中国石油大学（华东）
CHINA UNIVERSITY OF PETROLEUM

本 科 毕 业 设 计（论文）

题　目：基于主题模型与深度学习的微博文本分类及地理可视化

——以地面塌陷事件为例

学生姓名：党　辉

学　　号：1401060203

专业班级：测绘工程 14-2 班

指导老师：王海起

2018 年 6 月 21 日

基于主题模型与深度学习的微博文本分类及地理可视化

摘　要

随着大数据时代的到来，以及自然语言处理技术的飞速发展。使用文本数据挖掘来进行灾害事件的分析与预测已经成为一种趋势。文本分类作为自然语言处理中的基础和热点，在其中也起着至关重要的作用。

本文针对微博文本的特点，通过抓取与路面塌陷有关的文本数据，使用主题模型 BTM 作为无监督的概率统计模型在文本语义挖掘上进行训练从而取得良好的效果。通过主题聚类的思路进行文本的语料库构造，从而使得文本数据的标签更加准确可控。随着 word2vec 词向量学习工具的提出，深度学习近几年在自然语言处理领域取得了前所未有的发展，文本的特征序列与神经网络完美结合，在情感分析、文本分类等各领域都取得了长足的进步。模型的准确度已经不仅仅是评价性能的唯一指标，时间成本已经被考虑在内。Facebook 提出 FastText 模型，实验结果表明，在短文本处理以及时间成本上来讲，均要优于常规的 CNN、RNN 模型。文本采用该模型在通过主题聚类得到语料库的基础上进行模型训练，取得了很好的效果。同时也利用 GIS 技术对事件进行了相关空间分析，从而为后续的热点事件挖掘以及政府决策提供有力支撑。

实验表明，本文的方法在微博短文本分类上取得了良好的效果。能够有效的实现非监督模型基础上的文本分类，并且可以达到很高的工作效率。

关键词：网络爬虫；文本分类；BTM；可视化；神经网络

Weibo Text Classification and Geographic Visualization Based on Topic Model and Deep Learning

Abstract

With the arrival of the era of big data, and the rapid development of natural language processing technology. Using text data mining to analyze and predict disaster events has become a trend. As a basic and hot point in natural language processing, text categorization also plays a crucial role in it.

Based on the characteristics of Weibo texts, this paper uses the topic model BTM as an unsupervised probabilistic statistical model to train textual semantic mining to obtain good results by capturing text data related to road collapse. The corpus of text is constructed through the idea of clustering topics, which makes the tags of text data more accurate and controllable. With the development of the word2vec word vector learning tool, deep learning has achieved unprecedented development in the field of natural language processing in recent years. The feature sequence of the text is perfectly combined with the neural network, and it has made great achievements in various fields such as sentiment analysis and text classification. progress. The accuracy of the model is no longer the only indicator to evaluate performance, the time cost has been taken into account. Facebook proposed the FastText model. The experimental results show that both the short text processing and the time cost are better than the conventional CNN and RNN models. The model was trained on the basis of the subject's clustering of texts using this model, and achieved good results. At the same time, it also uses GIS technology to conduct relevant spatial analysis of events, so as to provide strong support for subsequent hot-spot event mining and government decision-making.

Experiments show that the method of this paper has achieved good results in the classification of microblog short texts. The text classification based on the unsupervised model can be effectively implemented, and high work efficiency can be achieved.

Keywords： Web crawler; Text classification; BTM; Visualization; FastText

目 录

第 1 章　绪论

随着移动互联的广泛普及，尤其是近几年来互联网大数据、信息技术的迅速发展，人们无时无刻不在产生着海量的数据信息；并且人们获取和交换信息的渠道也愈来愈丰富。在日常生活，人民通过手机短信、微博、网站评论等形式交换信息表达观点。这些信息具有信息篇幅短、表达灵活、数据规模大、用词不规范等特点，它们存在于我们生活的各个方面，因此提高这类信息的处理能力变得尤为重要，文本分类是信息处理其中的一种最基本的形式，分类对事件分析决策、热点事件发现、危险事件预测等都具有非常重要的价值。同时结合 GIS 空间分析技术，对短文本进行分析更具有深刻的意义。

1.1 研究背景

地面塌陷是在自然或人为因素作用下所导致的地表岩土下沉，从而形成地面陷坑的一种动态地质现象。它具有隐蔽性、突发性、群众性和多方面的特点。在人口密度大的城市，路面坍陷是威胁城市安全运行的重要隐患。造成路面塌陷的因素有地下水位变化、冻融、公路建设质量、水管渗漏、车辆超载、地下工程施工、暴雨等。虽然城市坍塌的规模很小，但由于人口的集中，对城市造成的破坏是巨大的。它作为城市地质灾害，往往会成为一个突发事件，影响城市生活。因此掌握路面塌陷的类型与分布对路面塌陷灾害的治理具有一定的必要作用。

随着移动互联的快速发展，像微博、twitter、网站评论等短文本信息开始充斥着人们的日常生活。我们能够第一时间获取到最新信息，其中很大一部分都是短文本。这些短文本具有长度较短、数量庞大等特点，其中包含人们对社会事件的阐述与看法。作为文本挖掘领域中的一个基本方向和重要方向，文本分类是涉及数据科学、信息科学与计算机科学的一个问题。其是指在已知的分类系统的前提下，根据主题或根据其他属性对文本进行分类的过程。文本分类在信息处理方面占有举足轻重的作用，用户可以根据分类结果高效的找到自己想要的信息，更快的完成对信息的提取与总结。本文就是实现对其海量短文本的分类体系评价，从而借助于 GIS 可视化技术进行分析。

1.2 国内外研究现状

1.2.1 短文本分类研究现状

近年来，随着移动互联技术的飞速发展，诸如网站评论、社交软件等各类文本信息的爆炸式增长，特别是像微博、twitter、朋友圈这种社交文本，每天更是以数以亿计的量在增长。如何在海量的文本中挖掘有价值的信息就变成整个信息处理工作的主要任务。短文本是指上述表达信息简洁（通常不超过 140 个字符）以及包含信息量较少的文本数据。短文本在其形式特征和内容表达方面都有与长文本数据不同，因此传统的分类方法在短文本分类方面的性能不同于长文本，效果不够理想，不能直接并有效地应用于短文本的分类处理中。

针对短文本分类问题，国外学者分别提出多种方式对短文本进行分类处理（比如：短文本相似度计算、Web 核函数、频繁词集挖掘、潜在语义索引等）以提高短文本分类的效果。M Sahami,TD Heilman 等人通过利用网络搜索结果来为短文本提供更大的背景，从而测量短文本片段（即使没有任何重叠术语的片段）之间的相似性[1]，Danesh Irani 等人通过用户发布在 twitter 端上的话题趋势与主题相关性热度来对文本进行分类[3]，Xuan-HieuPhan 等人利用大规模数据集合中发现的隐藏主题，提出了构建分类器的一般框架，该分类器可以处理短而稀疏的文本和 Web 段[2]。当然还有一些学者使用深度学习算法对短文本分类的尝试。TextCNN 是利用卷积神经网络对句子进行分类的算法，是由 Yoon Kim 于 2014 年提出的算法[4]。

中文短文本任务不同于英文短文本任务，它们在很多方面是存在着很大差异的，比如表达形式、结构、编码方式等；具体的中文中会出现大量不同的地名、人名以及网络新词。所以英文短文本分类处理的方法亦或改进方法都不能很好的应用在中文短文本的分类处理相关任务上，其根本原因是词语的使用更为灵活，中文短文本中所包含的信息很少，词的意义也更加模糊，文本表示的矩阵较为稀薄，其特征空间的维数较大，很难从其中抽取准确而关键的特征用于分类；同时互联网短文本大多都是实时更新的，速度快且数量庞大；另外短文本用词不规范，造成的特征噪声非常多，这就增加了中文短文本处理的难度。目前国内的学者也做了很多有关中文短文本分类的工作，有一些学者研究短文本的特征提取扩展方式，例如 wang 等利用抽取具有依存关系的词对以扩充短文本特征集合增加文本的特征[5]，从而完成分类；ning 等借助知网从语义方面通过衡量词的相似度，提出基于领域词语本体的短文本分类[6]；zhang 等通过

主题进行上下文区分，同时通过主题关联来减少稀疏性，采用 K 近邻方法对短文本进行分类[7]。这些方法的背后都依赖于大规模的数据支持，处理背景文本库需要消耗大量的精力。在中文微博短文本的研究方法上，虽然截止到 2018 年第一季度，微博月活跃用户已经超过 4 亿，但是由于国内关于短文本的研究起步时间较晚，并且国内各微博服务商目前没有开放微博数据获取方面的服务，因此微博的研究还处于探索阶段。然而，微博这一社交平台对人们日常生活的影响广泛而深刻，并且目前针对微博的研究内容大多都主要涉及文本的情感分类、观点识别等一些二分类问题上。因此针对微博这类中文短文本的多类别分类是非常有意义的。

1.2.2 主题模型研究现状

主题建模是一种经常使用在自然语言处理领域的文本挖掘工具，用于发现文本主体中隐藏的语义结构。它有助于发现文档中存在的隐藏的主题模式，根据这些主题可以注释文档；使用这些注释来组织、搜索和汇总文本。主题模型可以被描述为从一组文档中查找一组词汇（主题）的方法，这些文档最能代表该集合中的信息。它也可以被认为是文本挖掘的一种形式—一种在文本材料中获得重复出现的单词模式的方式。主题模型也被称为概率主题模型，是将高维度的文档-单词向量空间映射到低维度的文档-主题和主题-单词空间，最终得到文档-主题和主题-单词的概率分布。主题模型的起源是潜在语义分析,该方法早期自然语言处理常用到向量空间模型（Vector Space Model，VSM）[8]，这种方法简洁明了易实现，但对于法处理自然语言的模糊性的问题（包括一词多义和多词近义）空间向量也需要克服。举例来讲，“小米”一词，从一个角度来讲可以认为是手机品牌，从另一个角度讲也可以表示一种谷物，在特定语境中也能用来表示某一宠物的昵称。人类可以通过上下文轻松的识别这些词的多种不同含义，不过对于计算机来说，是相对困难的。如果用向量空间模型来表示文本，则会将同一词在不同语境下的不同用法与含义视为完全相同的。对于多词近义问题，向量空间模型中这些词各占一维，被视为完全无关，无法表达其语义上的相近关系，比如“大概”、“大约”、“大体”、“大致”、“大抵”、“或许”、“好像”、“可能”、“差不多”、“兴许”、“约摸”等，这些词都可以用来表示推测或估计。以上一词多义和多词近义在自然语言处理问题中相当常见，向量空间模型也制约着计算机对文档的处理和理解。人在探索着各类能够刻画词汇语义关联的文本表示方法的同时，主题模型就应运而生。

Deerwester 等学者提出潜在语义分析模型（Latent Semantic Analysis，LSA）[9]，该算法采用隐含的高阶结构在术语与文档的关联中的优势（“语义结构”）将词汇映射到潜在的语义空间。LSA 能够除去噪声，捕获常在同一篇文档中出现的近义词，被广泛应用于信息检索领域。但 LSA 不能很好地解决一词多义的问题，也无法解释在奇异值分解中所产生的负值的含义。Hofmann 等学者提出一种基于计数数据因子分析的统计潜在分类模式的自动文档索引新方法（Probabilistic Latent Semantic Analysis，PLSA）[10]。PLSA 是从潜在的语义分析发展而来的一个更健全的概率模型。PLSA 解决了一词多义和多词近义的问题，但可能出现过拟合。Blei 等学者提出一个文本和其他离散数据集合的生成概率模型—潜在狄利克雷分配模型（Latent Dirichlet Allocation，LDA）[11]，LDA 模型引入了主题和词汇的潜在狄利克雷分布，较好的解决了 PLSA 模型的过拟合问题，是一个三级分层贝叶斯模型。Yan 等学者提出 LDA 变种主题模型（Biterm Topic Model，BTM）[12]，BTM 模型通过直接建模整个语料库中词共现模式的生成来学习主题；它使用整个语料库中的聚合模式进行主题学习，一定程度上解决了短文本特征稀疏对主题建模所造成的问题。同样，它也是第一个通用于并面向短文本分类的主题模型。

主题模型作为一种挖掘抽象主题的机器学习方法，不仅可以挖掘语料潜在语义信息，最重要的是可以对语料库进行有效降维（如图 1-1 所示）。即可以将高维的无明显语义的语料矩阵转化为低维的有明显语义的矩阵，最终形成“文档-主题”、“主题-词汇”概率分布矩阵。另外，主题模型也具有很强的灵活性，它可以通过增强学习来实现主题挖掘。因此主题模型在文本、图像等领域得到了广泛应用。

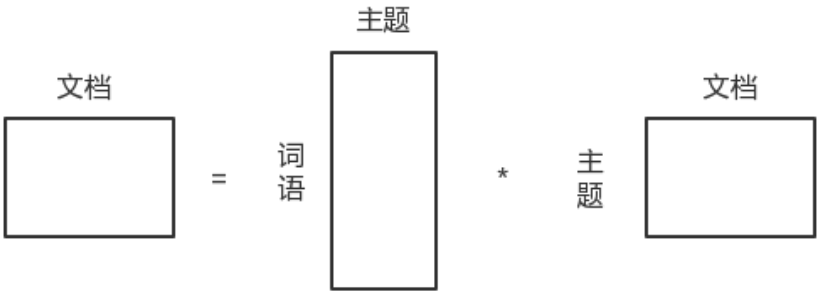

图 1-1 主题模型降维示意图

1.2.3 基于深度学习的文本分类研究现状

目前常用的文本分类模型有 TF_IDF、K 最近邻算法（KNN）、人工神经网络、决策树（Decision Tree）、支持向量机（SVM）、潜在的语义索引等。由于特征的提取对文本分类的效果具有非常大的影响，因此基于传统分类模型的文本分类方法的工作重点主要集中在特征提取和选择上，常用方法有 TF-IDF 文档相似度、文档频次、N-gram 特征、词频等。

近年来，随着深度学习在自然语言处理领域的应用，学者们将文本分类研究的重点转向了基于人工神经网络的分类模型的研究。特别是随着 Google 开源了基于深度学习的词向量学习工具 word2vec，基于深度学习的文本分类方法逐渐成为目前文本分类的主流。神经网络处理文本分类的优点之一就是不需要花费大量的时间在特征提取与选择上，并且将词的分布式表示作为特征输入到网络中，神经网络可以自动提取出相关有价值的信息。因此基于深度学习的文本分类模型得到快速发展。目前基于深度学习的文本分类研究主要有两个方向：一个是基于神经网络的文本表示方法（词向量）的研究，一个是基于神经网络结构上的研究。

词向量或者词嵌入（word embedding）[13]是自然语言处理（NLP）中的一组语言建模和特征学习技术的集体名称，是来自词汇表的单词或短语被映射到实数的向量。从概念上讲，它涉及从每个单词一个维度的空间到具有更低维度的连续向量空间的数学嵌入。单词和短语嵌入作为基础输入表示时，可以提高 NLP 任务的性能，如句法分析和情感分析等。生成这种映射的方法包括神经网络、词共现矩阵的降维、概率模型以及词的出现环境的显式表示。由于神经网络结构较为复杂而且灵活，普通的方法表示 N-gram 时，随着 N 的增大而使总数呈指数级增长，但是神经网络通过共享参数的形式，使得只用线性倍数的参数即可以表示 N-gram，由此神经网络可以表示复杂的上下文。Bengio[14]等人于 2003 年首次用神经网络的词向量表示应用到统计语言模型中，2010 年 Turian[13]等人通过词向量表示应用到半监督学习中，2013 年 Mikolov 等人先后提出了基于 word2vec 方法的 CBOW(Continuous Bag of words)模型和利用来自上下文的 concurrence 的信号来诱导词嵌入的 Skip-gram 模型[15][16][17]，从而可以快速从大型语料库学到准确的词向量。文献[18][19]从字词联合的角度学习词向量，文献[20][21][22][23]从字符级别的角度学习词向量。文献[24]提出上下文滚动矩阵和目标单词的词向量的拼接来代表这个单词，这样的词向量更具有隐含的语义。文献[25]提出一种简单而高效的文本分类和表征学习方法 FastText，该结构是一个学习由 Facebook 的 AI Research（FAIR）实

验室创建的词嵌入和句子分类的库，其模型架构类似于 word2vec 中的 CBOW 模型，区别在于 CBOW 模型通过上下文来预测中间词，而 FastText 模型可以直接输入文本来预测标签，而且 FastText 模型加入了 N-gram 特征，具有很高的运算效率，类似于卷积操作。还有 Yook 提出的基于卷积网络的分类方法[26]，Liu 等的基于 RNN 的文本分类方法[27]，Yang 等的基于 rnn 和 attention 模型的分类方法[28]和 Lai 的基于 rcnn 的文本分类方法[29]等。这些方法在文本分类任务中都取得了很好的性能。

如图 1-2 所示展示了基于深度学习的文本分类过程。基于深度学习的文本分类问题的重点是解决文本表示（文本的表示一定程度上影响着分类的精度以及运算效率），然后利用人工神经网络结构自动获取特征表达能力，最终实现解决文本分类问题的目的。

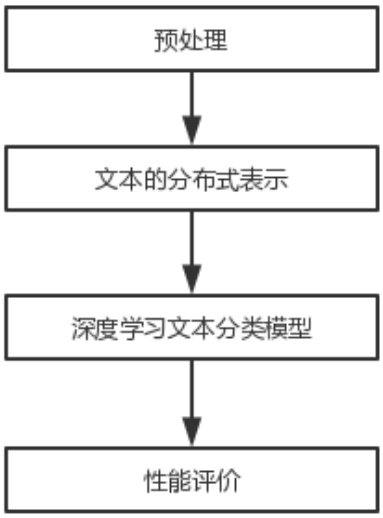

图 1-2 基于深度学习的文本分类过程

1.3 技术路线及研究内容

1.3.1 技术路线

基于网络爬虫微博关键词获取原始数据，以微博文本数据为研究对象，采用 BTM 模型挖掘微博短文本的各主题文档之间的最优相似度，并通过对其进行聚类获得微博文本数据的各类别。以此来构建文本语料库，产生训练集与测试集。采用基于深度学习平台的 FastText 模型对所产生的训练集进行模型训练，完善模型性能；并对主题类

别与分类结果进行可视化分析，从而获取微博文本主题的空间分布模式。技术路线如图 1-3 所示。

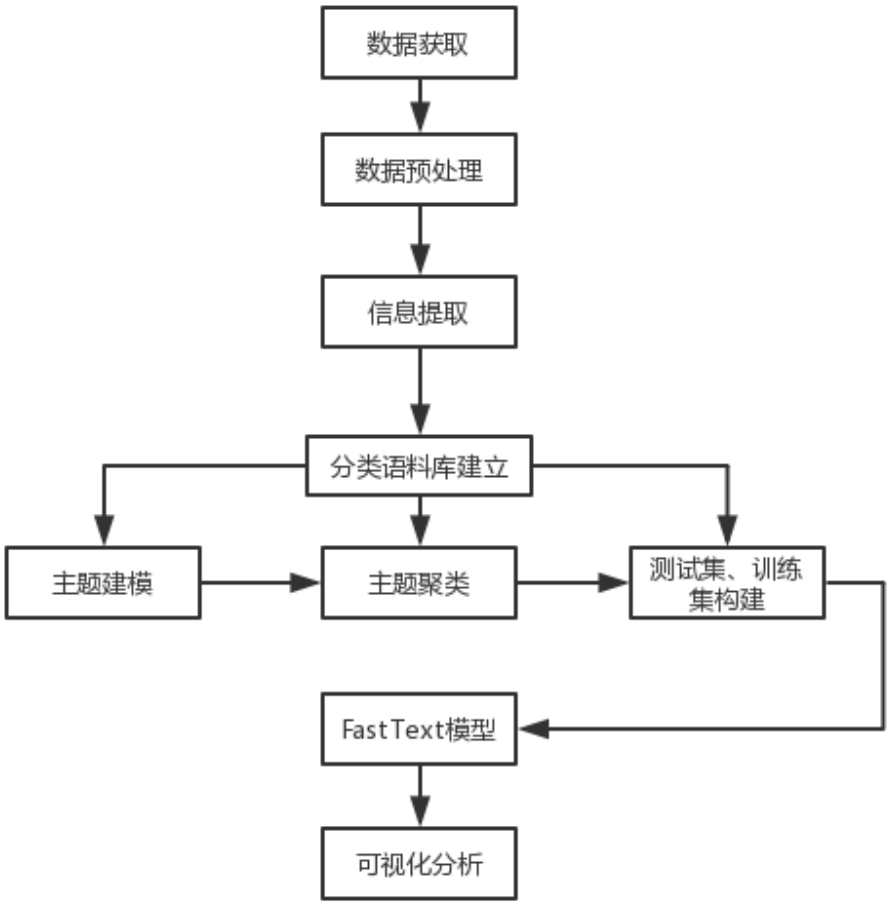

图 1-3 技术路线

1.3.2 研究内容

本文的研究内容主要分为三部分：

第一部分：微博文本数据的获取。首先选择了八爪鱼等多类网络采集工具，多次抓取会出现身份验证的问题。然后采用网络爬虫的方法，利用 selenium 测试工具，模拟微博登录；设置了自动身份验证的判断。抓取了在微博页面可以直观看到的文本数据，同时也获取到了每条文本所对应的时间、设备等各类信息。为后续的工作提供了数据基础。

第二部分：分类语料库的建立。首先比较了 LDA 模型与 BTM 模型在处理文本分类问题上的差异。微博文本数据属于短文本，在多个数据集上结果表明 BTM 在这方面更有优势。本文应用 BTM 模型对数据进行文本主题建模。对文档—主题相似度进行聚类从而完成分类，从而构建语料库。

第三部分：FastText 模型的训练及检验。FastText 模型是由 Facebook 的 AI Research（FAIR）实验室创建的词嵌入和句子分类的库。该模型还可以用来获取单词的矢量表示。它提供了一种简洁且高效的文本分类和表征学习的方法，其性能可以与深层神经网络媲美，而且其速度也会更快。因此本文选用 FastText 模型，以便取得了良好的效果。

1.4 本文组织结构

本文内容一共包括七章，根据研究内容，其安排如下：

第一章：绪论。主要介绍了本文的研究背景和研究内容，分析了国内外相关研究的现状以及前沿进展，并进一步阐述了本文的技术路线及主要研究内容。

第二章：相关算法原理与方法。主要介绍在本文中所用到的相关算法原理，包括主题模型 BTM、神经网络 FastText 模型、K-means 聚类等算法，以及中文分词等原理。

第三章：数据预处理。主要介绍了文本数据的获取、清洗、信息提取以及中文分词等过程。

第四章：主题模型与语料库的建立。将每条微博文本视为一篇文档，基于 BTM 模型进行微博短文本主题建模，训练主题模型从而使其达到最佳性能。针对 BTM 主题类型，在此基础上聚类，获取与该主题相关的充分的用于深度学习分类训练的语料库，构建测试集与训练集。

第五章：基于 FastText 浅层神经网络的文本分类。筛选与该主题相关的微博文本数据，利用获取的文本类别语料库训练 FastText（卷积神经网络）模型，之后借助训练好的模型对微博文本的词向量进行类别划分。

第六章：可视化分析。采用 GIS 可视化的方法对微博文本主题分类结果进行空间分析；并采用词云的方式对主题词进行展示，从而分析其内在联系。

第七章：总结和展望。总结全文，分析其中的不足，并提出在今后的研究工作中需要改进的方向。

第 2 章　相关算法原理与方法

2.1 主题模型概述

主题模型是对文本中潜在的主题进行建模的一种方法，可以以此挖掘文本各抽象主题。主题可以看作是词项的概率分布（即类似单词的集群），其中的每个词是经过这个词选择某一主题的概率的大小来决定的。各主题模型中，LDA 主题模型由于其最少的基本假设，而应用最为广泛。

2.1.1 BTM 主题模型概述

随着近几年互联网的广泛普及，以微博、网站评论、短新闻等为媒介越来越受欢迎，对应的短文本也呈现显著增长的趋势。LDA 等原始主题模型均以文档为基础进行建模，对长文本可以取到不错的效果。但将其处理方法直接应用于短文本，常常无法达到预期效果。

BTM 模型[12]作为 LAD 的变体模型，与 LDA 的区别就是通过对文档的整个语料库进行建模，从而实现主题学习，该方法从一定程度上避免了短文本所造成的数据稀疏。BTM 模型通过建模词共现模式来强化学习，在整个学习过程不需要任何其他外部数据的帮助。同时这也是第一个通用于短文本的主题模型。

2.1.2 BTM 主题模型定义

设语料库为 $D=\{ d_1,d_2,\ldots,d_{ND} \}$，其中 d_i 为一篇文档，语料中涉及的双词集合为 $M=\{m_1,m_2,\ldots,m_{Nm}\}$，其中 $b_i=(\omega_{i,1}, \omega_{i,2})$，为一个双词。$z \in |1, K|$表示某个主题，$K$ 表示主题数目，θ 表示语料库对应的主题分布，是狄利克雷分布，其先验参数为 α；φ 表示主题的词汇分布，也是一个狄利克雷分布，其先验参数为 β。

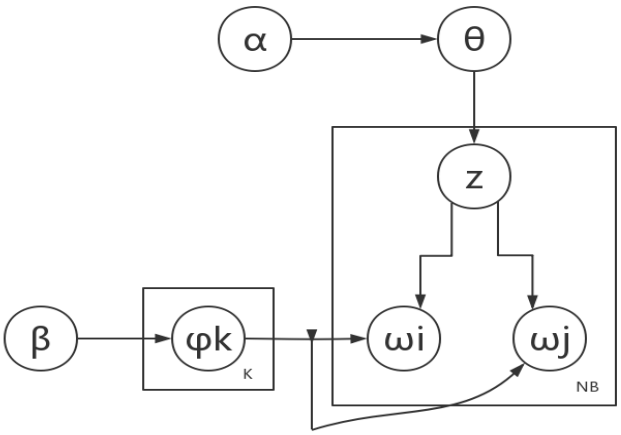

图 2-1 BTM 模型图

根据 BTM 模型，概率分布文档的产生过程见算法 1:

算法 1：BTM 模型文档的制作过程。

输入：α、β、K

输出：文档

1.对整个语料采样主题分布 $\theta \sim Dir(\alpha)$

2.对每个主题 $\kappa \in |1, K|$

采样一个词汇分布 $\varphi_k \sim Dir(\beta)$

3.对每个双词 $m_i \in M$

采样一个主题 $z_i \sim Mult(\theta)$

根据主题，独立采样两个词 $\omega_{i,1}, \omega_{i,2} \sim Mult(\varphi_{zi})$

2.1.2 BTM 主题模型参数估计

在统计学中，当直接采样困难时，吉布斯采样（Gibbs Sampling）是可用于从特定的多变量概率分布中获得一系列近似的观察序列的方法。与其他马尔科夫链蒙特卡罗理论（Markov Chain Monte Carlo，MCMC）算法一样吉布斯采样（Gibbs Sampling）也是用来生成样本的马尔可夫链，每个样本斗鱼附近的样本相关。通常利用吉布斯采样对 BTM 主题模型做参数估计，依据链式法则：

$$\boldsymbol{P}\ (\boldsymbol{z_i}|\boldsymbol{z}_{\neg \boldsymbol{i}},\boldsymbol{B}\ ) = \frac{\boldsymbol{P}(\boldsymbol{z},\boldsymbol{M})}{\boldsymbol{P}(\boldsymbol{z}_{\neg \boldsymbol{i}},\boldsymbol{B})} \propto \frac{\boldsymbol{P}(\boldsymbol{M}|\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})}{\boldsymbol{P}(\boldsymbol{M}|\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})} \tag{2-1}$$

¬表示去除当前词的影响，式(2-1)表明双词 i 的主题可以根据其他词汇的主题得到，反复迭代这一过程，循环采样，直至收敛。

式(2-1)中的四个因子可按照式(2-2)-(2-5)计算：

$$\begin{aligned} P(M|z) &= \int P(M|z,\Phi)P(\Phi)d\Phi \\ &= \int \left(\prod_{i=1}^{N_M} P(m_i|z_i,\varphi_{zi}) \right) P(\Phi)d\Phi \\ &= \int \prod_{k=1}^{K} \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \prod_{w=1}^{W} \varphi_{k,w}^{n_{kw}+\beta-1} d\varphi_k \right) \\ &= \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^K \prod_{K=1}^{K} \frac{\prod_{w=1}^{W}\Gamma(n_{kw}+\beta)}{\Gamma(n_k+W\beta)} \end{aligned} \tag{2-2}$$

$\Gamma(.)$是标准 Gamma 函数，n_{kw}表示主题为 k 的词中，词 ω 出现的次数，n_k为语料中主题为 k 的双词数目，同理可得：

$$P(z) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^k} \frac{\prod_{K=1}^{K}\Gamma(n_k+\alpha)}{\Gamma(N_B+K\alpha)} \tag{2-3}$$

$$P(B_{\neg \boldsymbol{i}}|z_{\neg \boldsymbol{i}}) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W}\right)^K \prod_{K=1}^{K} \frac{\prod_{W=1}^{W}\Gamma(n_{kw,\neg \boldsymbol{i}}+\beta)}{\Gamma(N_{k,\neg \boldsymbol{i}}+W\beta)} \tag{2-4}$$

$$P(z_{\neg \boldsymbol{i}}) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^k} \frac{\prod_{K=1}^{K}\Gamma(n_{k,\neg \boldsymbol{i}}+\alpha)}{\Gamma(N_M-1+K\alpha)} \tag{2-5}$$

将式(2-2)-式(2-5)代入式(2-1)，化简可得：

$$P(z_{\boldsymbol{i}}|z_{\neg \boldsymbol{i}},M) \propto (n_{k,\neg \boldsymbol{i}}+\alpha) \frac{(n_{kw_{i,1},\neg \boldsymbol{i}}+\beta)(n_{kw_{i,2},\neg \boldsymbol{i}}+\beta)}{(n_{k,\neg \boldsymbol{i}}+W\beta)^2} \tag{2-6}$$

在 BTM 模型中，语料的主题分布 θ 和主题的词汇分布 φ 可按公式(2-7)和公式(2-8)估计：

$$\varphi_{kw} = \frac{n_{kw}+\beta}{n_k+W\beta} \tag{2-7}$$

$$\theta_k = \frac{n_k+\alpha}{n_M+K\alpha} \tag{2-8}$$

其中 n_m 为双词总数。

在 BTM 主题建模过程中，可以通过文档中的双词主题来计算整个文档中每个主题的概率分布。假设文档 d 包含 N_d 个双词$\{m_{ij}|j \in [1, N_d]\}$，则该文档中主题 k 的比例为：

$$P(z=k|d) = \sum_j^{N_d} P(z|m_{i_j})P(m_{i_j}|d) = \sum_j^{N_d} \frac{\theta_k \phi_{k,w_{i,j},1}\phi_{k,w_{i,j},2}}{\theta_{k'}\phi_{k',w_{i,j},1}\phi_{k',w_{i,j},2}} \frac{n_d(b_{i_j})}{N_d} \tag{2-9}$$

其中，$\frac{n_d(b_{i_j})}{N_d}$表示文档 d 中的双词 m_{ij} 出现的次数。

2.2 深度学习 FastText 模型概述

FastText[25]是 Facebook 的科学家 Tomas Mikolov 在 16 年开源的一个文本分类器。 它提供了一种简单而高效的方法对文本进行分类和表征，分类效果可以与深度学习模型相媲美，而且速度更快效率更高。与支持向量机（SVM），Logistic 回归（Logistic Regression）和神经网络（neural network）等其它文本分类模型相比，FastText 不仅具有更好的分类性能，并且大大缩短了训练时间。

2.2.1 FastText 模型架构原理

FastText 模型主要包含模型架构、层次 Softmax 函数 和 N-gram 特征三个部分。FastText 是一个用于文本分类与表示的库，它将文本转换为连续的向量，可以用于任何与语言相关的任务。具体的，它可以将一个词（一段文本或者一句话)作为输入，将单词作为平均文本的特征表示，然后将其反馈给线性分类器。进一步映射出句子所对应的标签，最终得到该序列所对应的类别。它使用 Softmax 函数来计算预定义类中的概率分布，该模型在多个 CPU 上使用随机梯度下降和线性衰减学习速率进行异步训练。FastText 只能在 CPU 上工作以获得可访问性。这就是说，FastText 已经在可以在 GPU 上运行的 caffe2 库中实现。

2.2.2 FastText 模型架构

FastText 模型与 word2vec 中的 CBOW 模型都是由 Facebook 的 AI Research（FAIR）实验室的科学家 Tomas Mikolov 提出来的，模型架构类似；而且确实 FastText 也算是由 word2vec 所衍生出来的。不同之处在于，FastText 模型是以整个文档作为输入数据，并有特定的输入格式（如：凌晨 小区 建筑工地 基坑 发生 坍塌 致使 东门外 发生 大面积 大坑 深有 这件 事故 小区 东门 暂时 封闭 小区 院内 停放 多辆 趴窝 业主 郁闷 分享 小区 _label_zero），最终结果是预测标签，而 CBOW 模型要以上下文作为输入，最终结果是预测中间词。

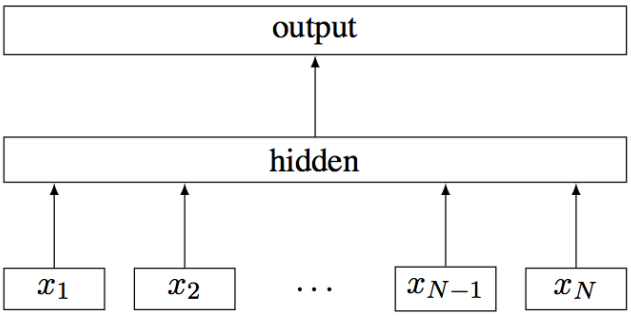

图 2-2 FastText 模型结构

2.2.3 改善运算效率——Softmax 层级

当数据集的数量很大时，计算线性分类器的计算量就会很大。FastText 使用了分层分类器（而不是扁平化分类器），不同的类别被集成到树形结构中（设想一个二叉树而不是一个列表）。由于大多文本分类的任务是属于多类别的，并且计算线性分类器很繁琐。为了优化运行时间，提升运算效率，FastText 模型采用了基于霍夫曼编码的分层技术。当搜索最可能类别时，分层 Softmax 技术在测试时间上也是很有利的，该技术是基于霍弗曼编码的，利用霍夫曼编码的结构可以大大地减少模型预测目标的数量，从而达到改善运算效率的作用。

一般来说，在进行分类问题时，某些类别的样本数要远大于其他样本，这就容易造成某些类别出现频率比其他类别更大。FastText 模型使用 Huffman 算法来建立用于表示文档类别的树形结构。霍夫曼编码首先计算每个字符出现的次数，排序后将其转换为二叉树，即可发现出现频率越多其层级数就越高，编码也越短；出现频率越少其层级数就越低，编码也越长。从而相对出现频率大的类别所对应的树形结构的深度要比出现频率小的类别的树形结构的深度要小，可以大幅度提高无损压缩的比例。这使得进一步的计算更加有效。

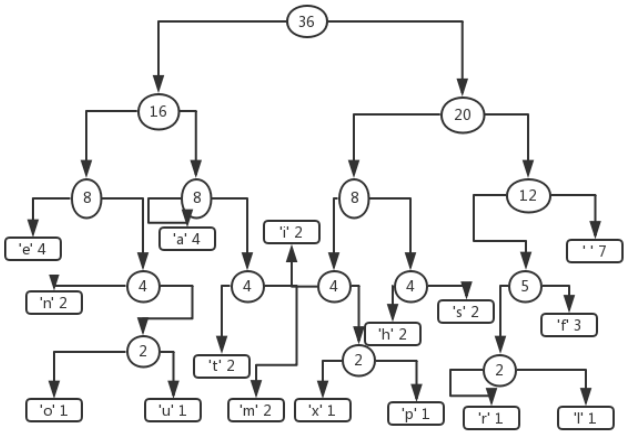

图 2-3 霍夫曼编码

2.2.4 N-gram 特征

在自然语言处理的相关任务中，词袋模型是经典的文本表示方法。在此模型下，像是句子或是文档这样的文字都可以用一个袋子的方式来表现。这种表现方式不考虑文档以及词的顺序，所以词袋模型对于各个词之间的顺序无法进行有效的区分，所涉及到的计算成本是非常昂贵的。相反，使用 N-gram 作为附加功能来捕捉有关词序的部分信息将会取得很好的效果。词序的表示对大多数文本分类问题来说是非常重要的，词序不同，文本就会产生不同的含义，比如“我想你”与“你想我”。而 FastText 模型充分考虑到了这一问题，使用向量表示单词 N-gram 特征，考虑到了句子的局部词序信息，通过对低频的 N-gram 进行过滤。实验表明，运算效率确实得到了进一步的提升。

2.3 中文分词原理概述

分词是文本挖掘数据预处理至关重要的一步。根据汉语分词的定义,中文分词就是将一段中文句子根据一定的词语组合切分成一个个词语的过程。例如“我要毕业了”的切分结果为“我\要\毕业\了”。自从 1983 年,北京航空航天大学的研究人员实现了我国第一个有较高切分精度、可实用的现代书面汉语自动分词系统 CDWS[32]。由于中文基本文法的独特性，汉语分词还有很长的路要走。国内的各相关领域的学者也都做了很多研究,并且提出了许多有效的算法。这些算法大致可分为以下几类，包括基于字符串匹配的、基于统计的、基于知识理解的分词方法。本研究采用结巴（jieba）中文分词组件。

2.3.1 结巴（jieba）分词模式

结巴（jieba）分词具有三种模式（精确模式、全模式、搜索引擎模式）。本研究采用第三种分词模式—搜索引擎分词。该模式是指在精确的将句子切分开之后，再次对长词进行切分，从而提高查全率，对本文的实验数据（基于微博关键词搜索的文本）比较适用。

2.3.2 结巴（jieba）分词原理

（1）jieba 分词采用了一种基于 Trie 树结构的算法。分词器通过整个语料库中根据 dict.txt 生成 Trie 树结构，同时将每个词出现的次数转换成了频率。基于 Tire 树生成了有向无环图（DAG），在 DAG 中记录的是某个词的开始位置，从而以便通过全模式来进行分词。

通过分析 jieba 分词的源码可以发现，jieba 分词本身就包含了一个有 2 万多词条的词典，词典中的词即为分词的基础依据。基于 Trie 结构的扫描就是将这些词条放到 Trie 的树结构之中，一旦扫描到的词条中和该词典中的词条具有相同的前缀，那么就实现了快速查找，从而实现分词。

（2）首先搜索将要进行切分的分词句子中已经切分好的词语，如果字典中没有该词，就将字典中最小频率附给它。在 jieba 分词中，还采用了动态规划的方法来查找最大概率路径。该方法对句子反向计算概率，最终得到最大概率的切分组合。

通过分析 jieba 分词的源码可以发现，jieba 分词不仅仅将字典中的词条放到 Trie 树结构中，同时，还将每个词的出现次数转换为了频率。

（3）除了分词器中自带的词典，jieba 分词器采用了 Viterbi 算法为处理未出现的词，并且将用于汉字成词的 HMM 模型也应用其中。

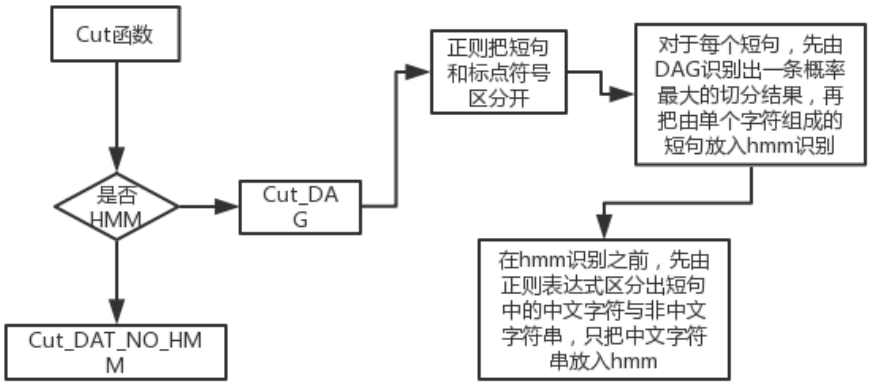

图 2-4 结巴（jieba）分词流程图

2.4 K-means 聚类算法

聚类算法（clustering algorithm）是探索性数据挖掘领域中很重要的任务。聚类分析可以通过各种算法（如：层次聚类、DBSCN 聚类等）来实现，聚类是把类似的对象通过静态分类的方法组成不同类别的组级或者更多的组级，使得在同一类别中的对象具有相似的特征，在二维空间上通常具有更短的距离。进行聚类分析的算法有很多，其中 K-means 聚类算法在众多聚类算法里属于比较简单的，属于基于一种基于距离的无监督聚类算法，实现起来也比较容易，效果也能够满足要求，所以具有广泛的应用。

K-means 聚类算法的实现过程见算法 2:

算法 2： K-means(K 均值)聚类。

输入： N，K

输出： K 个聚类

---

1. 指定 K 为聚类中心

2. 计算数据集中各个数据对象与各聚类中心之间的距离（欧式距离、余弦距离等），将这些数据对象按照各距离的大小进行类别划分

3. 根据聚类结果，计算每个类群中新的聚类中心。计算方法为取所包含的数据对象的平均值

4. 按照新的聚类中心，对数据对象重新聚类

5. 循环流程 2 到 4 步，直到聚类中心不再变化，停止运算

上述算法中，N 表示原始数据集中的数据对象的个数，K 表示预先规定的聚类个数。输入原始数据对象，规定聚类个数为 K，经过运算输出满足误差平方和准则函数的 K 个聚类。

第 3 章　数据预处理

3.1 数据获取

本文使用 selenium 测试工具模拟浏览器行为登录微博，结合 PhantomJS，在谷歌浏览器的载体，分析 DOM 节点，通过 Xpat 找到登录框对节点信息进行获取，实现与关键词（路面塌陷）有关文本的捕获，并存储至 Excel 中。 获取的微博信息包括：博主昵称、博客主页、微博内容、发布时间、微博地址、微博来源、转发、评论、赞等。共 7017 条数据，时间范围从 2017 年 1 月 1 日至 2018 年 4 月 14 日。

3.1.1 网络爬虫原理

网络爬虫，又称网络机器人，可以通过浏览互联网页面进行数据爬取。一个网络爬虫从一个被请求访问的网页地址，它可以识别出页面中的所有超链接并将这些链接添加到要访问的 URL 列表中。抓取工具即可按照预先设计好的规则，自动获取网络上的网页信息的脚本。由于互联网上的页面数量非常庞大，所以不同的网页都会有特定的标志，即每个网页都有自己特有的统一资源定位符（URL）。网络爬虫也可以验证超链接和 HTML 代码，从而用于网络抓取。

网络爬虫的工作原理：

（1）通过与服务器建立连接，向服务器发送一个请求，请求一个网页；

（2）服务器接收到 http 请求，从而做出响应，并将该响应返回给网络采集脚本的源码；

（3）爬虫机器人通过分析网页，获得网页中所包含的统一定位符，并将其加入队列；

（4）若队列中仍有未被抓取的，返回步骤继续执行。

3.1.1 网络爬虫主要流程

（1）用 selenium 模拟鼠标和键盘登录微博主页；

（2）访问微博搜索页面，模拟点搜索框，输入搜索关键词、时间范围、博主信息等；

（3）保存当页为 html 文件，模拟点击下一页；

（4）直到没有下一页，处理保存下来的 html。在页面有内容的前提下提取微博文本所对应的博主昵称、博主主页、微博内容、微博地址等信息，并写入到 Excel 中。

3.2 信息提取

本文采用 zhang 等[30]提出的基于角色标注的中文机构名识别算法，提取各条文本中的地址名。在 zhang 等搭建的中科院 NLPIR 汉语分词系统实体抽取系统上，系统能够智能的识别出文本中所出现的人名、地名等各类实体关键词，对本文的地名实体抽取具有很重要的作用。本文通过该系统对获取的微博文本数据进行地名实体抽取，并基于百度地图 API 采用 Geocoding 批量将文字地址转化为坐标地址，为后面的地理可视化提供基础。

3.3 数据清洗及中文分词

3.3.1 数据清洗

由于本文研究仅限国内地区的路面塌陷事件。为此需要去掉原始文本数据中与描述路面塌陷无关的文本以及所描述的事件发生在国外的文本；由于微博文本中，有较多文本很短，并不能很好的对事件进行描述表达，为此，本文进行中文分词后剔除掉了各条文本中词向量小于 3 的微博文本。最终共得到 5586 条微博文本作为实验数据。

3.3.2 中文分词

分词是自然语言处理中最基础的工作，词语分割的精度对之后任务的精度有很大影响。因此文本或句子的结构表征是语言处理的核心任务。本文使用结巴（jieba）中文分词器对实验文本数据进行分词，并去除停用词，取得了良好的效果。

第 4 章　主题模型与语料库的建立

4.1 主题提取

4.1.1 困惑度

聚类算法的质量评价通常两种方法，一种是使用具有分类标签的测试数据集，然后使用特定的算法（比如 Normalized Mutual Information，Variation of Information distance 等算法）来判断聚类结果与真实结果的差距；另一种是使用没有分类标签的测试数据集，用训练的模型来运行测试数据集，然后在测试数据集的基础上来计算一篇文档属于某主题的不确定程度，即就是 perplexity 指标[33]。这个指标是聚类质量的一个通用指标，不需要先验分布，它最初应用于语言的建模。它被定义为给定模型的几何倒数或者其等价物的测试语料库的可能性均值。其公式如下：

$$P(\widetilde{\mathcal{W}}|M) = \prod_{b=1}^{B} P(\widetilde{w}_{\bar{b}}|M)^{-\frac{1}{N}} = exp - \frac{\sum_{b=1}^{B} logP(\widetilde{w}_{\bar{b}}|M)}{\sum_{b=1}^{B} N_b} \tag{4-1}$$

其中，B 是指模型训练好之后的参数，在 LDA 中，它等同于是 theta 和 phi 或者其替代物；在 BTM 中，它等同于是 pd_z 和 pz_w 或者其替代物。例如 collapse Gibbs Sampler 所获取的状态（w 向量，z 向量），中间式子中指数-1 / N 的 N 指的是 Epsilon（Nm），注意这个公式独立于 LDA、BTM 模型，它可以应用于任何聚类模型，当然也可以应用于主题模型。

4.1.2 迭代次数

迭代是一种重复运算并反馈的过程，通常其目的是为了逼近所预期的目标以及结果。每一次对过程的反复就称为一次“迭代”，而每次迭代所得到的结果将会是下一次迭代的初始值。一般来说模型训练迭代次数越多，则越收敛。在主题模型中，迭代次数是越小越好，但是必须能满足迭代收敛，具体的次数一般较少讨论。

对于处理后的数据集，我们运行了 3 个不同主题数的 BTM 模型，其中每个主题数都是作者随机分配开始的。若每一个都运行了固定的 2000 次迭代次数。对于数据集和 100 个主题解决方案，吉布斯采样器的 2000 次迭代在标准的 2.6 GHz 64 位工作站上花费 3 小时的挂钟时间。对于 300 个主题的解决方案，花费 6 个小时进行 2000 次迭代；对于 500 个主题解决方案，花费了 10 个小时进行 2000 次迭代。数据量并非巨大，但却要花费大量的时间成本。

基于前者，本文利用模型在测试文档上的困惑度指标来评估模型的性能何时开始稳定。该指标表示模型预测数据时的预测能力，但如果一味追求模型的预测能力，会出现过拟合现象，同时也会增加计算成本。从而作者同样选择 3 个不同主题的 BTM 模型，其中每个主题数都是作者随机分配开始的。分别基于 BTM 模型来计算困惑度，从而选择最优的迭代次数。在本文中，我们参数 α 和 β 在下面描述的每个实验中分别被固定在 50 /t 和 0.01。

在下图 4-1 中可以看出，基于 BTM 主题模型的性能稳定得相当快，在测试文档的困惑度方面，大约在 200 次迭代之后就可以达到一个稳定的状态。这个收敛测试在之后的主题稳定性与主题解释上都很具有说服力。

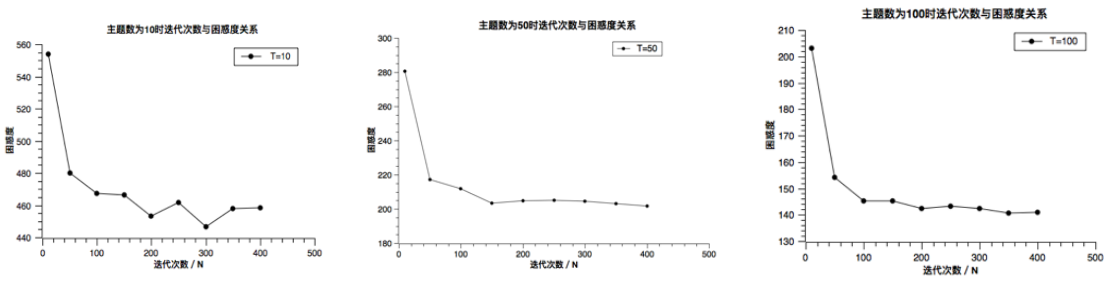

图 4-1 各主题数下迭代次数与困惑度关系

4.1.3 主题数确定

主题概率模型中的主题数同样深刻影响模型的性能。本文利用上述困惑度（Perplexity） 指标来确定 BTM 主题模型的主题数目。该指标可以用来表示模型预测数据时的模型稳定性（即文档属于哪个主题有多不确定），其取值越低，说明模型性能越稳定，主题分布效果越好。但无限小会出现过拟合现象，同时也会增加计算成本。

为提高训练效率，采用上文所讨论得到的迭代次数。将 BTM 模型的主题数目设置为5、10、25、50、100、150、200、250、300、400、500，三次实验的困惑度随主题数变化情况如图4-2所示。随着实验所设置的主题数目依次增加，困惑度指标大的大小呈逐渐下降趋势，当达到主题数目达到300个时，下降趋势不再明显。由于模型主题数目越多，预估参数就越多，计算成本就越大 ，为此本文选择T = 300确保模型最优。

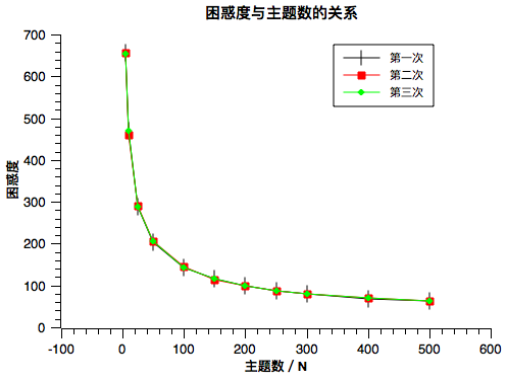

图 4-2 困惑度与主题数关系图

4.2 主题聚类

前者所确定的主题数并非最终所确定的文本分类的类别数。要确定文本的分类类别，本文通过对主题的相似性进行度量采用了 k-means 算法进行聚类，对于 k-means 算法来说，由于此时聚类的对象是各主题所对应的相似度所组成的相似性矩阵，所以采用通用的 k-means 算法即可。

聚类分析(cluster analysis)是指以文本主题之间相似性为基础把相似的对象通过静态分类的方法聚集成为不同的子集。其特点是基于相似性、并且制定多个聚类中心。K-Means[34]（又叫 K-均值）算法表示以空间中 K 个点为簇进行聚类，对靠近各中心并且小于某阈值的对象进行归类。K-Means 聚类算法的缺点是需要指定聚类个数。为此对于主题聚类的第一步是需要确定聚类个数。

作者采用斯坦福大学的 Robert 等教授提出了 Gap Statistic 方法[31]。文中定义 Gap 值为：

$$\mathrm{Gap_n(k)} = \mathrm{E_n^*(log(W_k))} - \mathrm{logW_k} \tag{4-2}$$

上式（4-2）中，因为 W_k 的值可能很大，所以做了取对数的处理。通过这个公式来找出 W_k 跌落最快的点，Gap 最大值对应的 k 值就是最佳聚类数。将各主题相似度做如上运算得到表格 4-1。

表 4-1 Gap Statistic 方法各聚类数对应指标

N	logW	E.logW	gap	SE.sim
1	2.762349	3.690860	0.9285114	0.002648199
2	2.748765	3.672360	0.9235947	0.002579454
3	2.736872	3.659458	0.9225859	0.002550605
4	2.725951	3.650173	0.9242212	0.002675857
5	2.713957	3.642500	0.9285424	0.002571045
6	2.703323	3.635520	0.9321970	0.002538346
7	2.700785	3.628850	0.9280648	0.002509393
8	2.688959	3.622621	0.9336624	0.002555760
9	2.676235	3.616545	0.9403105	.002503768
10	2.671802	3.610722	0.9389207	0.002538170

为了便于表达，作者将聚类范围缩小到 1—10 类。由 Gap 指标以及图 4-3 所示，当类别数为 9 时，GAP 值达到最高峰。所以 9 可以作为最佳聚类数。从而进行 k-means 聚类运算，对原始文本做标签处理。

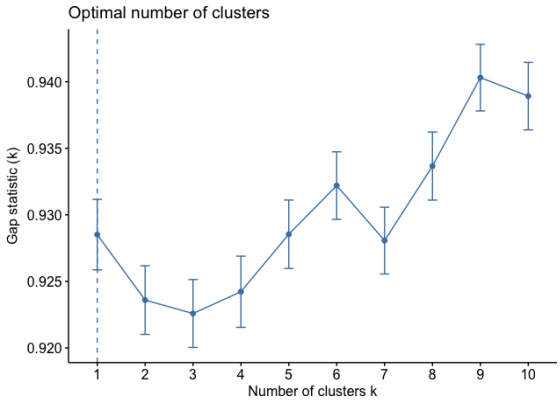

图 4-3 聚类个数与 Gap 值关系

第 5 章　基于 FastText 浅层神经网络的文本分类

FastText 是一个可伸缩解决方案的资料库，其目的是帮助创建文本表征和分类，并且可以为 294 种语言提供训练模型。深度学习 FastText 模型主要应用于构建词向量与文本分类，在许多文本分类问题上都有非常不错的表现，不仅具有良好的分类效果，而且具有优秀的分类效率，所以适用于大批量的数据样本。在各个方面都表现出了很强的能力（如：标签的预测、情感分析等），尤其是对于短文本分类具有很好的性能。同时微博文本就是典型的短文本数据，为此作者选用 FastText 模型以取得良好的效果。

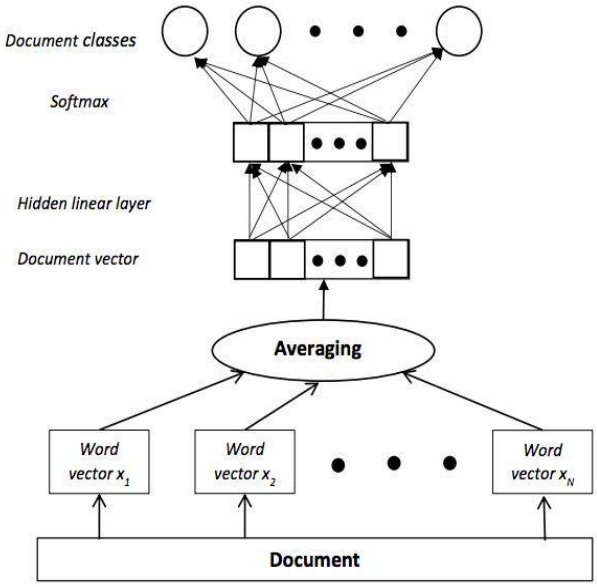

图 5-1 线性的 BOW 分类器

在本文中，使用以上步骤对原始数据（共 5586 条）进行人工标记后，将这些数据划分为九子类，选取其中 50%作为模型训练集，50%作为模型测试集。文本文件的每一行都包含一个标签列表，其后是相应的文档。所有标签都以__label__前缀开始，这就是 FastText 如何识别什么是标签或什么是一个词。然后对模型进行训练，以预测文档中给定单词的标签。

首先使用训练数据集对 FastText 模型进行训练，然后用测试数据集对分类结果进行检验，使用准确率与召回值来评估分类器对测试数据的好用程度，最终得到数据分类的 P 值（准确率）和 R 值（召回率）。其中 P 值代表某个类别中预测正确的数据数

目与该类别中预测结果总数目的比值，R 值代表某个类别中预测正确的数据数目与该类别中测试集数目的比值，F 度量值表示 P* R * 2 / (P + R)。

表 5-1 模型各类检验指标

类别	P	R	F
CLASS_ZERO	0.868421	0.970588	0.916667
CLASS_ONE	0.684132	0.686186	0.685157
CLASS_TWO	1.000000	1.000000	1.000000
CLASS_THREE	1.000000	0.956522	0.977778
CLASS_FOUR	0.875000	0.700000	0.777778
CLASS_FIVE	0.857143	1.000000	0.923077
CLASS_SIX	0.750487	0.738964	0.744681
CLASS_SEVEN	0.882353	0.937500	0.909091
CLASS_EIGHT	0.749285	0.748571	0.748928

由上表 5-1 可知，测试集的平均准确度达到了 74%，其中 class_two、class_three 具有百分之百的准确率，对于实验数据（数据量较少、质量相对较差的短文本数据）来说已经相当可观，由此可见针对短文本确实具有很高的效率。

在预测精度方面不仅可以通过对预处理数据进一步进行处理；还可以进行模型调参，比如改变时代的数量（使用选项-epoch，标准范围[5 - 500]）；改变学习率（使用选项-lr，标准范围[0.1-1.0]）；使用单词 n-grams（使用选项-wordNgrams，标准范围[1 - 5]）。在本文上述结果中，参数设置如下：lr(学习速率)、DIM（词向量大小）、Epoch（迭代次数）分别选择 0.1、100、200。

FastText 仅仅由一层神经网络构成，准确的说应该属于 Shallow Learning。但是作为有监督的算法，FastText 的效果很优秀。并且在具备预测速度快特性（实验数据 5586 条数据 2 秒内就可以给出分类结果）的同时，还要比一般的神经网络模型的精度还要高（FastText 使用词的 embedding 叠加获得的文档向量，向量的距离可以用来衡量词之间的语义相似程度）。因此 FastText 模型在同近义词挖掘、各类语言的文本分类系统都会得到很信赖的应用。

第 6 章　可视化分析

6.1 数据点分布情况

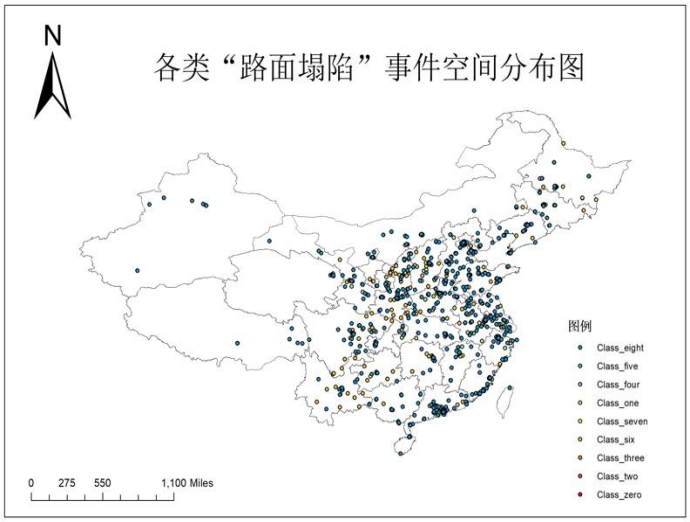

图 6-1 各类“路面塌陷”事件空间分布图

如上图 6-1 所示，路面塌陷事故的发生地主要集中在中部地区以及东南沿海地区。图中淡蓝色和浅黄色点分布较为广泛，即第一类、第六类、第八类在整个文本数据所对应的事件中占大多数。

如下图 6-2 所示，即为各文本数据所对应各类别事件的空间分布状况。同样第一类、第六类、第八类占大多数；第四类、第七类也相对较多。

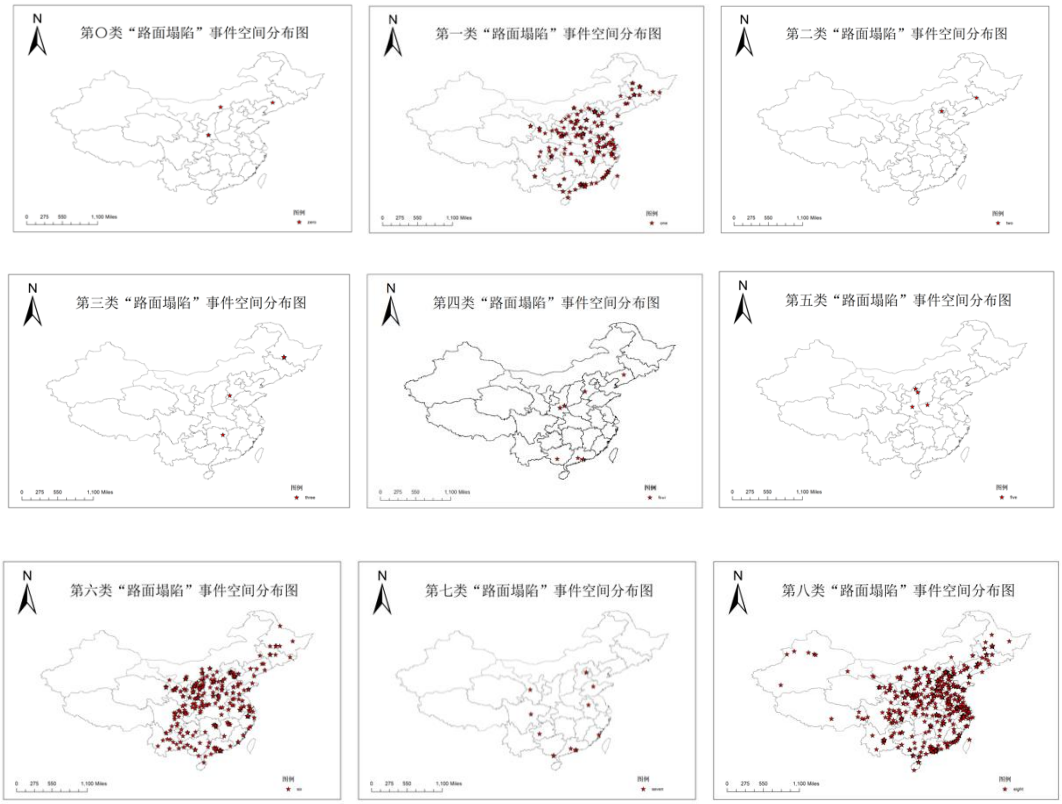

图 6-2 各类“路面塌陷”事件空间分布图（分类）

6.2 主题分布情况

图 6-3 各类词云图

基于分类结果分别利用 R 语言调用 wordcloud2 包对各类别文本数据进行词云展示。由分类词云图可见，第 0 类是大多与小区建设塌陷有关的文本；第一类是地铁施

工所造成的有关塌方事故的文本；第二类是北京西二旗路所对应的事件文本；第三类是劳斯莱斯轿车深陷大坑的事件；第四类是与货车有关的路面塌陷的事件文本；第五类是塌陷事件所对应的民警预警文本；第六类是为塌陷事件进行路面抢修的事件；第七类是电动车掉入大坑的事件文本；第八类是由于暴雨、积水等所造成的路面塌陷的事件文本。

6.3 结果分析

上述结果中，第零类、第二类、第三类，由地理可视化，这几类包含不同事件的数量很少；由主题词云展示，这几类均各指向同一类事件。由于这几类所包含的各事件具有相当数量，所以被分为了一类。查阅原始文本可知，第零类为发生在陕西某小区的路面塌陷事件；第二类为发生在北京市海淀区西二旗路的路面塌陷事件；第三类为发生在黑龙江的劳斯莱斯轿车掉入路面塌陷所产生的大坑内的事件。由主题词云展示可知，第一类大多表示与地铁建设导致的塌陷事件；第四类表示与货车有关的塌陷事件；第五类表示与交警、预警有关的塌陷事件；第六类表示与道路抢修有关的塌陷文本；第七类表示与电动车有关的塌陷事件；第八类表示与道路施工所导致的塌陷事件。

第 7 章　总结和展望

7.1 总结

本文研究基于微博短文本的分类技术。从无监督分类-BTM 主题模型入手，通过主题聚类为后续利用 FastText 模型进行分类提供基础数据。在多个实验数据上表明，该分类思路能够取得目前最好的分类效果。同时，在任何短文本分类任务上都适用。本文开展的主要工作包括：

（1）微博文本数据的获取

通过利用网络爬虫技术，模拟用户登录。基于关键词（路面塌陷）搜索，设置文本范围，从而获取文本以及其所对应的时间、用户、设备等各类信息。为获取微博数据提供了新思路，同时对于数据量不足的实验条件提供了解决办法。

（2）基于 BTM 主题模型与 K-means 构建语料库

基于 BTM 模型对微博文本语料进行建模，以文档-主题相似度分布矩阵作为文本特征的表示方法。利用困惑度指标对模型性能进行评价，从而确定最佳模型效果。同时在分类任务中，利用了 K-means 聚类的方法，以 Gap 指标作为确定聚类个数的依据。取得了很好的分类效果。

（3）基于 FastText 模型对微博文本进行分类

在文本分类的实验结果上表明，FastText 模型在大文本、短文本的实验结果上优于 CNN、RNN 等常规深度神经网络模型；并且具有非常高的运算效率。因此 FastText 更适用于本文的实验数据。

7.2 展望

虽然上述研究思路与方法对于微博文本数据都取得了相当可观的研究效果，但其中还是存在一些不足的地方，需要在之后的研究中继续改善。现总结如下：

（1）本文使用的实验数据是来自于微博关键词爬取的文本数据。由于当每条微博文本达到一定长度时，微博页面会将多余的文字隐藏，所以获取到的数据有相当部分是不能将事件描述完全的；又因为微博文本会出现转发的情况，导致会造成大量重复的文本；同时由于数据量总数少、数据质量差（短文本数据质量）等也导致实验结果不能达到完美。

（2）中文文本是文本挖掘工作的基础。在进行后续文本处理之前需要进行词语分割，分词的效果会直接影响后续的研究精度。比如：“吃饭/的/和尚/未/吃饭/的/人”，并未取得预期的效果。所以可以采用一些专业词汇的词典来配合分词会相应取得较好的结果。

（3）本文在深度学习训练上采用了常规的神经网络模型。后续可以采用一些更为优化或改进的模型（如：TextCNN），来进行类似的多文本分类工作。

致　谢

时间飞逝，经过几个月的努力，最后完成了论文的写作。借毕业论文的机会，特此向在本科学习与生活中给予我帮助的老师和家人表示深深的感谢。无论是为人还是治学，你们都是我的榜样！

感谢我的论文指导老师。首先感谢您在本科论文阶段给了我良好的学习环境和科研资源。其次感谢您带领我进入自然语言处理的美妙世界。老师很有耐心，很负责任地指导学生发现问题、排除问题、反思问题，让学生得以解决毕业设计中遇到的学术问题和实践上的困难。一直以来，老师坚持定期与我们进行学术研讨，了解学生的论文进展。从选题开始老师就开始教学生如何阅览文献，总结文献中遇到的问题；到论文的展开阶段，老师不厌其烦地聆听学生的思路和实验结果，并给出了很多非常有建设性的建议和宝贵的指导；到论文撰写阶段，您指导我如何在写作的过程中清晰地阐述问题、表达思路、总结问题。在此，再次对老师表示忠心的感谢和祝福，祝老师工作顺利，家庭幸福！

感谢学院的其他诸位老师。感谢你们曾给我的关心和帮助，老师们的言传身教使我慢慢建立起了正确的人生观和价值观，老师们的帮助和鼓励我更是感激不尽。

感谢我的父母，感谢我的家人，是你们让我有了我家的温暖与亲情的珍贵，让我在为人生奋斗的过程中无后顾之忧；是你们给予我精神与物质上的支持，让我不惧失败勇往直前；是你们对我学业的鼓励，让我对未来胸有成竹，充满信心。你们是我生命中最坚强的后盾，爱你们。

感谢我的同学们、朋友们。是你们让我的大学本科生活多姿多彩、是你们让我的大学本科变得尤为珍贵。与你们的交流，让我产生了很多新的思路与想法，这对于我顺利完成毕业论文具有很大的帮助。感谢你们，衷心祝福各位同学前程似锦。

感谢论文引用中的各位作者，你们让我有了灵感。感谢在我本科学习生活中给予我帮助的其他高校科研院所的老师，是你们让我更有动力。同时祝福培育我四年的母校，祝福母校更加辉煌。

致谢。

参考文献

[1] M Sahami,TD Heilman,A web-based kernel function for measuring the similarity of short text snippets [J],WWW 2006 May 23-26:377-386

[2] Phan X H, Nguyen L M, Horiguchi S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]// 国际万维网大会. 2008:91-100.

[3] Irani D, Webb S, Pu C, et al. Study of Trend-Stuffing on Twitter through Text Classification[C]// 2010.

[4] Kim Y. Convolutional Neural Networks for Sentence Classification[J].Eprint Arxiv,2014.

[5] 王鹏,樊兴华.中文文本分类中利用依存关系的实验研究[J].计算机工程与应用, 2010,46(3):131-133.

[6] 宁亚辉,樊兴华,吴渝.基于领域词语本体的短文本分类[J].计算机科学,2009, 36(3):142-145.

[7] 张志飞,苗夺谦,高灿.基于隐含狄列克雷分配的短文本分类方法[C]//全国青年计算语言学会议.2012.

[8] Salton G.A vector space model for automatic indexing[J].Communications of the ACM,1975,18(11):613--620.

[9] Deerwester S,Dumais S T, Furnas G W,et al.Indexing by latent semantic analysis[J].Journal of the Association for Information Science and Technology, 1990,41(6):391-407.

[10] Hofmann T.Probabilistic latent semantic indexing[C]//The 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.New York,NY,USA:ACM,1999:50-57.

[11] Blei D M, Ng A Y,Jordan M I.Latent dirichlet allocation[J].The Journal of Machine Learning Research,2003,3:993-1022.

[12] Yan X,Guo J,Lan Y,et al.A biterm topic model for short texts[C]// International Conference on World Wide Web.New York,NY,USA:ACM, 2013:1445-1456.

[13] Turian J,Ratinov L,Bengio Y.Word representations:a simple and general method for semi-supervised learning[C]// ACL 2010, Proceedings of the, Meeting of the Association for Computational Linguistics, July 11-16, 2010,Uppsala, Sweden.DBLP,2010:384-394.

[14] Bengio Y,Ducharme R,jean,et al.A neural probabilistic language model[J]. Journal of Machine Learning Research,2003,3(6):1137-1155.

[15] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[16] Mikolov T,Sutskever I, Chen K, et al. Distributed Representations of Words and Phrases and their Compositionality[J]. Advances in Neural Information Processing Systems, 2013, 26:3111-3119.

[17] Mikolov T,Yih S W,Zweig G.Linguistic Regularities in Continuous Space Word Representations[C]// 2013:296-301.

[18] Xiao Y,Cho K.Efficient Character-level Document Classification by Combining Convolution and Recurrent Layers[J]. 2016.

[19] 来斯惟.基于神经网络的词和文档语义向量表示方法研究[D].中国科学院大学, 2016.

[20] Kim Y,Jernite Y, Sontag D,et al.Character-Aware Neural Language Models[J]. Computer Science, 2015.

[21] Conneau A, Schwenk H, Barrault L, et al. Very Deep Convolutional Networks for Natural Language Processing[J]. 2016.

[22] Zhang X,Lecun Y.Text Understanding from Scratch[J]. Computer Science, 2015.

[23] Zhang X,Zhao J,Lecun Y. Character-level Convolutional Networks \for Text Classification[J]. 2015:649-657.

[24] Lai,Siwei, et al."Recurrent Convolutional Neural Networks for Text Classification." AAAI. Vol. 333. 2015.

[25] Joulin A,Grave E,Bojanowski P,et al.Bag of Tricks for Efficient Text Classification[J]. 2016.

[26] Kim Y. Convolutional Neural Networks for Sentence Classi- fication[J]. Eprint Arxiv, 2014.

[27] Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[J]. arXiv preprint arXiv:1605.05101, 2016.

[28] Yang Z, Yang D, Dyer C, et al. Hierarchical attention net- works for document classification[C]//Proceedings of NAACL-HLT. 2016: 1480-1489.

[29] Lai S, Xu L, Liu K, et al. Recurrent Convolutional Neural Networks for Text Classification[C]//AAAI. 2015, 333: 2267-2273.

[30] 俞鸿魁,张华平,刘群.基于角色标注的中文机构名识别[C]//Advances in Computation of Oriental Languages--Proceedings of the, International Conference on Computer Processing of Oriental Languages.2003.

[31] Tibshirani R, Walther G, Hastie T. Estimating the Number of Clusters in a Data Set via the Gap Statistic[J]. Journal of the Royal Statistical Society, 2001, 63(2):411-423.

[32] 梁南元. 书面汉语自动分词系统—CDWS[J]. 中文信息学报, 1987, 1(2):46-54.

[33]Heinrich G. Parameter Estimation for Text Analysis[J]. Technical Report, 2008.

[34] Hartigan J A, Wong M A. Algorithm AS 136: A K-Means Clustering Algorithm[J]. Journal of the Royal Statistical Society, 1979, 28(1):100-108.