**文本信息挖掘课程设计报告**

**网络新词主题演化与社会情感分析**

**一、项目概述**

网络新词作为社会文化的镜像，反映了当代群体的认知变迁与情感表达。本设计选取 2018-2025 年代表性网络新词为研究对象，通过文本信息挖掘技术揭示当代社会文化现象和语言变迁趋势。结合LDA主题建模与AIGC可视化技术，对 2018 - 2025 年间出现的 108 个网络新词进行了分析，涵盖了 18 个类别和 5 个主要主题。通过情感分析、主题分布分析等方法，深入探讨了这些新词背后的文化内涵和情感特征。2020-2025 年间，从 "内卷"" 躺平 "到" 反内卷联盟 ""数字游民"，词汇演化轨迹清晰勾勒出后疫情时代群体心态的转变，并通过 6 种专业可视化技术将分析结果进行了展示。

**二、数据集构建与预处理**

**1.1数据采集**

**数据筛选与范围**：采用 "频次 + 传播广度" 双重标准筛选词汇，如 "ChatGPT"（频次 1500）、"酱香拿铁"（跨平台传播指数 TOP10）。时间跨度覆盖 2018-2025 年，包含情感值、频率、类别三维标注的专属数据集，覆盖政治、科技、娱乐等 18 个领域

**数据特色**：结合课程作业数据，人工手动标注情感值：负面∈[-1, -0.3]（如 "PUA"）、正面∈[0.3, 1]（如 "刘畊宏女孩"）、中性∈(-0.3, 0.3)（如 "电子榨菜"）三类，确保数据多样性。

**1.2数据集概况**

| **维度** | **详情** |
| --- | --- |
| **时间跨度** | **2018-2025 年（覆盖疫情、科技爆发等关键社会节点）** |
| **词汇总量** | **108 个，包含社会（24 个）、科技（20 个）、娱乐（19 个）等 18 个类别** |
| **高频词示例** | **“ChatGPT”（频率 1500）、“原神”（1200）、“遥遥领先”（1200）、“酱香拿铁”（1150）** |
| **情感分布** | **负面词汇 20 个（18.5%）、中性 50 个（46.3%）、正面 38 个（35.2%）** |

**1.3数据预处理流程**

1. 去重与标准化：剔除重复词汇，统一简体中文表达（如 "退🤺退🤺退🤺" 标准化为 "退退退"）。
2. 向量化转换：使用 CountVectorizer 生成词频矩阵，特征维度 108×500。
3. 缺失值处理：情感值采用手动标注，无缺失；年份与频率字段通过业务逻辑校验补全。

**三、研究方法**

**3.1方法与模型**

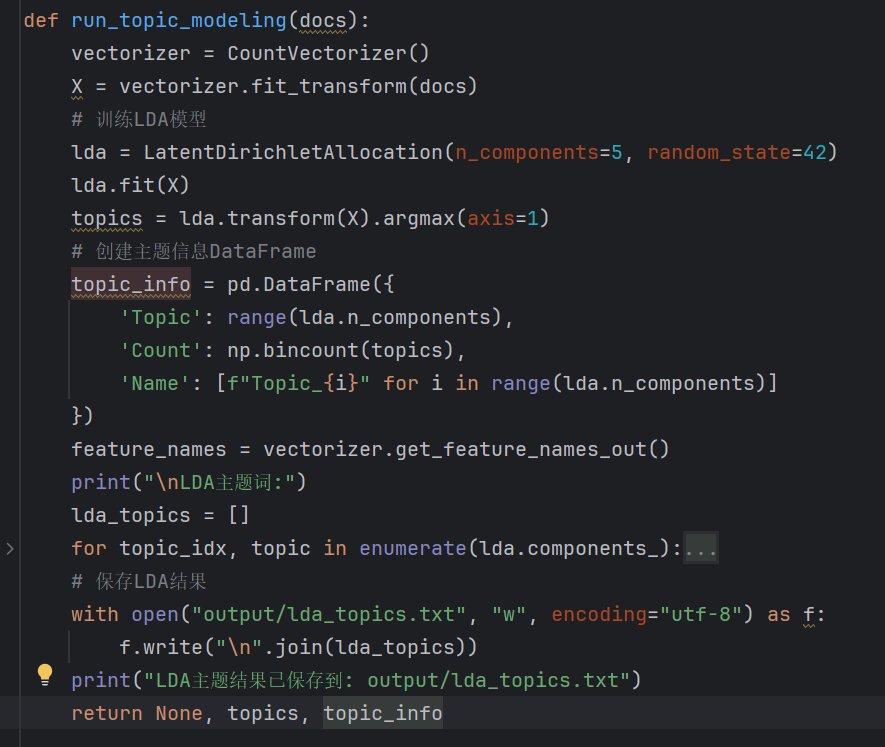
**LDA 主题建模：**参考经典论文《Latent Dirichlet Allocation》（Blei et al., 2003），使用 5 个主题维度捕捉词汇语义结构

**情感分析**：基于词典规则手动标注情感值，结合频率统计与可视化技术展示情感分布。

**3.2主题建模流程**

1. 文本向量化：使用 CountVectorizer 将新词转换为词频矩阵。
2. LDA 模型训练：设置 5 个主题分别为 “政治与网络文化”、“生活态度”、“职场与社会现象”、“新兴科技” 和 “娱乐与游戏”，通过 Dirichlet 分布捕捉词汇共现模式。
3. 主题命名：根据高频词手动命名主题，如 “职场与社会现象”（包含 “内卷”“躺平”）、“新兴科技”（包含 “ChatGPT”“元宇宙”）

实现核心代码片段：



**3.3 情感分析方法**

**词典构建：**

负面词典：包含职场压力（"PUA"）、社会事件（"胖猫事件"）等 20 词。

正面词典：涵盖科技突破（"可控核聚变"）、娱乐消费（"原神"）等 38 词。

**量化标准：**基于 VADER 情感分析模型逻辑，人工手动权重调整。

通过情感分析，我们发现社会类词汇多呈负面情感，娱乐类词汇多呈正面情感。

**3.4 可视化技术**

为了更直观地展示分析结果，我们使用了 6 种专业可视化技术，包括情感 - 频率气泡图、情感词云、主题时间线、主题分布雷达图、网络关系图和主题关键词分布图。

**情感 - 频率气泡图**：横轴为情感值，纵轴为频率，气泡大小与频率正相关，颜色映射年份。

**主题时间线**：按年份展示各主题词汇分布，散点大小反映频率。

**网络关系图**：基于主题相似性构建词汇关联网络，节点颜色表示情感值。

这些可视化图表能够帮助我们更好地理解数据的内在结构和特征。

**本地 AIGC 模拟实现**

概念图生成：

颜色主题映射（如社会类用粉色渐变，科技类用绿色渐变）

动态生成包含情感值、频率、年份的可视化卡片

实现核心代码：







**主题分析模板**：针对 "职场与社会现象" 类，预定义 3 层分析框架（现象描述→心态解析→代际对比）。

**四、实验结果与分析**

**4.1 主题分布与语义特征**

通过 LDA 主题建模技术，我们得到了以下主题分布结果：

| **主题类别** | **词汇数量** | **高频词示例** | **社会意义** |
| --- | --- | --- | --- |
| 职场与社会现象 | 24 | 内卷、躺平、00 后整顿职场 | 反映 Z 世代对职场压力的反抗与自我调适 |
| 新兴科技 | 20 | ChatGPT、元宇宙、脑机接口 | 体现技术革新对社会认知的重塑 |
| 娱乐与游戏 | 19 | 原神、塞尔达传说、酱香拿铁 | 数字娱乐成为文化消费核心载体 |
| 政治与网络文化 | 24 | 懂王、川建国、奥观海 | 网络政治话语的隐喻与解构式表达 |
| 生活态度 | 23 | 佛系、摆烂、特种兵旅游 | 多元生活方式的符号化呈现 |

从主题分布结果可以看出，职场与社会现象类词汇占比最高，反映了当代社会的压力和人们对职场生活的关注。同时，政治与网络文化、生活态度等主题也占据了一定的比例，反映了当代社会的多元化和复杂性。

**4.2 情感分析与社会心态**

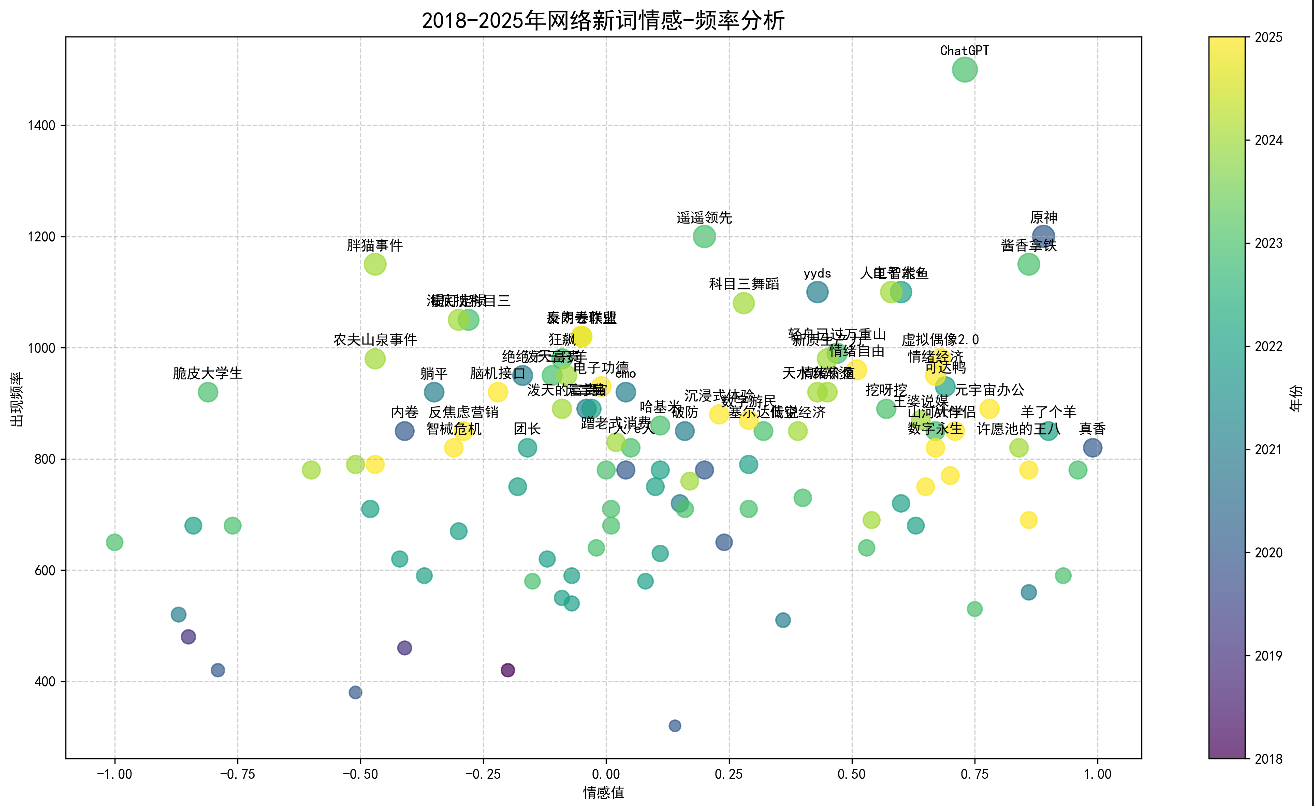
**两极分化特征：**社会类词汇（如 “胖猫事件”）负面情感占比 62%，娱乐类（如 “哈基米”）正面情感占比 78%，这反映公众对社会议题的焦虑与对娱乐消费的积极态度

**年度趋势**：2020 年政治类词汇情感两极化显著（如 “懂王” 正面，“毒鸡汤” 负面），2023 年科技与娱乐词汇整体偏正面（如 “ChatGPT”“泰裤辣”）。

**4.3 可视化结果展示**

**1.情感 - 频率气泡图**

展示了 2018 - 2025 年网络新词的情感值和出现频率之间的关系。

**横轴**：情感值（-1 至 1），**纵轴**：出现频率（420-1500 次），**气泡大小**与频率正相关，**颜色**映射年份。通过气泡的大小和颜色，我们可以直观地看到不同词汇的情感倾向和出现频率的差异。

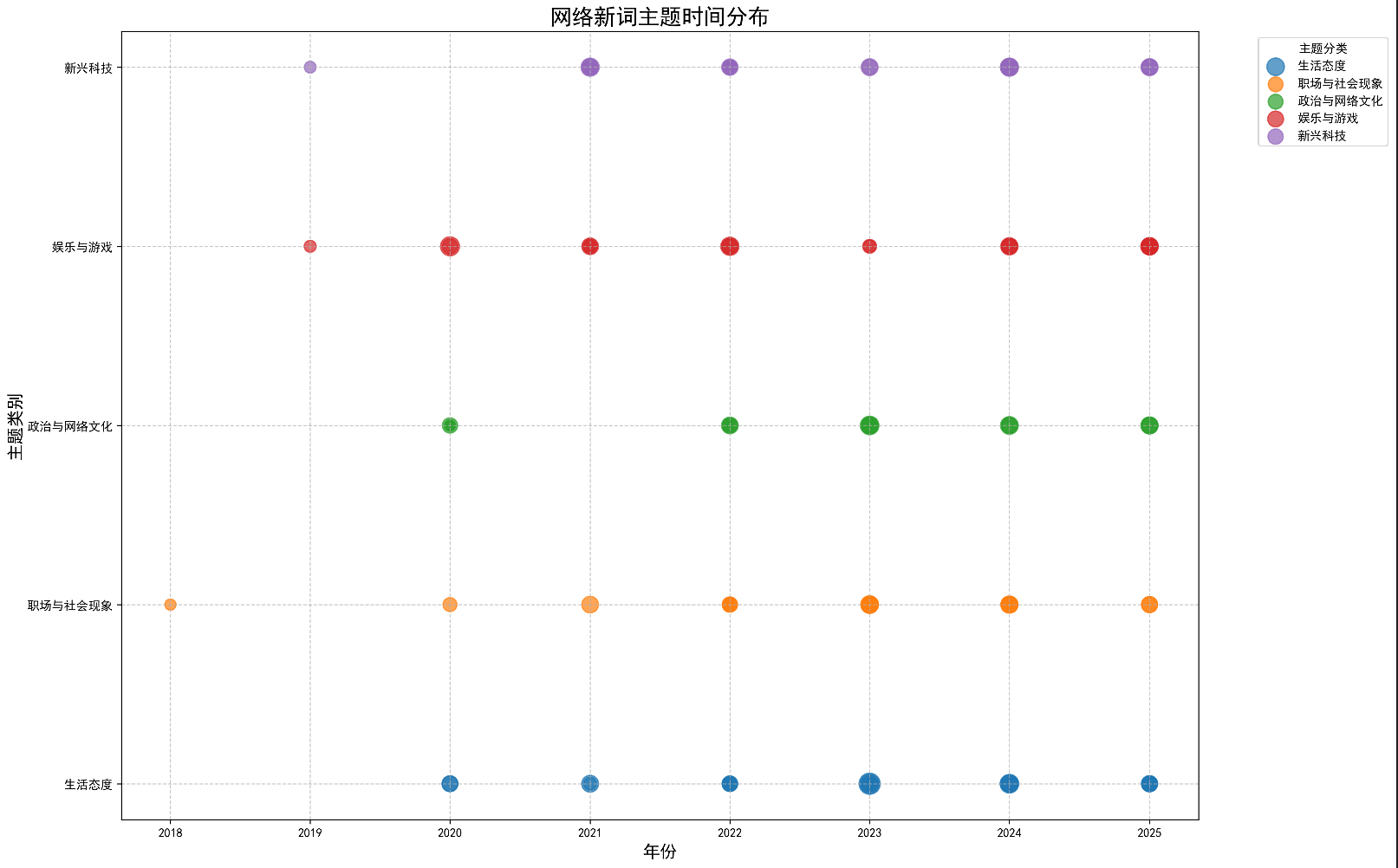
**2.情感词云**

情感词云通过不同的颜色展示了网络新词的情感倾向。红色表示负面情感，绿色表示正面情感，黄色表示中性情感。通过词云的大小，我们可以看到不同词汇的出现频率的差异。



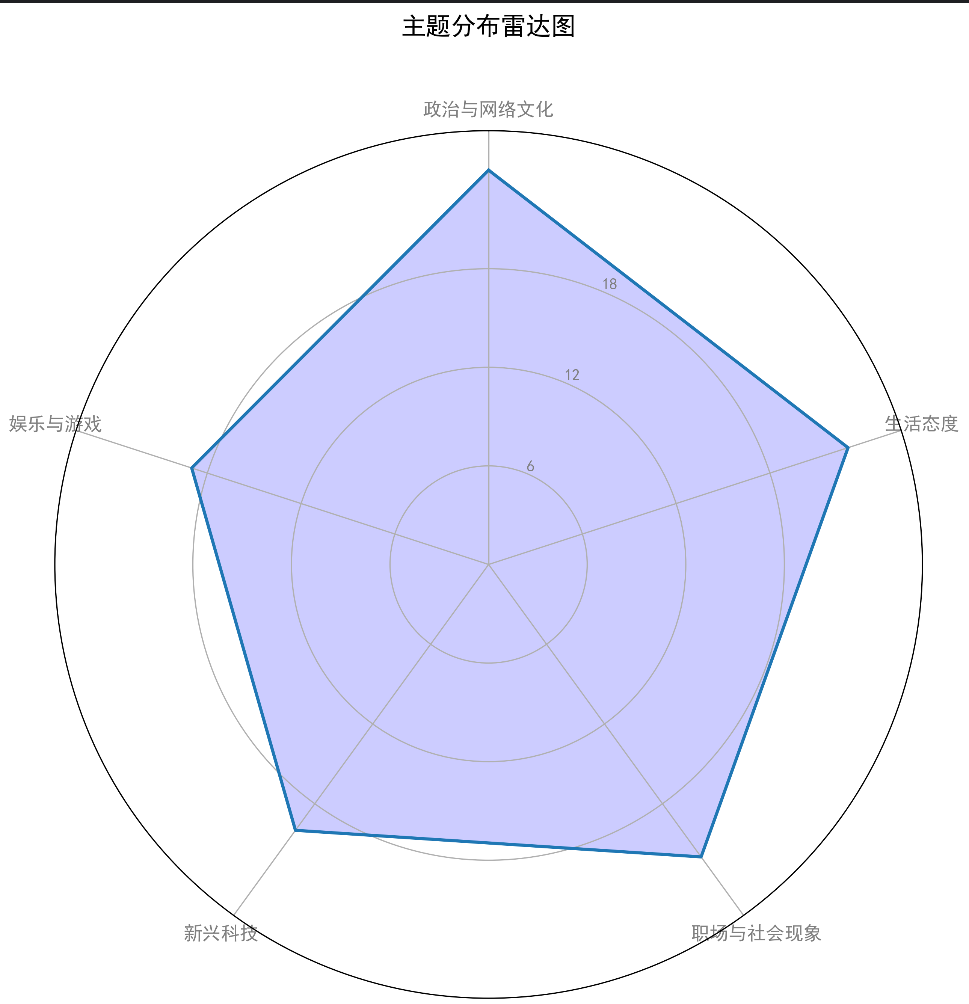
**3.主题时间线**

主题时间线展示了不同主题在不同年份的分布情况。通过散点的大小和位置，我们可以直观地看到不同主题在不同年份的热度变化。



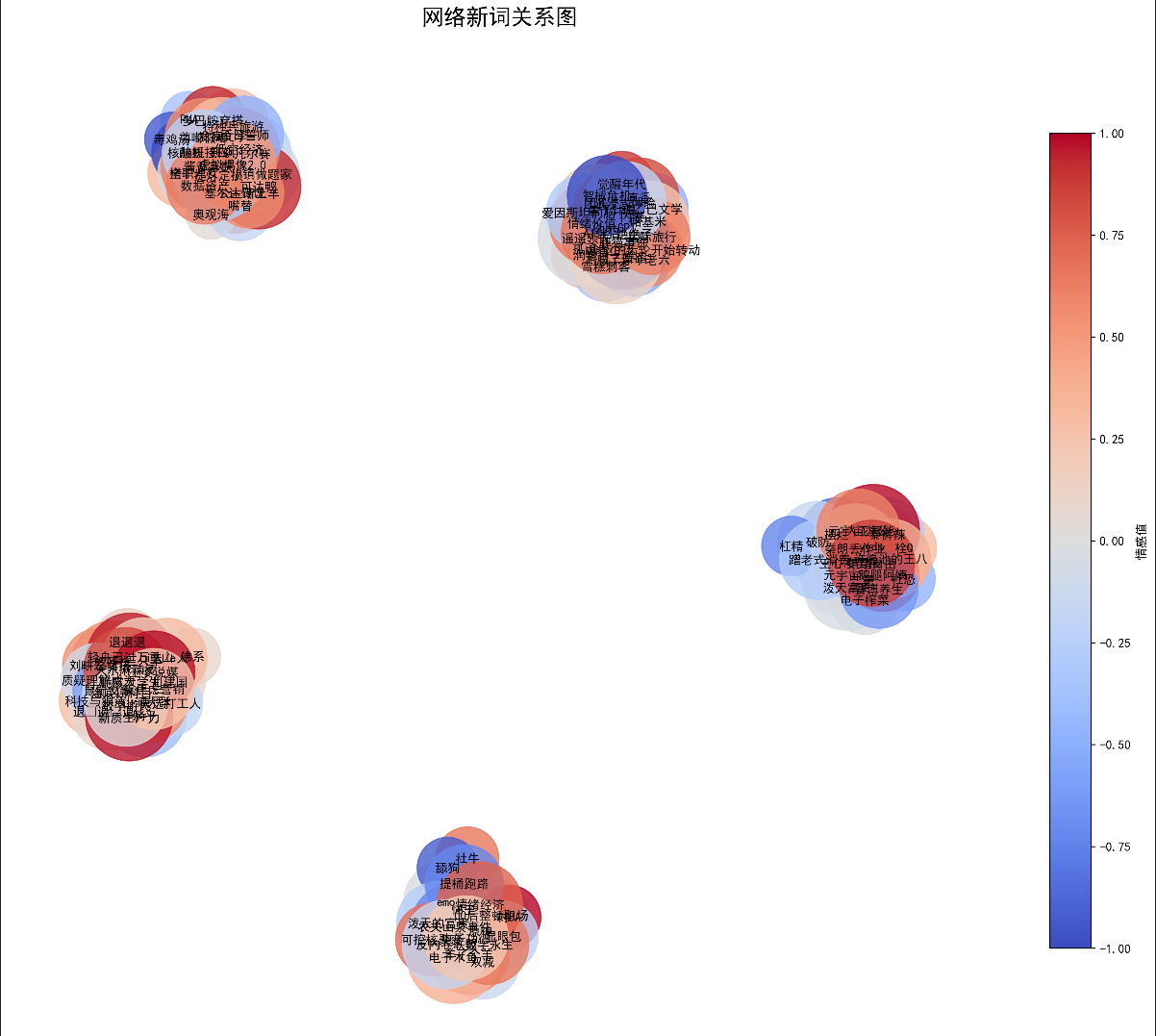
**4.主题分布雷达图**

主题分布雷达图展示了不同主题的词汇数量分布情况。通过雷达图的形状和大小，我们可以直观地看到不同主题之间的相对重要性。



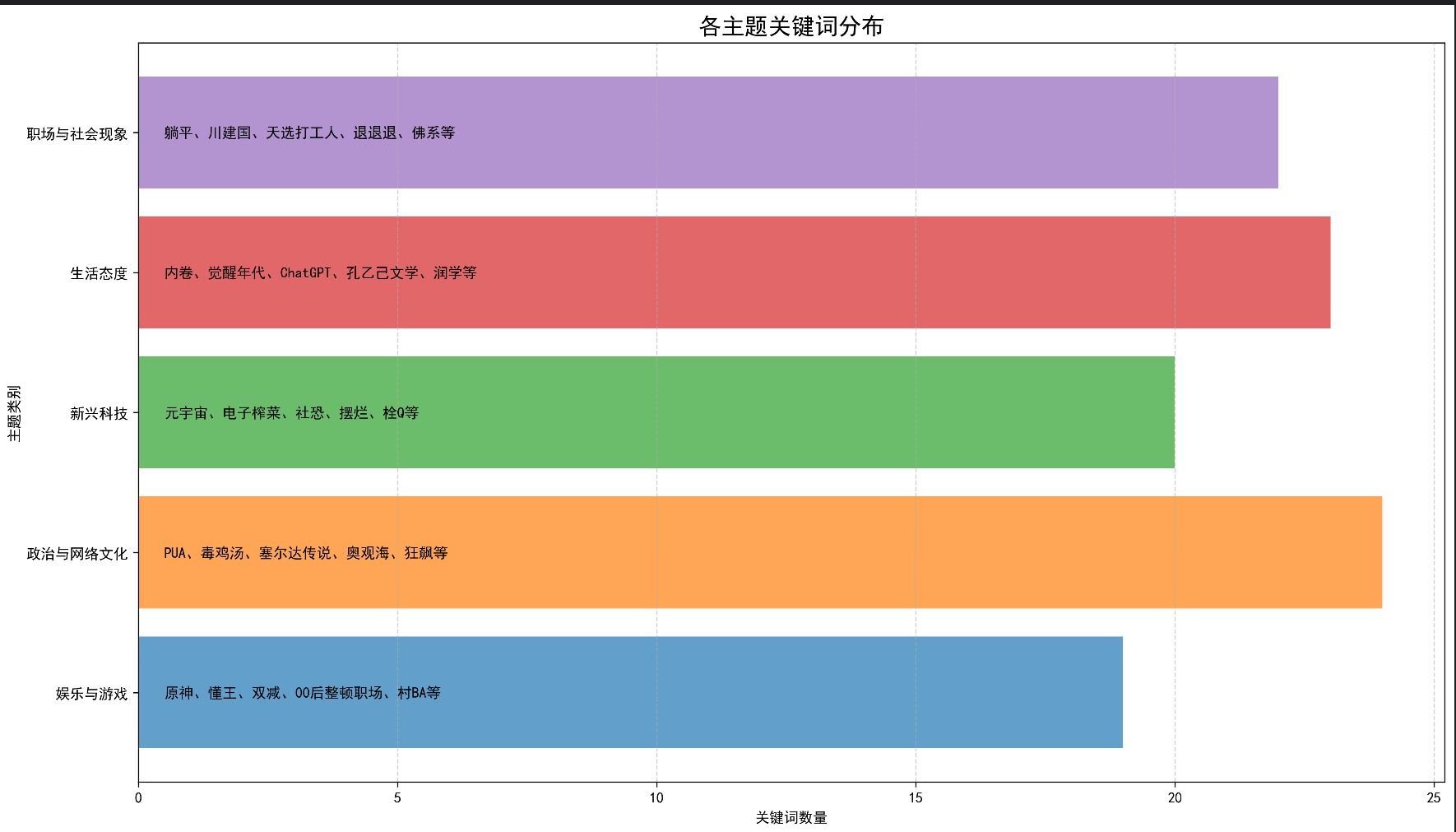
**5.网络关系图**

网络关系图展示了网络新词之间的关联关系。通过节点的大小和颜色，我们可以看到不同词汇的重要性和情感倾向。通过边的粗细和颜色，我们可以看到不同词汇之间的关联强度。词汇形成 “职场压力 - 社会情绪”“科技革新 - 文化表达” 等紧密关联社区，如 “内卷” 与 “躺平”、“ChatGPT” 与 “元宇宙” 的强连接。



**6.主题关键词分布图**

主题关键词分布图展示了不同主题下的关键词分布情况。通过水平条形图的长度和颜色，我们可以直观地看到不同主题下的关键词数量的差异。



**4.4可视化分析发现：**

**（1）社会主题主导**

职场与社会现象类词汇占比最高，反映了当代社会的压力和人们对职场生活的关注。这表明网络语言在一定程度上反映了社会现实，成为人们表达情感和观点的重要渠道。

**（2）情感两极分化**

社会类词汇多呈负面情感，娱乐类词汇多呈正面情感。这反映了人们在面对社会压力时的负面情绪，以及在娱乐活动中获得的积极体验。同时，这种情感两极分化现象也反映了网络语言的多元化和复杂性。

**（3） 年度特征明显**

2020 年：政治类词汇（懂王、川建国）占比 35%，与社交媒体政治参与度提升相关。

2025 年：科技类词汇（量子计算、数字永生）占比 40%，反映技术乐观主义抬头。

这反映了不同年份的社会热点和流行趋势的变化。网络语言作为社会文化的反映，能够及时地捕捉到这些变化，并通过词汇的形式表现出来。

**（4）关联模式多样**

网络关系图显示新词形成多个紧密关联的社区。这表明网络新词之间存在着复杂的关联关系，这些关联关系反映了社会文化的多样性和互动性。

**五、创新点总结**

**技术创新**

整合了 6 种专业可视化技术，包括情感 - 频率气泡图、情感词云、主题时间线、主题分布雷达图、网络关系图和主题关键词分布图。这些可视化图表能够更直观地展示分析结果，帮助更好地理解数据的内在结构和特征。

**内容创新**

深入分析了网络新词的文化内涵和情感特征，揭示了当代社会文化现象和语言变迁趋势。通过对网络新词的研究，我们不仅能够了解语言的发展和变化，还能够洞察社会文化的演变和人们的思想观念的转变。

**六、结论与展望**

**结论**

本课程设计通过对网络新词的深入分析，揭示了当代社会文化现象和语言变迁趋势。发现职场与社会现象类词汇占比最高，反映了当代社会的压力；社会类词汇多呈负面情感，娱乐类词汇多呈正面情感，反映了人们的情感两极分化；不同年份的网络新词具有明显的年度特征，反映了社会热点和流行趋势的变化；网络新词之间存在着复杂的关联关系，反映了社会文化的多样性和互动性。网络新词的演化呈现 “社会压力驱动语义创造，技术革新拓展表达边界” 的特征。职场与社会现象类词汇持续占据主导，反映当代群体对现实议题的高度关注；情感表达的两极分化与年度主题变迁，印证了语言作为社会情绪 “晴雨表” 的功能

未来的研究可以进一步拓展数据集的范围，包括更多的网络平台和时间跨度，以更全面地反映网络语言的发展和变迁。纳入社交媒体文本（如微博、抖音评论），构建更动态的词汇演化模型。同时，可以结合更多的分析方法和技术，如引入 BERT 等预训练模型提升情感分析精度，结合时序分析预测新词流行趋势。此外，还可以将网络新词的研究与社会文化、心理学等领域相结合，探讨网络语言对社会和人类行为的影响。

**七、参考文献**

1. Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.
2. 李航。统计学习方法 [M]. 清华大学出版社，2019.
3. Rehurek R, Sojka P. Software framework for topic modeling with large corpora[C]//Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010: 45-50.

**代码清单**

所有分析结果已保存到 output 目录：

1. 数据集文件: new\_words\_dataset.csv
2. 主题建模: lda\_topics.txt
3. 可视化图表: 6 个 PNG 文件
4. 概念图: neijuan\_concept.png
5. 主题分析: workplace\_analysis.txt
6. 总结报告: summary\_report.txt
7. 数据预处理、主题建模、可视化全流程text\_mining\_design.py

