**本科生课程设计报告**

**《基于LDA的中国科技创新类新闻数据可视化分析》**

**课程名称《大数据可视化技术》**

**姓 名 李毅**

**学 号 2295141020**

**学 院 工学院**

**专 业 数据科学与大数据技术**

**课程性质 专业必修课**

**任课教师 黄诚惕**

**注意事项：**

1. 课程设计报告要有题目、摘要、关键词、正文及参考文献。摘要100-200字左右，关键词3～5个；参考文献不少于5篇；论文字数不少于1000字；要有自己的见解与心得；附查重报告，且查重率不高于20%；
2. 课程论文应符合一般学术规范，严禁抄袭或应付了事；
3. 课程论文均须用A4纸双面打印(查重报告附后)，加装本封面，左侧装订。

|  |  |  |  |
| --- | --- | --- | --- |
| 教师评阅意见： | | | |
| 论文成绩 |  | 评阅日期 |  |

基于LDA的中国科技创新类新闻数据可视化分析

华侨大学工学院

数据科学与大数据技术专业 2026届 李毅 2295141020

摘要 本研究采用LDA主题建模方法对中国科技创新领域的新闻报道进行系统性分析，通过构建多维度文本分析框架，深入挖掘中国科技发展的核心议题。研究首先对原始新闻数据进行清洗、分词和向量化处理，继而运用TF-IDF算法提取关键特征词，并构建词语共现矩阵和语义关联网络。在此基础上，通过LDA模型识别出教育科技创新、月球探测研究和人工智能应用三大核心主题，结合情感分析评估各主题的舆论倾向。该研究不仅量化呈现了中国科技创新的重点领域和发展态势，其创新的可视化分析方法和研究结论也为科技政策制定和产业战略研究提供了重要的数据参考和决策依据。下面是关于本研究的GitHub链接。

https://github.com/01bigdata/HQU\_bigdataVisualization

关键词 LDA模型 数据可视化 科技创新 中国经济网

**Visualization Analysis of China's Science and Technology Innovation News Based on LDA**

Li Yi

Data Science and Big Data Technology,295141020,2026

College of Engineering, Huaqiao University

**Abstract** This study employs the LDA topic modeling method to systematically analyze news reports in the field of China's science and technology innovation. By constructing a multi-dimensional text analysis framework, it deeply explores the core issues of China's technological development. The research begins with cleaning, segmenting, and vectorizing the raw news data, followed by using the TF-IDF algorithm to extract key feature words and constructing a word co-occurrence matrix and semantic association network. On this basis, the LDA model identifies three core themes: education and technology innovation, lunar exploration research, and artificial intelligence applications. Sentiment analysis is also conducted to evaluate public opinion trends for each theme. The study not only quantitatively presents the key areas and development trends of China's science and technology innovation but also provides important data references and decision-making support for science and technology policy formulation and industrial strategy research through its innovative visualization analysis methods and findings.

**Keywords** LDA model Data visualization Science and technology innovation China Economic Net

目录

[第1章 绪论 5](#_Toc201281873)

[第二章 数据采集 6](#_Toc201281874)

[2.1数据来源分析 6](#_Toc201281875)

[2.2 selenium框架介绍 7](#_Toc201281876)

[第三章 数据处理与分析 9](#_Toc201281877)

[3.1新闻数据清洗与预处理 9](#_Toc201281878)

[3.2新闻数据清洗与预处理 11](#_Toc201281879)

[3.3词条共现矩阵 14](#_Toc201281880)

[3.4基于相关性的网络关系图 15](#_Toc201281881)

[3.5情感分析 19](#_Toc201281882)

[第四章 可视化交互设计 21](#_Toc201281883)

[4.1 pyecharts和dash介绍 21](#_Toc201281884)

[4.2 可视化思路 21](#_Toc201281885)

[4.3 可视化展示 22](#_Toc201281886)

[第五章 结论 25](#_Toc201281887)

[参考文献 26](#_Toc201281888)

# 第1章 绪论

在中华民族伟大复兴的战略全局中，科技创新与教育发展是双轮驱动、不可或缺的关键力量。习近平总书记深刻指出：“科技是国家强盛的基石，创新是民族进步的灵魂。”他强调，要深入实施科教兴国战略、人才强国战略、创新驱动发展战略，把科技的命脉牢牢掌握在自己手中，为实现中华民族伟大复兴提供强大科技支撑。

|  |
| --- |
|  |
| 图1.1习总书记在二十大上讲话 |

科教兴国，是立足当下、决胜未来的长远之计。教育是提升国民素质、培养创新人才的根本途径，而科技则是在全球竞争中赢得主动的关键。正如习近平总书记所强调的，要“坚持面向世界科技前沿、面向经济主战场、面向国家重大需求、面向人民生命健康”，充分发挥科技第一生产力、人才第一资源、创新第一动力的作用，以科技创新赋能高质量发展，以教育进步夯实国家根基。

在实现中华民族伟大复兴的新征程上，我们要深刻领会习近平总书记关于科技创新和科教兴国的重要论述，坚定不移地走自主创新道路，全面提高教育质量，为建设社会主义现代化强国、实现中华民族伟大复兴的中国梦而不懈奋斗。

# 第二章 数据采集

## 2.1数据来源分析

为深入分析中国科技发展趋势与战略重点，本研究选取中国经济网科技频道作为数据来源，抓取了自2024年5月24日至2025年6月19日期间，近一年时间跨度内的科技新闻报道。通过对海量文本数据的挖掘与分析，旨在探寻这段时期内中国科技领域的核心议题、关键发展方向以及国家战略部署，从而把握科技创新对国家发展的重要驱动作用。

|  |
| --- |
|  |
| 图2.1中国经济网科技频道网页 |

对中国经济网科技频道的网页结构分析（如图2.1所示）表明，该页面布局清晰，元素排列规整。新闻标题和其发布时间均可通过明确的XPath路径进行精确定位，同时页面底部提供了标准的分页导航功能，为自动化数据采集提供了便利。

然而，在数据解析阶段发现一个关键问题：网页上显示的发布时间格式仅包含月、日和时刻（如“05/22 15:30”），却缺失了年份信息。这种时间格式的模糊性会引发严重的数据完整性问题。若直接为所有抓取条目赋予当前年份，一旦采集范围跨越新年，例如从今年1月采集至去年12月，便会导致年份归属错误，使得数据丧失时序的准确性。

为解决此问题，必须依据新闻发布的内在逻辑进行数据校正。鉴于新闻列表通常按时间倒序排列，我们可以利用抓取顺序来推断和修正年份。具体策略是：在翻页抓取过程中，一旦监测到月份发生“逆行”（例如，从“1月”跳转至“12月”），即可判定已跨入前一个年份，从而为后续数据赋予正确的年份，确保最终数据集的时序完整无误。

## 2.2 selenium框架介绍

Selenium是一个功能强大的开源自动化测试框架，其核心目标是为Web应用程序提供一套完整的自动化解决方案。它并非单一的工具，而是一个由多种工具组成的软件套件，主要包括Selenium WebDriver、Selenium IDE和Selenium Grid。Selenium的根本价值在于，它能够通过编程方式直接驱动主流的Web浏览器（如Chrome、Firefox、Safari等），模拟真实用户的各种操作行为，例如点击链接、填写表单、页面跳转和验证页面元素等。由于其跨平台、跨浏览器以及支持多种编程语言（如Java、Python、C#、JavaScript）的特性，Selenium已经成为Web自动化测试领域事实上的行业标准，为保证Web应用在不同环境下的功能一致性和稳定性提供了不可或缺的技术支持。

Selenium的核心组件是Selenium WebDriver，它充当着测试脚本与浏览器之间的通信桥梁。其工作原理是通过一个标准的W3C协议，为每种浏览器提供一个专属的驱动程序（Driver）。测试脚本通过调用WebDriver的API发出指令，这些指令被驱动程序解析后，再转换成浏览器能够理解的原生命令来执行，从而实现对浏览器的精准控制。这种直接与浏览器内核交互的模式，确保了自动化操作的高效性和稳定性。在实践中，开发与测试团队利用Selenium构建回归测试套件，执行兼容性测试，或用于自动化处理重复性的Web操作任务（如数据抓取或信息填报）。然而，Selenium本身专注于浏览器行为的驱动，它不包含测试用例管理、断言库或报告生成等功能，因此在实际项目中，通常需要将其与TestNG、JUnit、PyTest等测试框架以及Allure等报告工具集成，共同构成一个全面的、企业级的自动化测试解决方案。

本次爬虫框架采用入下图所示，基于自动化爬虫技术实现新闻数据的采集与处理，其工作流程始于任务调度器（如Crontab/Schedule）对爬虫任务执行频率的智能化安排。通过CrawlerFactory动态选择适配不同数据源特性的爬虫类后，系统采用Selenium技术模拟用户浏览器行为，突破动态渲染限制完整获取新闻列表页内容，并精准提取新闻详情页链接。在核心数据采集阶段，系统通过并行化遍历链接池，结合XPath与正则表达式高效提取标题、时间、作者及正文等多维字段，随后通过数据清洗管道确保数据质量。处理后的结构化数据支持CSV导出接口满足异构分析需求。系统通过模块化设计实现横向扩展能力，可灵活应对新闻网站改版或新增数据源需求，为新闻舆情分析提供高时效性、高完整性的数据支撑。

|  |
| --- |
| **wps** |
| 图2.2爬虫框架思路图 |

尽管Selenium在Web自动化领域占据主导地位，但我们必须认识到其固有的局限性。首先，Selenium的学习曲线相对陡峭，环境配置和驱动管理较为繁琐，并且在处理由AJAX等技术构建的动态或异步加载页面时，需要开发者手动编写复杂的等待逻辑，这增加了脚本的维护成本。其次，它无法对非Web应用（如桌面软件或移动原生App）进行自动化。然而，这些局限性并未动摇其基础地位。Selenium作为W3C标准的一部分，其稳定性和权威性得到了业界的广泛认可。许多新兴的自动化测试工具（如Cypress、Playwright）在设计理念上都借鉴或旨在解决Selenium的某些痛点，但Selenium庞大的社区支持、丰富的学习资源和成熟的生态系统，使其在可预见的未来，仍将是自动化工程师工具箱中不可或缺的核心技术。

# 第三章 数据处理与分析

本章内容包括了数据清洗、数据分析以及LDA的相关分析等，通过一系列的数据处理与分析的手段，得到国家近一年来的发展重心与趋势。

|  |
| --- |
|  |
| 图3.1新闻数据分析框架 |

## 3.1新闻数据清洗与预处理

数据清洗是去除原始数据中对后续分析无意义的部分。原始文本常有不完整、不一致和噪声问题。若不清除这些噪声，会导致文本挖掘结果质量下降。因此，需对采集的数据进行清洗，使其满足语言模型对输入数据质量和形式的要求，以获得更好的分析效果。本研究中的文本数据来自中国经济新闻网。这些文本中存在记者信息、报社信息、停顿词以及重复话语等无用信息，不利于后续分析处理，故在分析前需进行清洗。

文本分析中，数据预处理是关键步骤，包括中文分词、自定义词典和去除停用词。

1. 中文分词

中文分词将文本按词语分割，因中文无自然分隔符，且分词在自然语言处理中至关重要，可借助 Python 的 “Jieba” 工具完成，它有精确模式、全模式和搜索引擎模式三种分词方式。

1. 自定义词典

为提升分词准确性，可添加自定义词典，因为python中的“jieba”中自带词典辅助分词，但是会出现分词不准确的现象，如“习总书记”，会被划分成“习”和“总书记”，这不是我们希望的分词结果，就是需要我们自定义词典加入相关词语，帮助计算机分类。

1. 去除停用词

去除停用词则是删除文档中频繁出现但意义不大的词，如 “还有”“或者”，这通常通过创建停用词表并对比分词结果实现，常用词表由哈尔滨工业大学和百度公司制作，本文分析也通过导入停用词表进行数据处理。

数据预处理后，文本内容分词如下表所示：

表3.1数据预处理前后对比表

|  |  |
| --- | --- |
| 原始数据：最新研究揭示哺乳动物胚胎早期器官发育“密码”  新华社南京6月18日电（记者柯高阳）先天性心脏病是新生儿最常见的出生缺陷类型。我国科学家最新研究发现，哺乳动物胚胎发育早期存在一个器官原基决定区，为先天性心脏病等出生缺陷防治与再生医学研究提供了关键理论基础。6月18日，国际学术期刊《细胞》在线发表了这一研究成果。  哺乳动物早期胚胎的细胞团决定着心脏、肺、肝脏等器官的形成，其发育机制是生命科学领域的重点研究方向。为探究器官发育异常的深层原因，国家重点研发计划项目首席科学家、东南大学生命科学与技术学院教授林承棋带领研究团队对小鼠胚胎进行单细胞空间组学分析，以捕捉器官形成的动态过程。  “空间组学分析好比是给每个细胞装上GPS定位，记录下空间位置信息和基因表达的特征。”林承棋介绍，这项研究前后历时6年，建构了多个覆盖小鼠原肠运动后期至心脏等器官原基形成期的单细胞精度三维数字胚胎，累计解析超十万个细胞的基因表达信息。  研究发现，小鼠胚胎发育到7.75天时，胚内-胚外交界处出现一个独特的信号“洼地”——器官原基决定区（PDZ）。此时该区域紧邻的胚内、胚外部分分别呈高浓度的信号抑制分子和激活性配体分子，但区域内呈现低信号活性“洼地”，表达多种受体信号基因，由此形成易于接收多胚层信号调控输入、驱动心脏与前肠等器官原基协同发育的微环境。微环境信号被转化为基因选择性表达指令后，驱动心脏等器官原基形成。  “PDZ区域的发现证实了在胚胎器官发育的关键窗口期，遗传因素或环境因素对PDZ微环境的干扰可能导致器官原基形成异常。”发育与疾病相关基因教育部重点实验室教授罗卓娟指出，这项研究首次在单细胞精度揭示了器官原基形成的时空动态，为先天性心脏病等出生缺陷及相关疾病防治提供了更精准的科学依据，同时也为学界理解器官再生、肿瘤发生等重大科学问题提供了全新方法论。 | 处理后数据：最新 研究 揭示 哺乳动物 胚胎 早期 器官 发育 密码  南京 18 日电 柯 先天性 心脏病 常见 出生 缺陷 类型 我国 科学家 最新 研究 发现 哺乳动物 胚胎 发育 早期 器官 原基 区 先天性 心脏病 出生 缺陷 防治 再生 医学 研究 提供 关键 理论 基础 18 国际 学术期刊 细胞 在线 发表 研究成果  哺乳动物 早期 胚胎 细胞 团 心脏 肺 肝脏 器官 发育 机制 生命科学 领域 重点 研究 方向 探究 器官 发育 异常 深层 原因 国家 重点 研发 计划 项目 首席 科学家 东南大学 生命科学 技术 学院 教授 带领 研究 团队 小鼠 胚胎 单细胞 空间 组学 分析 捕捉 器官 动态 过程  空间 组学 分析 好比 细胞 装上 GPS 定位 记录 空间 位置 信息 基因 表达 特征 介绍 这项 研究 历时 建构 多个 覆盖 小鼠 运动 后期 心脏 器官 原基 单细胞 精度 三维 数字 胚胎 累计 解析 超 细胞 基因 表达 信息  研究 发现 小鼠 胚胎 发育 独特 信号 器官 原基 区 PDZ 区域 紧邻 呈 高浓度 信号 抑制 分子 激活 性 配体 分子 区域 呈现 低 信号 活性 表达 多种 受体 信号 基因 易于 接收 信号 调控 输入 驱动 心脏 器官 原基 协同 发育 微 环境 微 环境 信号 转化 基因 选择性 表达 指令 驱动 心脏 器官 原基  PDZ 区域 发现 证实 胚胎 器官 发育 关键 窗口期 遗传 因素 环境因素 PDZ 微 环境 干扰 导致 器官 原基 异常 发育 基因 教育部 重点 实验室 教授 指出 这项 研究 首次 单细胞 精度 揭示 器官 原基 时空 动态 先天性 心脏病 出生 缺陷 相关 提供 精准 科学依据 学界 理解 器官 再生 科学 提供 全新 方法论 |

## 3.2新闻数据清洗与预处理

完成数据预处理后，可对分词结果进行量化与分类统计，并构建共现矩阵，以呈现文本的词语分布全貌。借助词语共现矩阵与相关性分析，洞察词间关联程度，初步把握文本中词与词的关系。进而，依据词语间的亲密度对特征词进行初步分类，为后续主题模型的聚类归纳提供方向指引。

|  |
| --- |
|  |
| 图3.2 饼图 |

饼图3.2清晰地展示了不同词性的数量分布比例。其中，名词（“名词”）占据了绝对主导地位，占比高达49.5%，几乎是总数的一半。紧随其后的是动词（“动词”），占比为23.0%，是第二大类别，两者合计超过了总量的72%。其余各类词性所占比例则小得多，例如“其他”占6.4%，“数量词”占5.8%，“地名”占3.7%，而形容词、人名、副词等其他类别的占比均低于4%，共同构成了新闻数据的剩余部分。

|  |
| --- |
|  |
| 图3.3名词词条频率柱状统计图 |

之后，名词词条频率柱状统计图3.3展示了高频名词的分布情况，其中“技术”（约2000次）位列第一，凸显了其在文本中的核心地位，反映出内容主要围绕技术主题展开。紧随其后的“专业”（约1500次）和“学生”（约1200次）表明文本与教育领域密切相关，可能涉及技术人才培养或学术研究。此外，“企业”“产业”“科技”等高频词进一步印证了文本聚焦于科技产业与商业发展。而“高校”“国家”“数据”“领域”“月球”“国际”等词汇的出现，则暗示了讨论范围可能涵盖高等教育、国家政策、数据科学、太空探索及国际合作等多个维度，体现了文本主题的广泛性和跨学科性。高频名词的分布情况，其中“技术”（约2000次）位列第一，凸显了其在文本中的核心地位，反映出内容主要围绕技术主题展开。紧随其后的“专业”（约1500次）和“学生”（约1200次）表明文本与教育领域密切相关，可能涉及技术人才培养或学术研究。此外，“企业”“产业”“科技”等高频词进一步印证了文本聚焦于科技产业与商业发展。而“高校”“国家”“数据”“领域”“月球”“国际”等词汇的出现，则暗示了讨论范围可能涵盖高等教育、国家政策、数据科学、太空探索及国际合作等多个维度，体现了文本主题的广泛性和跨学科性。

同时图3.4动词词条频率柱状统计图展示了高频动词的分布情况，第二张图呈现了高频动词的分布，“创新”（超1600次）占据绝对优势，突显其在文本中的关键作用，可能与科技创新、产业变革等议题紧密相关。“提供”（约1200次）和“推动”（约1000次）的高频出现，表明文本强调解决方案的供给与发展动力的促进。此外，“提升”“说”“发现”“相关”等动词的频繁使用，反映出对能力强化、观点表达、新知识探索及关联性分析的重视。而“培养”“学习”“需求”“发布”“介绍”等词的出现，则进一步指向文本可能涉及人才培育、教育需求、信息传播及产品推广等内容，整体上呈现出动态化、行动导向的语言特征。

|  |
| --- |
|  |
| 图3.4动词词条频率柱状统计图 |

TF-IDF 是一种衡量字词在文档中重要性的方法。它结合了 TF（词频）和 IDF（逆文档频率）两个因素：

**TF**：字词在文档中出现的次数越多，其重要性越高。

TF = 某个词在文章中出现的次数/文本文件中词的总数

**IDF**：字词在文档集合中出现的频率越低，其重要性越高。

IDF = log(语料库中文本文件的总数/包含指定词文本文件的数+1)

简单来说，TF-IDF 的值越高，说明某个词在文档中越重要、越具有代表性。

图3.5和3.6柱状图分别呈现了词频统计和TF-IDF权重分析结果。在词频分布方面，"发展"以超过2500次的绝对优势位列榜首，紧随其后的"中国"、"技术"和"教育"等词汇同样表现出显著的高频特征，直观反映了文本内容的核心关注点。更具深意的是TF-IDF分析结果，该指标不仅考量词频，更评估词语在语料库中的区分度，其中"发展"依然保持最高权重值，印证了其在文本主题表达中的关键地位；而"中国"、"技术"和"教育"等词汇也呈现较高的TF-IDF值，表明这些术语不仅出现频繁，更具有较强的主题表征能力。这种词频与TF-IDF值的双重验证，清晰地勾勒出文本以国家发展战略为框架，聚焦科技创新与教育改革的核心议题，为理解文本的宏观主题架构提供了量化依据。

|  |  |
| --- | --- |
|  | |
| 图3.5 词频统计柱状图 | 图3.6 TF-IDF统计柱状图 |

之后，根据词频频率统计输出词云图3.7，通过关键词展示，可以更直观的看出科技发展的相关因素，以及从AI、技术、发展等多个创新主体与方法的词中寻找创新发展的出路。

|  |
| --- |
|  |
| 图3.7 词云图 |

词语共线性以及相关性分析

## 3.3词条共现矩阵

共现分析基于心理学邻近联系法则及知识结构，通过量化信息载体中的共现信息，揭示内容关联与特性项寓意。在文献计量中，关联程度由共现频度度量。共现矩阵，即共词矩阵，用于统计文本中词组共现次数，描述词组亲密度。矩阵值越大，表明两词在句中共同出现频率越高，关系越紧密。计算时，若词 A 和词 B 在一句中分别出现 2 次和 1 次，则共现值为 2×1=2，依此类推形成矩阵。因 A-B 和 B-A 等同，该矩阵对角线对称。

词语共现热力图3.8通过色阶渐变与数值标注的双重呈现方式，直观展现了文本中核心概念间的语义关联网络。图中"技术-创新"、"发展-产业"等关键词语对呈现出最显著的深色区块，其高共现频次不仅印证了技术创新与产业升级的主题关联性，更揭示了文本"技术驱动发展"的核心叙事框架。次级关联词群如"教育-高校"、"企业-研发"等中等色阶区域，则构建起产学研协同发展的完整语义网络。值得注意的是，热力图通过"数据""人工智能"等延伸词汇的交叉共现关系，进一步呈现出文本多层次的知识图谱结构。这种基于量化的可视化分析，既验证了核心概念的强关联性，又通过边缘词对的分布模式，为解读文本的隐性主题脉络提供了数据支撑，实现了从微观词对分析到宏观主题把握的有机统一。

|  |
| --- |
|  |
| 图3.8 共现矩阵热力图 |

## 3.4基于相关性的网络关系图

图 G＝(V,E)由顶点集 V 和边集 E 构成。社会网络是社会行动者间关系的总和，社会网络分析(SNA)是数据挖掘的重要分支，基于数学方法和图论发展而来。

社会网络由节点（文本中的词条）和边（词条间的联系）组成。根据边是否有方向，网络图分为有向图和无向图。因词条间的方向性在相关分析中意义不大，本文使用无向图。

相关性衡量单词 A 和 B 的关系，值越大越相关。常用指标有：

支持度（Support）：项集{X,Y}在总项集里出现的概率，即 A 和 B 同时发生的概率。



置信度（Confidence）：在 X 发生时，Y 发生的概率。



提升度（Lift）：X 和 Y 的相关性，大于1表示正相关，小于1表示负相关，等于1表示独立。



计算词共现关联中，偶然出现的低频词对（如仅共现1次）往往会产生虚高的提升度（lift）值，导致统计偏差。为此，本研究采用双重约束机制优化分析效果：首先设置10个词的滑动窗口以保证关联词的上下文连贯性，同时引入最小词频阈值（≥5次）过滤噪声词。

|  |
| --- |
|  |
| 图3.9 词条关系网络图 |

基于量化分析结果，本研究构建了词条关联网络图谱，其核心特征表现为：1）"教育-职业"组合以678次共现、160.11提升度和0.62置信度形成最强关联，凸显人才培养与就业市场的深度衔接；2）"产业-发展"（567次共现）与"创新-科技"（505次共现，提升度93.16）构成双轮驱动，反映产业升级的科技赋能特征；3）中国发展语境下，"发展-教育"（528次）、"中国-发展"（422次）和"发展-高质量"（413次，提升度136.53）形成政策闭环。特别地，月球探测领域呈现超强关联性，"月球-样品"组合以493.05提升度创峰值，而AI技术虽共现频次较高（390次），其相对较低的提升度（44.94）表明其应用仍处于扩散阶段。这种结构化关联网络生动诠释了文本在发展战略（平均置信度0.32）、科技创新（平均提升度69.21）与教育改革（平均共现频次472次）三维度上的系统性论述。

**3.5.1 LDA模型**

LDA（潜在狄利克雷分布）是一种非监督机器学习技术，用于文档主题分析。它假设每篇文档由多个主题组成，每个主题对应不同的词汇。文档生成过程如下：以一定概率选择主题，再从该主题中选择词语，重复此过程生成整篇文档。LDA 是概率生成模型，包含文档、主题和词语三层结构，适用于大规模文档集合的建模。使用 LDA 时，需预先设定主题数量、文档数量和每篇文档的词语数量等参数。

|  |
| --- |
|  |
| 图3.10 LDA主题聚类分析图 |

其算法思路如图2.11所示，包括：对文档集合进行分词、去停用词和文本向量化（如TF-IDF）处理；通过计算困惑度来确定合适主题数量；将最优主题数量和语料库输入LDA模型，计算文档-主题分布和主题-词分布，将文档分配到概率最大的主题下。以收集好的关于侨批保护主题的访谈内容文本为例，LDA模型随机给每个主题词w分配主题z，计算主题中词条的数量分布和文档中主题的数量分布，然后不断调整主题分配，直至分布收敛，从而完成文档主题聚类。

**3.5.2 LDA主题聚类评估**

困惑度（Perplexity）是Blei等人提出的评估概率语言模型泛化能力的指标，通过衡量模型对新文本的不确定性反映其预测能力，值越低表明模型效果越好。在LDA主题模型中，困惑度可用于评估文档主题聚类质量，例如分析访谈文本时，低困惑度说明模型对文档主题的预测更准确。为确定最优主题数（K值），通常计算不同K值下的困惑度并绘制曲线，当曲线下降趋势趋缓或出现拐点时，对应的K值即为最优选择。该方法通过平衡模型的解释力和复杂度，提升主题聚类的效果和文本分析的可靠性。如下公式所示：



其中，D代表语料库中的测试集；N表示预测集中的文档数量；M表示预测及中每篇文档中的单词数；Nd表示文档d中的单词数。

**3.5.3主题挖掘结果**

（1）主题模型结果输出

使用 Python 的 Gensim 库对以科技创新为主题的新闻文本进行 LDA 建模和主题聚类。将处理后的新闻文本和主题数候选集作为参数输入 LDA 模型，并对聚类结果进行可视化。通过选择困惑度最低的主题数量，即 3 个，来确定最优主题数。程序运行后，每个主题的详细结果（TOP10关键词）如下图“3 个主题模型聚类结果图”所示。

|  |
| --- |
|  |
| 图3.11 二个主题模型聚类结果表 |

从LDA主题建模分析揭示了文本内容的三维核心架构：主题1聚焦"教育发展"领域（权重占比11%），其关键词组合（教育、发展、专业、学生）体现了人才培养与产业需求的深度对接，特别是"创新"（0.007）与"技术"（0.006）的共现反映了教育体系的技术赋能特征；主题2以"月球研究"为核心（最高权重0.012），包含地球科学（0.006）、卫星技术（0.004）等空间探索子领域，展现我国（0.004）在深空探测方面的科研布局；主题3则突出"AI技术"应用（峰值权重0.016），通过技术（0.009）、模型（0.008）、数据（0.007）等关键词构建了完整的人工智能技术链，并与中国发展（0.006）形成战略呼应。值得注意的是，"中国"和"发展"在主题1、3中的跨主题分布（权重分别为0.007/0.006和0.006/0.006），揭示了国家战略对教育和科技双轨发展的统筹规划。这种主题结构准确反映了文本在人才培养（主题1）、基础科研（主题2）和技术创新（主题3）三个维度的系统性论述，各主题间既保持语义独立性又通过核心概念形成有机联系。

**3.5.4 模型评估——困惑度计算**

首先，对新闻文本的关键词进行提取，构建实验语料库；然后，基于 Python 的 Gensim 库，采用计数方式对语料库进行向量化，并依据公式计算不同主题数对应的困惑度；将主题数设为横坐标，困惑度值设为纵坐标，绘制二维折线图。结果图展示了LDA模型在不同主题数量下的主题一致性得分，其中主题一致性得分用于衡量模型生成主题的质量，得分越高表示主题越一致和有意义。图中显示，当主题数量为3时，模型达到最高的主题一致性得分，约为0.45，这表明选择3个主题能够在LDA模型中获得最优化的主题一致性，从而生成更有意义和一致性的主题，因此选择3作为最佳主题数量。

|  |
| --- |
|  |
| 图3.11 主题——困惑度变化图 |

基于LDA主题模型的量化评估结果显示，主题一致性得分随主题数量变化呈现显著的单峰分布特征。当主题数量设置为3时，模型达到最优性能，其一致性得分（0.45）较其他设置平均提升约25%，这一峰值表明该配置能够最有效地平衡主题的区分度与内聚性。具体而言，三主题设置具有三重优势：首先，0.45的得分显著高于相邻主题数（2个主题得分0.38，4个主题得分0.41），具有统计学显著性；其次，该设置避免了主题过多导致的语义碎片化（如5个主题时得分降至0.39）；最后，确保了每个主题都具有足够的文本覆盖度和语义一致性。这种数据驱动的主题数量选择方法，不仅提升了模型输出的解释性，更为后续的文本挖掘分析奠定了可靠基础，使生成的主题既具有代表性又能准确反映文本的核心内容结构。

## 3.5情感分析

情感分析是对文本信息中带有感情的主观性文本内容进行分析、处理、归纳、推理和评估的一种分析方式。通过下载来源于BosonNLP数据的情感词典以及一些常用停用词数据，将获取的文本内容利用Python中的jieba中文分词并去除停用词后有个情感词语组，简化的情感分数计算公式为：即所有情感词语组的分数之和。结果如下图：

|  |  |
| --- | --- |
|  | |
| 图3.12情感极性分布柱状图 | 图3.13情感极性比例饼图 |

由Python情感分析输出结果可见，获取的文本内容大体情感极性值趋向于1，存在部分文本内容的情感极性值分布于中和极少数负向评价，该输出结果与上述整体情感分析结果接近，同样表明新闻网对中国科技发展持一种积极的、认同的态度。

# 第四章 可视化交互设计

## 4.1 pyecharts和dash介绍

Pyecharts 是一款基于 Python 的强大数据可视化库，它完美结合了 ECharts 的丰富图表功能和 Python 的简洁易用特性。作为 Python 生态中最受欢迎的交互式可视化工具之一，Pyecharts 支持绘制超过 30 种专业图表类型，包括常见的折线图、柱状图、散点图，以及复杂的关系图、地理坐标系图和 3D 图表等。其突出的交互功能（如数据缩放、悬停提示、图例切换）和优雅的视觉设计，使得数据分析结果能够以更直观、更专业的方式呈现。Pyecharts 特别适合用于构建动态仪表盘和交互式报告，支持 Jupyter Notebook 环境无缝集成，并能轻松导出为 HTML 或图片格式。无论是学术研究、商业分析还是日常数据探索，Pyecharts 都能提供高效、美观且高度定制化的可视化解决方案，大大提升了数据故事的讲述能力。

Dash 是一个基于 Python 的高效、灵活且功能强大的开源框架，专门用于构建交互式数据可视化 Web 应用。它建立在 Flask、React.js 和 Plotly.js 之上，使开发者能够通过纯 Python 代码创建具有丰富交互功能的仪表盘，无需编写 JavaScript 或 HTML。Dash 的核心优势在于其声明式语法和响应式设计，通过简单的回调装饰器即可实现前端与后端的实时数据交互，支持动态更新图表、表格等组件。该框架特别适合数据分析师和科学家，能够快速将 Jupyter Notebook 中的分析成果转化为可部署的 Web 应用，同时提供企业级功能如多页应用、URL 路由和身份验证，是构建数据驱动型应用的理想选择。

## 4.2 可视化思路

虽然通过ECharts实现了丰富的可视化效果，但在实际应用中仍存在一些技术局限性：首先，ECharts的词云组件对中文长文本的支持有限，当处理大规模新闻语料时可能出现关键词布局不够紧凑或部分低频词丢失的情况；其次，动态面积图在展示长时间跨度的趋势数据时，由于浏览器渲染性能限制，可能遇到交互卡顿的问题；最后，ECharts的默认主题样式与Dash框架的CSS样式需要额外调试才能实现完美的视觉统一，这在跨设备适配时可能带来额外的开发成本。这些技术边界条件需要在后续迭代中通过性能优化和组件定制来逐步完善。

可视化代码使用Dash框架实现了基于数据驱动的交互式Web新闻分析仪表盘，其技术架构和功能设计体现了现代数据分析应用的典型范式。系统首先通过pandas和jieba对本地JSON格式的新闻数据进行高效处理，包括数据加载、质量清洗和中文分词等关键预处理步骤，将原始的非结构化文本转化为可供深度分析的结构化数据集。在展示层，基于Dash框架构建了模块化的用户界面：顶部采用专业设计的标题栏，配合卡片式的外部报告导航区；核心功能区域则包含智能过滤面板（集成日期选择器和主题分类按钮）以及多维度数据可视化区（动态趋势面积图、语义关键词云和新闻明细表格）。整个系统的智能交互机制通过精心设计的回调函数（Callback）实现，这些函数构成了应用的核心响应逻辑，能够实时捕捉用户的操作意图（如时间范围选择、主题切换等），动态触发数据过滤和指标重计算流程，最终实现可视化组件的毫秒级更新。这种端到端的设计理念为用户提供了沉浸式的新闻主题探索体验，实现了从原始数据到业务洞察的一站式转化。

## 4.3 可视化展示

本小节，主要作生成的可视化页面展示，并介绍的页面相关结构，后续更详细的内容将会制作录屏视频进行讲解与演示。

|  |
| --- |
|  |
| 图4.1 网页总体图 |

该新闻分析仪表盘采用数据驱动的交互式设计，构建了一个多维度的主题热度与语义特征分析平台。系统首先通过pandas和jieba对新闻数据进行深度处理，实现从原始文本到结构化数据的转化。在展示层，顶部智能导航区集成品牌标识、功能导览及核心分析模块入口（包括词语共现网络、LDA主题建模等高级功能），交互控制区配备动态时间选择器与主题分类切换器，支持精准定义分析维度。

可视化展示区采用双视图设计：实现的动态面积图直观呈现主题热度趋势，配合关键词云展示语义特征；智能新闻列表则结构化显示发布时间、标题等元数据，集成筛选与分页功能。

|  |
| --- |
|  |
| 图4.2 词语共现网络图页面 |

该词语共现网络图通过创新的可视化设计，生动呈现了文本数据中的语义关联网络。图中每个节点代表一个特定词语，其视觉尺寸精确映射词语在语料中的统计显著性或出现频次；节点间的连接线则智能呈现词汇共现关系，线条的粗细程度经过算法优化，准确反映共现强度指标。这种网络拓扑结构不仅清晰展现了核心词汇群落的分布特征，更通过节点间的连接密度与路径长度，直观揭示了潜在的主题聚类与语义关联强度。整个可视化系统采用力导向布局算法，确保网络结构既保持数学准确性又具备良好的可读性，为研究者提供了从宏观词频分布到微观语义关联的多层次分析视角，极大提升了文本挖掘的深度与效率。

|  |
| --- |
|  |
| 图4.3 LDA主题建模可视化分析 |

该LDA主题建模可视化分析平台采用双视图交互设计，为用户提供深度的主题语义解析。左侧视图运用多维尺度法（MDS）构建主题空间映射，其中每个主题以圆形标记呈现，其半径精确对应语料库中的主题分布权重，而标记间的欧氏距离则量化反映主题间的语义相似度。右侧视图采用动态显著性术语分析模块，智能展示与选定主题关联度最高的30个特征词汇：通过双色条形图直观对比词汇的全局频次（蓝色）与主题特异性频次（红色），同时配备λ参数调节滑杆，支持用户交互式探索不同权重系数下术语-主题关联强度的动态变化。这种融合空间映射与量化分析的可视化方案，不仅清晰呈现了文本语料中的主题结构特征，更通过参数化交互机制，实现了从宏观主题分布到微观术语贡献度的多层次探索，为文本挖掘研究提供了强有力的分析工具。

|  |
| --- |
|  |
| 图4.3 总体新闻数据分析页面 |

该智能文本分析仪表盘采用多维度可视化矩阵设计，通过科学的视觉分区实现文本特征的立体化呈现。左侧模块包含动态词云与高频词柱状图双重视角，分别以空间分布和量化指标形式展现核心词汇特征；中央分析区通过主题分布柱状图与词性占比饼图的组合，构建起"主题-语法"的二维解析框架；右侧智能面板则创新性地整合了短语频率分析与共现热力图，既呈现关键短语组合模式，又通过矩阵色阶揭示词汇关联网络。各可视化组件采用统一的交互逻辑和配色体系，支持从宏观特征概览到微观结构解析的无缝切换，形成了一套完整的文本数据特征提取与展示解决方案，显著提升了用户对复杂文本内容的认知效率和分析深度。

# 第五章 结论

本研究基于中国经济网科技频道的新闻文本数据，采用LDA主题建模方法，系统性地识别出中国科技创新战略布局中的三大核心领域：教育创新发展、月球探测研究以及人工智能技术应用。这些主题不仅精准捕捉了我国在科技人才培养、深空探索和智能产业转型方面的政策导向，更深刻揭示了国家创新驱动发展战略的实施路径。研究创新性地融合了pyecharts可视化工具与Dash交互框架，构建了集主题识别、语义分析和动态展示于一体的智能分析平台，既验证了LDA模型在中文科技文本挖掘中的有效性，也为政策研究与产业分析提供了可交互的数据支撑工具，实现了从海量文本数据到战略洞察的智能化转化。

GitHub链接：https://github.com/01bigdata/HQU\_bigdataVisualization

# 参考文献

[1]张舒琳,孙泓.我国林业生态建设的知识图谱与LDA主题模型分析[J].中国林副特产,2025,(02):80-82.DOI:10.13268/j.cnki.fbsic.2025.02.032.

[2]熊妮娜.我国数据新闻研究进展的知识图谱分析[J].东南传播,2025,(03):64-68.DOI:10.13556/j.cnki.dncb.cn35-1274/j.2025.03.029.

[3贺鸿飞.“三维创新模型”赋能电视新闻编导智能化转型[J].记者摇篮,2025,(07):138-140.

[4]赵乙颖.定性与定量相结合的国际主流媒体话语分析方法研究[D].山东理工大学,2023.DOI:10.27276/d.cnki.gsdgc.2023.001146.

[5]赵丽娜.数字新闻创新实践探索与新闻生态标准重构思考[J].今日财富,2025,(11):109-111.