第八届泰迪杯
数据挖掘挑战赛

“智慧政务”中的文本挖掘应用

摘要

近年来，随着网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统对提升政府的管理水平和施政效率具有重要意义。

针对问题一，该题为中文数据文本多分类问题，本文提出了 **Word2vec+Logistic Regression 模型**以及 **Text-CNN 模型**对留言内容进行分类，前者为经典机器学习模型，后者是基于深度学习的卷积神经网络模型。主要步骤是将原始数据进行划分和数据清洗，接着对每条留言进行分词和**词向量化**等一系列操作；最后送入分类器中得到结果，并对两个模型进行评估。实验数据表明，通过比较发现 Text-CNN 模型的 F1-Score 要高于 Logistic Regression 的经典模型。总体来说，该分类模型更优，能够投入实际应用。

针对问题二，该题为热点信息文本挖掘问题，本文首先进行数据初始化，转换成易处理的数据格式，然后使用基于 **CRF** 的实体命名识别进行关键词识别，本文提出一种**基于词频矩阵和简单匹配系数的组内聚类算法**，能够将留言信息有效分类。对于留言的信息类别，本文设计了一套热度分配算法 **HAM**，算法考虑了留言类别中的留言数、同义信息以及评论信息，基于此，提出了一个热度矩阵 M，能够有效对留言信息进行热度评价。最终对 M 进行降序排列，得到了热度留言归类排序。总体来说，该算法的评价结果能够得到符合实际的结果。

针对问题三，实验从答复意见的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，建立了**答复意见质量的评价指标和评价模型**。良好的答复意见提取了内容完整性、有条理、相关且专业性这三个特征，较差的答复意见主要特征是套用模板或回复相似度较高。结合预设的各项评价指标权重，采用加权的方法，可以得到相关部门答复意见质量评价分数。

关键词：**智慧政务　Word2vec　答复意见质量评价模型　Text-CNN　HAM**

Abstract

In recent years, as the online questioning platform has gradually become an important channel for the government to understand public opinion, gather people's wisdom, and gather public opinion, the amount of text data related to various social conditions and public opinion has been increasing, giving the past the main reliance on manual to divide the message and hot spot The work of the department has brought great challenges. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of a smart government system based on natural language processing technology is of great significance to improve the government's management level and governance efficiency.

Aiming at problem 1, the title is multi-classification of Chinese data text. This article proposes the Word2vec + Logistic Regression model and the Text-CNN model to classify the message content. The former is a classic machine learning model, and the latter is based on deep learning convolutional neural Network model. The main steps are to divide and clean the original data, and then perform a series of operations such as word segmentation and word vectorization on each message; finally, it is sent to the classifier to get the results, and the two models are evaluated. The experimental data shows that the F1-Score of the Text-CNN model is higher than the classic model of Logistic Regression through comparison. Overall, the classification model is better and can be put into practical use.

Aiming at problem 2, the topic is hot spot information text mining. This paper first initializes the data, converts it into a manageable data format, and then uses CRF-based entity name recognition for keyword recognition. This paper proposes a word frequency matrix and simple matching Coefficient clustering algorithm can effectively classify message information. For the message category of the message, this paper designs a set of heat distribution algorithm HAM. The algorithm considers the number of messages, synonymous information and comment information in the message category. Based on this, a heat matrix M is proposed to effectively heat the message information. Evaluation. Finally, M is sorted in descending order, and the hot messages are sorted. Overall, the evaluation results of the algorithm can get results that are in line with reality.

In response to question 3, the experiment gives a set of evaluation schemes for the quality of the response opinions from the perspectives of relevance, completeness, and interpretability of the response opinions, and establishes evaluation indicators and evaluation models for the quality of the response opinions. Good response comments

extract the three characteristics of content integrity, organization, relevance, and professionalism. The main feature of poor response comments is the application of templates or the similarity of responses. Combining the preset weights of various evaluation indicators and using the weighted method, the quality evaluation scores of the opinions of relevant departments can be obtained.

Keywords: Smart government affairs Word2vec word vectorization reply quality evaluation model Text-CNN HAM

目录

1 问题重述

1.1 问题背景

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。请利用自然语言处理和文本挖掘的方法解决下面的问题。

1.2 要解决的问题

1.2.1 群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

1.2.2 热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映 入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关 部门进行有针对性地处理，提升服务效率。请根据附件 3 将某一时段内反映特定地点或特定 人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果。

1.2.3 答复意见的评价

针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

2 问题分析

2.1 问题一的分析

为了能够解决群众留言的标签分类问题，本文提出了 Word2vec+Logistic Regression 模型以及 Text-CNN 模型来解决这个文本多分类问题，问题一的操作流程图如图 2-1 所示，前者为经典机器学习模型，后者是基于深度学习的卷积神经网络模型。主要步骤是根据附件 2 所提供的数据，对数据集进行划分和数据清洗，接着对每条留言进行分词和词向量化等一系列操作；后送入分类器中得到结果，并计算准确度，召回率以及 F1 值等评测指标，对两个模型进行评估选择最优的分类模型。最后所选择的训练模型得到分类结果。

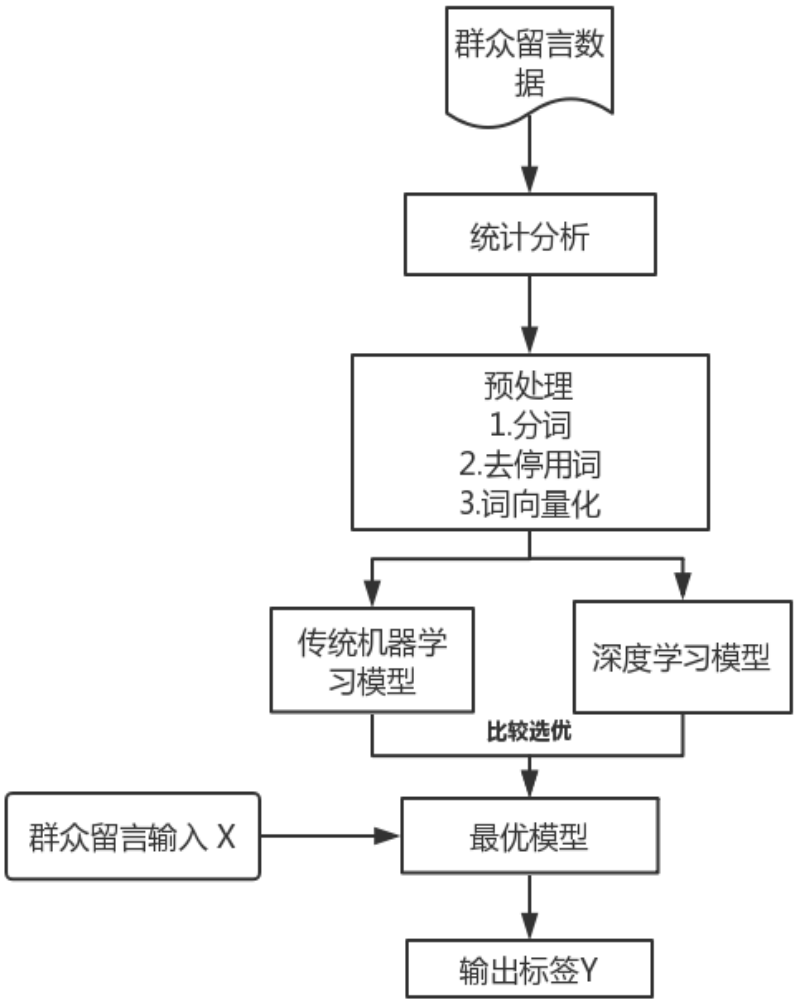

图 2-1 问题 1 流程图

2.2 问题二的分析

该题为热点信息文本挖掘问题，本文首先进行数据初始化，转换成易处理的数据格式，然后使用基于 **CRF** 的实体命名识别进行关键词识别，本文提出一种基于词频矩阵和简单匹配系数的组内聚类算法，能够将留言信息有效分类。对于留言的信息类别，本文设计了一套热度分配算法 **HAM**，算法考虑了留言类别中的留言数、同义信息以及评论信息，基于此，提出了一个热度矩阵 M，能够有效对留言信息进行热度评价。最终对 M 进行降序排列，得到了热度留言归类排序。

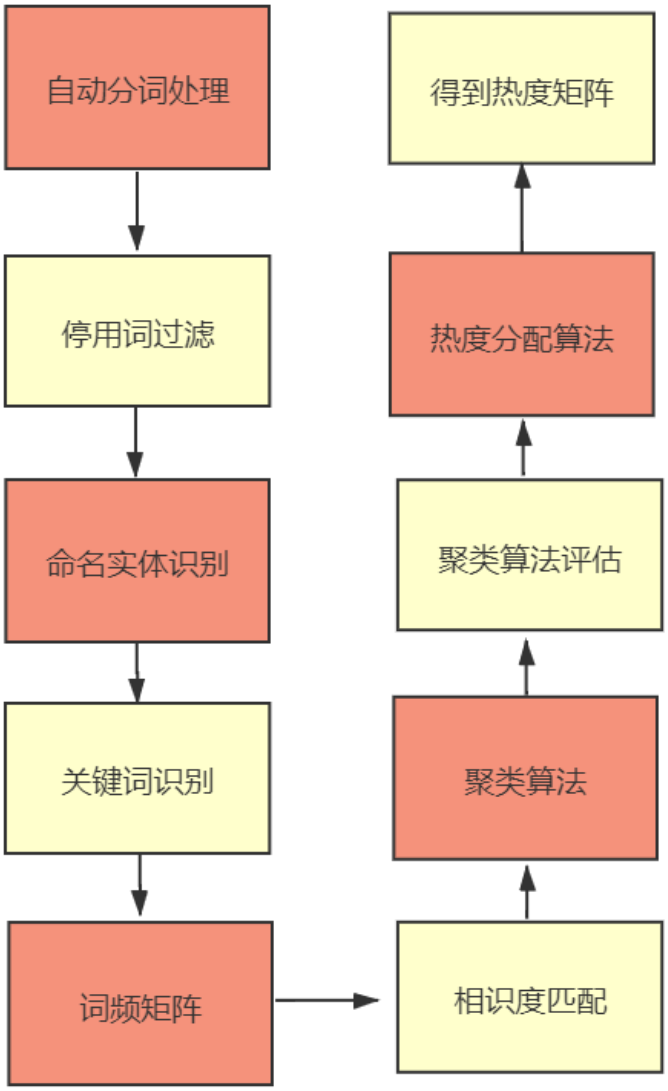

图 2-2　热点问题挖掘流程图

2.3 问题三的分析

如何衡量相关部门答复意见的质量和回复态度，是网络问政平台发展亟需解决的重要问题。根据附件 4 的数据，我们对相关部门对留言的答复意见质量优劣的特征进行归纳，得到以下优劣评判特征。

1、良好答复意见的参考特征：

(1) 答复意见和留言详情以及留言主题内容词语的语义相似度高；

(2) 答复意见内容充实，有条理。

(3) 答复意见内容引用相关规定，例 如，“根 据 x x x 条例或规定，…”；

2、较差回复的参考特征：

(1) 答复意见内容过于简短，只答复了部分问题；

(2) 答复意见内容套用固定模板，例如 “您反应的问题已转交 x x x 单位…”等；

(3) 对不同留言使用同一回复， 或答复意见内容的相似度很高。

3 分析方法与过程

3.1 问题 1 分析方法与过程

3.1.1 数据统计与分析

根据附件 2 的数据进行相关统计，如表 3-1 所示，群众留言共计 9210 条，群众留言数据分类占比如下图 3.1 所示，7 个一级标签包括：城乡建设，环境保护，交通运输，教育文体，劳动和社会保障，商贸旅游以及卫生计生。通过对数据的统计，我们发现其中留言数量最多的是城乡建设占比 22%，留言数量较少的是交通运输，占比 7%，可以发现 7 个一级标签的占比较为平均。

表 3-1　群众留言数据统计结果

一级标签	留言数量
城乡建设	2009
环境保护	938
交通运输	613
教育文体	1589
劳动和社会保障	1969
商贸旅游	1215
卫生计生	877
总计	9210

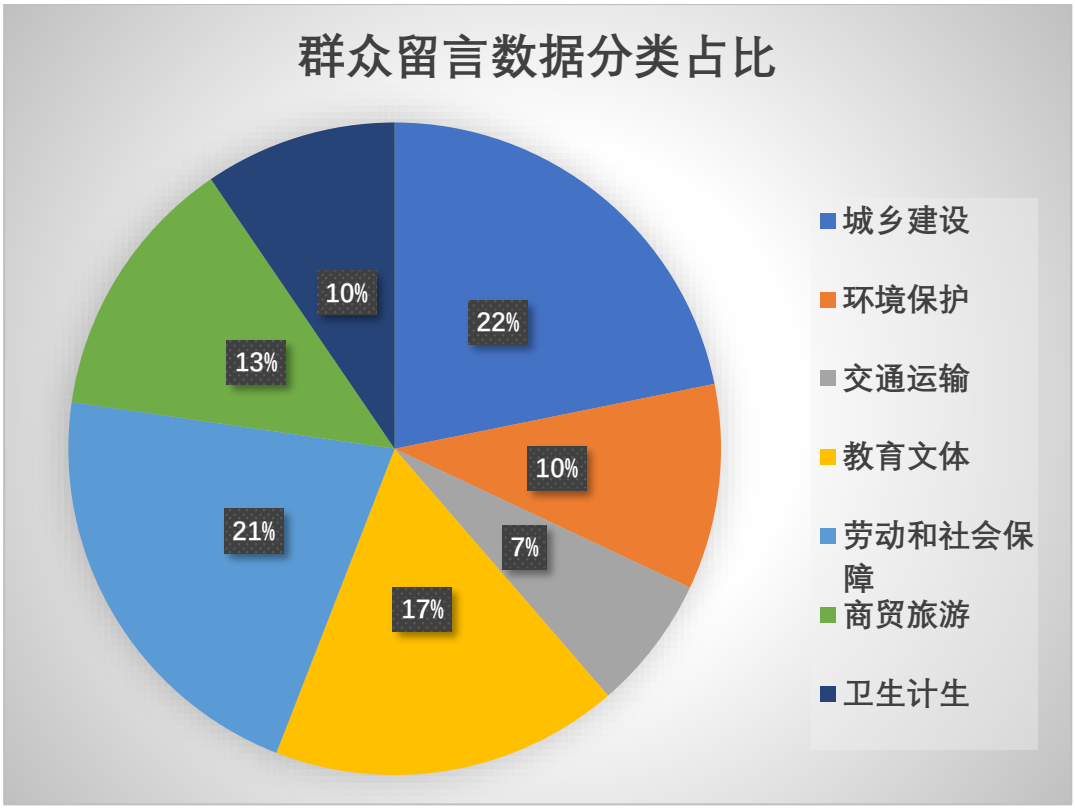

图 3-1 群众留言数据分类占比

3.1.2 文本预处理

进行文本分类之前，首先要对文本进行预处理，由于在分词过程中，留言的标题对于分类的结果非常重要，我们将留言主题和留言详情合并成一列。通过对数据浏览能够看出，文本信息相对比较工整，无空白数据，对于该文本来说，主要是分词和去停用词。

（1）分词

由于中文文本的特点是词与词之间没有明显的界限，从文本中提取词语时需要词，本文采用 Python 开发的一个中文分词模块 jieba 分词，对问题和回答中的每一句话进行分词进行中文分词。

jieba 分词用到的算法为最短路径匹配算法，该算法首先利用词典找到字符串中所有可能的词条，然后构造一个有向无环图。其中,每个词条对应图中的一条有向边，并可利用统计的方法赋予对应的边长一个权值，然后找到从起点到终点的最短路径，该路径上所包含的词条就是该句子的切分结果。分词结果如下图 3-2 所示。

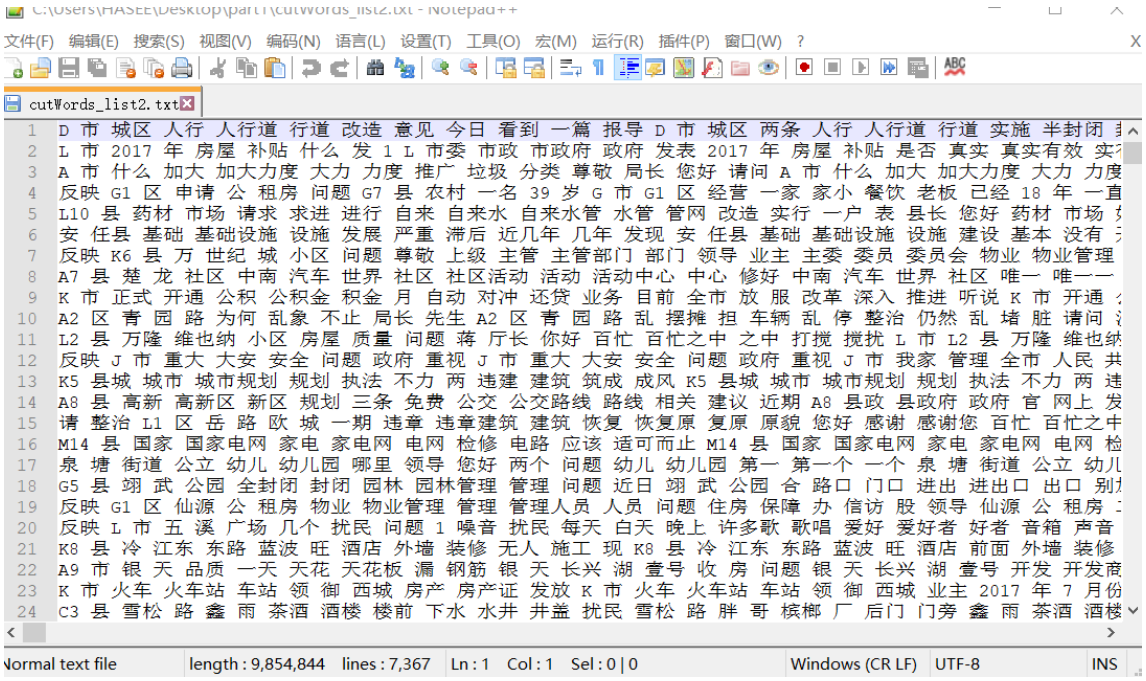

图 3-2 分词结果

（2）去停用词

在文本预处理中，停用词是指那些功能极其普遍,与其他词相比没有什么实际义的词，它们通常是一些单字，单字母以及高频的单词,比如中文中的“第二、一番、一个”等，英文中的“the，this，an，a，of”等。对于停用词一般在预处理阶段就将其删除，避免对文本，特别是短文本,造成负面影响。本文采用的是哈工大停用词表。表内容如下图 3.3 哈工大停用词表所示。

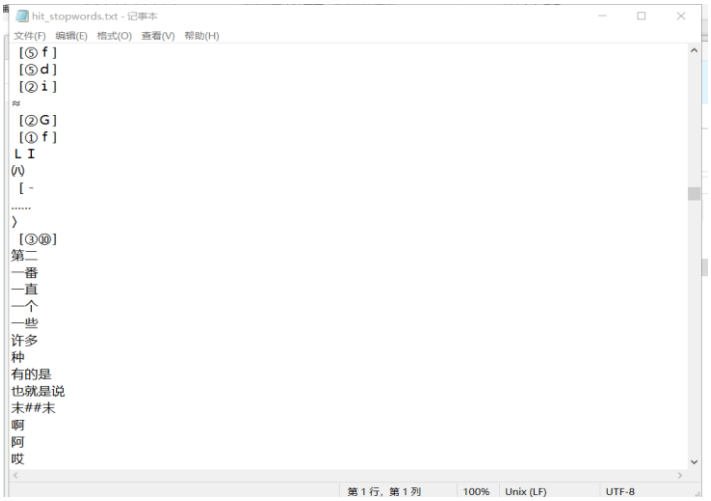

图 3-3 哈工大停用词表

3.1.3 文本数值化

文本的数值化，即使用数字代表特定的词汇，因为计算机无法直接处理人类创造的词汇。为了让计算机能够理解词汇，我们需要将词汇信息映射到一个数值化的语义空间中，这个语义空间我们可以称之为词向量空间。主要的方法有 One-hot 编码以及 Word2vec。

（1）One-hot 编码

一个自然的想法是把每个词表示为一个很长的向量。这个向量的维度是词表大小。其中绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的词。这就是独热编码形式(One-hot)。独热编码虽然方便易懂，但也有显而易见的不足；首先，One-hot 编码的维数由词典长度而定，过于稀疏，存在降维难问题，给计算造成了很大不便；其次，One-hot 编码下任意两个词之间都是孤立的，丢失了语言中的词义关系。

（2）Word2vec

Word2vec 是 Mikolov 在 2013 年提出的用于快速有效地训练词向量的模型，作者的目标是要从海量的文档数据中学习高质量的词向量，该词向量在语义和句法上都有很好地表现，已经广泛应用于自然语言处理的各种任务中。Word2vec 包含了两种训练模型，分别是 CBOW 和 Skip-gram 模型，如图 3.4 所示。中 CBOW 模型利用上下文预测当前词,而 Skip-gram 模型利用当前词预测其上下文。

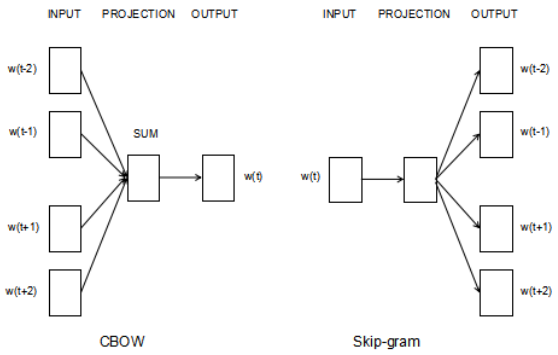

图 3-4　两种 word2vec 算法网络示意图

其中，Skip_gram 模型的训练目的就是是得下式得值最大：

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c}\log p\left(w_{t+j}\mid w_t\right) \tag{3-1}$$

其中，c 是窗口的大小，T 是训练文本的大小。基本的 Skip_gram 模型计算条件概率，如下所示：

$$p\left(w_O\mid w_I\right)=\frac{\exp\left(v'^{T}_{w_O}v_{w_I}\right)}{\sum_{w=1}^{W}\exp\left(v'^{T}_{w}v_{w_I}\right)} \tag{3-2}$$

其中 Vw 和 Vw'是单词 w 的输入和输出向量，W 是词典的大小。

3.1.4 文本分类模型的选择

文本分类模型，可以大体上分为基于传统机器学习的文本分类模型和基于深度学习的文本分类模型。文本分类通常采用基于统计的文本分类方法,主要有贝叶斯、最近邻方法、支持向量机和逻辑回归等方法。这些方法实现机制比较简单，文本分类效果良好。

目前基于深度学习模型的文本分类模型已经成为了主流，下面主要介绍本文所用到的逻辑回归模型以及基于 CNN 的文本分类模型。

（1）逻辑回归模型

如图 3.5 所示，直观来说，用一条直线对一些现有的数据点进行拟合的过程，就叫做回归。Logistic 分类的主要思想：根据现有数据对分类边界建立回归公式，并以此分类。建立拟合参数的过程中用到最优化算法，这里用到的是常用的梯度上升法。

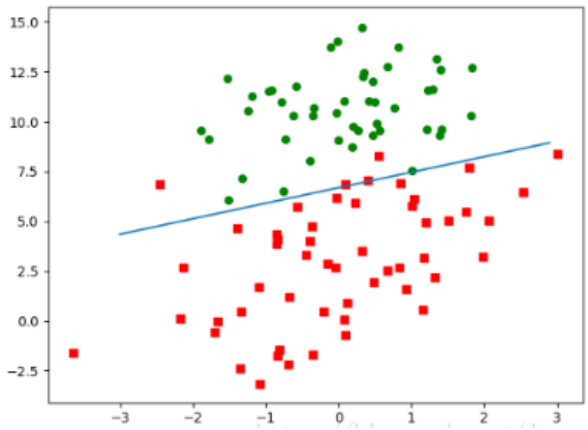

图 3-5　逻辑回归模型

（2）Text-CNN 模型

Text-CNN 模型是由 Yoon Kim 在论文(2014 EMNLP) Convolutional Neural Networks for Sentence Classification 提出。将卷积神经网络 CNN 应用到文本分类任务，利用多个不同 size 的 kernel 来提取句子中的关键信息（类似于多窗口大小的 n-gram），从而能够更好地捕捉局部相关性。

Text-CNN 模型的整体网络架构如图 3-6 所示，整个模型由四部分构成：模型主要分为 4 层，嵌入层，卷积层，池化层以及全连接层。

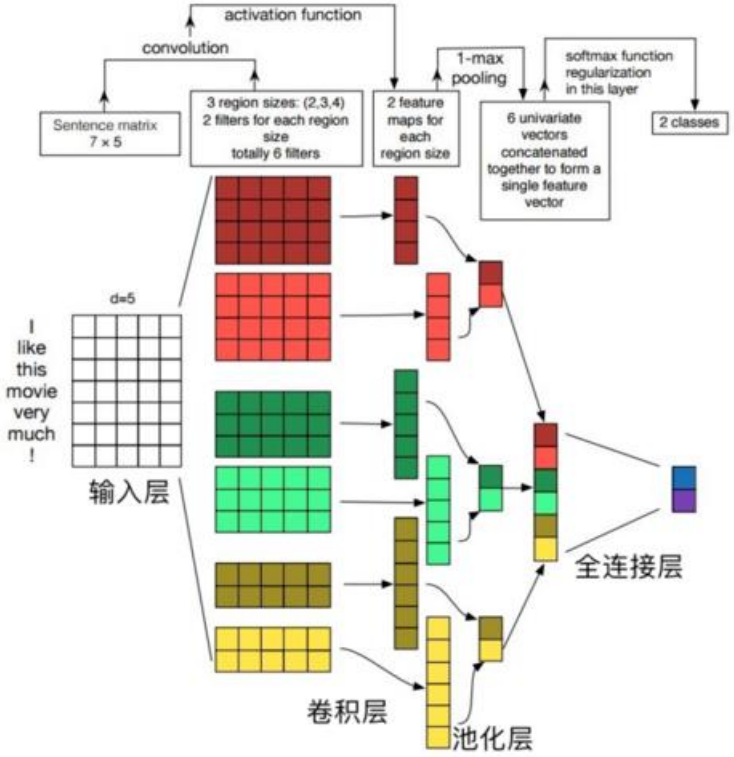

图 3-6　Text-CNN 模型网络架构图

（1）嵌入层

Text-CNN 模型的输入层需要输入一个定长的文本序列，我们需要通过分析语料集样本的长度指定一个输入序列的长度 L，比 L 短的样本序列需要填充，比 L 长的序列需要截取。最终输入层输入的是文本序列中各个词汇对应的词向量。

（2）卷积层

在 NLP 领域一般卷积核只进行一维的滑动，即卷积核的宽度与词向量的维度等宽，卷积核只进行一维的滑动。

在 Text-CNN 模型中一般使用多个不同尺寸的卷积核。卷积核的高度，即窗口值，可以理解为 N-gram 模型中的 N，即利用的局部词序的长度，窗口值也是一个超参数，需要在任务中尝试，一般选取 2-8 之间的值。

（2）池化层

在 Text-CNN 模型的池化层中使用了 Max-pool（最大值池化），即减少了模型 的参数，又保证了在不定长的卷基层的输出上获得一个定长的全连接层的输入。

卷积层与池化层在分类模型的核心作用就是特征提取的功能，从输入的定长文本序列中，利用局部词序信息，提取初级的特征，并组合初级的特征为高级特征，通过卷积与池化操作，省去了传统机器学习中的特征工程的步骤。

（4）全连接层

全连接层的作用就是分类器，原始的 Text-CNN 模型使用了只有一层隐藏层的全连接网络，相当于把卷积与池化层提取的特征输入到一个 LR 分类器中进行分类。

至此，Text-CNN 的模型结构就算大体了解了，有人把深度学习模型看作一个黑盒子，知道格式化的输入，我们就可以利用别人搭建好的模型框架训练在自己的数据集上实现一定的功能。但是在不同的数据集上，模型的最佳状态也不唯一，这就需需要我们在新的数据集上需要进行调优。

为了比较两者之间的差异，本文使用机器学习的 word2vec+LogisticRegression 模型和深度学习的 Text-CNN 模型进行比较。并且选择效果更优的分类模型。

3.1.5 实验结果和分析

本文将附件 2 的数据集进行 82 分割，如下表 3-2 所示，分成 7366 条训练集和 1844 条测试集，总计 9210 条。

表 3-2　附件二分割情况

	Train	**Test**	**Total**
城乡建设	1607	402	2009
环境保护	938	188	938
交通运输	490	123	613
教育文体	1271	318	1589

劳动和社会保障	1575	394	1969
商贸旅游	972	243	1215
卫生计生	701	176	877
总计	7366	1844	9210

为了比较两者之间的差异，本文使用机器学习的 word2vec+LogisticRegression 模型和深度学习的 Text-CNN 模型进行比较。评价结果如下图 3.8 和图 3.9 所示。

Label	Precision	Recall	F1	Support
交通运输	0.707547	0.609756	0.655022	123
劳动和社会保障	0.887218	0.898477	0.892812	394
卫生计生	0.840237	0.806818	0.823188	176
商贸旅游	0.782427	0.769547	0.775934	243
城乡建设	0.802469	0.808458	0.805452	402
教育文体	0.884146	0.911950	0.897833	318
环境保护	0.904040	0.952128	0.927461	188
总体	0.839650	0.841649	0.840257	1844

图 3.8　word2vec+逻辑回归模型评价指标

Label	Precision	Recall	F1	Support
交通运输	0.723077	0.764228	0.743083	123
劳动和社会保障	0.858770	0.956853	0.905162	394
卫生计生	0.898204	0.852273	0.874636	176
商贸旅游	0.887701	0.683128	0.772093	243
城乡建设	0.848101	0.833333	0.840652	402
教育文体	0.893939	0.927673	0.910494	318
环境保护	0.887755	0.925532	0.906250	188
总体	0.863989	0.862798	0.860869	1844

图 3.9　Text-CNN 模型评价指标

表 3-3 两模型的评测指标

	Precision	Recall	F1
word2vec+逻辑回归模型	0.8367	0.8416	0.8402
Text-CNN 模型	0.8639	0.8627	0.8608

为了得到最优模型，我们将两个分类模型的评测值标进行对比，如表 3-3 所示，并且根据实验数据绘制了模型的评价指标对比图，从中我们发现，Text-CNN 模型的结果要优于 word2vec+逻辑回归模型，前者 F1 值只有 0.8402 而后者 F1 值达到 0.8950，Text-CNN 该模型在数据集中表现了更好的性能，说明该模型可以投入到实际应用中。

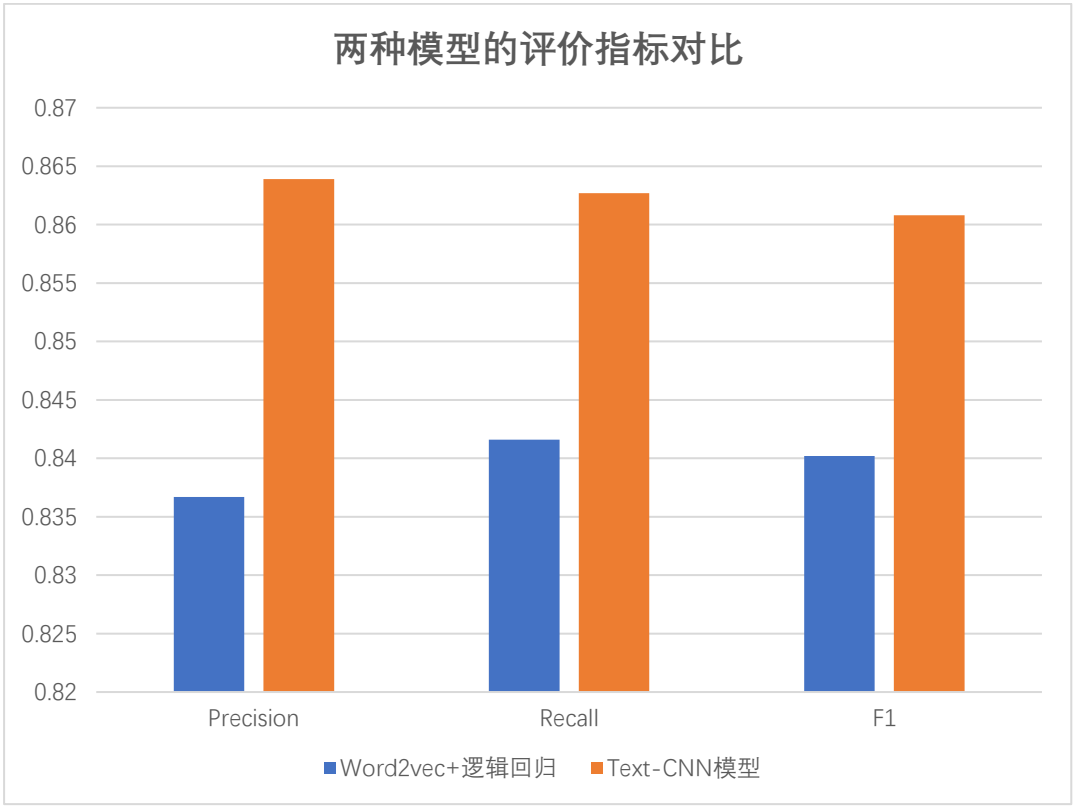

图 3-10　两模型的评测指标

3.2 问题 2 分析方法与过程

3.2.1 数据初始化与关键词识别

附件 3 中数据词条共为 4326 条，我们首先将其初始化过滤得到所需要的数据格式，首先使用工具 jieba 来完成分词操作，然后实际生活中常用的不能用于区分文档之间关系的一些词语，如“的”，“你”，“我”，“他”等，过滤这些停用词完成初始化数据。

下一步是进行命名实体识别，命名实体一般指的是文本中具有特定意义或者指代性强的实体，通常包括人名、地名、组织机构名、日期时间、专有名词等。实体命名识别系统就是从非结构化的输入文本中抽取出上述实体，并且可以按照业务需求识别出更多类别的实体，比如地名、商家店名等，在本文中使用条件随机场（CRF）模型用于命名实体识别。由于该方法简便易行，而且可以获得较好的性能，因此受到业界青睐，已被广泛地应用于人名、地名和组织机构等各种类型命名实体的识别，并在具体应用中不断得到改进，可以说是命名实体识别中最成功的方法。基于 CRF 的命名实体识别把命名实体识别过程看作一个序列标注问题。其基本思路是：将给定的文本首先进行分词处理，然后对人名、简单地名和简单的组织机构名进行识别，最后识别复合地名和复合组织机构名，而在留言主题中最为明确的正是实体的名字。基于 CRF 的命名实体识别方法属于有监督的学习方法，因此，我们利用已训练好的 CRF 工具包进行识别，确定特征模板。特征模板一般采用当前位置的前后 n（n≥1）个位置上的字（或词、字母、数字、标点等，不妨统称为“字串”）及其标记表示，即以当前位置的前后 n 个位置范围内的字串及其标记作为观察窗口：（…w－n/tag－n，…，w－1/tag－1w0/tag0，w1/tag1，…，wn/tagn，…）。考虑到，如果窗口开得较大时，算法的执行效率会太低，而且模板的通用性较差，但窗口太小时，所涵盖的信息量又太少，不足以确定当前位置上字串的标记，因此，一般情况下将 n 值取为 2～3，即以当前位置上前后 2～3 个位置上的字串及其标记作为构成特征模型的符号。并将所识别的实体名都纳入关键词。

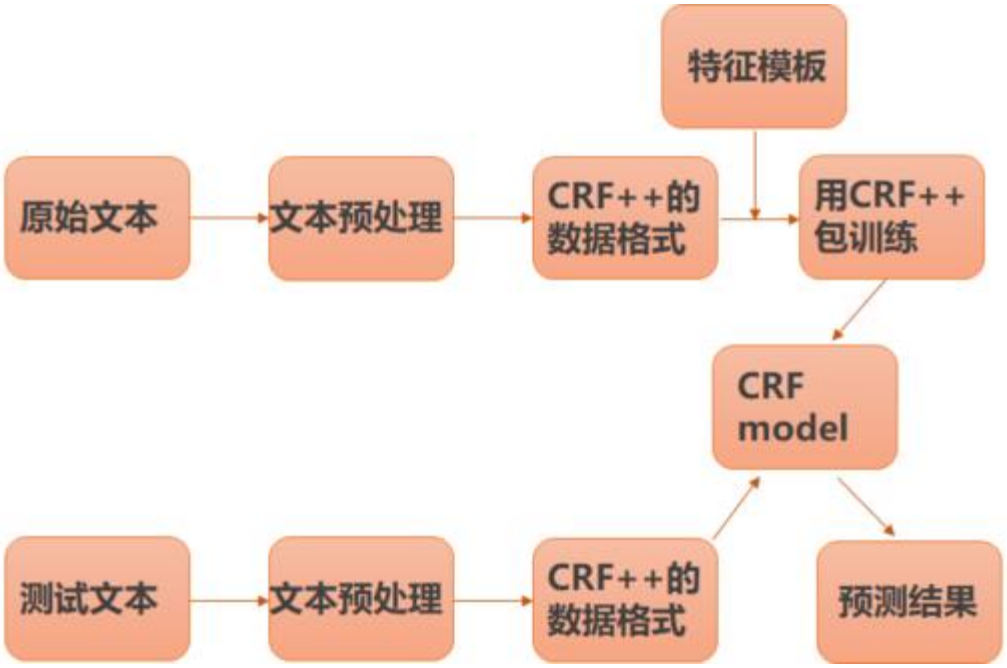

图 3-11. 基于 CRF 的命名实体识别

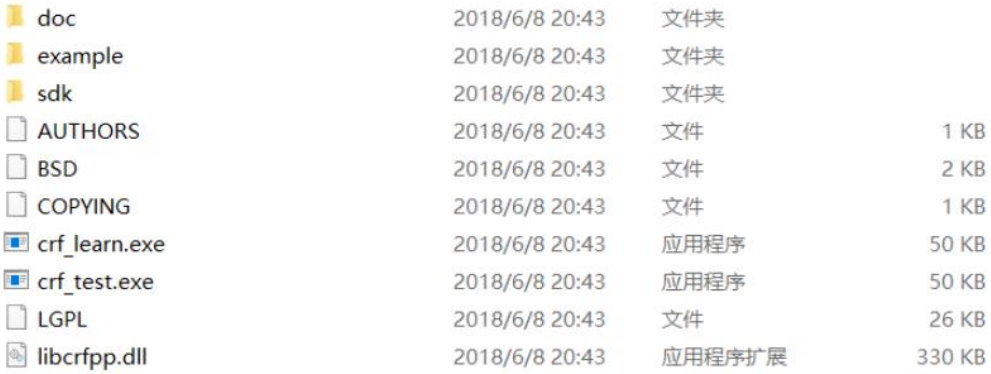

图 3-12 CRF 工具包

关键词是文本特征中非常重要的属性，体现了文本词条的主要内容。关键词出现的频次越高，越能反映该领域当前的研究重点和热点。检索附件 3 中数据词条共为 4326 条，以这些词条的关键词构成本文研究所用数据。将 Excel 中相应留言信息的关键词构建文献关键词的词频矩阵，以此矩阵作为聚类分析的数据来源。

	预留医院用地	**A 市地铁**	**市南路**	**…**
留言 1	0	1	0	…
留言 2	0	0	1	…
留言 3	1	0	0	…

…	…	…	…	…

图 3-13 词频矩阵

3.2.2 聚类分析模型建立

文本聚类主要是依据著名的聚类假设：同类的词条相似度较大，而不同类的词条相似度较小。将上述词频矩阵运用 python 进行聚类分析，系统聚类方法选择聚类效果较好的组内联结法。两个类中所有样例两两之间的距离平方和的均值。而组间连接只计算不同类中样品的距离，同类样例间的距离就不计算了。由于本文所构建词频矩阵具有二值矩阵的特点，因此采用简单匹配系数度量词条之间的相似性。对于文献词频矩阵来说，简单匹配系数是两篇文章在所有关键词上取值相同的情况出现的频率，简单匹配系数越大，两条词条就越相似。简单匹配系数 M 可以表示为

$$M=\frac{A+D}{A+B+C+D} \tag{3-3}$$

A 表示两条词条中同时出现的关键词数量；D 表示两条词条中都不包含的关键词数量；B 和 C 表示两条词条中包含而另一篇文章不包含的关键词数量；A+B+C+D 表示全部关键词的数量，即词频矩阵的变量数。因此，简单匹配系数不仅计算公式简洁，而且可以有效利用这种词频矩阵的稀疏性来衡量相似性。各类别之间在词条数量上存在差距。具体体现为：根据匹配规则，我们将所有词条聚类划分为 986 类。其中涵盖了邻里矛盾、教育文体、地区建设等话题，其中，以对地铁规划的留言最为广泛。

图 3-14 关键词识别

3.2.3 热度分配

留言信息中有大量表达大意相近的留言，我们将其称之为词条的近义关系，本文算法将同义关系作为算法的重要因素，按照上述聚类分析结果，根据每个类别的关键词进行热度评价。并且每条词条都存在点赞数和反对数，这都是热度评价的重要依据。本节是关于留言类别在总类别中的权重计算，可根据其权重计算结果对其进行降序排序，排序越靠前表明越重要。留言类别之间的近义关系可以用一个有向网络来描述。被评论的意见用集合 D＝｛d1,d2,…,dm｝表示；意见分类用集合W＝｛w1，w2，…，wj｝表示，每个意见分类中每条留言用 Wi=｛l1,l2,…,lx｝表示。

本文提出一种热度分配算法（Hot allocation method，HAM）方法，详细过程如下：

（1）我们提出一个热度矩阵 M，M 有 j 行 j 列，我们计算 W 中的每个意见类别留言的含有留言数量 N：

$$Nj_{x,y}=\mathrm{n} \tag{3-4}$$

其中 x，y 为矩阵中的行号和列号。

（2）我们计算 W 中的每个意见类别留言的同义权重 Qj，其中，rP 表示 Wx 和 Wy 之间是否有同义关系，有同义关系取 1，不相关取 0。

$$Qj_{x,y} = \sum rP_{x,y} \tag{3-5}$$

然后计算评论热度矩阵 A，其元素是 Aj，表示 D 中的意见的热度含量,赞同数的权值为 3，反对数的权值为 1，都可以表示评价热度。Aj 的计算公式如下：

$$Aj_{x,y} = \sum \alpha Z + \beta O \tag{3-6}$$

其中，α 和 β 为权重系数，经过试验验证，α=3，β=1，Z 为该条留言的点赞数，O 为该条留言的反对数。

（3）对于数据集中的每条词条，然后将 W 中每条词条的留言数量与同义热度矩阵和评论热度矩阵全部累加，即

$$M = \sum Nj_{x,y} + Aj_{x,y} + Qj_{x,y} \tag{3-7}$$

最终得到最终热度矩阵 M。

（4）通过对 M进行降序排序，我们就能得到特征词的热度排名。

依照热度排名得到的排名前 5 的热点问题如下图所示，相应热点问题对应的留言情况详细见附件 2：

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	86	2019/2/11 至 2019/10/22	A 市地铁	地铁建设不合理投诉与咨询
2	2	57	2019/7/3 至 2019/11/8	区教育局	A 市教育局的学区规划问题
3	3	30	2019/2/10 至 2019/6/24	香樟路口	香樟路口的道路规划与周边环境差问题
4	4	22	2019/4/23 至 2019/6/20	A 市木莲中路	公交路线不合理以及周边环境差
5	5	21	2019/3/11 至 2019/7/21	A 市北辰三角洲	小区房屋品质与整改与小区环境问题

图 3-15:　排名前 5 的热点问题

3.3 问题 3 分析方法与过程

3.3.1 答复意见质量的评价方法

相关部门答复意见的质量由很多种因素决定，可以从答复的相关性、完整性、可解释性等角度对其进行评价。首先预处理附件 4 中的数据，剔除有问题的答复意见数据；然后从以下几个方面对答复意见质量进行量化评价。

1、重复留言的处理

在附件 4 的留言详情和答复意见中，有部分用户重复留言，且相关部门给出的答复意见都相同，这类情况我们将之视为文本数据中的冗余数据。在数据预处理中，需要将冗余数据剔出，只保留一份留言详情和答复意见的数据即可。

以下对相关部门答复意见的评价都是基于预处理后的文本数据。若答复意见内容信息满足良好答复意见的标准，则计正分；若满足较差答复意见的标准，则计负分。本文引入权重系数，对相关部门所有答复意见的计分进行加权求和，得到相关部门答复意见质量的评价得分。

在海量数据之下，如何基于这些特征利用用自然语言处理和文本挖掘的方法对相关部门答复意见进行评价仍然是一个极具研究价值的问题。下面从答复的相关性、完整性、可解释性等角度对其质量进行量化描述，从而建立评价模型。

2、答复意见的量化方法

(1) 答复意见内容是否相关且专业

相关部门需要围绕着留言主题以及留言详情的内容来进行答复，因此可以根据留言的内容来考虑回复内容的相关度。此外，问政平台正成为一个为社会公众关注的热点话题“代言”的平台。住房、环保、教育、交通、医疗、城市管理等等这些大众所关注的重点领域，也正是在问政平台上的留言中提到的热门话题，相关部门答复意见需要围绕这些重点领域充分考虑回复内容的专业度。统计各重点领域的相关责任单位，如表 3-4 所示。

表 3-4 重点领域的相关责任单位

问政领域	领域相关责任单位
交通	公安交通主管部门、市交通运输局、市城乡规划局、市政局、城乡建设局、市城管局
教育	教育局、市政局、城乡建设局、区规划分局
医疗	省卫计委、人力资源和社会保障局、市卫计委
环保	卫生局、环保部门、卫生和计划生育局
住房	房屋交易管理中心、自然资源和规划局、市住建局、人民政府
城市管理	市工务局、市城管执法局、财政局、市工商局
……	……

根据留言主题以及留言详情的内容、表 1 中重点领域的相关责任单位可以判断相关部门答复意见内容的相关度和专业度，可以认为答复意见语句中所用的词语和留言内容的词语的语义相似度越高，且涉及相关的责任单位词汇，则答复意见内容质量越好。

为此，需要引入工具 word2vec 来描述留言中词语 w_i 与答复意见中的词语 w_j 的相似度 WORSIM(w_i,w_j)，将留言内容的关键词与答复意见的信息进行对比分析。

与传统的词向量相比，word2vec 训练出的词向量[1]避免了词语表示 的“维数灾难”和“高稀疏”的问题，实现了连续 bag-of-word 模型[2]和计算词向量的 skip-gram 结构[3]来将文本中的词语表示为词向量。这些词向量可以作为词的特征应用到自然语言处理问题中[4]。对附件 4 文本数据进行分词和 word2vec 训练，得到文本数据的词向量和对应词语的相似度。

利用 word2vec 可以方便地计算出留言内容的关键词与答复意见的词语间的相似度 WORSIM(w_i,w_j)。

$$WORDSIM(w_i, w_j) = \frac{\sum_{i=1}^{n}(x_{i1} \times x_{i2})}{\sqrt{\sum_{i=1}^{n} x_{i1}^2} \times \sqrt{\sum_{i=1}^{n} x_{i2}^2}} \tag{3-8}$$

其中，两个 词 语 w_i 和 w_j 的词向量表示为：

$$w_i = (x_{11}, x_{21}, x_{31}, \cdots, x_{n1}); \tag{3-9}$$

$$w_j = (x_{12}, x_{22}, x_{32}, \cdots, x_{n2}); \quad (3\text{-}10)$$

其中 n 表 示用 word2vec 训练词向量时设定的词向量的维度。 当 $w_i = w_j$ 时，可以通过系数$1+\lambda'$增加相似度。由此，我们建立如下回复内容的专业程度的评价项 F_1：

$$F_1 = \begin{cases} WORDSIM(w_i, w_j), & w_i \neq w_j \\ (1+\lambda')WORDSIM(w_i, w_j), & w_i = w_j \end{cases} \quad (3\text{-}11)$$

(2) 答复意见是否有理有据

答复意见的内容需要有严密的逻辑和准确的表达，即回复信息是否有理有据。假设剔除部分冗余数据后的剩余文本数据中共有 N 条比较规范的回复数据。将“经查办”、“核实”、“现作如下回复”以及引用相关条文规定时出现的（如“根据 x x x 规定实施”）等关键词和答复意见中的文本进行语义匹配。若在答复意见的文本中匹配到相应的关键字，则认为该条答复意见对问题进行了阐明和解决。统计答复意见内容引用关键词的回答数目 N_i ，计算出现频率，即 N_i 与 N 的比值。该频率值可以作为“是否有理有据”的评价项，记为 F_2：

$$F_2 = \frac{N_i}{N} \quad (3\text{-}12)$$

(3) 答复意见内容是否充实

答复意见内容的详细程度与回复的文本长度有直接的关系。简短内容的回复信息量一般不够，评分应该较低；同时，较长文本的回复评分不应该过高。因此，可以考虑使用对数函数来量化回复的文本长度与评分的关系，建立“答复意见内容是否充实”的评价项 F_3：

$$F_3 = \frac{1}{N}\sum_{i=1}^{N} \log_m L_i \quad (3\text{-}4) \quad (3\text{-}13)$$

其中，L_i 为针对第 i 个留言的答复意见的文本长度，m为常数。

(4) 回复套用模板或相似度较高

有些较差答复意见中会出现很多固定词语，如“网友：您好！留言已收悉”、“您反映的问题转相关部门进行处理”等等。本文建立了一个较差答复意见的关键词组成的集合 T_{key}。对答复意见进行关键词匹配，若一条回复中出现 T_{key}，则认为该条回答为较差回答。统计此类回复的数量 N_T，以其出现频率作为评价项 F_4：

$$F_4 = \frac{N_T}{N} \tag{3-14}$$

3.3.2 答复意见质量评价模型

评价相关部门答复意见的关键在于如何建立对答复质量的量化评分模型，以在附件 4 的数据下利用自然语言处理和文本挖掘的方法对其进行质量评分。根据上述量化方法，给出答复意见评价模型，其流程如图 3-16 所示。

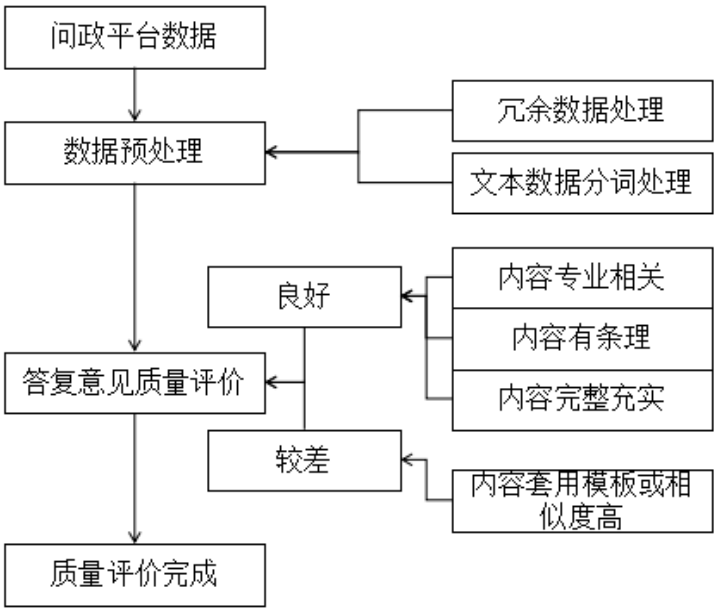

图 3-16 答复意见评价模型流程图

下面对 4 项量化指标进行整合，以计算答复意见信息的得分情况，建立答复信息的质量评价函数 F，即构造如下质量的评价函数 F：

$$\begin{cases} F(K) = M_k \cdot \lambda^T \\ \lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4), M_K = (F_1, F_2, F_3, F_4) \end{cases} \tag{3-15}$$

其中，λ^T 为 λ 向量的转置向量，向量 λ 和 M_K 为反映答复意见信息不同侧面的权重向量和得分向量。

3.3.3 实验结果和分析

本文实验数据来源于附件 4 的数据。数据中有 2816 条留言回复，其中留言主题有 2766 个，留言详情有 2783 条，答复意见有 2751 条，即都有重复数据。实验采用 python 编程语言。

1、答复意见质量评价模型的计算步骤

为了给出质量评分，模型的具体步骤如下：

(1) 剔除重复数据。

(2) 计算答复意见的各项得分，表示为向量 $M_K=(F_1,F_2,F_3,F_4)$

(3) 设置权重向量为 $\lambda=(\lambda_1,\lambda_2,\lambda_3,\lambda_4)$，计 算 $F(K)=M_K\cdot\lambda^T$ 作为最终的答复意见质量评价函数 $F(K)$。

根据相关性、完整性、可解释性这 3 项判断答复意见质量的评分，而这 3 项对于评估一个律师回复质量的重要性不同，因此本文给予这 3 项的权重分别为 0.3，0.35，0.25。若多次答复意见相似度过高或回复套用模板，则要扣分，本文给予其权重为 0.1。这样分配的权重不仅满足权重之和为 1，也符合实际的权重分布。如此将律师的回复质量权重取定后，就可以得到律师的回复质量评价函数 $F(K)$ 为：

$$F(K)=0.3F_1+0.35F_2+0.25F_3-0.1F_4 \quad (3\text{-}7)$$

2、答复意见质量评价模型的结果与分析

把 4 项答复意见质量评价函数的 F 值全部映射到[0,1]后的结果如 3-5 所示。

表 3-5 各项答复意见质量评价函数值

评分项	得分
F_1	0.8359
F_2	0.8971
F_3	0.8831
F_4	0.0893

从表 3-3 中就可以得到相关部门答复意见的质量评价分数：

$$F(K)=0.3F_1+0.35F_2+0.25F_3-0.1F_4=0.7766 \quad (3\text{-}16)$$

实验结果表明，该模型可以很好地评价一个相关部门对留言的答复意见的质量。该模型研究也有助于建设“智慧政务”的网络平台，促进优质和高效的问政回复，帮助平台提供更好的用户体验。

参考文献

[1] TANG M,ZHU L,ZOU X C. Document V ector R epresentation Based on W ord2 Vec[J]. Computer Science, 2016, 43(6):214-217 ,269. (in Chinese)唐明,朱磊,邹显春.基于 Word2Vec 的一种文档向量表示[J].计算机科学, 2016,43(6) :214-217 ,269.

[2] MIKOLOV T ,CHEN K,CORRADO G,et al. Efficient estima-tion of word representations in vector space[J]. arXiv preprintarXiv:1301. 3781,2013.

[3] MIKOLOV T,SUTSKEVER I,CHEN K ,et al. Distributed rep-resentations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems. 2013:3111-3119.

[4] LI Y P,JIN C,J1 J C. A Keyword Extraction Algorithm Basedon Word2vec [J]. E-science Technology &. Application ,2015，6(4) :54-59. (in Chinese)李跃鹏,金翠,及俊川.基于 word2vec 的关键词提取算法[J].科研信息化技术与应用, 2015, 6(4):54-519.

[5] 石凤贵.基于 TF-IDF 中文文本分类实现[J].现代计算机,2020(06):51-54+75.

[6] 朱梦. 基于机器学习的中文文本分类算法的研究与实现[D].北京邮电大学,2019.

[7] 吴萍萍.基于信息熵加权的 Word2vec 中文文本分类研究[J].长春师范大学学报,2020,39(02):28-33.

[8] Kim, Yoon. "Convolutional Neural Networksfor Sentence Classification." Proceedings of the 2014 Conferenceon Empirical Methods in Natural Language Processing (EMNLP). 2014.

[9] Yang, Zichao, et al. "Hierarchicalattention networks for document classification." Proceedings of the2016 Conference of the North American Chapter of the Association forComputational Linguistics: Human Language Technologies. 2016.

[10] 刘海峰, 苏展, 刘守生. 一种基于词频信息的改进 CHI 文本特征选择[J]. 计算机工程与应用, 2013, 49(22): 110-114.

[11] 张敏, 罗梅芬, 张艳. 国际文本挖掘研究主题群识别与演化趋势分析[J]. 图书馆学研究, 2017(2): 15-21.

[12] Hao T Y, Chen X L, Li G Z, et al. A bibliometric analysis of text mining in medical research[J]. Soft Computing, 2018, 22(23):7875-7892.

[13] Feldman R, Dagan I. knowledge discovery in textual databases(KDT)[C]// Proceedings of the 1st International Conference onKnowledge Discovery and Data Mining.Palo Alto: AAAI Press,1995: 112-117.

[14] 杨丽华, 戴齐, 杨占华. 文本分类技术研究[J]. 微计算机信息,2006(15): 209-211.

[15] 肖建国 . 试论文本挖掘及其应用[J]. 图书馆学研究, 2008(4):22-24.

[16] 谌志群, 张国煊. 文本挖掘研究进展[J]. 模式识别与人工智能, 2005, 18(1): 65-74.

[17] 郑双怡. 文本挖掘及其在知识管理中的应用[J]. 中南民族大学学报(人文社会科学版), 2005(4): 127-130.