**2024年第三届燕山大学可视化与可视分析挑战赛**

**（YSU Data Visualization Challenge 2024）**

**作品说明文档**

参赛队名称：黑项目歌词

作品名称：歌曲的视觉语言

作品主题关键词： 情感变化，歌手关系，歌曲类别占比，歌曲年代变化

团队成员： 张丽平，燕山大学，2151555078@qq.com ，队长

刘子硕，燕山大学，[2484246430@qq.com](mailto:2476618985@qq.com)

徐国庆，燕山大学，1936326577@qq.com

刘恩铭，燕山大学，3456495067@qq.com

魏溪语，燕山大学，1320373732@qq.com

郭栋梁，燕山大学，121162433@qq.com，指导老师

使用的分析工具或开发工具：Excel，MySQL，echarts，Vscode，PyCharm

共计耗费时间（人天）： 60人天

本次比赛结束后，我们是否可以在网络上公布该文档与相关视频（是或否）：是

**一、作品简介：请围绕作品主题、要解决的问题\场景、目标用户\读者、应用价值等方面简要介绍作品**

本作品以中文歌词数据可视化分析为核心，围绕歌词数据的多维特征展开探索，揭示华语音乐文化中的深层信息。通过数据可视化技术，直观呈现歌词的语言、情感和创作特征，构建一个兼具趣味性与实用性的分析平台。

针对中文音乐领域缺乏系统性分析工具的问题，本作品旨在帮助用户从海量歌词中提炼有效信息，解决文化与社会洞察、创作支持以及教育与研究的需求。例如，通过歌词中不同年代流行词汇的词云分析，可挖掘社会热点和时代特征；通过歌曲的地域分布统计，揭示各地音乐风格和创作特色；情感变化曲线则反映了音乐作品在不同年份中情感表达的趋势，帮助理解社会情绪的演变。此外，歌词主题类别的可视化展示可为音乐创作者提供创作规律的参考，激发新的灵感；人物关系网络的构建则揭示了音乐创作背后的人际合作模式，为研究歌词创作群体和音乐产业网络提供数据支持。

本作品的目标用户包括音乐爱好者、创作者与教育者、研究学者以及音乐产业从业者。音乐爱好者可通过流行词云和情感曲线发现歌词中的文化故事；创作者与教育者能利用地域分布和主题分析获得灵感与教学素材；研究学者则可利用情感与合作网络分析展开多学科研究；音乐产业从业者可借助趋势洞察与流行特征分析，辅助市场定位和决策。本作品通过文化传承与认知、趋势洞察与预测、创意启发与学习以及产业支持等方式，帮助用户全面了解华语音乐的历史与地域特色，提升创作与商业价值。



图1.1 可视化界面整体展示图

**二、数据介绍：请围绕数据来源、数据格式、数据严谨性、数据清洗等方面简要介绍**

1.本项目使用的中文歌词数据来源于分为两部分：

①.网络收集和整理，构建了一个涵盖绝大多数华语歌手2019年之前歌曲的综合性数据库。数据库包含102197首歌曲，涉及4019位歌手，其中作品数超过20首的有1086人，超过100首的有233人，平均每位歌手拥有25.4首歌曲。这些数据经过初步筛选和分类，确保了基础信息的完整性和代表性。

数据以JSON文件的形式存储，便于解析和处理，整体被分为五个文件，分别记录了不同歌手的歌曲信息，并按作品数量降序排列。每首歌曲的数据结构包含三个主要字段：name表示歌曲名称，singer表示演唱者，lyric存储歌词文本。此外，还包括一些辅助数据文件，例如词频统计文件（data/words.json）、句首词统计文件（data/first\_words.json）和押韵表（data/rhymes.json），为更深入的分析提供了丰富的基础素材。

为了提高数据的严谨性和实用性，本项目对原始数据进行了全面的清洗和整理。在清洗过程中，我们统一了歌名和歌手名称的格式，修正了数据中存在的拼写错误和异常字符。同时，对歌词文本进行了优化，移除了多余的空格和特殊字符，使其更加规范化。此外，还核对了数据的逻辑一致性，确保歌名、歌手与歌词内容的匹配，进一步增强了数据的准确性。这些处理步骤为后续的数据分析和可视化奠定了可靠的基础。

②.使用基于Python爬虫技术，通过构建Music类对网易云音乐平台("https://music.163.com")进行系统化数据采集，采用fake\_useragent实现随机User-Agent切换以避免反爬，并通过lxml进行HTML解析。数据采集过程按照歌手分类和字母顺序进行遍历，获取歌手基本信息、简介、歌曲列表及其详细信息（包括歌词、热度数据和发布时间）。所有数据以JSON格式存储在music\_data目录下，包含独立的歌手文件（singer\_{id}.json）和汇总文件（all\_singers.json），每个歌手的数据结构包含id、姓名、分类、简介和歌曲详情等字段。通过合理的延时机制和异常处理确保了数据采集的稳定性和完整性，最终构建了一个包含12021首歌曲、涉及4300位歌手的json文件，为音乐文本分析提供了丰富的数据基础。

文本

描述已自动生成屏幕上有字

描述已自动生成

图2.1 部分数据展示图

2.本项目对数据预处理部分的实现和功能：

为了实现了对网易云音乐歌词的多维度分析和分类。系统首先构建了两个关键词词典：包含快乐、悲伤、愤怒、恐惧、平静五种情感类别的emotion\_keywords，以及包含古典、流行、摇滚、民谣、电子、说唱、儿童七种风格的style\_keywords。通过analyze\_lyrics函数对歌词进行处理，使用defaultdict数据结构记录情感和风格关键词出现频次，并结合jieba分词器提取高频词汇。在process\_data函数中，系统按照年份组织数据，解析JSON格式的时间戳，并为每首歌曲生成包含歌名、风格、发布时间、关键词和情感类型的结构化数据。最终，系统将处理结果以JSON格式保存，并通过异常处理机制确保数据处理的健壮性。

文本

描述已自动生成

图2.2 歌曲情感色彩分类图

**三、分析任务与可视分析总体流程**

**1. 流行词汇分析：揭示社会文化趋势的词云展示**

通过对歌词中出现的高频词汇进行统计，我们设计了词云球的展示方式，让用户可以直观地看到在不同时间段内的流行词汇变化。这种可视化方式不仅帮助我们识别出流行词汇，还能揭示不同历史时期的社会文化趋势。通过切换不同时间段，用户可以发现与社会热点、文化潮流相关的关键词，进一步分析社会文化变迁。

**2. 情感分析：探索歌词情感的时间演变**

通过提取歌词中的关键词语，我们对每首歌的情感进行了分类，并结合歌曲的发布年份进行了归类。情感分析的目的是帮助我们了解不同歌手、不同歌曲类型的情感表达特点，进一步分析这些情感如何随时间变化。为此，我们设计了动态折线图，通过年份对情感的变化进行展示，同时提供柱状图切换功能，允许用户根据具体年份放大查看数据，使展示更清晰、直观。

**3. 主题挖掘：基于高频词汇与情感分析的歌曲分类**

为了更深入地挖掘歌词的主题和情感，我们对歌词中的高频词汇进行了筛选，并将这些高频词汇与情感分类结合，以便更全面地了解歌曲的主题方向。通过饼状图统计，我们展示了不同情感类型的歌曲所包含的高频词汇，进一步揭示了情感倾向与歌词内容之间的联系。这一分析为我们提供了对歌曲主题和情感的双重视角。

**4. 地域特色分析：基于歌手地区与创作风格的地图展示**

通过爬取的歌手数据，我们分析了不同地区歌手的创作风格和情感特征。每个地区的歌手在创作中展现出的风格和情感都有其独特性，我们通过地图的方式，将这些地域特征与歌手的创作情感进行关联，清晰地展示出各地区的创作风格差异。

**5. 歌手风格分析：地区与歌手特色的可视化联动**

通过联动排行榜和地图，我们展示了不同地区歌手的创作特色。用户可以通过地图查看每个地区发布歌曲最多的四位歌手，并进一步了解哪些歌手最能代表该地区的音乐文化特色。这种联动展示为我们提供了对地区文化的深入理解，并有助于发掘歌手在地域文化中的影响力。

**6. 合作网络分析：探索歌手间的创作合作关系**

为了进一步分析歌手之间的合作模式，我们构建了一个基于歌词创作的合作网络。当用户点击某一歌手时，系统会隐藏与该歌手没有合作关系的歌手，突出显示该歌手与其他歌手之间的合作联系。通过这一互动式的合作关系网，用户可以更直观地了解歌手间的创作合作网络。

**7. 时代变迁分析：通过时间轴展示歌词内容的演变**

为了观察歌词内容随着时间的变化，我们在多个可视化展示中加入了年代时间轴。通过这一功能，用户可以清晰地看到歌词内容在不同历史时期的演变，进而反映出社会价值观、生活方式的变化。这不仅帮助我们识别出歌词内容的变化趋势，还能深入探讨社会背景和文化潮流对歌词创作的影响。

**四、数据处理与算法模型**

**1. 数据采集与预处理**

项目使用Python编写爬虫程序，从多个来源（包括公开的音乐数据库和歌词平台）抓取相关的歌词数据。为了确保数据的全面性和多样性，我们特别关注了包括歌手信息、歌曲名称、歌词文本、发布时间、情感标签等多维度数据。这些数据被以JSON格式存储，便于后续处理。

爬取到的原始数据往往包含许多无关信息或格式不规范的字段，因此我们需要进行数据清洗。通过Python中的json模块和pandas库，我们对爬取的JSON字符串进行了格式化处理，剔除无用字段并将其转换为标准化的结构。这一过程中，我们剔除了冗余字段，并确保每条数据的时间戳、情感标签以及歌词内容等关键信息按需正确整理。

**2. 数据存储与管理**

数据库存储：为了更方便地管理和查询数据，我们将清洗后的数据导入到关系型数据库中（如MySQL）。通过设计合理的数据表结构，我们确保了歌词数据的高效存取和扩展性。歌曲、歌手、歌词、情感标签和发布时间等数据被分配到不同的表中，并通过外键关系进行了关联。这不仅方便了数据的快速查询，也为后续的数据分析和可视化提供了高效支持。

**3. 数据处理与算法实现**

为了对歌词的情感倾向进行分类，我们采用了基于文本的情感分析算法。首先，利用jieba分词工具对歌词文本进行分词处理，然后结合情感词典和机器学习模型（如朴素贝叶斯分类器或支持向量机SVM）对歌词文本进行情感倾向判断。根据分析结果，每首歌的情感被分类为“正面”、“负面”或“中性”，并标注到数据库中。该分析模型通过训练数据的迭代优化，确保了较高的准确性。

为了从歌词中提取高频词汇并进行后续分析，我们使用了TF-IDF（词频-逆文档频率）算法。通过对歌词文本进行向量化表示，计算每个词汇在歌词中的重要性和频率。TF-IDF算法不仅帮助我们识别出歌曲中的高频词汇，还能通过分析歌词中的关键词揭示歌曲的主题和情感。最终，我们通过词云图呈现高频词汇，以便用户更直观地理解歌词内容。

**五、可视化与交互设计**

本系统通过多种可视化组件展示中文歌词数据，并提供互动设计，以帮助用户深入分析歌词的趋势和变化。词云图展示了歌词中出现频率较高的关键词，词汇的大小和颜色反映了其频率，用户可以直观地看到哪些词汇在不同时间段内更为突出，帮助理解歌词主题的演变以及歌曲随着年代变迁产生的变化。用户还可通过下拉框选择时间段，系统会根据选择展示相应的情感趋势和词汇分析结果，使得数据探索更具灵活性。



图5.1 不同年代歌曲词云图

为了挖掘歌词中常见的主题和题材，我们通过对歌词的分类和聚类进行数据处理，最后通过南丁格尔玫瑰图，来展示不同类别在总体中的占比，图中每个扇形代表一个类别，扇形的大小表示该类别在总体中的百分比。图例显示了不同颜色对应的类别名称，如“人”、“爱情”，“心”等，以及他们在图中所占的比例。

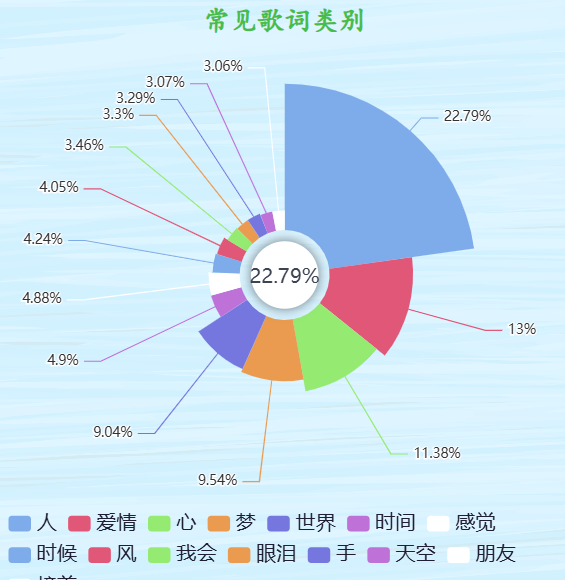


图5.2 歌曲类别统计图

通过折线图展示不同年份歌曲的情感分布，用户可观察到情感变化的趋势，如“快乐”和“悲伤”的情感在不同年代的表现差异。观察歌词内容随时间的变化，反映社会价值观、生活方式的变迁。通过查阅资料，将歌词的情感主要分为五个类别，分别为：快乐、悲伤、愤怒、恐惧、平静。用户可自行选择折线图，柱状图，纯数据以及放大某些年份的数据来观察，随着时间的变迁，歌曲的情感变化情况。同时，我们也设置了最高值和最低值来方便用户去观察不同年代的歌曲请改变化趋势。

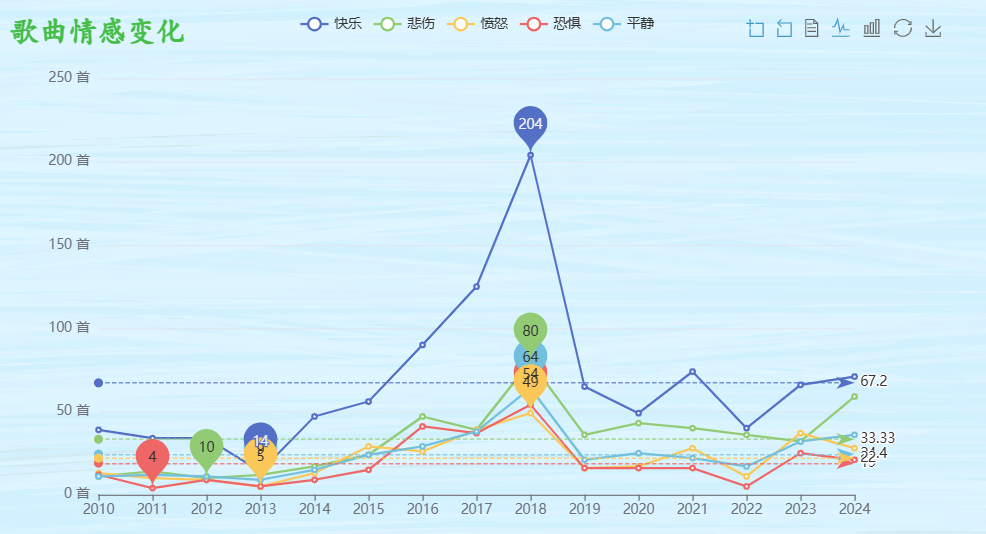


图5.3 歌曲情感变化图

系统还通过力导向图展示歌手之间的合作关系，每个节点代表一个歌手，连线表示合作。通过节点的大小和颜色，用户能够识别出重要歌手和其合作网络。在人物关系图中，用户可点击歌手节点查看其合作网络，系统支持“展示所有节点”功能，以全面展示歌手的合作关系。点击节点时会动态更新相关歌手的合作关系，增强了交互性。



图5.4 歌手人物关系图

此外，地图可视化展示了不同省份歌曲发布量的分布，用户可以直观查看各地区的歌曲发布情况，各地区歌曲的风格特点，并通过点击省份查看该地区的歌手排名。此外，排行榜模块通过动态渲染和实时更新，帮助用户查看不同省份歌手的歌曲发布情况。

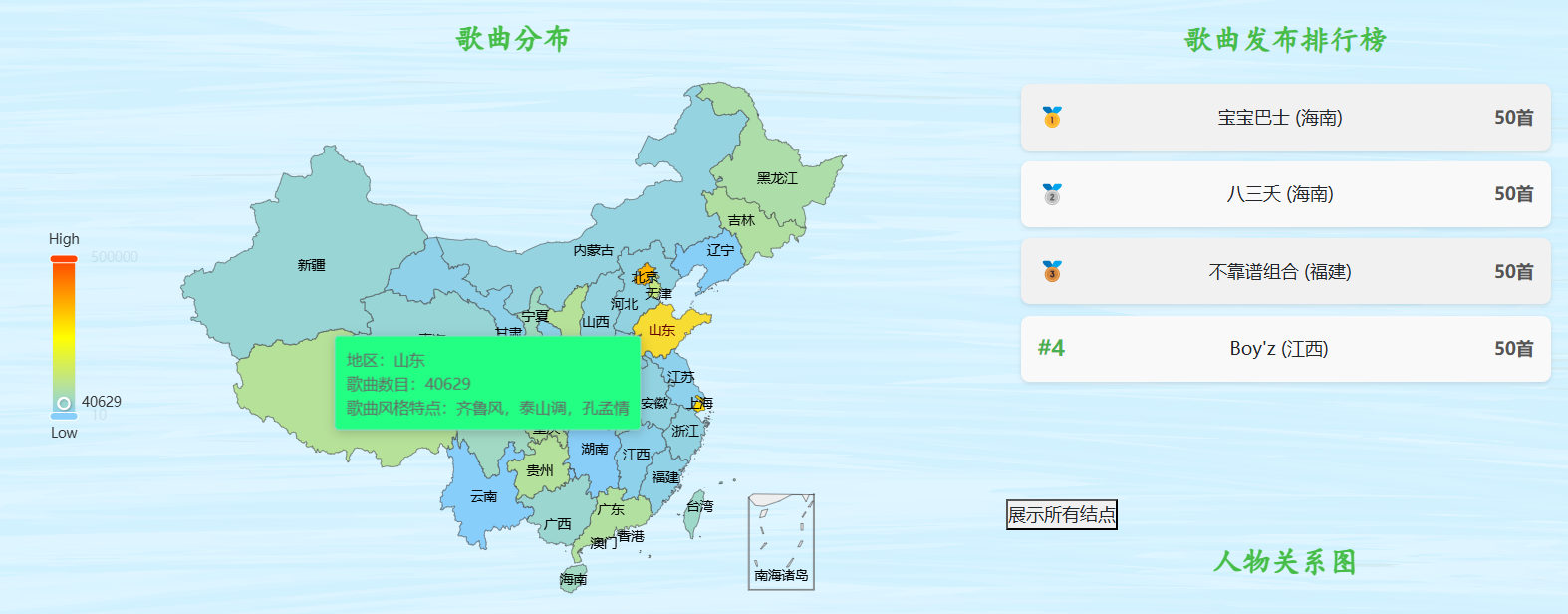


图5.4地图各地区歌曲分布及歌曲发布排行榜图

系统使用响应式布局确保页面在不同设备上的良好显示，无论是在PC端还是移动端，用户都能够轻松访问和互动。使用Bootstrap框架优化了界面的适配性，使得所有的图表和交互组件能够自动调整，以适应屏幕大小的变化。

**六、可视化结果分析**

1.不同年代词云结果分析

在可视化大屏的左上角，我们显示的不同年代的关键词组成的一个3D立体词云，在这里我们可以通过切换下方的下拉列表选择从1980到2024期间每个年代的时间，从而达到词云切换的效果。我们通过观察歌词内容随时间的变化，反映社会价值观、生活方式的变迁。从关键词的数量上来说，在这里我们可以看到80,90年代这时候的关键词比较少，到了00、10、20年代之后关键词逐渐变多，这也表明随着我们国家的不断发展，人民的生活水平不断提高，歌曲也在不断向着多元化发展。从词云的内容上来说，我们可以看到每个年代都有爱国这个关键词，这也展示了我们每个人对于国家的热爱。同时我们也可以看到2015-2024年国风、古风他们的频繁出现，越来越多的人喜欢中国式的音乐，也反映出了当代我们的文化自信，道路自信。同时我们也能看到20世纪80年代的歌曲可能以浪漫和悲伤为主要情感，而2000年以后，随着社会变革，歌曲中表达的情感可能更多地偏向于积极和励志。



图6.1 不同年代歌曲对比图

2.常见歌词类别结果分析

在可视化大屏的左下角，这是一个南丁格尔玫瑰图，为了挖掘歌词中常见的主题和题材，我们通过对歌词的分类和聚类进行数据处理，图中每个扇形代表一个类别，扇形的大小表示该类别在总体中的百分比。这个图揭示了歌词中不同主题的普遍性，其中“人”以22.79%的占比成为最常提及的主题，紧随其后的是“梦”和“时候”，分别占9.54%和9.04%，而“心”和“爱情”也较为常见，占比11.38%和13%。相对较少被引用的类别包括“眼泪”、“时间”、“感觉”等，而“朋友”、“天空”和“手”则更为罕见。这些数据反映出歌词内容多聚焦于情感表达、人际关系及个人体验，同时涉及自然元素和抽象概念，为我们提供了对歌词创作趋势和主题偏好的深刻洞察。

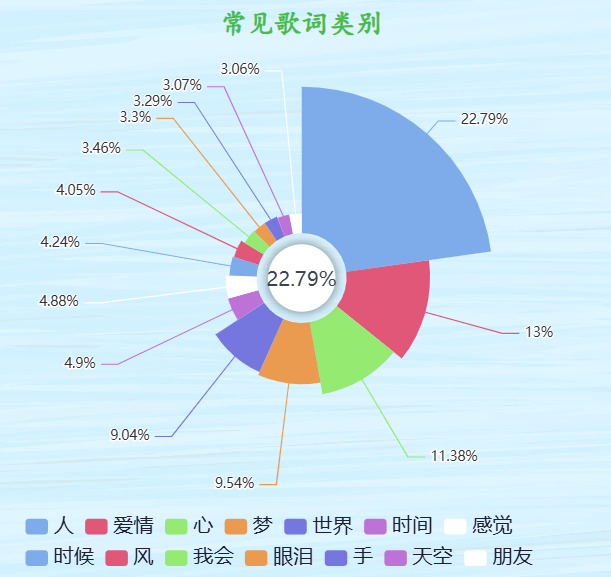


图6.2 常见歌词类别占比图

3.常见歌词类别分类结果分析

在可视化大屏的下方中间部分是一个折线图，用来观察歌词内容随时间的变化，反映社会价值观、生活方式的变迁。通过查阅资料，我们将歌词的情感主要分为五个类别，分别为：快乐、悲伤、愤怒、恐惧、平静。用户可自行选择折线图，柱状图，纯数据以及放大某些年份的数据来观察，随着时间的变迁，歌曲的情感变化情况。同时，我们也设置了最高值和最低值来方便用户去观察不同年代的歌曲请改变化趋势。在这里我们可以观察到不同情感在各个年代的表现差异。图中将歌词情感分为五大类别：快乐、悲伤、愤怒、恐惧和平静，并通过折线图的形式，直观地呈现了这些情感随时间的变迁。特别是“快乐”情感在2018年达到了204首的高峰，而“悲伤”情感在2016年也有显著的增长，达到80首。这些情感的波动不仅反映了歌词内容的变化，也可能映射出社会价值观和生活方式的演进。用户可以通过选择不同的图表类型，如柱状图或纯数据视图，以及放大特定年份的数据，来深入观察和分析情感变化的趋势。此外，图表中还设置了最高值和最低值标记，帮助用户更清晰地识别不同年代歌曲情感的显著变化。

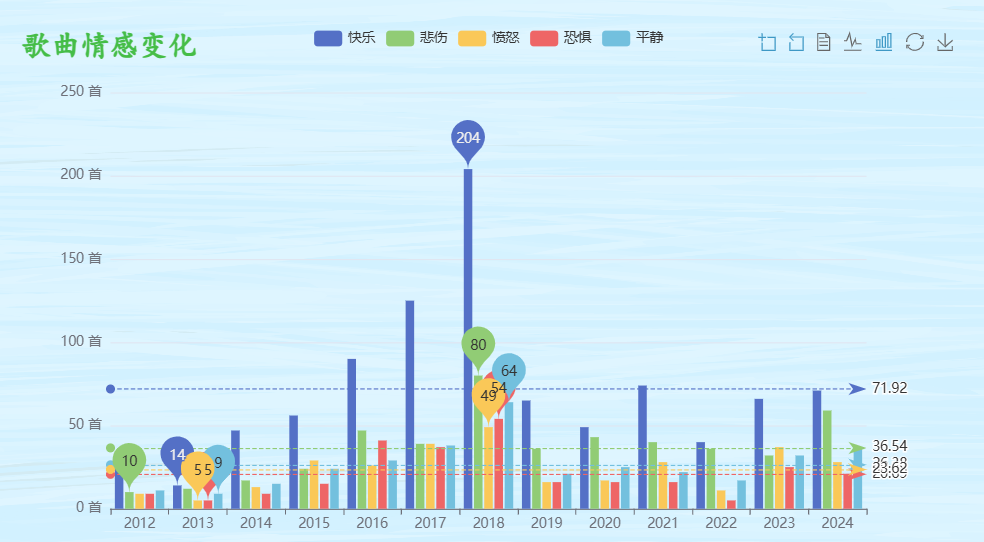


图6.3 歌曲情感变化图

4.人物关系图结果分析

在可视化大屏的右下角，我们通过力导向图展示歌手之间的合作关系，每个节点代表一个歌手，连线表示合作。用户可以点击任一歌手节点，查看与该歌手有合作关系的其他歌手，探索其合作网络。在这里歌手之间的合作不仅限于共同演唱歌曲，还可能涉及词曲创作、音乐制作等多个方面。通过分析歌手的合作关系，能够揭示出音乐产业内的网络结构，以及哪些歌手与其他人合作较为频繁，哪些歌手可能是行业中的“核心”人物。在这里我们以林俊杰和刘若英两位歌手为例进行分析，可以看到，林俊杰合作的歌手较多，而刘若英更加偏向于独自创作或在歌曲方面合作较少等信息。



图6.4 不同歌手的合作关系对比图

5.地区与歌曲关系结果分析

由于中文歌曲在不同地区的歌词创作上展现出明显的地域文化差异。某些地区的歌曲往往带有特定的方言或文化色彩，这种文化差异可以通过对歌词的地域分析来揭示。所以在可视化大屏的中间我们设置了地图，展示不同地区的歌曲发布量和文化特点。例如，北方的歌曲可能更注重情感表达和直接性，而南方的歌曲可能更注重意境和细腻感。再比如，内蒙古的歌曲特色为草原风、牧歌情，蒙古调。通过区域文化分析，用户可以发现，不同地区的歌手在创作歌词时，往往会融入本地的文化元素。例如，南方歌手的歌词中可能频繁出现与“江南”或“山水”相关的词汇，而北方歌手的歌词可能更多涉及“奋斗”或“英雄”之类的主题。这种分析还可以帮助音乐产业从业者更好地理解不同区域的音乐偏好，从而调整市场策略，推出更符合当地文化特色的音乐作品。同时在大屏的左上角我们，我们设置了歌曲发布排行榜，在这里我们可以通过点击地图上的某一个地区，来显示该地区歌手的发布歌曲的排行榜，在这里我们就可以看到哪位歌手的“产量”比较高。



图6.5 地图各地区歌曲分布及歌曲发布排行榜图

1. **讨论与总结**

本系统通过可视化技术展示中文歌词数据，为用户提供了多维度的分析视角。在可视化设计方面，词云图、情感分析折线图、歌手合作关系图以及区域分布地图等图表，成功将复杂的歌词数据转化为易于理解的视觉信息。通过这些图表，用户能够直观地把握歌词主题的变化、社会情感的变迁以及歌手间的合作网络，从而深入理解中文歌曲在不同时期的文化背景和社会影响。

系统的交互设计进一步提升了用户体验。用户可以通过下拉框选择不同时间段，探索各时期歌词的情感变化与主题演变，动态更新的排行榜和人物关系图让用户能够灵活查看不同维度的数据。特别是在人物关系图中，通过点击节点查看合作网络的功能，使得用户能够深入探索歌手间的合作模式和网络结构。这种互动性不仅提高了数据的可探索性，还帮助用户更好地理解数据背后的复杂性。

然而，系统的可视化展示依然存在一定的局限性。例如，随着数据量的增大，某些图表可能出现加载速度较慢的问题，尤其是在展示较复杂的网络关系时。此外，词云图和情感分析折线图虽然能够展示总体趋势，但在面对极为细化的情感变化时，可能无法精准地反映个别歌曲的情感特征。因此，未来可以考虑优化数据加载与渲染速度，并引入更细粒度的数据分析与可视化方式，以更准确地呈现复杂的情感分析结果。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 小组名称 | 黑项目歌词 | | | | |
| 小组人员 | 张丽平 | 刘子硕 | 徐国庆 | 刘恩铭 | 魏溪语 |
| 贡献度 | 25% | 20% | 25% | 15% | 15% |
| 项目主题 | 歌曲的视觉语言 | | | | |