绍系统中使用的 Word2Vec 跳字模型（Skip-gram）。

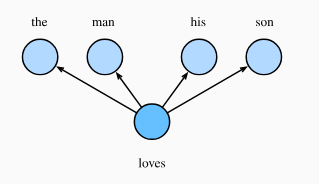
### 4.1.1 Word2Vec 的背景

在 Word2Vec 模型出现之前，自然语言处理领域面临两大难题：维度灾难和词汇鸿沟。维度灾难指的是高维稀疏向量表示导致的计算复杂性和效率低下问题，而词汇鸿沟则是指传统方法难以捕捉词语之间的语义关系。为了解决这些问题，研究者们致力于寻找一种能够将高维词向量映射到低维空间的方法。Word2Vec 的提出正是为了应对这些挑战。它通过将每个词语表示为一个定长的稠密向量，使得这些向量能够较好地表达词语之间的相似性和类比关系，从而显著提升了文本表示的效果。

### 4.1.2Word2Vec 的跳字模型

跳字模型假设基于某个词来生成它在文本序列周围的词。举个例子，假设文本序列是“the”“man”“loves”“his”“son”。以“loves”作为中心词，设背景窗口大小为2。如图10.1所示，跳字模型所关心的是，给定中心词“loves”，生成与它距离不超过2个词的背景词“the”“man”“his”“son”的条件概率，即

假设给定中心词的情况下，背景词的生成是相互独立的，那么上式可以改为：



图跳字模型关心给定中心词生成背景词的条件概率

在跳字模型中，每个词被表示成两个 d 维向量，用来计算条件概率。假设这个词在词典中索引为 i ，当它为中心词时向量表示为，而为背景词时向量表示为 Rd。设中心词在词典中索引为c，背景词在词典中索引为o，给定中心词生成背景词的条件概率可以通过对向量内积做softmax运算而得到：

其中词典索引集。假设给定一个长度为T的文本序列，设时间步 t的词为。假设给定中心词的情况下背景词的生成相互独立，当背景窗口大小为 m时，跳字模型的似然函数即给定任一中心词生成所有背景词的概率

### 4.1.3模型实现

为了训练 Word2Vec 模型，首先需要对原始文本数据进行预处理。本研究以《全唐诗》作为训练数据，具体预处理步骤如下：从文本文件（data/全唐诗2.txt）中读取数据，文件内容为全唐诗的文本。使用正则表达式移除所有非中文字符（如标点符号、数字等），仅保留汉字。将每一行文本拆分为单个汉字，形成句子列表。每个句子由一个汉字列表表示，例如 ['春', '眠', '不', '觉', '晓']。在预处理过程中，每处理 100 行数据时输出当前进度，以便实时跟踪处理状态。

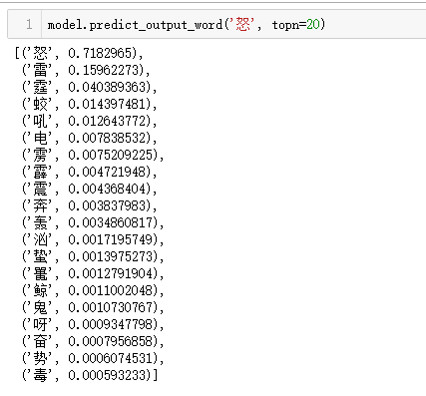
在完成数据预处理后，我们使用 Gensim 库中的 Word2Vec 模型进行训练。

模型参数设置：词向量的维度设置为 100，上下文窗口大小为 10（以捕获更多的上下文信息），忽略出现次数少于 10 次的低频词，使用 Skip-gram 模型进行训练，使用负采样方法（设置负采样数为 10），训练轮数为 20 轮，使用多线程加速训练（线程数为 CPU 核心数）。

模型保存与加载：如果模型已经存在，则直接加载；否则，训练模型并保存到指定路径。这样再次运行程序可以直接加载已训练好的模型，减少不必要的时间浪费。

## 4.2相关字

基于已训练好的字嵌入 Word2Vec 模型，我们可以将高维空间（维度为所有单字的数量）映射到一个低维连续向量空间中。在这一过程中，每个单字被表示为实数域上的向量。通过将单字转化为向量，我们可以利用向量之间的余弦相似度来计算单字之间的相关性。具体而言，求解与单字 A 最相关的单字，可以通过计算单字 A 的向量与其他单字向量之间的余弦距离，并选择距离最近的向量对应的单字来实现。实验结果如图所示。



图

在唐诗中，某些情绪词汇（如“悲”、“忧”、“愁”、“怒”、“惧”）频繁出现。基于训练好的 Word2Vec 模型，我们提取了这些情绪词汇的相关字，并将其整理为情绪词典。该词典将用于后续的诗词情绪分类任务，为情绪分析提供基础数据支持。

## 4.3情感预测

在上一节中，我们完成了 Word2Vec 模型的训练，并基于该模型实现了相关字功能。为进一步扩展功能，我们希望在单字级别的基础上，实现诗句级别的情感分析。由于已构建了情绪词典，我们首先将《全唐诗》按诗句进行切分，并对每句诗进行情感标注。接着，将诗句转化为向量表示，并将情感标签编码为数值形式。在此基础上，我们可以使用多种机器学习模型（如逻辑回归、神经网络、随机森林和支持向量机等）进行训练。通过对比各模型的性能指标（如准确率、召回率等），选择表现最佳的模型用于最终的情感预测任务。

## 4.4情感预测模型

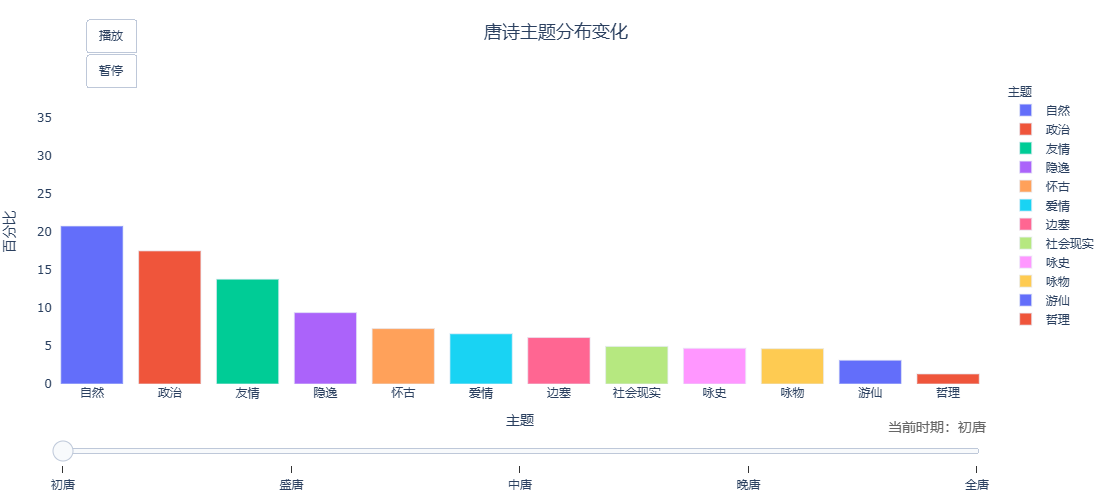
### 4.4.1模型构建

### 4.4.2模型评估

# 第五章可视化展示

本章将系统展示从《全唐诗》数据中分析与挖掘的有效信息，并通过图表等可视化形式直观呈现。

## 5.1主题变化



图

初唐时期（618-712）：自然主题占比最高（20.74%），其次是政治题材（17.47%）和友情主题（13.76%）。这一时期政治题材的比重相对较高，反映了初唐统一后的政治关注度。

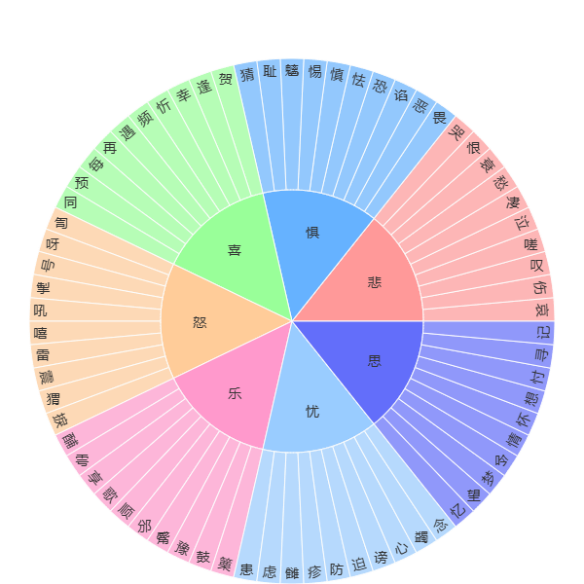
盛唐时期（713-765）：友情主题大幅上升至33.26%，成为主导题材。自然（14.75%）和隐逸（14.69%）题材并列第二，反映了盛唐诗人对山水田园生活的向往。政治题材降至4.73%，表明盛唐诗歌更注重抒情与写景。

中唐时期（766-835）：友情主题保持领先（26.38%），但比例有所下降。自然（17.42%）和隐逸（14.68%）主题保持稳定。社会现实题材上升至9.62%，反映了这一时期诗人对现实问题的关注增加。

晚唐时期（836-907）：友情（20.50%）、隐逸（19.69%）和自然（17.11%）三大主题比例接近。社会现实题材维持在较高水平（9.26%）。政治题材降至3.81%，显示晚唐诗人更倾向于抒发个人情感。

整体发展趋势：友情主题始终占据重要地位，在盛唐达到顶峰后逐渐回落。自然与隐逸主题贯穿始终，反映了唐代文人对自然生活的持续向往。政治题材从初唐的高位逐渐下降，表明诗歌主题从政治转向个人情感的发展趋势。社会现实题材在中晚唐时期明显上升，反映了诗人对现实关注度的提高。

## 5.2情感词系统

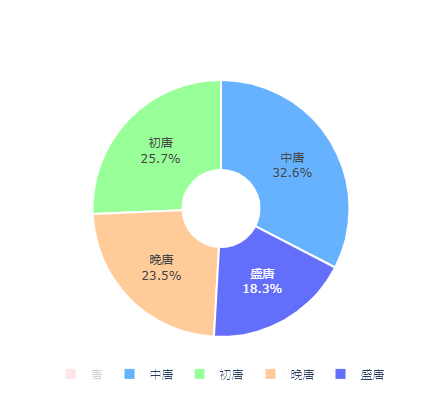


图

情感词系统展现了唐诗中丰富的情感表达方式。通过轮盘图的层级结构，可以清晰地展示不同类别的情感词及其关联词。

情感词的多样性不仅反映了唐诗在情感表达上的丰富性和细腻度，还揭示了诗人情感的复杂性与多层次性。不同类别的情感词共同构成了一个完整的情感表达系统，充分展现了唐诗的抒情特色与艺术魅力。

## 5.3诗词时期分布

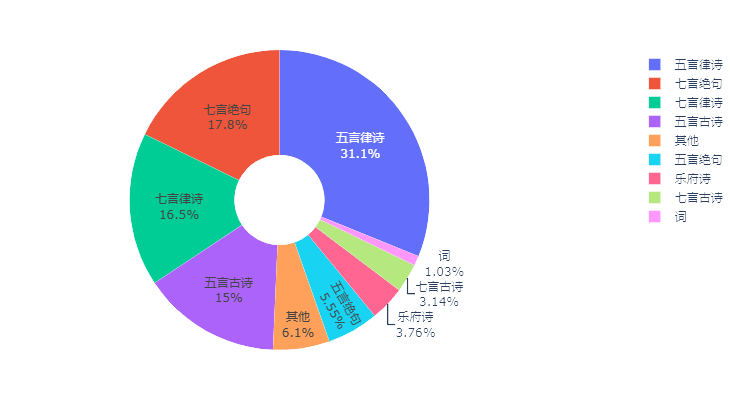


图

从时期分布来看，中唐时期的诗词占比最高，达到 32.6%，成为四个时期中的主导时期。初唐时期次之，占比为 25.7%，位列第二。晚唐时期占比 23.5%，与初唐比例接近。盛唐时期占比 18.3%，在四个时期中占比最小。

中唐诗歌占比接近三分之一，可能反映了这一时期文学创作的繁荣与活跃。初唐和晚唐的比例相近（分别为 25.7% 和 23.5%），表明这两个时期的创作活跃度相当。值得注意的是，尽管盛唐被公认为诗歌的黄金时期，但其诗词数量占比却最低。这一现象可能表明，盛唐诗歌更注重艺术质量而非数量，体现了这一时期诗歌创作的高度精致化与艺术追求。

## 5.4诗歌体裁分布

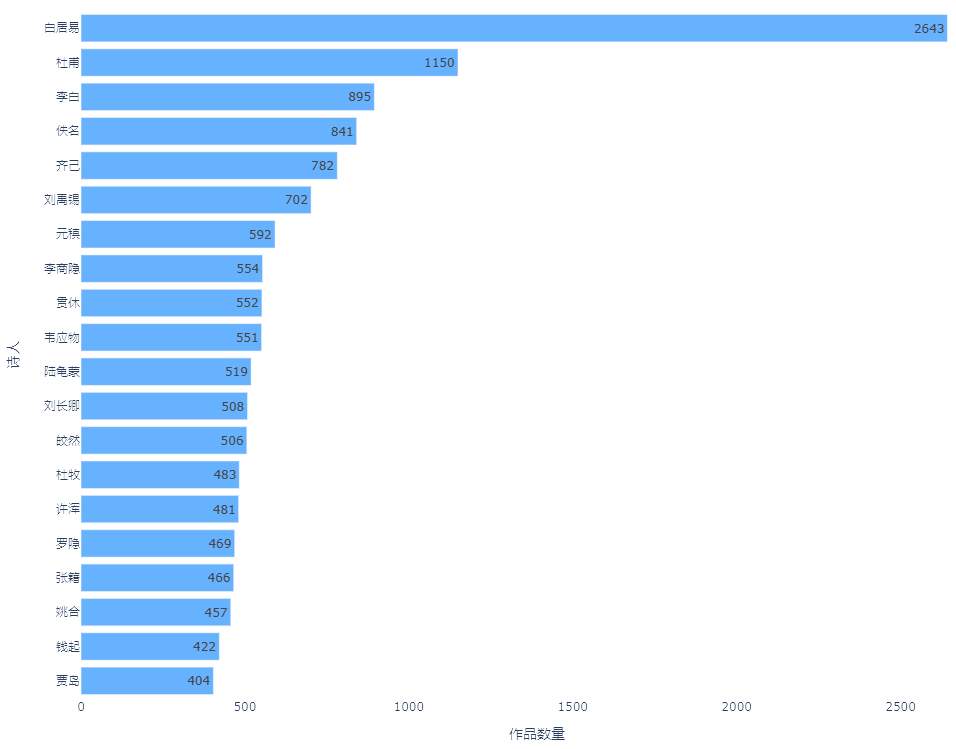


图

从诗歌体裁分布来看，五言律诗占比最高，这可能反映了其在唐代诗歌创作中的成熟度与普及性。五言律诗以其对仗工整、声律和谐的特点，成为宫廷诗歌及文人创作的主流形式。七言绝句占比位居第二，其短小精炼、易于传诵的特点使其深受喜爱，可能与其适合即兴表达和记录日常生活密切相关。

相比之下，古体诗的占比较低，这可能与其形式自由、不受格律限制的特点有关，导致创作的规范性较低。词作为一种后来兴起的文学形式，在唐代更多处于附属地位，尚未成为主流。此外，乐府诗占比偏低，可能与其以采集民间歌辞为主的性质有关，在文人创作中的比重相对有限。

## 5.5诗人作品排名



图

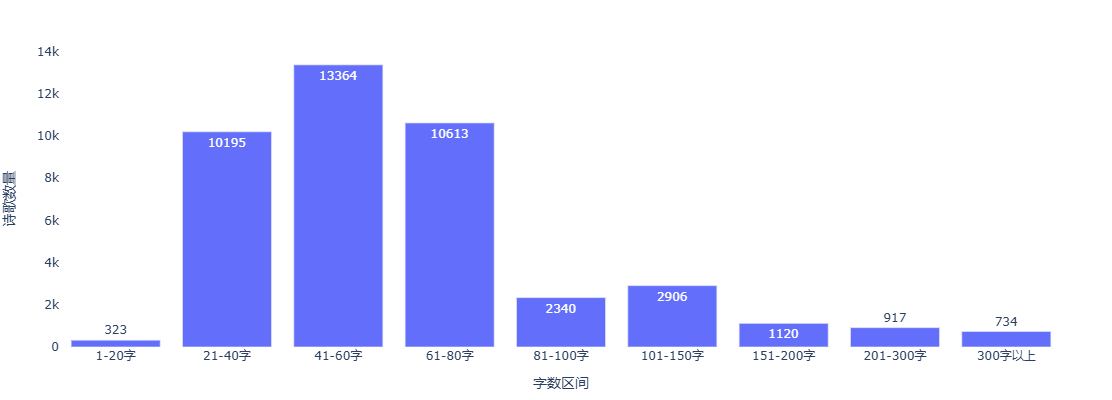
从诗人作品数量排名来看，白居易的创作高产性尤为突出。这可能与其长期在仕途和民间生活的丰富经历有关，他不仅创作了大量抒情诗，还撰写了众多针砭时弊的讽喻诗，展现了其对社会现实的深刻关注。

杜甫与李白的对比也值得关注。杜甫的创作数量高于李白，这可能与其坎坷的人生经历和持续创作的动力密切相关。相比之下，李白的创作虽数量较少，但以其卓越的艺术成就和独特的风格著称，体现了“以质取胜”的特点。

此外，佚名诗歌的数量接近 900 首，占据了相当比例。这反映了唐代文人或民间无署名作品在诗歌文化中的重要地位，同时也为研究唐代诗歌的多样性和民间文化提供了宝贵资料。

需要指出的是，作品数量并不完全等同于文学成就。例如，杜甫、李白等人的作品数量虽不及白居易，但其文学影响力却更为深远，体现了诗歌创作中质量与数量的辩证关系。

## 5.6诗歌长度分布



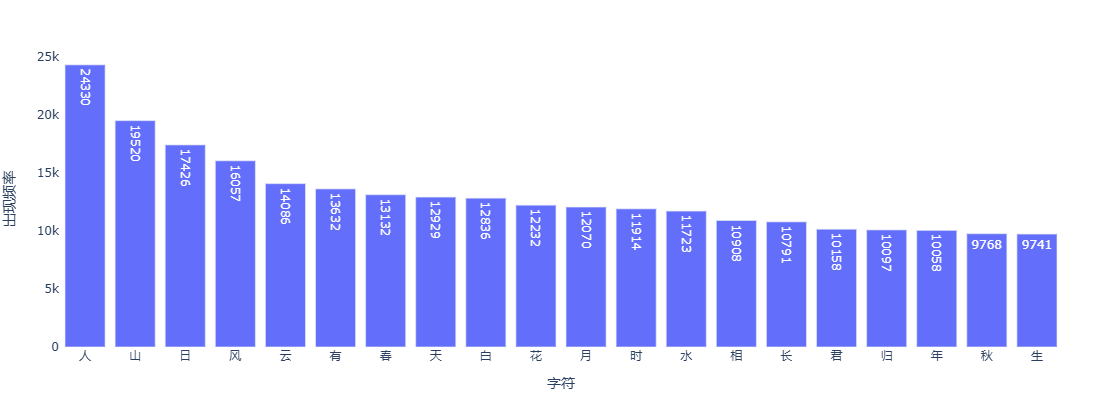
图

从诗歌长度分布来看，中等长度的诗歌占据主导地位，其中 40-80 字 的诗歌数量最多。这一现象与唐代常见的诗歌形式（如律诗、绝句）的标准字数相符，体现了唐代诗歌创作的规范化特点。

相比之下，长篇诗歌（超过 100 字）的数量显著减少，反映出唐诗创作更倾向于简洁凝练的表达方式，而非冗长的铺陈。同时，极短诗歌（少于 20 字）的数量极少，表明过于简短的形式并非唐诗的主要创作方向。

整体来看，诗歌长度分布呈现明显的正态分布特征，以 40-80 字 为中心的钟形分布进一步印证了唐代诗歌在长度上的规范性与集中性。这种分布特征不仅反映了唐代诗歌形式的成熟，也揭示了诗人对诗歌结构的严谨追求。

## 5.7字频统计



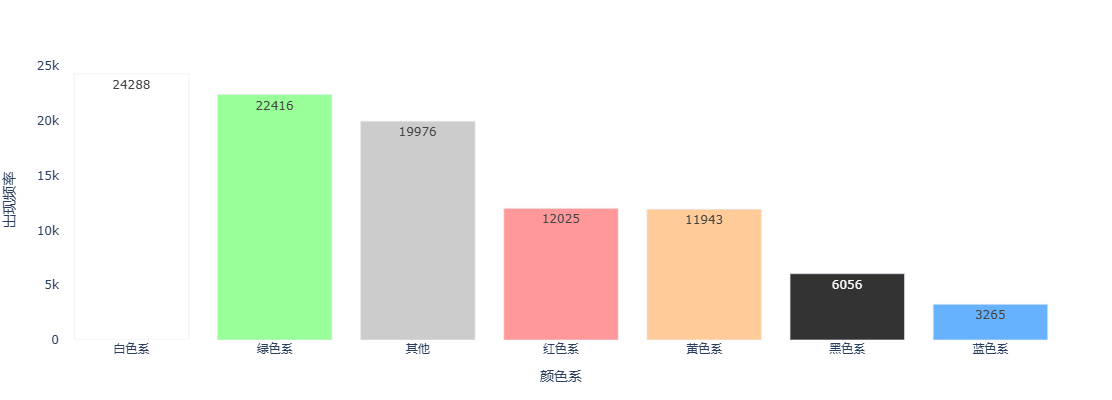
图

从字频统计结果来看，“人” 字的出现频率最高，达到约 24,330 次，这一现象反映了唐诗对人事的关注及其浓厚的人文情怀。“山” 字位居第二，出现约 19,520 次，展现了自然景物在唐诗中的重要地位，体现了诗人对自然的热爱与赞美。此外，“日” 和 “风” 字的出现频率分别约为 17,426 次 和 16,957 次，成为描写自然现象的常用字，进一步凸显了唐诗与自然的紧密联系。

高频字中，自然景物相关词汇占据了主要地位，反映了唐诗与自然的密切关系及其对自然意象的广泛运用。与此同时，“人” 字的高频出现体现了唐诗对人文主题的深刻关注。此外，与时间相关的字（如“日”）出现频率较高，显示了唐代诗人对时间流逝的敏感与哲思。

这些高频字共同构成了唐诗典型的意象系统，涵盖了自然、人事、时空等多个维度，不仅展现了唐代诗歌的丰富内涵，也揭示了其独特的审美倾向与文化特征。

## 5.8颜色词系统分布



图

从颜色词的使用频率来看，白色系的出现次数最多，约为 24,288 次，这可能与唐诗中月光、雪、云等自然景象的频繁描写有关。绿色系位居第二，出现约 22,416 次，主要用于描绘春季景色、山林植被等，展现了唐诗对生机勃勃的自然景象的偏爱。红色系出现约 12,025 次，多用于描写花卉、晚霞等，为诗歌增添了鲜艳的色彩。黄色系出现约 11,943 次，常用于描绘秋景、落叶等，体现了对季节变化的细腻捕捉。黑色系出现约 6,056 次，可能用于描写夜色、墨色等，为诗歌增添了深沉与神秘感。蓝色系出现频率最低，约为 3,265 次，主要用于描写天空、水面等，展现了唐诗对广阔自然空间的描绘。

白色和绿色的高频使用，反映了唐诗对明亮、清新意象的偏好，同时也体现了诗人对自然景观的细致观察与热爱。颜色词的使用频率与自然景观的描写密切相关，例如绿色与春季、黄色与秋季的关联，展现了唐诗对四季变化的敏感捕捉。此外，某些颜色词可能带有特定的文化象征意义，例如白色象征纯洁、红色象征喜庆等，这些文化内涵也影响了它们在诗歌中的使用频率。

总体而言，颜色词的使用不仅丰富了唐诗的视觉表现力，还反映了唐代诗人对自然、季节及文化象征的深刻理解与艺术表达。

# 第六章问题与总结

本章将系统梳理在研究过程中遇到的主要问题及其解决方案，并对本课题的研究成果进行总结。

## 6.1问题一：诗文内容的完整性

在处理《全唐诗》数据时，无论是老师提供的参考文件，还是我们自行爬取的版本，均存在特殊字符问题。这些特殊字符可能影响数据的统一性和后续分析的有效性。为解决这一问题，我们探讨了以下方案：

方案一：通过人工方式逐一检查并替换每首诗词中的特殊字符，能够确保处理的精准性。然而，由于《全唐诗》数据量庞大（四万余首），人工处理不仅耗时耗力，且难以在短时间内完成，因此这一方案不具备可行性。

方案二：我们尝试通过爬取其他网站获取完整的诗文内容，但发现大多数网站同样存在特殊字符问题。此外，中国古典文献资源中的《全唐诗》多为繁体字，进一步增加了数据获取和处理的难度。因此，这一方案也难以满足需求。

为确保数据的格式统一并便于后续分析，我们最终决定采用过滤特殊字符的方式。通过编写脚本，自动识别并移除诗文中的特殊字符，从而在保证数据质量的同时，提高处理效率。

## 6.2问题二：诗词体裁与主题的获取

在《全唐诗》文件中，诗词的体裁和主题信息并未直接存储。为了对每首诗进行准确标注，我们探讨了以下三种方案：

方案一：在理想情况下，人为标注能够确保标注结果的高度准确性，并充分契合诗词内容。然而，面对四万多首诗词的庞大规模，完全依赖人力进行标注将耗费大量的人力、物力和时间资源，因此这一方案在实际操作中并不可行。

方案二：通过编写程序，利用诗词的平仄规则、诗句长度、关键字等特征进行自动判断。这一方案在速度上具有显著优势，但其准确性较低，主要原因是程序难以捕捉文字之间的深层语义关系，导致标注结果不够精确。

方案三：大模型处理。鉴于传统程序在处理复杂语义问题上的局限性，我们采用现有的先进大模型来解决这一挑战。通过访问大模型的官方接口文档，学习并调用其接口，利用大模型的强大语义理解能力对诗词数据进行处理。这一方案在准确性和效率之间取得了较好的平衡，是前两种方案的折中与优化。

最终，我们选择方案三作为主要解决方案，以兼顾标注的准确性与处理效率。

## 6.3总结

本文全面系统地介绍了全唐诗数据分析的全过程，涵盖了数据处理、探索性分析、数据挖掘及可视化等多个环节。通过深入探讨各种解决方案的可行性，最终成功构建了全唐诗数据分析系统。数据处理作为整个系统的核心环节，其充分的数据信息与规范化的数据格式为后续分析奠定了坚实的基础。大模型的引入不仅体现了现代科技的发展，更展现了科技与现实问题的深度融合，通过先进技术有效解决了传统问题。机器学习模型的应用，使我们能够深入挖掘诗词中蕴含的深层次信息，进一步拓展了研究的深度与广度。最后，借助可视化技术，复杂的数据信息得以直观、生动地呈现，增强了分析结果的可理解性与传播力。

全唐诗数据分析系统的构建，不仅为我们在现代科技背景下挖掘中华优秀传统文化提供了有力工具，更展现了科技与文化结合的独特价值。通过科技手段，我们得以从全新的视角审视文化瑰宝，深入挖掘其中蕴含的宝贵财富。这种科技与文化的交融，不仅有助于推动传统文化的传承与弘扬，也为未来的文化研究开辟了新的路径。

# 参考文献

指导教师评语

成绩：

指导教师签名：

年 月 日

# 附录: 主要源程序

1.\*\*功能

代码格式：五号Times New Roman，单倍行距

2.\*\*功能