

学 号：_____

密 级：_____公开_____

合肥工业大学

Hefei University of Technology

本科毕业设计（论文）

UNDERGRADUATE THESIS



类 型：_____设计_____

题 目：_____社交网络国民安全话题检测及词云可视化_____

专业名称：_____

入校年份：_____

学生姓名：_____

指导教师：_____

学院名称：_____计算机与信息学院（人工智能学院）_____

完成时间：_____2023 年 05 月_____

合 肥 工 业 大 学

本科毕业设计（论文）

社交网络国民安全话题检测及词云可视化

学生姓名：_____

学生学号：_____

指导教师：_____

专业名称：_____

学院名称：计算机与信息学院（人工智能学院）

2023 年 05 月

A Dissertation Submitted for the Degree of Bachelor

**Detection of National Security Topics in Social
Networks And Word Cloud Visualization**

By

Hefei University of Technology

Hefei, Anhui, P.R.China

May, 2023

毕业设计（论文）独创性声明

本人郑重声明：所呈交的毕业设计（论文）是本人在指导教师指导下进行独立研究工作所取得的成果。据我所知，除了文中特别加以标注和致谢的内容外，设计（论文）中不包含其他人已经发表或撰写过的研究成果，也不包含为获得合肥工业大学或其他教育机构的学位或证书而使用过的材料。对本文成果做出贡献的个人和集体，本人已在设计（论文）中作了明确的说明，并表示谢意。

毕业设计（论文）中表达的观点纯属作者本人观点，与合肥工业大学无关。

毕业设计（论文）作者签名：

签名日期： 2023 年 5 月 18 日

毕业设计（论文）版权使用授权书

本学位论文作者完全了解合肥工业大学有关保留、使用毕业设计（论文）的规定，即：除保密期内的涉密设计（论文）外，学校有权保存并向国家有关部门或机构送交设计（论文）的复印件和电子光盘，允许设计（论文）被查阅或借阅。本人授权合肥工业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，允许采用影印、缩印或扫描等复制手段保存、汇编毕业设计（论文）。

（保密的毕业设计（论文）在解密后适用本授权书）

学位论文作者签名：

指导教师签名：

签名日期：2023 年 5 月 18 日

签名日期：2023 年 5 月 18 日

摘 要

随着互联网技术和社交网络平台的迅猛发展，人们越来越依赖社交网络这个窗口进行信息的浏览与传播。准确、高效地检测与“国民安全”相关的话题对舆情控制、维护稳定的社会公共空间有着重要的意义。本文基于话题检测问题，主要工作包含以下两个方面。

(1) 设计并实现基于 TF-IDF 特征的 K-means 微博文本话题检测，并对检测到的主题进行词云可视化。TF-IDF (Term Frequency-Inverse Document Frequency) 用于衡量一个词对于一个文档集合的重要程度，通过计算词频和逆文档频率，对词的重要性和区分性进行量化。该模型的整个流程涉及爬虫实时获取微博文本数据、数据筛选及数据预处理、提取微博文本的 TF-IDF 特征、基于 TF-IDF 特征进行 K-means 聚类以及提取话题词进行词云可视化。经过实验，该模型可以较为准确地检测出国民安全类话题，生成的词云图直观且具有较高可解释性。

(2) 针对微博文本的稀疏性、提取单一 TF-IDF 特征无法提取微博文本的上下文语义信息的问题，本文设计并实现通过 LDA (Latent Dirichlet Allocation) 模型对文档-主题和主题-词汇两个生成过程进行建模，从文本中提取主题特征；使用 Word2Vec 模型将词汇映射为低维稠密实值向量来表示词汇的语义特征，捕捉到邻近词汇之间的语义关联；将 LDA 文档-主题分布特征和加权 Word2Vec 词向量特征融合起来，构建出微博短文本的融合特征。基于文本的融合特征使用 K-means 聚类算法对其进行文本主题检测以期提高话题检测的效果。经过实验对比验证，基于主题模型和加权词向量的融合特征的话题检测模型的准确率、召回率、F1 值均显著高于基于单一 TF-IDF 特征的话题检测模型。

关键词：话题检测；主题模型；K-means 算法；词云可视化

ABSTRACT

With the rapid development of Internet technology and social networking platforms, people increasingly rely on social networks as windows for browsing and disseminating information. Accurate and efficient detection of topics related to "national security" is of great significance for public opinion control and maintaining a stable social public space. This dissertation focuses on the topic detection problem and includes the following two aspects:

(1) Proposed a k-means-based topic detection approach for Weibo text using TF-IDF features, and visualized the detected topics using word clouds. TF-IDF (Term Frequency-Inverse Document Frequency) is used to measure the importance of a word in a document collection by quantifying its term frequency and inverse document frequency. The entire process of this model involves real-time web scraping of Weibo text data, data filtering and preprocessing, extraction of TF-IDF features from Weibo text, k-means clustering based on TF-IDF features, and extraction of topic words for word cloud visualization. Experimental results show that this model can accurately detect topics related to national security, and the generated word clouds are intuitive and highly interpretable.

(2) Addressing the issues of sparsity in Weibo text and the inability of single TF-IDF features to capture contextual semantic information, this article proposes a method that models the document-topic and topic-word generation processes using Latent Dirichlet Allocation (LDA) to extract topic features from the text. It also utilizes the Word2Vec model to map vocabulary into low-dimensional dense real-value vectors, capturing semantic relationships between adjacent words. The LDA document-topic distribution features and weighted Word2Vec word vector features are then combined to construct fused features for Weibo short texts. The k-means clustering algorithm is applied to the fused text features for topic detection, aiming to improve the effectiveness of topic detection. Through experimental comparison and verification, the topic detection model based on the fusion of topic modeling and weighted word vectors achieves significantly higher accuracy, recall, and F1 score than the topic detection

model based on a single TF-IDF feature.

KEYWORDS: topic detection; TF-IDF; topic model; K-means;

目 录

1 绪论.....	1
1.1 研究背景与意义.....	1
1.2 国内外研究现状.....	1
1.2.1 话题检测研究现状.....	1
1.2.2 主题模型研究现状.....	2
1.3 论文组织结构.....	2
2 相关理论与技术.....	4
2.1 爬虫技术.....	4
2.2 文本预处理.....	5
2.2.1 中文分词.....	6
2.2.2 文本去停用词.....	6
2.3 文本表示.....	7
2.3.1 基于向量空间模型的文本表示.....	7
2.3.2 基于主题模型的文本表示.....	8
2.3.3 基于词嵌入模型的文本表示.....	9
2.4 聚类方法.....	12
2.5 本章小结.....	13
3 基于 TF-IDF 和 K-means 的国民安全话题检测.....	14
3.1 基于 TF-IDF 和 K-means 的话题检测框架.....	14
3.2 微博数据获取及筛选.....	15
3.3 微博数据预处理.....	15
3.3.1 微博数据清洗.....	15
3.3.2 中文分词及去除停用词.....	15
3.4 TF-IDF 特征提取.....	16
3.5 基于 TF-IDF 特征的 K-means 主题检测.....	17
3.6 安全话题词抽取及词云可视化.....	19
3.7 实验与结果分析.....	20

3.7.1 实验环境.....	20
3.7.2 实验数据.....	20
3.7.3 实验结果与分析.....	21
3.8 本章小结.....	24
4 基于主题模型和词向量融合的国民安全话题检测.....	25
4.1 基于主题模型和词向量融合的国民安全话题检测框架.....	25
4.2 微博数据获取及处理.....	25
4.3 主题模型和加权词向量融合特征提取.....	25
4.3.1 文本浅层特征提取.....	26
4.3.2 词汇语义特征提取.....	26
4.3.3 特征融合.....	27
4.4 基于融合特征的 K-means 主题检测.....	29
4.5 实验与结果分析.....	29
4.5.1 实验数据.....	29
4.5.2 评价指标.....	30
4.5.3 实验结果与分析.....	31
4.6 本章小结.....	32
参考文献.....	33
致谢.....	35

插图清单

图 2.1 传统网络爬虫工作流程	4
图 2.2 聚焦网络爬虫工作流程	5
图 2.3 CBOW 模型和 Skip-gram 模型	10
图 2.4 Transformer-encoder 结构示意图	11
图 3.1 基于 TF-IDF 和 K-means 的话题检测框架	14
图 3.2 5 月 4 日爬取到的微博文本	20
图 3.3 预处理后的微博文本	21
图 3.4 部分微博文本聚类结果	22
图 3.5 部分微博文本聚类结果	22
图 3.6 5 月 4 日话题检测得到的词云图之一	23
图 3.7 5 月 4 日话题检测得到的词云图之一	23
图 3.8 5 月 4 日话题检测得到的词云图之一	24
图 4.1 基于主题模型和词向量融合的国民安全话题检测框架	25
图 4.2 LDA 主题模型	26
图 4.3 微博文本添加话题标签号	30

表格清单

表 2.1 K-means 算法	12
表 3.1 实验开发环境	19
表 4.1 话题检测效果对比	31

1 绪论

1.1 研究背景与意义

随着互联网技术以及社交网络的迅猛发展，在全国乃至全世界范围内发生的事情都会第一时间在社交网络上散布开来。人们也越来越依赖社交网络这个载体来了解最新的要闻以及通过社交网络来抒发自己对各个事件的看法。根据新浪微博 2022 一季度财报显示，截止一季度末，微博月活跃用户达到 5.28 亿，日活用户达到 2.52 亿。拥有如此庞大的用户群，微博这个社交网络平台产生的海量信息，对于新兴话题的检测具有很高的价值。

由此可见，与人们息息相关的话题会以最快的速度在社交网络上传播开来，迅速遍布各大社交网络平台。如果不进行检测和干预，任其发展，很有可能造成事实的歪曲，造成人们的恐慌。如果仅仅是从社交网络通过逐条阅读人们发布的相关言论，不仅耗时耗力，也很难发现主导话题的舆情，导致后续工作无法开展。面对微博平台每天产生的海量文本，如果能够通过自然语言处理技术快速检测话题，并以最直观的词云图方式展现出来，就能为相关部门提供第一手资讯，为其发布权威信息提供依据。

在所有话题中，与国民安全相关的突发话题更是人们关注的重点，如云南保山地震、巴厘岛中国情侣遇害案、甲流病毒感染等。因此对这些国民安全类的话题进行检测、分析，进而协助相关部门对舆情的控制，这对维护社会公共安全、维护人民切身利益具有重要的意义。

1.2 国内外研究现状

1.2.1 话题检测研究现状

话题检测是话题检测与跟踪下的一个分支。话题检测与跟踪（TDT）是一种自动化的自然语言处理技术，其主要目的是对大规模文本数据进行分析，并自动检测和识别出其中的主题或话题，随着时间的推移，TDT 技术能够对这些话题的变化和发展进行跟踪。而话题检测可以帮助用户从海量文本中提取出与特定主题相关的文本信息，为用户提供更好的信息检索和分析服务。

国内外学者对于话题检测的研究主要集中在两个方面。一是如何更加全面提取

文本特征及算法的改进。文本特征的提取对于后续文本聚类算法的效果表现起到至关重要的作用，而对话题进行检测的核心技术是文本聚类。国内外很多学者已经对此做出了一些研究成果，下面主要对这个方面进行介绍。

Bao 等人的研究通过使用跨越不同的在线平台的多媒体信息流，探索了如何进行话题检测。为此，他们提出了一种名为 RCPMM-CC(Robust Cross-Platform Multimedia Co-Clustering)的方法，并通过对该方法进行定性和定量评估，验证其在话题检测方面的有效性。Qian 及其团队提出了一种新方法，利用主题图网络来进行话题检测，并采用马尔可夫决策过程进行主题修剪，从而在保留语义信息的同时有效地提高了模型性能。研究结果显示，相比使用概率主题模型，这种方法表现更为优越。

在微博话题检测方面，传统的文本聚类手段占主导，主要包含图论聚类、层次聚类等。谢修娟及其团队借鉴了密度算法的思想，对传统的 K-means 算法进行了改进，通过优化初始聚类中心的选择来提高算法的效率。最终，他们将改进后的算法成功应用于新浪微博话题发现^[1]。檀娟娟提出了一种基于微博数据特点的多属性无向加权图聚类算法，旨在实现对微博热点事件的检测^[2]。方一向使用基于谱聚类的多视图聚类算法，对微博文本进行聚类，并从聚类结果中提取具有代表性的关键词，用以描述话题簇^[3]。

1.2.2 主题模型研究现状

LDA 主题模型的主要用途以概率分布的形式推测文档集中每篇文档的主题分布。通过分析文档的主题分布，可以更好地理解文档的主题结构，进而进行有效的信息提取和数据分析。在主题模型领域，LDA 模型具有举足轻重的地位。

路荣等根据微博文本的特殊性，利用 LDA 模型进行隐主题检测，并采用隐主题分析技术计算微博文本之间的相似度，实现微博话题聚类。为了更好地利用微博的特征信息如点赞、评论和转发等，YeY 等提出了 MF-LDA 新主题模型。陈珊珊提出了一种利用 LDA 主题模型进行隐含主题检测的方法，并利用主题信息对文本进行表示。王亚民等针对微博的语体特征，提出了 BTM 模型用于词对建模，并结合优化后的 TF-IDF 算法计算文本相似度，实现微博聚类。

1.3 论文组织结构

根据本文的研究内容，论文的组织结构如下：

第 1 章，绪论。本章首先叙述了微博话题检测的研究背景与意义，然后分析了国内外研究现状，最后总结本文的主要研究内容并提出了本文的组织结构。

第 2 章，相关理论与技术。本章首先叙述了话题检测的大致流程，然后介绍了热点话题发现所涉及到的理论知识与关键技术，及几种常见的文本聚类算法，并对其基本原理进行简要阐述。

第 3 章，基于 TF-IDF 和 K-means 的国民安全话题检测。本章首先使用爬虫实时爬取新浪微博平台上的微博文本，再根据微博文本的特点，对微博进行数据预处理。然后提取文本的 TF-IDF 特征将文本表示为向量空间模型，通过计算每个单词在文档中的词频（TF）以及在整个文档集中的逆文档频率（IDF），将文本转换为数值向量表示；接着根据文本的 TF-IDF 特征进行 K-means 文本聚类。经过在真实的新浪微博数据集上进行实验，验证了本文提出的基于 TF-IDF 特征和 K-means 主题检测模型的有效性。

第 4 章，基于主题模型和词向量融合的国民安全话题检测。本章在前一章对微博实现主题聚类结果的基础上，通过分析微博短文本缺乏上下文关联以及提取单一 TF-IDF 特征的不足，提出一种基于主题模型和加权词向量的融合特征的话题检测模型。经过对比实验验证了基于融合特征的话题检测效果显著优于基于单一 TF-IDF 特征的话题检测模型。

2 相关理论与技术

2.1 爬虫技术

大数据时代下，互联网上的海量信息资源，如果仅仅依靠人力去搜索、整理，不仅效率极低，而且耗费的人力物力成本都很大。在自然语言处理等领域，数据的获取是极为重要的一环，而爬虫技术是一种自动化的信息采集技术，它可以按照人为规定的规则进行有规律、有方向的采集数据，在这个过程中制定的规则称为网络爬虫算法。

爬虫技术主要分为四个类别，分别是通用网络爬虫、聚焦网络爬虫、增量式网络爬虫和深度网络爬虫。在实际的应用中，可以将其结合起来使用。如图 2.1 所示是通用网络爬虫的工作流程。

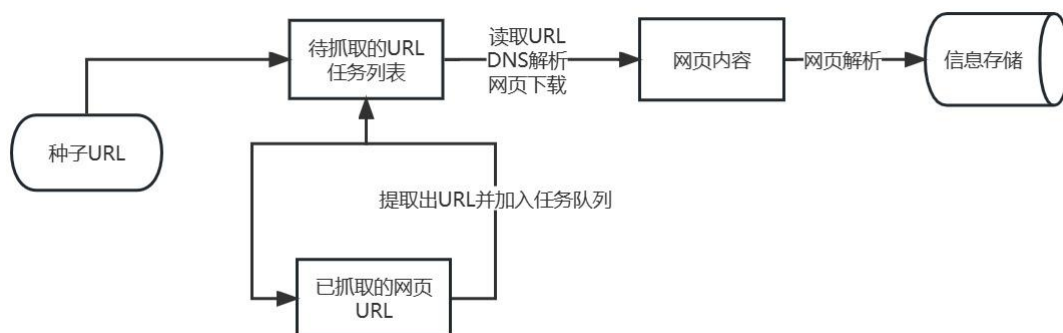


图 2.1 传统网络爬虫工作流程

在通用网络爬虫中，种子 URL 是由人为规定，或由指定的初始网页指定。对该 URL 发送请求并爬取页面上的数据，将爬取到的网页数据存储进数据库中。按照一定的规律找到新的 URL，将其放入待抓取的 URL 列表中。重复以上过程，从待抓取的 URL 列表中读取尚未爬取的 URL，爬取该页面并存储，同时再寻找新的 URL。往复循环，直到达到人为规定的结束条件为止。

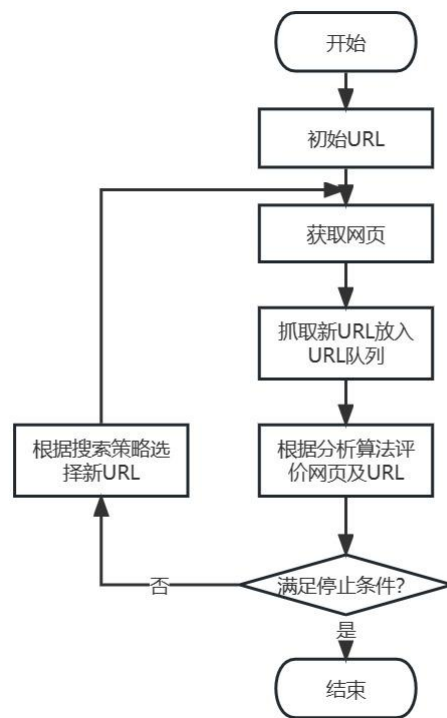


图 2.2 聚焦网络爬虫工作流程

如图 2.2 所示是聚焦网络爬虫的工作流程，聚焦网络爬虫的工作流程相对复杂，它依赖人工指定的分析算法对网页链接进行精准过滤，仅将符合条件的链接纳入 URL 抓取列表中。此外，所有的网页链接都会被系统贮存，建立索引，以便后续的检索。

增量式网络爬虫通过更新已下载网页及只爬取新产生或变化的网页来保证爬取尽可能新的页面。

深度网络爬虫是一种能够自动遍历互联网并收集大量数据的程序。与传统的网络爬虫不同，深度网络爬虫能够在网页的深层链接中进行遍历，获取更多的信息和数据。这种爬虫使用了一些高级算法和技术，例如自然语言处理、机器学习和人工智能等，使得其能够在收集数据的同时进行数据的自动分类、分析和处理，能够快速高效地发现和获取所需的信息。

2.2 文本预处理

微博平台里的文本信息是较为繁杂的，通常会带有标签、标签符号等噪声。且由于微博发布的随意性，导致微博文本的结构不明晰、语言口语化，所以对文本进行预处理是非常重要的。文本预处理的效果通常决定着后续算法的执行效果。文本

预处理通常包括数据清洗、分词、去除停用词等操作^[4]。

2.2.1 中文分词

分词技术是自然语言处理中的一项基础技术，其主要目的是将连续的文本划分为有意义的词汇单元，使得后续的文本处理能够更加高效、准确。在英文文本中，单词之间用符合语言使用习惯的空格隔开，这为分割和处理提供了便利。然而，中文文本缺少这样明显的分隔符，因而需要进行特殊的分词处理。例如对“今天来了许多新同事”进行中文分词的结果是“今天/来/了/许多/新/同事”。目前主流的中文分词技术如下所示。

（1）基于规则的分词方法依赖人工定义的语法规则和词典，用于对文本进行分词处理。通过识别文本中的词性、语法结构等信息，进行分词处理。虽然这种方法的准确性较高，但是难以处理一些特殊情况，且这种方法的效率和准确度较低。

（2）基于统计的方法则根据大量的语料库数据，通过分析不同词汇之间的出现频率和相关性等统计特征，来实现更加准确的分词。这种方法是基于大量的语料库数据进行分析，主要分析的是不同词汇之间的出现频率和相关性等统计特征。

（3）随着深度学习技术的发展，近年来也出现了基于深度学习的分词方法，这种方法利用深度神经网络模型对中文文本进行分析，通过学习文本的语义和上下文信息，实现更加准确的分词效果。这种方法通常需要大量的数据训练和调参，但是其分词效果在一定程度上超过了传统的方法。

分词技术在信息检索、自然语言处理、机器翻译等领域中都有广泛的应用。例如，在搜索引擎中，对查询语句进行分词可以提高搜索的准确性和召回率；在机器翻译中，对源语言文本进行分词可以提高翻译的准确性和流畅度。总之，分词技术是自然语言处理中非常基础的技术，其准确性和效率对后续的文本处理任务至关重要，而不同的分词方法也可以根据具体任务的需要进行选择和优化。

随着我国对中文分词技术的深入探究，许多中文分词技术涌现出来。其中 jieba 分词因为其速度和准确性高，得到了较为广泛的应用。因此，本文采用的是 jieba 分词技术。

2.2.2 去除停用词

停用词是指那些在语句中频率高但没有实际含义的字词，如语气助词和非语素词等。这些词汇在文本分析中对于实际意义的把握并不重要，而且还会占用存储空间和影响分析效果。因此，去除停用词成为了对文本进行预处理的必不可少的步骤之一。

为了方便去除停用词，人们通常会构建停用词表，将其中的词语与文本中的词汇进行对比，将相同的词汇去除。目前国内已经有多个针对中文文本的标准停用词表，如百度停用词表、哈工大停用词表等^[5]，这些停用词表可以有效地去除文本中的无意义词汇，提高文本分析的准确度和效率。

2.3 文本表示

文本是无法直接作为输入传进计算机里的，因为计算机无法直接理解自然语言，它只能处理二进制数据。通过将文本转换为固定长度的向量空间，可以选择低维或高维的维度来实现，这取决于不同的映射方式。将文本以恰当的方式转化为向量或矩阵形式是必要的步骤，以便于机器学习和自然语言处理任务的处理和分析。

文本表示模型选择得合适与否，直接决定了后续话题检测的效果。目前主流的文本表示方法主要分为三类，分别是基于向量空间模型的文本表示，基于主题模型的文本表示，基于词嵌入模型的文本表示^[6]。

2.3.1 基于向量空间模型的文本表示

基于向量空间模型（Vector Space Model, VSM）^[7]的文本表示是一种经典的方法，被称为词袋模型（bag-of-words model）。词袋模型是一个集合，它包含了文本中所有出现的词汇向量，相当于一个装满词汇的袋子。每个向量的每个维度表示一个词汇在文本中出现的频率。因此，向量空间模型可以看作是一个基于单词出现频率的高维向量空间，其中每个维度对应一个单词。如果两个文本在这个向量空间中的向量非常相似，那么它们在语义上也是相似的。

假设有文本集合 $p = \{p_1, p_2, \dots, p_n\}$ ，每个文本 p_i 可以由 $T = \{t_1, t_2, \dots, t_m\}$ 个特征词表示，由于每个特征词在其文本中的重要性不同，可以为其分配不同的词权重 w_i ，则文本 p_i 可以表示为 $p_i \{t_{i1}:w_{i1}, t_{i2}:w_{i2}, \dots, t_{im}:w_{im}\}$ 。其中表示词权重应用最为广泛的方法是 TF-IDF，它可以为每个词分配词权重，TF-IDF 值越大，其表征文本的能力就越强。

向量空间模型的主要优点是简单易用，可以方便地将文本表示为数值向量，而且对单词之间的语法和顺序不敏感。同时，向量空间模型也可以支持一些基本的文本分类和相似度计算任务。

然而，向量空间模型也存在一些缺点。首先，它忽略了单词之间的关系和语法信息，无法表示词汇之间的复杂关系。其次，向量空间模型中的高维向量可能非常稀疏，导致计算和存储成本很高。另一个方面，向量空间模型没有考虑到同义词和多义词的问题，因此可能导致词汇的歧义和语义不准确。

2.3.2 基于主题模型的文本表示

基于主题模型的文本表示方法的核心思想是假设文本由多个主题组成，每个主题又包含了一些相关单词。主题模型假设文本包含了多个主题，每个主题由一些相关单词组成。例如，在一个新闻文章中，可能会包含关于文化、政治、经济、环境等多个主题，每个主题又包含了一些相关的词汇，比如政治主题可能包含“投票”、“选举”、“领导人”等单词，经济主题可能包含“GDP”、“通货膨胀”、“股市”等单词。

主题模型的统计建模过程通过训练数据来估计每个主题和每个单词的概率分布。具体地，给定一个包含 N 个文本的数据集，每个文本由 M 个单词组成。主题模型假设每个文本由多个主题组成，每个主题由一些相关单词组成。通过训练数据可以估计每个主题的概率分布和每个单词在每个主题中的概率分布。得到主题和单词的概率分布后，便可以用这些概率分布来表示文本。具体来说，将每个文本表示为一个主题分布向量，其中每个元素表示文本中某个主题的概率。例如，一个包含 N 个主题的主题分布向量可以表示为一个 N 维向量，其中第 i 个元素表示文本中第 i 个主题的概率。通过这种方式，将文本表示为一个高维向量，其中每个元素都反映了文本中某个主题的重要程度。

LDA (Latent Dirichlet Allocation)^[8] 是一种广泛应用的主题模型。LDA 主题模型旨在通过统计方法挖掘文本中的主题结构，其原理是：假设有一组文档，每个文档都由多个单词组成，希望找到这些文档的潜在主题。LDA 主题模型假设文档是由多个主题混合而成的，每个主题由多个单词组成。通过统计每个文档中每个主题的出现概率，可以得到该文档的主题分布向量。同样地，通过统计每个主题中每个单

词的出现概率，可以得到该主题的单词分布向量。这样，可以将文档和主题都表示为概率分布，从而将文本表示为主题分布向量的形式。

LDA 主题模型的具体过程是，首先随机初始化每个单词所属的主题。然后，对于每个单词，计算其在不同主题下的概率，然后以一定概率重新分配该单词的主题，直到收敛为止。在这个过程中，每个单词所属的主题和每个主题所包含的单词都会不断被调整。最终，便可以得到每个文档的主题分布向量和每个主题的单词分布向量。

主题模型的优点是它能够更好地捕捉文本的语义信息。相比于传统的词袋模型，主题模型能够将文本表示为一个更加抽象和语义化的向量，从而在文本分类、聚类、信息检索等任务中取得更好的性能。此外，主题模型也可以用于发现数据中的隐藏结构和模式，帮助我们更好地理解数据。

主题模型的缺点是它需要大量的计算和存储资源。训练主题模型需要处理大量的文本数据，并且需要对大量的概率分布进行估计和计算。此外，主题模型的结果可能受到超参数的影响，需要对超参数进行调整。

2.3.3 基于词嵌入模型的文本表示

Word Embedding 最初由 Rumelhart 等人提出^[9]，是一种在自然语言处理中广泛应用的方法，它将单词转换为向量表示，同时考虑了词汇之间的语法关系和语义信息。相比传统的文本表示方法（如 TF-IDF），Word Embedding 在处理短文本上表现更好。目前主流的 Word Embedding 模型为 Word2vec 和 BERT。Word2vec 是一种基于神经网络的模型，可以生成高质量的词向量。BERT 是一种预训练的深度双向 Transformer 模型，它可以从大规模语料库中学习到更丰富的语义信息，并用于下游自然语言处理任务。

以下是两种主流的 Word Embedding 模型。

（1）Word2Vec 模型

Word2Vec 是 Word Embedding 的方式之一，它基于神经网络的模型，用于生成高质量的词向量。在 Word embedding 之前占据主流的文本表示方法主要是 one-hot 编码，但是 one-hot 编码有一个不可忽视的缺点缺点在于无法表达词语之间的相似性

关系和维度灾难问题。而 Word2Vec 可以将单词表示为低维度向量，同时保留语义和语法上的相关性，从而更好地表达单词之间的关系。

Word2Vec 本质上是一个神经网络，由输入层、隐藏层和输出层构成。输入层输入的是一个由 One-Hot 向量表示的单词，其长度等于语料库中单词的数量，其中只有一个元素为 1，其他元素都为 0。隐藏层包括若干个神经元，其数量通常在几百到几千之间。隐藏层将输入层的 One-Hot 向量转换为较低维度的向量表示，并捕捉单词之间的语义和语法关系。输出层通常使用 softmax 函数，将隐藏层的向量表示映射为每个单词的概率分布。这样便可以通过概率分布来预测单词的上下文或目标单词。

Word2Vec 模型包含两种不同的算法，分别是 CBOW^[10]和 Skip-Gram^[11]。CBOW 模型使用上下文单词来预测目标单词，而 Skip-Gram 使用目标单词来预测上下文单词。如图 2.3 所示是 CBOW 和 Skip-Gram 的模型图示。

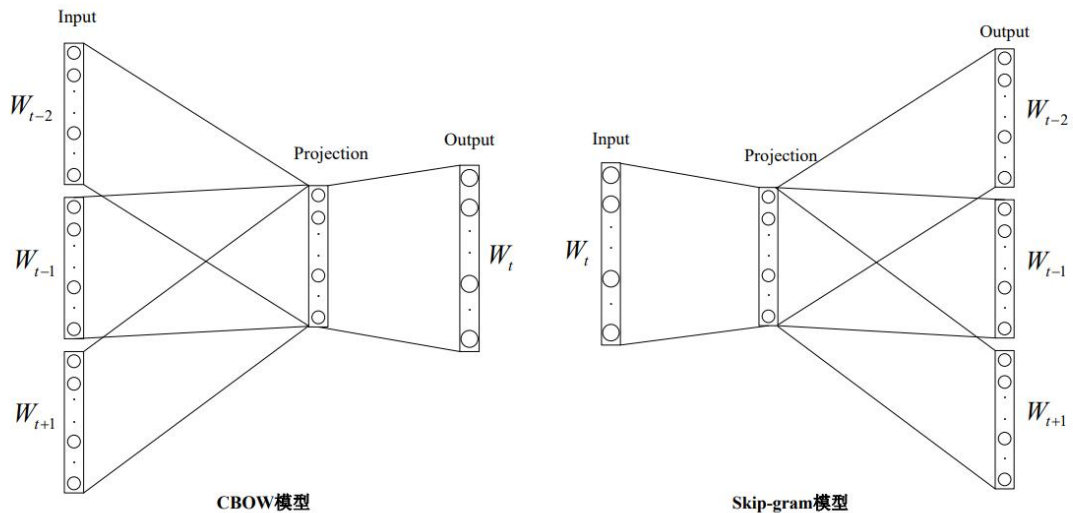


图 2.3 CBOW 模型和 Skip-gram 模型

(2) BERT 模型

BERT 是在 2018 年由谷歌 AI 团队提出的预训练语言模型^[12]，通过对超大规模无标注语料进行模型参数学习。相较于先前的预训练模型，BERT 整合了更多的语法和词法知识，实现了真正双向的上下文信息获取。在多个 NLP 任务上表现出色。

BERT 以 Transformer 编码器作为核心结构，Transformer 由 6 个编码器和 6 个解码器组成。这个模型具备强大的文本表征和并行计算能力。BERT 主要借助

Transformer 的 Encoder 组件进行语言模型训练，获得富含语义信息的文本向量表示。每个单元包含两个子层，即自注意力机制和前馈神经网络，有效地捕捉输入序列的上下文信息。这种设计使得 BERT 能够有效捕捉文本中的关联信息。如图 2.4 所示，是 Transformer-encoder 的示意图。

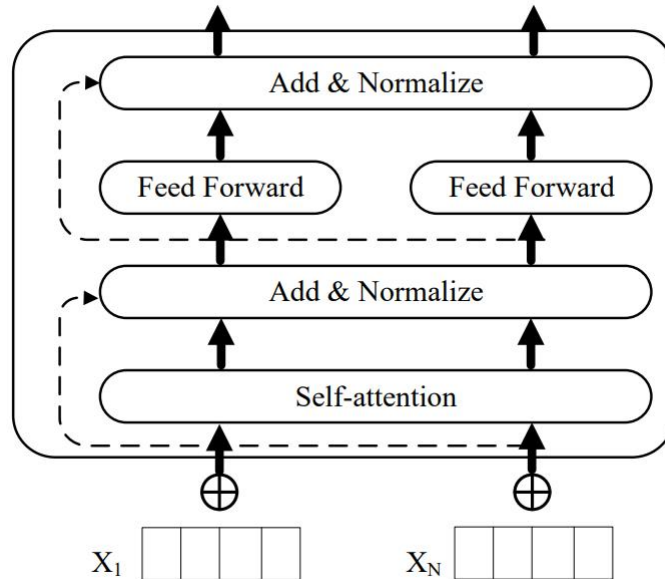


图 2.4 Transformer-encoder 结构示意图

在 BERT 模型中，Transformer-encoder 中的自注意力机制是非常重要的组成部分之一。自注意力机制是一种注意力机制的形式，其中每个输入元素都可以作为查询、键和值同时进行处理，从而捕捉输入序列中的上下文关系。

具体来说，在 Transformer-encoder 中的自注意力机制中，每个输入序列的每个元素都会得到三个向量表示：一个查询向量，一个键向量和一个值向量。然后，每个查询向量与所有键向量进行点积操作，得到一个注意力分布向量，表示每个查询与键之间的关注程度。最后，所有值向量根据这个注意力分布向量进行加权求和，从而得到一个聚合向量表示整个序列。

自注意力机制具有明显优势，它使模型能够有选择地聚合和捕捉输入序列中的上下文信息。这种机制能够有效地提取关键信息，从而改善模型对文本的理解能力。同时，由于每个元素都可以作为查询、键和值，所以可以捕捉到更加复杂的序列关系，使得模型在处理自然语言任务时具有更强的泛化能力。

2.4 聚类方法

聚类是一种无监督学习方法，与有标签的分类任务不同。它将一组对象分成多个簇或类别，确保同一簇内的对象具有较高的相似度，而不同簇之间的相似度较低^[13]。目前主流的聚类算法主要分为六类，分别是基于划分的聚类算法、基于层次的聚类算法、基于密度的算法、基于图论的聚类算法、基于网格的聚类算法和基于模型的聚类算法。下面主要介绍基于划分的聚类算法及它的代表经典算法。

基于划分的聚类算法是把所有的样本数据划分为 n 个类别，其中每个类别至少有 1 个样本数据且每个样本数据仅属于 1 个类别。**K-means** 算法^[14]属于基于划分的聚类算法，是一种经典、应用广泛的聚类算法。其通过制定规则的迭代将样本数据划分为 k 个类别，使得聚类结果得出的损失函数最小。**K-means** 聚类算法的损失函数的定义：每个样本数据聚类样本中心的误差平方和，公式表示为

$J(c, \mu) = \sum_{i=1}^k \|x_i - \mu_{c_i}\|^2$ ，具体的算法描述如表 2.1 所示。

表 2.1 K-means 算法

输入：聚类数目 k ，包含 n 个文本的文本集 D

输出： k 个簇划分

Begin

1. 在数据集 D 中任意选中 k 个初始聚类中心，记 $\{D_1, D_2, \dots, D_k\}$ ；
2. 计算每一个数据点 x (除去 k 个中心) 到初始聚类中心 D_j 的距离；
3. 将 x 分配到与其距离最近的聚类中心 D_j 所在类簇中；
4. 更新聚类中心点
5. 循环步骤 2-4，直到每个聚类中心都不再变化，则聚类算法结束。

End

K-means 算法对于初始聚类中心的选择非常敏感，不同的初始选择可能会导致不同的聚类结果。为了克服这个问题，通常会多次运行 **K-means** 算法，并从多个初始选择中选择具有最小误差的聚类结果^[15]。

2.5 本章小结

本章主要对话题检测涉及到的相关理论和技术进行了介绍。首先是爬虫技术，对通用网络爬虫和聚焦网络爬虫作了具体的介绍和对比；接着介绍了文本预处理技术，包括中文分词和去除停用词；然后对三种文本表示进行了介绍，包括基于向量空间模型的文本表示、基于主题模型的文本表示和基于词嵌入模型的文本表示。最后介绍了聚类算法，重点阐述基于划分的聚类算法和其经典代表算法 K-means 聚类算法。

3 基于 TF-IDF 和 K-means 的国民安全话题检测

本章研究基于 TF-IDF 特征的微博文本话题检测，并对提取到的主题进行词云可视化。本章将对话题检测的主要流程进行详细阐释，并对实验结果进行展示与分析。

3.1 基于 TF-IDF 和 K-means 的话题检测框架

本章提出的基于 TF-IDF 和 K-means 主题聚类框架如图 3.1 所示。主要是流程是利用爬虫技术从新浪微博平台上的热榜板块爬取微博文本，再进行有关“国民安全”相关的筛选；对筛选后的文本进行文本预处理，包括数据清洗、中文分词、去除停用词；接着提取文本的 TF-IDF 特征；使用 K-means 聚类进行文本聚类进而话题检测；最后提取安全话题词并进行词云可视化展示。

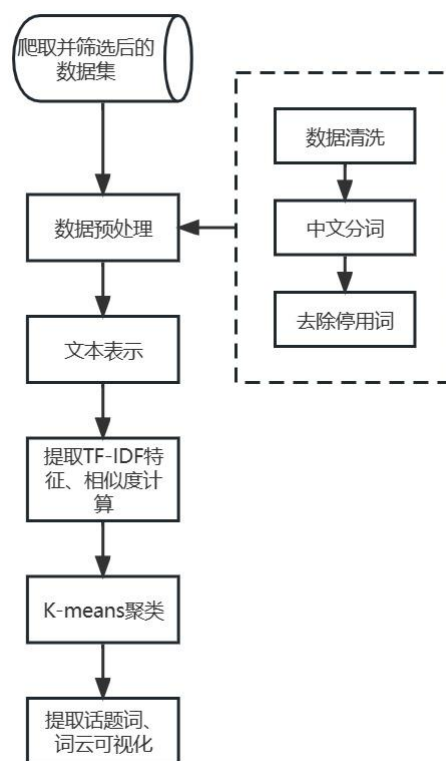


图 3.1 基于 TF-IDF 和 K-means 的话题检测框架

下面具体介绍其中的步骤。

3.2 微博数据获取及筛选

将微博热榜的 url 设为目标 url，手动构造 headers 参数，包括请求头 User-Agent 和 Cookie，向目标 url 发送请求得到页面的 HTML 文件。接着使用 Xpath 定位 HTML 文档中的目标元素——热榜话题标签，解析后将 50 条热榜话题标签爬取下来存储进列表中。

为了实现国民安全类话题的筛选，手动设置了一个国民安全话题相关的关键词列表，涉及到社会空间安全、网络空间安全、个人、国家等层面。将爬取到的 50 条热榜话题标签文本与该关键词列表进行比对，若热榜话题标签上的话题标签文本上含有关键词列表中的关键词，则判定为与国民安全话题相关，将其存储下来便于后续进一步深度得爬取微博文本。

接下来对过滤后的标签文本进行进一步微博文本爬取，这样可以保证爬取下来的微博文本皆是与“国民安全”话题相关。

3.3 微博数据预处理

从微博平台上爬取到的原始微博文本往往形式杂乱、结构随意，如果不对其进行文本预处理，会严重影响后续的话题检测。对原始文本数据进行预处理主要包含以下三个方面，分别是微博数据清洗、中文分词和去除停用词。

3.3.1 微博数据清洗

未经处理的原始微博文本存在大量噪声，包括网址、表情符号、广告链接等，需要对其剔除。

（1）去除长度过短的微博文本。字数少于 5 个字的微博往往没有什么实际内涵，因为字数过短的文本大多数来自“僵尸用户”，它们并不能够表达出具体的观点，也难以将其分类到具体的话题类别中，所以需要将其从文本数据集中过滤掉。

（2）将微博文本中的网址、表情符号、转发等字符剔除。微博文本中存在的网址链接、表情符号、转发字符对于话题的检测不仅没有实际价值，还会成为文本聚类的干扰，所以应将其从文本数据集中剔除。

3.3.2 中文分词及去除停用词

在对原始文本进行过数据清洗后，对其进行中文分词和去除停用词操作，使用

到的技术来自 2.2 节。

（1）jieba 分词

jieba 分词是一个流行的中文分词工具，它采用了基于字典和基于统计的混合分词策略。它的设计目标是将连续的中文文本切分成有意义的词语。使用 jieba 分词工具对数据清洗后的微博文本进行切分，得到分词后的结果。

（2）去除停用词

选取哈工大停用词表对分词后的结果进一步过滤，从而过滤掉文本中高频率出现但却没有实际意义的词语，如语气助词、介词、副词等。这一操作减少了无意义的词汇对后续话题检测带来的干扰。

3.4 提取 TF-IDF 特征

TF-IDF (Term Frequency-Inverse Document Frequency)^[16] 是一种常用的文本特征提取方法，用于衡量一个词语在文本中的重要性。它综合考虑了词频和逆文档频率两个因素，通过计算词语在文本集合中的频率和在整个语料库中的稀有程度来确定权重。

具体来说，TF-IDF 的计算步骤如下：

1. Term Frequency (词频)：词频表示一个词语在一个文本中出现的频率。它可以通过简单地计算词语在文本中的出现次数来获得。tf 的计算公式如下：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,i}} \quad (3-1)$$

2. Inverse Document Frequency (逆文档频率)：逆文档频率衡量了一个词语在整个语料库中的稀有程度^[17]。计算逆文档频率时，首先需要确定文档频率 (Document Frequency)，即包含某个词语的文档数量。然后通过对文档频率取倒数，并可以进行平滑处理，得到逆文档频率。idf 的计算公式如下：

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}|} \quad (3-2)$$

3. TF-IDF 权重计算

TF-IDF 权重可以通过将词频与逆文档频率相乘来计算。词频表示一个词语在一

个文本中的重要性，而逆文档频率表示一个词语在整个语料库中的重要性。通过将二者相乘，可以得到词语在文本中的相对重要性。综上所述，TF-IDF 的计算公式如下：

$$tfidf_{i,j} = tf_{i,j} * idf_i \quad (3-3)$$

提取文本的 TF-IDF 特征一方面将文本表示为向量空间模型，通过计算每个单词在文档中的词频（TF）以及在整个文档集合中的逆文档频率（IDF），将文本转换为了数值向量表示；另一方面 TF-IDF 特征可以用于选择最具代表性和区分性的特征词。通过计算单词的 TF-IDF 值，可以衡量单词在文档中的重要程度。高 TF-IDF 值的单词通常在该文档中频繁出现，但在整个文档集合中相对较少出现，因此这些单词可能具有更好的区分能力。

3.5 基于 TF-IDF 特征的 K-means 主题检测

K-means 算法是一种被普遍使用的聚类算法，它的的原理是通过一定规则的迭代将样本数据划分为 k 个类别，使得聚类结果得出的损失函数最小，从而实现簇内文本差异小，而簇间文本差异大的效果。

本文基于 TF-IDF 特征计算相似度得到相似度矩阵。基于 TF-IDF 特征，可以使用余弦相似度（Cosine Similarity）来计算文档之间的相似度。余弦相似度衡量了两个向量之间的夹角，它的值介于 -1 和 1 之间。对于两个文档向量，余弦相似度越接近于 1，表示它们越相似。余弦相似度的计算公式如下：

$$sim(x, y) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n (x_i)^2} \sqrt{\sum_{i=1}^n (y_i)^2}} \quad (3-4)$$

K-means 聚类算法中，初始聚类簇数 k 的值对结果的影响比较大，所以 k 值的选定是一个不容忽视的问题。本文采取了肘部法则来确定 k 值。肘部法则（Elbow Method）是一种常用的评估方法，用于选择 K-means 聚类算法中最佳簇数量 K 的值。该方法通过观察聚类结果的误差平方和（SSE）与不同 K 值之间的关系，找到一个拐点（即肘部），该拐点对应于最佳的 K 值。以下是肘部法则的具体步骤：

1. 执行 K-means 聚类：首先选择一个较大的 K 值范围，例如从 1 到 10，然后应用 K-means 算法对数据进行聚类，对每个 K 值计算相应的聚类结果。

2. 计算误差平方和（SSE）：对于每个 K 值，计算相应的聚类结果的误差平方和（Sum of Squared Errors, SSE）。SSE 指的是每个样本与其所属簇中心之间的距离的平方的总和。较小的 SSE 表示样本在聚类中心附近更加紧密，聚类效果较好。

3. 绘制肘部曲线：将不同 K 值下的 SSE 绘制成曲线图，横轴为 K 值，纵轴为 SSE 值。可以观察到 SSE 随着 K 值的增加而减小，因为较大的 K 值会使每个簇包含更少的样本，导致簇内的平方距离较小。随着 K 值的继续增加，SSE 的改善速度会减缓。

4. 寻找肘部拐点：在肘部曲线图中，寻找一个拐点，即 SSE 值开始显著减小的位置。这个拐点通常被认为是最佳的 K 值，因为它在保持较低 SSE 的同时，避免了过度拟合的情况。

5. 选择最佳 K 值：根据肘部拐点，选择对应的 K 值作为最佳的聚类数量。这个 K 值在一定程度上平衡了聚类效果和模型的复杂度，提供了相对合理的聚类结果。

综合以上，基于 TF-IDF 特征的 K-means 文本聚类的算法描述如下：

1. 整合文本数据集：每个微博文本文档表示为一个向量，其中每个维度表示一个分词，并使用 TF-IDF 权重进行编码。

2. 构建 TF-IDF 矩阵：使用 TF-IDF 算法计算每个单词在每个文档中的权重，构建 TF-IDF 矩阵。

3. 使用肘部法则确定 k 个聚类中心。

4. 迭代更新聚类中心：重复以下步骤直到达到停止条件：

- a. 使用余弦相似度计算每个文档与每个聚类中心之间的距离。
- b. 将每个微博文本分配给距离最近的聚类中心。
- c. 将每个聚类中所有微博文本向量的平均更新每个聚类的中心。

5. 当微博文本的聚类分配不再发生变化时，停止迭代。

6. 输出聚类结果：最终得到 k 个聚类，每个类别下的文本代表一个话题。

3.6 安全话题词抽取及词云可视化

通过 K-means 聚类算法对微博文本进行聚类后，微博文本依据 TF-IDF 特征被分成了 k 个类别，每个类别代表一个话题，每个类别下包含了该话题下的若干个文本。为了获得每个类别下的词频统计，对每个类别的文本进行词频统计：遍历每个文本，统计每个词语在该类别下的出现次数；接着对该类别下的所有文本的词频统计进行分类，从而获得该类别下总体的词频统计。高频出现的词语可视为代表该话题的关键词，将其抽取出来作为代表该类别的话题词。

词云可视化是一种常见的文本数据可视化方法，词云通过将词语按照重要性或频率进行可视化展示，可以直观地展示出文本中哪些词语出现得更为频繁或更为重要。较大和较突出的词语往往表示在文本中出现次数较多或具有较高的权重，使得用户可以快速捕捉到文本的关键主题或关注点。不仅如此，词云可视化可以提供整体概览。词云将大量的文本信息以图形化形式呈现。通过词云，用户可以一目了然地了解文本的关键词汇，不需要深入阅读每个词语，就可以快速获取整体的印象和洞察。在共享和传播方面，词云可以轻松地在报告、演示文稿、网站或社交媒体等平台上共享和传播。通过将文本信息以可视化的方式呈现，词云使得复杂的文本数据更易于理解和分享，有助于促进交流和传播关键信息。

本文采用 WordCloud 工具对提取出的话题词进行词云展示，为用户快速获取文本的关键主题、关注点和核心词汇，提供整体概览。

3.7 实验与结果分析

3.7.1 实验环境

本文从数据获取、筛选，到数据的处理，TF-IDF 特征的提取，聚类算法的编写均使用 python 语言编写。以下是实验环境说明，如表 3.1 所示。

表 3.1 实验开发环境

	名称	参数
硬件环境	内存	8.0GB
	处理器	Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz 1.99 GHz
	硬盘	1T 固态硬盘
软件环境	操作系统	Windows10 专业版 64bit
	开发语言	Python
	开发工具	Anaconda3

3.7.2 实验数据

本文使用的数据均是使用爬虫技术从微博平台获取到的真实数据。主要步骤是将微博热榜的 url 设为目标 url，手动构造 headers 参数，包括请求头 User-Agent 和 Cookie，模拟用户登录新浪微博，向目标 url 发送请求得到页面的 HTML 文件。接着使用 Xpath 定位 HTML 文档中的目标元素——热榜话题标签，解析后将 50 条热榜话题标签爬取下来存储进列表中。接着使用国民安全话题相关的关键词列表进行比对筛选。然后对筛选后得到的与国民安全相关的话题标签文本进行深度微博文本爬取，这一步爬取下来的文本数据作为本实验的原始数据。

在 5 月 4 日爬取到的微博文本数据如图 3.2 所示，共计 3014 条。

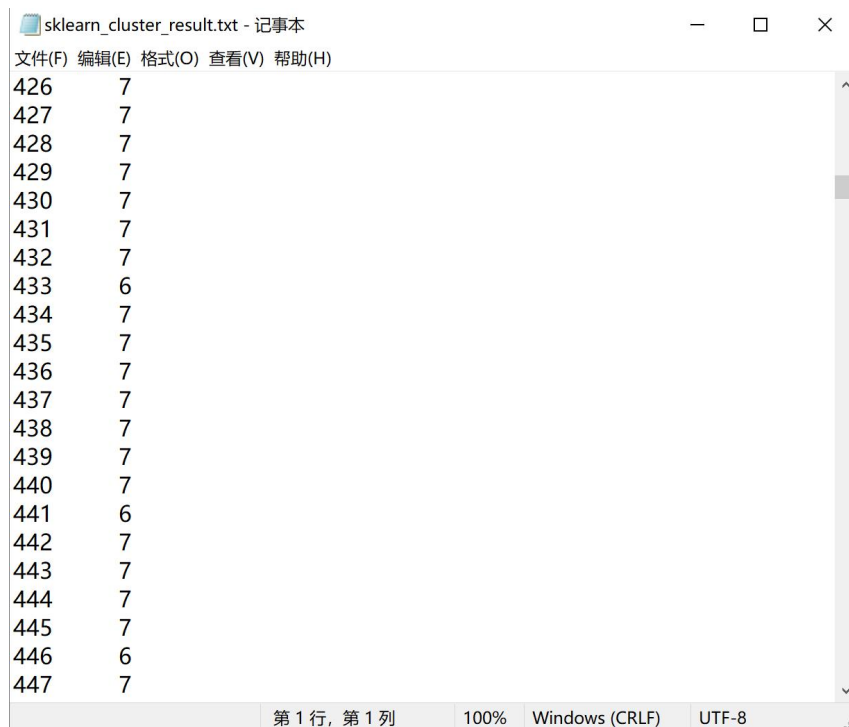
列表示聚类后属于的类别。



```

sklearn_cluster_result.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
0      0
1      0
2      0
3      0
4      0
5      0
6      0
7      0
8      0
9      0
10     0
11     0
12     0
13     0
14     0
15     0
16     0
17     0
18     0
19     0
20     0
21     0
    
```

图 3.4 部分微博文本聚类结果



```

sklearn_cluster_result.txt - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
426    7
427    7
428    7
429    7
430    7
431    7
432    7
433    6
434    7
435    7
436    7
437    7
438    7
439    7
440    7
441    6
442    7
443    7
444    7
445    7
446    6
447    7
    
```

图 3.5 部分微博文本聚类结果

如图 3.6 所示, 是对微博文本进行话题检测后得到的词云图之一。可以看到, 这张词云图直观得展示了“孩子在河边摸螺蛳不幸溺水身亡”的话题。



图 3.6 5 月 4 日话题检测得到的词云图之一

如图 3.7 所示, 这张词云图直观得展示了“唐山铁矿透水事故”的话题, 从中可以清楚看到该事故中死亡人数被县领导谎报的事实。



图 3.7 5 月 4 日话题检测得到的词云图之一

如图 3.8 所示, 这张词云图直观得展示了“二次感染新型冠状病毒”的话题, 从中可以看出, 群众发烧、嗓子疼、鼻塞等症状, 表达了难受、酸痛的感受。



图 3.8 5 月 4 日话题检测得到的词云图之一

3.8 本章小结

本章主要介绍了基于 TF-IDF 特征和 K-means 算法的话题检测的具体流程，以及话题检测的效果展示。可以看出算法较为准确地检测出了各个国民安全相关的话题且词云图直观地将话题词以不同重要程度的标记展示了出来。

4 基于主题模型和词向量融合的国民安全话题检测

针对微博文本的稀疏性及上下文语义欠缺、提取单一 TF-IDF 特征无法提取上下文语义信息的问题，本章将 LDA 文档—主题分布特征和加权 Word2Vec 词向量特征融合起来，构建出微博短文本的深层特征，并使用 K-means 聚类算法对其进行文本主题检测，提高话题检测的效果。

4.1 基于主题模型和词向量融合的国民安全话题检测框架

如图 4.1 所示是基于主题模型和词向量融合的国民安全话题检测框架。数据源来自新浪微博实时爬取，获取到文本数据后进行数据预处理操作，包括数据清洗、中文分词、去除停用词。接着提取微博文本数据的 TF-IDF 词汇权重特征和 Word2Vec 词向量，再将它们的特征权重加权，使用 LDA 模型对微博文本数据提取文本浅层特征，然后将它们进行向量拼接操作获得文本数据的融合特征。最后使用 K-means 聚类算法进行话题检测。

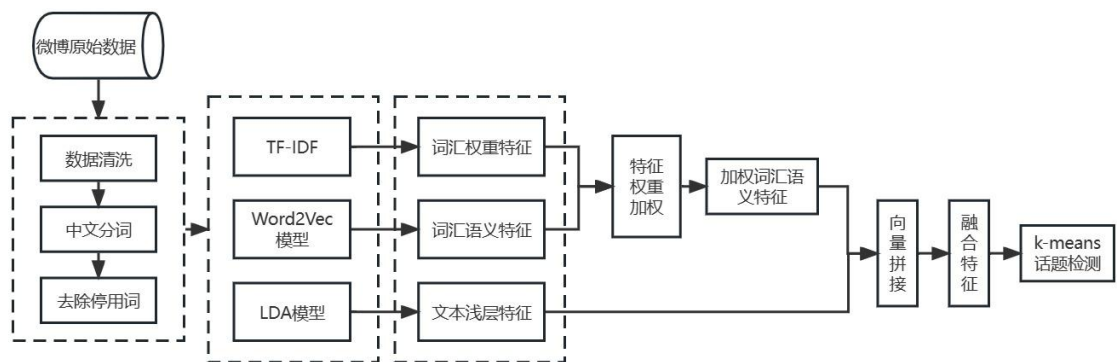


图 4.1 基于主题模型和词向量融合的国民安全话题检测框架

4.2 微博数据获取及处理

同本文的 3.2 至 3.3 节，使用爬虫实时爬取新浪微博平台上热榜的热门话题标签文本，然后对其使用“国民安全”关键词列表进行筛选，获得与国民安全话题相关的话题标签文本。接着使用爬虫对这些标签文本依次进行深度爬取，获得本章的原始微博文本数据。

4.3 提取主题模型和加权词向量融合特征

4.3.1 文本浅层特征提取

LDA 模型的核心概念是认为文档由主题的概率分布组成，而主题又由词汇的概率分布组成。因此，LDA 模型从文档-主题和主题-词汇两个层面对生成关系进行建模，描述了文档、词汇和主题之间的关联。具体而言，LDA 模型在生成文档的过程中，首先从主题的概率分布中选择一个主题，然后根据选定的主题从词汇的概率分布中选择一个词汇。这个过程不断重复，直到生成完整的文档。这样，每个文档就可以被看作是主题的概率分布，而每个主题又可以被看作是词汇的概率分布。

通过对文档-主题和主题-词汇两个生成过程进行建模，LDA 模型可以通过观察到的文档和词汇来推断出主题的概率分布。这样就可以通过 LDA 模型从文本中提取主题特征，帮助理解文本的语义和结构。

根据图 4.2，语料库中包含了多条微博，每条微博的词汇量为 N 。其中， α 和 β 分别服从狄利克雷分布。LDA 主题模型可以被看作是文档生成的逆过程。对于给定的微博 D ，首先从先验概率分布中抽样得到其在主题上的概率分布 θ ，然后根据文档-主题分布采样得到微博 D 中第 k 个词汇的主题 z 。同样地，对于给定的主题 z ，从先验概率分布 β 中抽样得到其词汇分布 ϕ ，并根据主题-词汇分布 ϕ 抽样生成词汇 $w^{[18]}$ 。

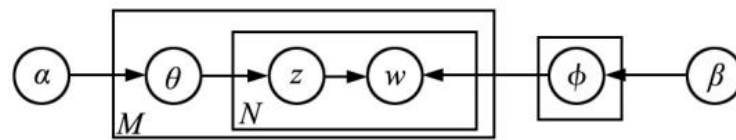


图 4.2 LDA 主题模型

对于微博文本 D ，其文档-主题特征表示如下： $P=[p_1, p_2, \dots, p_t]$ ，其中 p_t 表示在第 t 个主题下的概率。LDA 主题模型通过概率值估计和词汇共现信息来探索文本的主题分布特征，从而揭示文本的全局语义信息和特征表达。然而，在训练数据时，该模型将文档中的词汇视为相互独立，忽略了上下文词汇之间的语义联系，因此在本质上只能提供文本特征的表面表示。

4.3.2 词汇语义特征提取

Word2Vec 是 Word Embedding 的方式之一，是一种用于生成高质量词向量的基于神经网络的模型。Word2Vec 本质上是一个神经网络，由三层构成，分别是输入层、隐藏层和输出层。输入层接受一个 One-Hot 向量表示的单词，其长度等于语料库中

单词的数量。隐藏层将输入层的 One-Hot 向量转换为较低维度的向量表示，并捕捉单词之间的语义和语法关系。输出层通常使用 softmax 函数，将隐藏层的向量表示映射为每个单词的概率分布。这样便可以通过概率分布来预测单词的上下文或目标单词。

Word2Vec 模型和 LDA 主题模型在特征表达方面有不同的重点。Word2Vec 模型通过将词汇映射为低维稠密实值向量来表示词汇的特征，而 LDA 主题模型则关注整个文本集合的特征表达。通过 Word2Vec 生成的词汇特征向量，能够捕捉到邻近词汇之间的语义关联，从而弥补了短文本环境下特征表达中词汇语义的缺失。对于微博这种短文本，每条微博所包含的词汇量较少，导致目标词汇的上下文语义信息明显不足。因此，在本节中选择了 Word2Vec 模型的 Skip-gram 学习模式来生成微博语料集的词向量。在此基础上，对微博文本 D 进行 Word2Vec 词向量映射，该文本特征的表达如下：

$$A = \begin{bmatrix} a_{11}, a_{12}, \dots, a_{1t} \\ a_{21}, a_{22}, \dots, a_{2t} \\ \dots \\ a_{k1}, a_{k2}, \dots, a_{kt} \end{bmatrix} \quad (4-1)$$

其中第 k 行代表微博文本 D 中词汇 w_k 对应的词向量，t 是词向量的维度。

4.3.3 特征融合

Word2Vec 模型并未明确考虑词汇对主题的贡献，这可能导致非关键词对特征的语义表达产生影响。为了提高词向量在表达主题上的区分能力，可以使用 TF-IDF 值对 Word2Vec 词向量进行加权处理。通过这种方式，可以更好地捕捉关键词对主题的贡献，并减少非关键词对特征的干扰。这种综合使用 Word2Vec 和 TF-IDF 的方法能够提升特征的语义表达能力，更好地反映文本中的主题信息。

微博文本 D 中的词汇权重特征如下所示：

$$T = [tfidf_1, tfidf_2, \dots, tfidf_k] \quad (4-2)$$

其中 $tfidf_k$ 表示微博文本 D 中词汇 w_k 的 $tfidf$ 值，这个数值越高代表这个词汇的重要性和区分文本的能力越高。将词汇的 $tfidf$ 值与它对应的词向量相乘，得到微博文

本 D 的加权词汇语义特征向量 AT。

$$AT = A * T = \begin{bmatrix} a_{11} * tfidf_1, a_{12} * tfidf_1, \dots, a_{1t} * tfidf_1 \\ a_{21} * tfidf_2, a_{22} * tfidf_2, \dots, a_{2t} * tfidf_2 \\ \dots \\ a_{k1} * tfidf_k, a_{k2} * tfidf_k, \dots, a_{kt} * tfidf_k \end{bmatrix} \quad (4-3)$$

LDA 主题模型和 Word2Vec 模型在向量化微博短文本方面有不同的侧重点。LDA 模型通过主题分布向量全局描述文本特征，但由于使用词袋模型，无法很好地挖掘深层的语义信息。相比之下，Word2Vec 模型能深入了解序列词汇的语义关联，但更注重局部邻近词汇的关系，可能导致全局信息缺失。

简言之，LDA 模型提供了全局文本特征的描述，但无法捕捉深层语义信息。而 Word2Vec 模型揭示了词汇之间的语义关联，但更关注局部邻近词汇关系，可能忽视了全局信息。综合使用这两种模型可以弥补彼此的不足，提高微博短文本的向量化表示能力，并更全面地反映文本的语义特征。

因此，本章将 LDA 的文档主题分布向量和文本加权词向量纵向拼接，形成融合特征向量 ATL：

$$ATL = \begin{bmatrix} a_{11} * tfidf_1, a_{12} * tfidf_1, \dots, a_{1t} * tfidf_1 \\ a_{21} * tfidf_2, a_{22} * tfidf_2, \dots, a_{2t} * tfidf_2 \\ \dots \\ a_{k1} * tfidf_k, a_{k2} * tfidf_k, \dots, a_{kt} * tfidf_k \\ p_1, p_2, \dots, p_k \end{bmatrix} \quad (4-4)$$

通过在维度层面上将低维稠密的文本加权词向量 AT 和文档主题分布向量 L 进行纵向拼接，可以获得一个低维稠密的向量。这种方法有效地解决了短文本数据高维稀疏的问题。在语义层面上，融合特征的向量包含了文本的全局语义、词汇顺序信息和深层语义关联信息。通过对词向量进行加权，可以降低噪音词汇的干扰。通过这种融合特征的表征方式，弥补了 LDA 模型和词向量模型的缺点，丰富了短文本向量的语义信息。

简而言之，通过将加权词向量和文档主题分布向量进行融合，我们得到了一个维度较低、稠密且具有丰富语义信息的向量表示。这种融合方式解决了短文本的高维稀疏问题，同时保留了全局语义和词汇顺序信息，提高了短文本向量的语义表达

能力。

4.4 基于融合特征的 K-means 主题检测

K-means 算法属于基于划分的聚类算法，是一种被普遍使用的聚类算法。本章基于主题模型和加权词向量的融合特征计算相似度得到相似度矩阵。基于主题模型和加权词向量的融合特征，使用余弦相似度（Cosine Similarity）来计算文档之间的相似度。本章同样采取肘部法则来确定 k 值以解决初始聚类簇数 k 的值对结果的影响较大的问题。

综合以上，基于主题模型和加权词向量的融合特征的 K-means 文本聚类的算法描述如下：

1. 整合经过数据预处理后的微博文本数据集
2. 根据 4.3.3 节描述构建主题模型和加权词向量的融合特征的特征矩阵
3. 使用肘部法则确定 k 个聚类中心。
4. 迭代更新聚类中心：重复以下步骤直到达到停止条件：
 - a. 使用余弦相似度计算每个文档与每个聚类中心之间的距离。
 - b. 将每个微博文本分配给距离最近的聚类中心。
 - c. 将每个聚类中所有微博文本向量的平均值更新为每个聚类的中心。
5. 当微博文本的聚类分配不再发生变化时，停止迭代。
6. 输出聚类结果：最终得到 k 个聚类，每个类别下的文本代表一个话题。

4.5 实验与结果分析

4.5.1 实验数据

为了对比基于主题模型和加权词向量的融合特征的话题检测与基于单一 TF-IDF 特征的话题检测的效果，本章采用的实验数据与 3.7.2 节相同，均是 5 月 4 日爬取到的 3014 条原始微博文本。

根据新浪微博自带的话题标签文本以及人工标注，为这 3014 条微博文本标注了

话题标签，经过标注共有 8 个标签，分别是“二次阳”、“云南保山地震”、“巴厘岛身亡”、“性骚扰”、“江阴枪击事件”、“孩子摸螺蛳溺水身亡”、“唐山透水事故”和“美国 CIA 网络攻击”。如图 4.3 所示，在文件新添加了一列“label”为各个微博文本的话题标签。

A	B	C	D	E	F	G	H	I	J	K	L	M
页码	微博id	微博作者	发布时间	微博内容	转发数	评论数	点赞数	发布于	ip属地_城	ip属地_省	ip属地_国家	label
	2 4. 898E+15	OohBakery	5/4/2023 14:30	一次阳和二	0	4	1	海南	海口	海南	中国	0
	2 4. 898E+15	海海71590	5/4/2023 13:46	#二阳#我第一	0	0	0	河北	张家口	河北	中国	0
	2 4. 898E+15	wmx是我	5/4/2023 13:44	我室友刚刚测	0	0	0	广东	广州	广东	中国	0
	2 4. 898E+15	弦语聆听	5/4/2023 13:44	#新冠# 并不	0	1	6	北京	北京	北京	中国	0
	2 4. 898E+15	六月里的易	5/4/2023 12:59	我们家身体最	0	0	0	四川	成都	四川	中国	0
	2 4. 898E+15	时尚8銚銚	5/4/2023 12:33	嗓子疼 肿的	0	0	0	重庆	重庆	重庆	中国	0
	2 4. 898E+15	木啊木啊木	5/4/2023 12:07	同事：感觉最	0	0	0	浙江	温州	浙江	中国	0
	2 4. 898E+15	啧啧切	5/4/2023 12:01	记录一下二	0	6	1	湖北		湖北	中国	0
	2 4. 898E+15	一只北方	5/4/2023 11:10	2023. 4. 22	0	0	0	上海	上海	上海	中国	0
	2 4. 898E+15	爱吃奥凸凸	5/4/2023 10:46	想知道二	0	1	0	甘肃	天水	甘肃	中国	0
	3 4. 898E+15	带猎狗的兔	5/4/2023 10:34	朋友的哥 五	0	0	0	广东	深圳	广东	中国	0
	3 4. 897E+15	柳佳忆	5/3/2023 21:21	#阳了# 我20	13	72	136	江苏	南京	江苏	中国	0
	3 4. 898E+15	蒲英儿520	5/4/2023 10:18	朋友们，所以	0	0	0	江苏	连云港	江苏	中国	0
	3 4. 898E+15	补钙缺钙	5/4/2023 10:07	五一收假回来	0	3	0	广东	广州	广东	中国	0
	3 4. 898E+15	漫漫ui	5/4/2023 9:20	#新冠 查漏	0	18	29	山东	济南	山东	中国	0
	3 4. 898E+15	喂喂怪小李	5/4/2023 9:44	好久没运动突	0	1	0	河南	南阳	河南	中国	0
	3 4. 898E+15	用户64975	5/4/2023 9:29	不知道是不是	0	0	0	安徽	宿州	安徽	中国	0
	3 4. 898E+15	你算哪块	5/4/2023 9:28	#新冠 查漏	0	0	2	江苏		江苏	中国	0
	3 4. 898E+15	Slrius111	5/4/2023 9:24	#阳了#五一	0	0	1	广西	南宁	广西	中国	0
	3 4. 898E+15	开心雅尔	5/4/2023 8:54	#新冠 查漏	3	7	5	浙江	金华	浙江	中国	0
	4 4. 898E+15	Vc不拿铁	5/4/2023 8:48	今天发现我好	0	1	0	内蒙古	赤峰	内蒙古	中国	0
	4 4. 898E+15	勇往直前	5/4/2023 7:39	#新冠 查漏	0	1	4	安徽	滁州	安徽	中国	0
	4 4. 898E+15	慧小尤	5/4/2023 7:37	#新冠 查漏	0	2	5	陕西	西安	陕西	中国	0
	4 4. 898E+15	扮酸	5/4/2023 7:29	#分不清自己	0	1	2	江苏	徐州	江苏	中国	0
	4 4. 898E+15	风俱静	5/4/2023 6:58	靠，不会是二	0	1	0	湖北	武汉	湖北	中国	0
	4 4. 898E+15	以后少生气	5/4/2023 5:48	#新冠 查漏	0	1	3	广西	桂林	广西	中国	0
	4 4. 898E+15	星空放飞	5/4/2023 4:07	#阳了#二	0	3	6	陕西	榆林	陕西	中国	0
	4 4. 897E+15	一只渴望	5/4/2023 2:25	鼻子堵流清鼻	0	0	2	湖北	武汉	湖北	中国	0
	4 4. 897E+15	我是珺珺5	5/4/2023 1:28	#新冠 查漏	0	0	2	广东	东莞	广东	中国	0
	5 4. 897E+15	失屋外	5/4/2023 1:18	#阳了# 比二	0	1	0	福建	泉州	福建	中国	0
	5 4. 897E+15	無光享受T	5/4/2023 1:04	#阳了#真的二	0	0	0	湖南	永州	湖南	中国	0
	5 4. 897E+15	白羊不傲娇	5/4/2023 0:59	我怀疑我二	0	0	0	河南	郑州	河南	中国	0
	5 4. 897E+15	水瓶座的有	5/4/2023 0:47	二次阳能不	0	0	0	四川	成都	四川	中国	0
	5 4. 897E+15	小乐	5/4/2023 0:37	#阳了#最近	0	0	0	新疆		新疆	中国	0
	5 4. 897E+15	剧综学姐	5/4/2023 0:33	#费曼 为了唐	30	31	35	广东	湛江	广东	中国	0
	5 4. 897E+15	此生影随心	5/4/2023 0:29	#阳了# 公司	0	0	0	湖北	鄂州	湖北	中国	0

图 4.3 微博文本添加话题标签号

4.5.2 评价指标

精确率（Precision）、召回率（Recall）和 F1 值是用于评估分类模型性能的常见指标。它们主要用于衡量模型在预测正例和负例上的准确性和完整性。

1. 精确率（Precision）：精确率是指在所有被模型预测为正例的样本中，实际上属于正例的比例。它衡量了模型对正例的预测准确程度^[19]。在本节中，指预测为主题 i 的微博中实际主题为 i 的比例。精确率的计算公式如下：

$$precision = \frac{TP}{TP + FP} \quad (4-5)$$

其中，TP（True Positives）表示真正例的数量，FP（False Positives）表示假正例的数量^[20]。精确率的取值范围是 0 到 1，数值越高表示模型的预测准确性越高。

2. 召回率（Recall）：召回率是指在所有实际为正例的样本中，模型正确预测为正例的比例。它衡量了模型对正例的识别能力和查全率。在本节中指实际主题为 i 的微博中被预测为 i 的微博比例。召回率的计算公式如下：

$$recall = \frac{TP}{TP + FN} \quad (4-6)$$

其中 TP（True Positives）表示真正例的数量，FN（False Negatives）表示假反例的数量。召回率的取值范围也是 0 到 1，数值越高表示模型的识别能力和查全率越好。

3. F1 值：F1 值是精确率和召回率的调和平均值，综合考虑了两者的性能。它是一个综合指标，用于平衡精确率和召回率之间的权衡。F1 值的计算公式如下：

$$F1 = \frac{2 * (precision * recall)}{precision + recall} \quad (4-7)$$

本文采用综合评价指标 F1 值衡量模型的主题聚类效果。首先，分别计算每个主题的精确率和召回率，再利用宏平均求得整个模型的精确率和召回率，最后求得模型的 F1 值。

4.5.3 实验结果与分析

针对微博文本数据，分别使用第三章介绍的基于 TF-IDF 特征的 K-means 文件聚类和本章基于主题模型和加权词向量融合特征的话题检测对其进行话题检测，并进行结果的对比。其中基于 TF-IDF 特征的话题检测的 F1 值为 73.2，而基于融合特征的话题检测的 F1 值为 84.7。实验对比结果如表 4.1 所示。

表 4.1 话题检测效果对比

话题检测方法	精确率	召回率	F1 值
TF-IDF+K-means	71.6	73.8	72.6
融合特征+K-means	83.7	85.2	84.4

基于 TF-IDF 计算词汇权重的方法用于构建文本特征向量时，由于微博短文本中同一词汇的低频率，导致许多特征权重为零，进而引起高维稀疏性问题。此外，该方法未考虑文本的潜在语义信息，从而降低了主题聚类效果。

将多个特征融合并应用 **K-means** 主题聚类方法，可以得到更好的效果。通过该方法，精确率、召回率和 F1 值都能超过 80%。融合特征的方法在一定程度上解决了微博短文本的高维稀疏和语义缺失问题，能够更全面、准确地表达微博文本的信息。同时，将低维稠密词向量和低维语义空间向量结合在一起，并未引起特征维度的大幅增加。融合特征涵盖了文本的全局语义和词汇的深层语义信息。此外，通过对词向量进行 TF-IDF 加权，还能提高主题聚类的准确性。

4.6 本章小结

本章主要针对微博短文本的稀疏性及上下文语义欠缺、提取单一 TF-IDF 特征无法提取上下文语义信息的问题，本章采用了特征融合的方法：**LDA** 文档-主题分布特征和加权 **Word2Vec** 词向量特征。通过将这两种特征结合起来，构建了微博短文本的深层特征，并利用 **K-means** 聚类算法进行文本主题检测，以提高话题检测的效果。这种融合特征的方法能够充分考虑微博文本的语义信息，并在特征表示上更加全面和准确。

经过实验对比，基于主题模型和加权词向量的融合特征的话题检测的效果好于基于单一 TF-IDF 特征的话题检测模型。

参考文献

- [1] 谢修娟, 李香菊, 莫凌飞. 基于改进 K-means 算法的微博舆情分析研究[J]. 计算机工程 与科学, 2018, 40(01):155-158.
- [2] 檀娟伢. 中文微博的热点话题发现[D]. 安徽大学, 2014.
- [3] 方一向. 多视图微博话题检测方法研究[D]. 哈尔滨工业大学, 2012.
- [4] 杨开平. 基于语义相似度的中文文本聚类算法研究[D]. 西安: 西安电子科技大学, 2018.
- [5] 蒋斌. 基于停用词处理的汉语语音检索方法[D]. 哈尔滨工业大学, 2008.
- [6] 贾君霞, 王会真, 任凯, 康文. 基于句向量和卷积神经网络的文本聚类研究[J]. 计算机工程与应用, 2021, 10: 59.
- [7] Lu Yuchang, Lu Mingu, Li Fan, et al. Analysis and construction of word weighting function in VSM[J]. Journal of Computer Research & Development, 2002, 39(10):1205-1210.
- [8] Blei D M, Ng A, Jordan M I. Latent dirichlet allocation[J]. The Journal of Machine Learning Research, 2012, 3:993-1022.
- [9] Rumelhart D E, Hinton G E, Williams R J. Learning Representations by Back Propagating Errors[J]. Nature, 1986, 323(6088):533-536.
- [10] Kenter T, Borisov A, Rijke M D. Siamese CBOW: Optimizing Word Embeddings for Sentence Representations[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016.
- [11] Geoff, Hollis, Chris, et al. Extrapolating human judgments from skip-gram vector representations of word meaning[J]. Quarterly Journal of Experimental Psychology, 2016, 70(8):1-45.
- [12] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2021-2-15]. <https://arxiv.org/abs/1810.04805>.
- [13] 郝丽丽, 薛禹胜, K.P.WONG, 徐泰山. 关于电力系统动态仿真有效性的评述[J]. 电力系统自动化, 2020, 34 (10) : 1-7.
- [14] 李晓瑜, 俞丽颖, 雷航等. 一种 K-means 改进算法的并行化实现与应用[J]. 电子科技大学学报, 2017, 46(01):61-68.
- [15] 刘梦颖. 基于频繁词集的微博热点话题发现技术研究[D]. 北京: 北京工业大学, 2021.
- [16] Zhang Yun-tao, Gong Ling, Wang Yong-cheng. An improved TF-IDF approach for text classification[J]. Journal of Zhejiang University SCIENCE A, 2005, 6(1):49-55.
- [17] 聂昕. 基于图注意力网络的短文本分类研究[D]. 湖北: 华中科技大学, 2020.

- [18] 颜端武, 梅喜瑞, 杨雄飞, 朱鹏.基于主题模型和词向量融合的微博文本主题聚类研究[J].2021, 41 (10): 67-74.
- [19] 叶璞钰.基于缺失数据的理财产品推荐模型与算法[D].江苏: 南京大学, 2020.
- [20] 梅文娟.基于生鲜电商在线评论的情感分析及有用性预测[D].江苏: 南京财经大学, 2020.

致谢

本论文是在指导老师 xxxx老师的悉心指导下完成的，从毕业设计选题、构思到修改、定稿，感谢她孜孜不倦地指点。在整个过程中给我提出了很多有益的意见建议，不断督促我完善精进，其严谨严格的治学态度深深地影响着我。其次我还要感谢信息安全专业的所有老师，是他们为我奠定好扎实的专业知识，让我在今后的学术道路上能够走得更加自信从容。

特别感谢我的家人，把无私的爱奉献给了我。长久以来，他们在学业上默默支持我，在生活无微不至关心我。在我遇到挫折时，总能给予我无限的力量。

还要感谢我的同窗和亲爱的室友，一起求学的时光如此珍贵，将是我这辈子的美好回忆。

心存感念，胜过千言，唯有祝愿，永祈心间。

作者：

2023 年 05 月 18 日