基于深度学习方法的智慧政务处理研究

摘要

随着互联网技术的发展,网络问政平台逐渐成为政府了解民意、汇聚民智、凝聚民气的重要渠道。各类文本的数据量不断攀升，给以往的人工数据处理工作带来了极大挑战。而近年来自然语言处理技术(NLP)的发展能够帮助我们有效进行文本处理。本文基于深度学习方法,主要对留言数据进行分类、聚类、热点整理,对政府答复意见进行评价。

根据题目给出的具体问题,我们主要进行了以下三个方面的研究:

1. 构建基于深度学习的卷积神经网络(textCNN)模型进行文本分类

首先,对留言数据进行预处理,包括中文分词,去停用词等；其次,建立基于Word2vec 的词向量模型生成包含语义间相关性的词向量；然后,使用改进的TF-IDF 方法即 CTF-IDF 进行特征提取,词向量转化为句子向量；最后,使用卷积神经网络(textCNN)模型分别对留言主题和留言详情进行文本分类,并与KNN,NB,SVM,BI-LSTM 等分类方法比较,结论是:(1)基于留言详情分类比基于留言主题分类结果好,(2)textCNN 模型的分类效果最好,基于留言详情的 textCNN 分类模型 F1 值为 89.6%。

2. 使用 K-means 法进行文本聚类,定义热度评价指标,给出话题热度排序

首先使用 k-means 聚类方法对留言进行聚类,使用 CTF-IDF 方法对聚类后的每一类进行话题提取；其次,根据留言条数、话题持续时间、点赞数和反对数等因素构建热度评价指标:N(topic)、T-L(time-long)、A-O（agree-oppose)；最后,给出热度评价方法,计算各类的热度评分,进行话题热度排序。

3. 给出基于层次分析法(AHP)的政府答复意见的评价方案

通过层次分析法(AHP)构建答复意见评价指标体系,包括文本相似度(Cosine Distance)、回答中心度(ansCen)、语言多样性(lingDiv)、时间间隔(ansTime)、文本长度(length)、情感支持(emSup)六个指标,从相关性、完整性、回答时效性、可解释性四个角度对政府答复意见进行评价,给出评分结果。

关键词:Word2vec　CTF-IDF　卷积神经网络　k-means　热度评价　AHP

Abstract

With the development of Internet technology, the platform has gradually become an important channel for the government to understand the public opinion, gather the wisdom of the people and gather the people's spirit. The data volume of all kinds of text is rising, which brings great challenge to the manual data processing work in the past. And the development of natural language processing technology (NLP) in recent years can help us to carry out text processing effectively. Based on the deep learning method, this paper mainly classifies the message data, clusters and hot spots, and evaluates the government's reply opinions.

According to the specific problems given by the topic, we mainly carry out the following three aspects of research:

1. Constructing convolutional neural network (textCNN) model based on deep learning for text classification

Firstly, the message data is preprocessed, including chinese word segmentation, de-deactivating words, etc. Secondly, a word vector model based on Word2vec is established to generate word vectors containing semantic correlations. then, the improved TF-IDF method is used, that is, CTF-IDF feature extraction, word vectors are transformed into sentence vectors; Finally, the convolutional neural network (textCNN) model is used to classify message topics and message details respectively, and compared with KNN,NB,SVM,BI-LSTM and other classification methods.The conclusion is :(1) the classification based on message details is better than that based on message topic . (2) the classification effect of textCNN model is the best, and the F1 value of textCNN classification model based on message details is 89.6%.

2. Clustering text by K-means method, defining heat evaluation index and ranking topic heat

Firstly, the k-means clustering method is used to cluster the messages, and the CTF-IDF method is used to extract each category after cluster-ing.Secondly,According to the number of messages, the duration of the topic, the number of likes and the number of objections and other factors to construct the heat evaluation index: N, T-L (time-long), A-O (angle-oppose); Finally, give the heat evaluation method, calculate all kinds of heat score, and sort the topic heat.

3. Evaluation of government responses based on AHP

The evaluation index system of response opinion was constructed by AHP, including text similarity (Cosine Distance), response centrality (ansCen), language diversity (lingDiv), time interval (ansTime), text length (length), emotional support (emSup). From the four aspects of relevance, integrity, timeliness and interpretability, the government's response comments are evaluated and the results are given.

Keywords: Word2vec,CTF-IDF, convolutional neural network, k-means, heat evaluation, analytic hierarchy process

目录

1.引言

1.1 挖掘背景及意义

随着互联网技术的发展,近年来微信、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。通过网络渠道收集民意的方式方便了广大群众问政留言,也方便了政府及时了解民意。广大群众通过网络平台问政留言,发表自己关于各类事件的看法,观点等,但是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。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。而近年来自然语言处理技术(NLP)的发展能够帮助我们有效进行文本处理,本文构建基于深度神经网络的模型,对大量的问政留言数据进行分类,聚类,热度评价,对政府留言答复进行综合评价,对提升政府的管理水平和施政效率具有极大的推动作用。

1.2 挖掘目标

基于题目提出的问题,本文的挖掘目标主要有三个:一是根据群众的问政留言记录,建立关于留言内容的一级标签分类模型；二是发现热点问题,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；三是针对政府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给出一套评价方案。

1.3 本文主要研究内容

目前网络问政平台广泛应用,产生了大量的群众留言文本数据,给传统的人工处理工作带来了极大的挑战。本文利用自然语言处理技术(NLP)群众留言记录进行研究分析,建立了留言分类模型,设计热点问题评价方案并对政府相关部门留言答复建立了评价方案。

本文的主要研究内容如下:

(1)改进了传统的TF-IDF文本向量表示模型,加入了类与类之间的信息,建立了 CTF-IDF 模型。在对群众留言记录进行中文分词,去停用词处理之后,把中文分词结果用 one-hot 编码转化为词向量,然后建立 Word2vec 词向量模型,该模型加入了词语之间语义相关信息。把词向量转化为句子向量的过程中,对传统的 TF-IDF 权重模型进行改进,加入了类与类之间信息因素,建立了 CTF-IDF 模型,能够更好地表示句子向量。

(2)建立卷积神经网络模型(textCNN)对文本进行分类,与其他的文本分类模型 KNN,NB,SVM,BI-LSTM 进行比较分析。分别基于留言主题和留言详情进行分类,比较两者的分类效果。

(3)利用 k-means 对群众留言进行聚类分析,热度评价,整理热点问题。首先使用 k-means 聚类方法对留言进行聚类,使用 CTF-IDF 方法对聚类后的每一类进

行话题提取；其次,根据留言条数,话题持续时间,点赞数和反对数等因素构建热度评价指标:N(topic)、T-L(time-long)、A-O（agree-oppose)；最后,给出热度评价方法,计算各类的热度评分,进行话题热度排序。

(4)给出基于层次分析法(AHP)的政府答复意见的评价方案.通过层次分析法(AHP)构建答复意见评价指标体系,包括文本相似度(Cosine Distance),回答中心度(ansCen),语言多样性(lingDiv),时间间隔(ansTime),文本长度(length),情感支持(emSup)六个指标,从相关性,完整性,回答时效性,可解释性四个角度对政府答复意见进行评价,给出评分结果。

1.4 论文的结构安排

论文分为 5 个部分对群众问政留言进行研究,论文的结构安排如下:

第一章为引言部分,主要对本文的挖掘背景和意义进行介绍,同时说明了本文的挖掘目标,整体上说明了本文的主要研究内容,介绍了本文论文的结构安排

第二章介绍了数据预处理工作,介绍了中文分词,去停用词方法,研究了词向量训练 Word2vec 词向量模型,介绍了改进的 TF-IDF 权重模型--CTF-IDF。

第三章介绍了群众留言文本分类模型,主要介绍了卷积神经网络(textCNN)模型,并与其他的分类模型 KNN,NB,SVM,Bi-LSTM 等进行对比。

第四章介绍了使用 k-means 聚类方法对留言进行聚类,定义热度评价指标,进行话题热度排序,整理热点问题。

第五章介绍了对政府留言答复意见的评价方案,主要通过层次分析法构建评价指标体系进行综合评价。

最后,第六章对本文进行总结与展望。

2.数据预处理

由于计算机不能直接理解人类的自然语言,所以在对文本进行分析之前需要对原始语料进行预处理。通常的中文文本预处理流程包括分词、去停用词以及文本的向量化表示,将文本数据转化为计算机能理解的向量。

2.1 文本预处理

2.1.1 中文分词

中文语料数据为一批短文本或者长文本，比如：句子，文章摘要，段落或者整篇文章组成的一个集合。一般句子、段落之间的字、词语是连续的，有一定含义。而进行文本挖掘分析时，我们希望文本处理的最小单位粒度是词或者词语，所以这个时候就需要分词来将文本全部进行分词。本文采用 python 的中文分词模块 jieba 分词对原始语料进行分词。通过对原始预料的观察，发现留言中所用

地名做了脱敏处理。所以先通过正则表达式匹配出相应的区、市、县等名词。将匹配出的词语加入 jieba 分词模块，可以得到更好的分词结果。

Jieba 分词基于 Trie 树结构实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图(DAG)；并且采用了动态规划查找最大概率路径，找出基于词频的最大切分组合；对于未登录词，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。得到的分词结果如下图 2-1 所示,分词之后的词云图如图 2-2 所示:

```
A3区 保利 谷林语 桐梓 坡路 与麓 松路 交汇处 地铁 凌晨 施工 扰民
A7县 特立 东四 路口 高峰 太堵 调整 信号灯 配时
A3区 青青 家园 小区 乐果 零食 炒货 公共 通道 摆放 空调 扰民
拆除 聚美龙楚 西地省 商学院 宿舍 安装 变压器
A市 利保 壹号 公馆 项目 夜间 噪声 扰民
A市 地铁 号线 星沙 大道 地铁 出入口 设置 不合理
A4区 北辰 小区 非法 改商 何时能 解决
K3县 乡村 医生 卫生室 执业 许可证
A7县 春华 石塘 铺村 党员 家开 麻将馆
```

图 2-1 分词结果图

图 2-2 词云图

2.1.2 去停用词

停用词一般指对文本特征没有任何贡献作用的字词，比如标点符号、语气、人称等一些词。所以在一般性的文本处理中，分词之后，接下来一步就是去停用词。对于停用词，一般在预处理阶段就将其删除，避免对文本造成负面影响,本文使用哈工大停用词表。通过观察原始语料，发现群众留言中的地区名词对类别影响不大，所以将上述匹配出的区、市、县等名词加入哈工大停用词表中，在预处理阶段将其作为停用词删除掉

2.2 Word2vec 词向量模型

词向量也成词嵌入，是以词语为单位进行文本表示的方式。将词以向量的形式映射到实验空间,进行数值化处理的技术。传统的词向量表示的方法主要有(1)布尔模型(Boolean Model):最早、最简单的文本表示模型；(2)向量空间模型(VSM):又称词袋模型（BOW）:以句子为单位进行空间表示的方式。(3)概率模型。但是用这些模型表示词向量并没有表达出词语之间的关系,Word2vec 词向量模型可以表达出词语之间的内部结构和语义之间的相关性。

word2vec 是 Google 在 2013 年开源的一个用于生成词向量的神经网络算法，word2evc 用到两个重要模型——CBOW（Continuous Bag of Words）模型即连续词袋模型和 Skip-gram 模型，如图 2-3 所示。

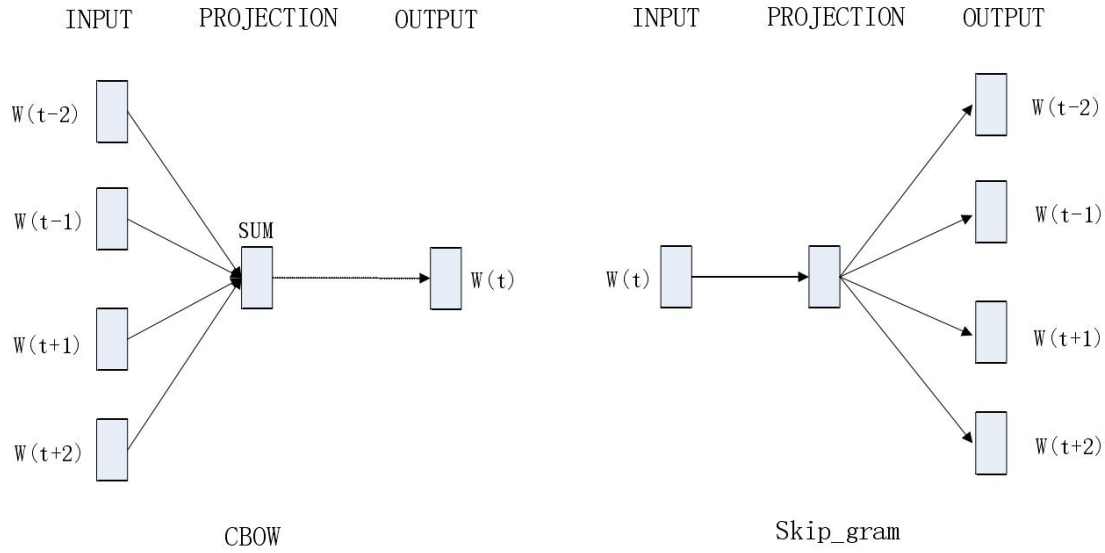

图 2-3 word2evc 算法示意图

根据图片中的神经网络可知，两个模型都包含三层：输入层、投影层和输出层。左边的 CBOW 模型是已知一段上下文的前提下预测中间词；而右边的 Skip-gram 模型是在已知一个词的前提下，预测这个词的上下文词。

word2vec 可以从语料库中训练出词向量，与传统的 One-Hot 词向量相比，word2vec 生成的词向量不仅维度更短，而且包含了语句的逻辑结构信息在其中，可以学到词汇的深层次语义信息，更好的提取词汇语义信息对提高文本分类的准确率具有重要意义。

使用 Python 的 g e n s i m 第三方软件包中的 word2vec 工具，可以很容易地实现对语料库的训练并生成对应的词向量。

2.3 基于改进的 TF -IDF 加权的文本表示方式

词向量可以很好的表示一个词语的语义信息，在一个文本中有很多词向量，不是所有的词向量都是同等重要的，所以在文本表示时通过计算特征词汇权值的方法进行特征提取，加权得出文本向量表示。比较常用的特征词汇权值处理方法有词频逆文档频率算法，布尔权重计算法，本文主要介绍 TF-IDF 算法，并对其进行改进，加入类与类之间的信息

2.3.1 TF-IDF 算法

TF -IDF（词频-逆文档频率）是一种统计方法，也是文本分类中经常用到的加权算法，可以估算特征词的重要程度，核心思想是出现频率越高的词语重要。

令 M(x)代表特征词 x 在文本中出现的次数，M 代表文本总次数，则有词频：

$$TF = \frac{M(x)}{M}$$

令N代表语料库中的文本总数目,N(x)代表语料库中含有词语x的文本总数，那么则有逆文档频率计算：

$$IDF(x) = \log \frac{N}{N(x)+1}$$

$$TF - IDF = TF \times IDF$$

通常在计算完 TF - IDF 值并标准化以后，就可以为文本分类提供特征并计算，TF -IDF 值与词频正相关，与文档频率负相关。

2.3.2 改进的 TF -IDF 算法-CTF-IDF

对文本分类来说词语对类别的影响更重要,而 TFIDF 衡量词语对某个文本的重要性并没有考虑词语在类内和类间分布情况，所以本文考虑在 TF-IDF 的基础上加入类别因素 c，提出新的权重确定方法 CTF-IDF，数学表达式为式:

$$CTF - IDF = c \cdot tf_{i,j} \cdot idf_i$$

$$c = \frac{p}{p+q}, p = \frac{n}{n+m}, q = \frac{k}{k+l}$$

类别因素 c，随着词语 t 在类 r 中出现频率 p 的加而增加；随着词语在非类别中出现频率 q 的增加而减小，理想情况下词语 t 都出现在某一个类别中，类别因素 c=1。n 表示出现词语且属于类别 r 的短文本数量；m 表示属于类别 r，但没出现词语的短文本数量；k 表示出现词语但不属于类别 r 的短文本数量 l 表示没出现词语也不属于类别 r 的短文本数量确定词向量权重算法 CTF-IDF 之后,采用加权求和的方法得到短文本的向量表示、数学表达式为:

$$Vd_j = \sum_{i \in d_j} v_i \cdot CTF - IDF$$

其中，Vd_j表示文本d_j的向量，v_i表示词语t_i的词向量。

3.基于深度神经网络的分类模型

3.1 文本分类概述

近年来随着社交网络的发展,互联网论坛、电商评论、微博等文本数据迅速增长。对这些短文本数据进行文本分类,从中提取有价值的信息就显得至关重要。由于文本分类在评论数据挖掘、信息检索、舆情分析等领域应用广泛,所以文本分类算法一直是自然语言处理领域的研究热点。

针对短文本分类问题,总的方法可以分为两种,一种是传统的文本分类模型,如支持向量机(Support Vector Machine, SVM)、朴素贝叶斯(NB)、K 近邻(K-Nearest Neighbor, KNN)、决策树等。另一种是基于深度学习的神经网络分类模型,如循环神经网络(Recurrent Neural Network, RNN)、卷积神经网络（Convolutional Neural Networks, CNN）,长短时记忆网络(LSTM)等。

本文主要用了改进的 textCNN 方法来进行文本分类,首先对数据集中的留言主题进行预处理,然后使用 Word2vec 训练词向量模型将文本用向量表示。再使用改进的 TF-IDF 权重进行加权,将句子向量化,最后使用改进的 textCNN 模型进行训练,训练好模型之后,用该模型进行留言类别的分类。

3.2 TextCNN 分类算法

卷积神经网络主要用来处理具有网络结构的数据，具有较强的特征提取能力，随着自然语言处理技术的不断发展，卷积网络在文本处理领域也可以取得不错的效果。卷积神经网络与全连接神经网络的整体架构较为相似，均有输入层、输出层以及若干的中间层组成,与普通全连接神经网络的主要区别在于它的卷积层和池化层。一个经典的卷积神经网络主要由输入层、卷积层、池化层、全连接层以及输出层组成，网络结构如下图 3-1 所示：

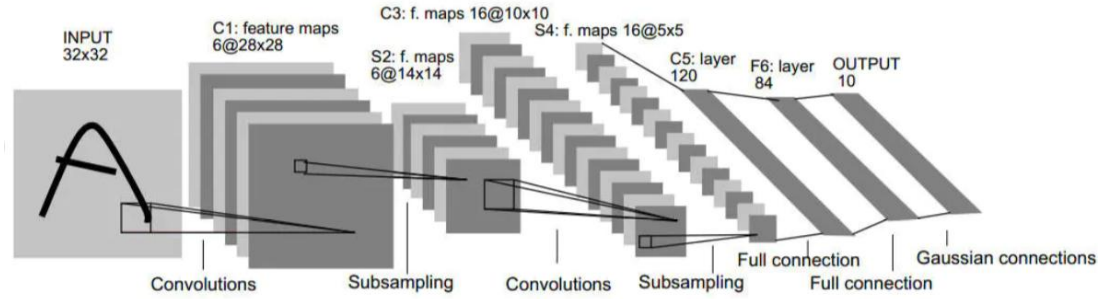

图 3-1 卷积神经网络结构图

输入层（input layer）:输入层接收具有网络结构的数据进行处理。图像数据本身由多个像素点组成，具有天然的网格结构；对于文本数据，通常将构成文本的词对应的词向量垂直拼接为一个二维矩阵作为输入数据。

卷积层（convolution layer）:卷积层是卷积神经网络中最重要的结构，具有稀疏连接、参数共享的机制。全连接神经网络由于各层网络之间每个结点都存在连续关系，参数数量巨大，而卷积神经网络中的卷积操作的稀疏连接和参数共享机制使得其参数数量极大减少。

卷积核又称过滤器（filter）,如输入数据为 n*n，过滤器大小为 f*f，在步幅为 1 的情况下执行卷积操作后生成的图像大小为(n-f+1)*(n-f+1)。为了减少尺寸变化带来的影响，CNN 中引入了 padding 操作对数据进行边界填充（zero-padding），从而防止卷积操作带来的尺寸变化问题。卷积步长指的是滤波器在图像矩阵上滑动幅度，步长过大时生成的特征举证相对较小。在引入了 padding 和 stride 后，卷积操作的输出矩阵大小如公式所示：

$$\left[\left(\frac{n+2p-f}{s}+1\right)\right]\times\left[\left(\frac{n+2p-f}{s}+1\right)\right]$$

池化层（pooling layer）:使用某特征的相邻输出的总体特征来代替网络在该位置上的输出，它的目的是为了保持平移不变性，卷积层得到的特征矩阵往往维度较大，且无法对特征的重要性进行区别池化层的通过池化操作可以提取出卷积层输出结果中最重要的一些特征减少参数数量。常用的池化操作包括最大池化（max poling）、平均池化（mean pooling）。

全连接层:将池化层的输出结果展开为一个一维向量，作为全连接层的输入，

输出层:对数据进行分类，在处理分类问题时常常采用 softmax 进行分类，并输出各个类别的概率。

本文中涉及的留言分类任务属于文本分类的范畴，TextCNN 是由 Kim 在 2014 年提出的一种用于文本分类的卷积神经网络，该网络采用多个不同大小的卷积核来提取文本序列中的关键信息，可以更好的获取序列中的局部相关性。其网络结构如图 3-2 所示。

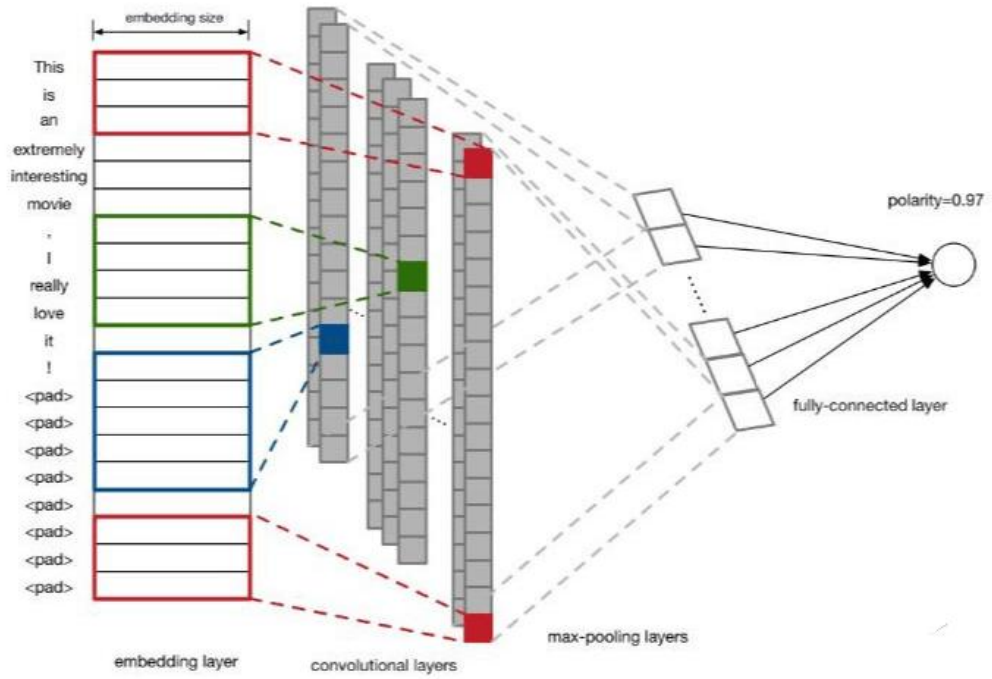

图 3-2 TextCNN 网络架构图

输入层为一个 n*k 的二维矩阵，每一行代表一个长度为 k 的词向量，n 表示样本集合的文档的最大长度，将句子中的所有词按照顺序进行垂直拼接 $X_{1:n}=X_1\oplus X_2\oplus X_3.....\oplus X_n$ 令卷积窗口大小为 h，卷积层采用 h*k 的卷积核对输入层数据进行局部特征提取，令提取后的新特征为 c_i，c_i 由第 i 到第 i+h-1 个词产生，即 $c_i=f(W\cdot x_{i:i+h-1}+b)$。图 3-3 为其详细过程原理图。

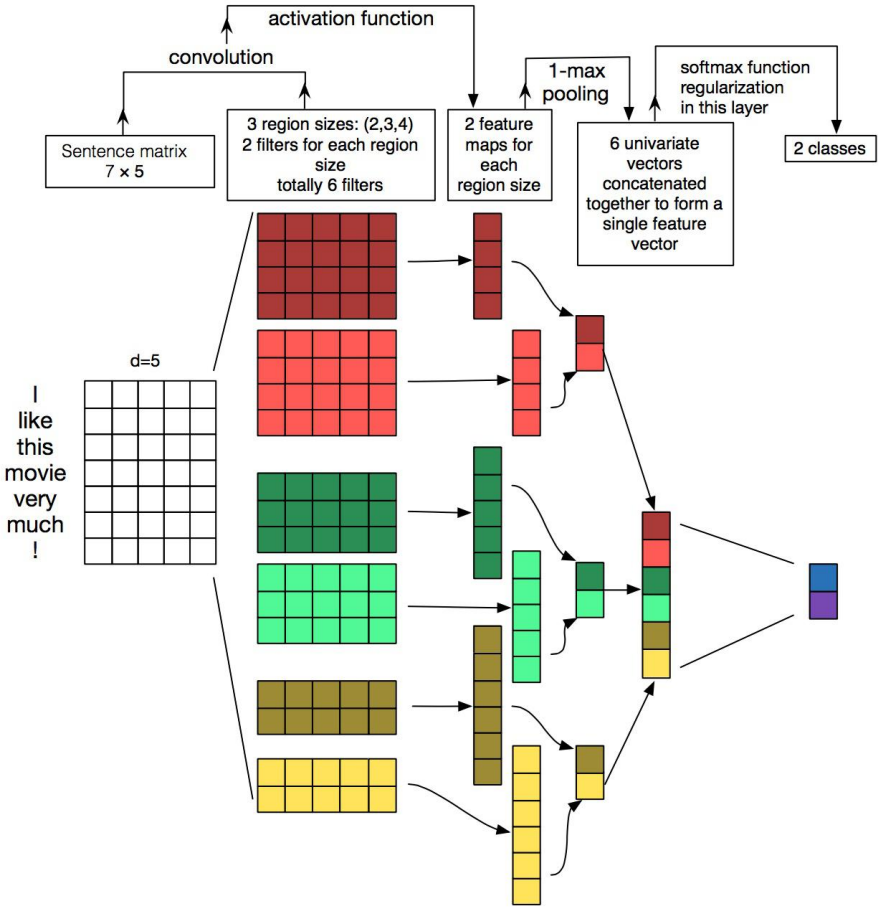

图 3-3 TextCNN 过程原理图

卷积层使用的卷积核为固定宽度，与词向量宽度相等。TextCNN 与 ngram 特征提取类似，在实际使用中需要使用多个不同尺寸的卷积核来提取不同宽度的局部特征，每个尺寸的卷积核又可以有多个，每个卷积核提取一种特征。上图中采用尺寸为 2*5、3*5、4*5 的三种卷积核各两个进行卷积操作，每个卷积核可以指定多个通道，在执行完卷积操作后生成一个长度为 n-h+1 的 feature map，图中卷积后得到 2*3 个不同大小的特征向量。然后经过池化层的 1-maxpooling 对特征进行抽象提取，由卷积层输出的 6 个大小不同的一维向量得到 6 个特征值，将提取后的特征拼接为一个一维特征向量，最后经过 softmax 层进行分类，输出每个类别的概率。

3.3 模型描述

本章对留言文本数据的分类处理分为以下步骤:

Step1:文本预处理；对留言数据进行预处理,包括中文分词,去停用词等;

Step2:文本表示;首先进行 one-hot 编码,建立基于 Word2vec 的词向量模型生成包含语义间相关性的词向量;

Step3:特征提取；使用改进的 TF-IDF 方法即 CTF-IDF 进行特征提取，词向量转化为句子向量；

Step4:分类器；使用卷积神经网络(textCNN)模型分别对留言主题和留言详情进行文本分类，并与 KNN，NB，SVM，BI-LSTM 等分类方法比较，

实验流程图如图 3-4 所示：

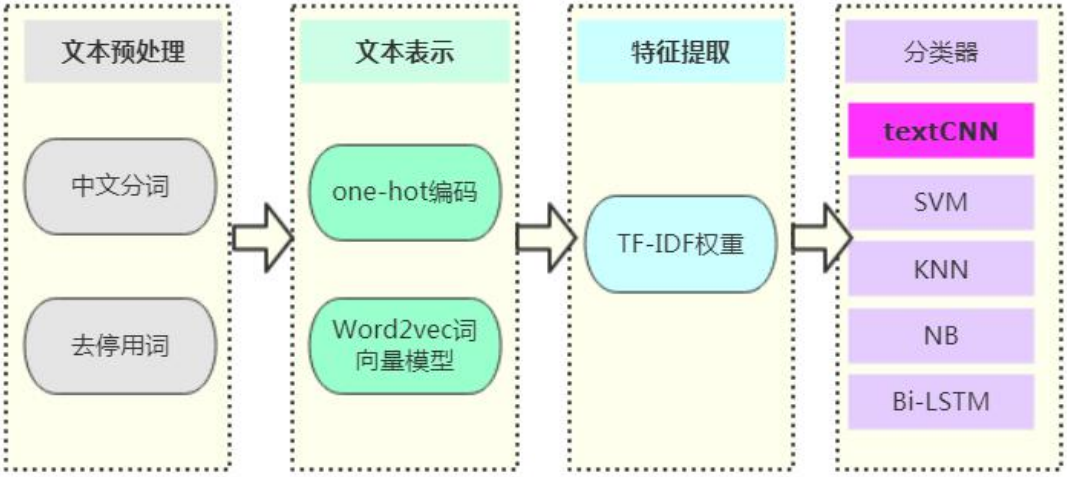

图 3-4 文本分类流程图

3.4 实验

3.4.1 实验环境

实验的顺利进行需要良好的实验硬件和软件环境，本文的实验环境配置如下表 3-1 所示：

表 3-1 实验环境配置

实验环境	具体信息
硬件环境	CPU:Intel(R)Core(TM)i7-9700KF 3.60GHz
	内存:16.0GB
软件环境	windows10 64bit Python 3.7.6
	tensorflow2.0

3.4.2 实验数据

本文的实验数据集为泰迪杯官网提供的实验数据集，改数据集包括 2010 年 11 月 2 日-2020 年 1 月 8 日的政务留言信息，一共有 9210 条评论数据，按照一级分类指标共 7 类，其中用于训练集数据 9607 条，测试集数据 2303 条，各类别的文本数量分布如表 3-2 所示，用条形图直观表示如图 3-4 所示：

表 3-2 各留言类别样本分布

类号	类别	总样本数	训练集	测试集
0	环境保护	938	703	235
1	城乡建设	2009	1507	502
2	卫生卫计	877	658	219
3	教育文体	1589	1192	397
4	商贸旅游	1215	911	304
5	劳动和社会保障	1969	1477	492
6	交通运输	613	460	153
	总体	9210	6907	2303

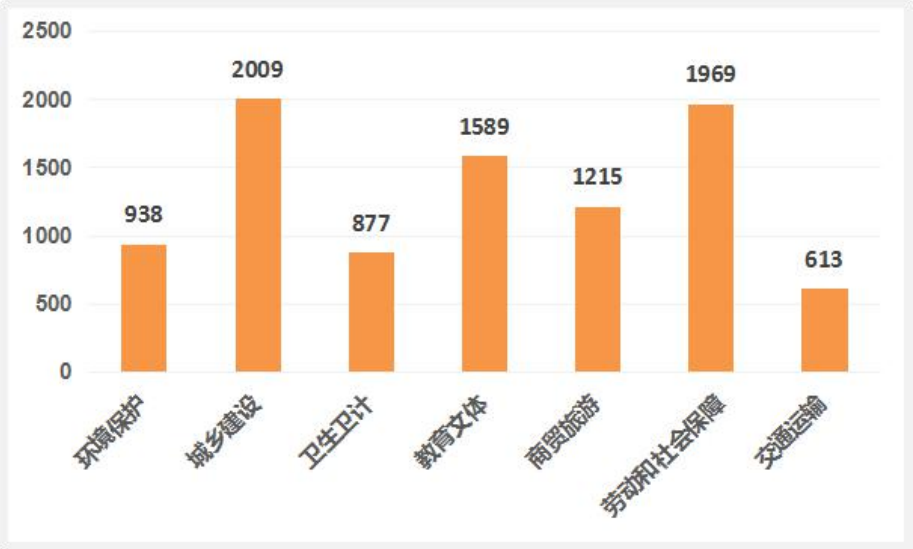

图 3-4 留言类别分布

3.4.3 评价标准

对于分类问题常用的实验评价指标包括准确率、查准率、查全率和 F1 值。各指标公式中的意义如下表 3-1 所示：

表 3-3 分类结果矩阵

	预测正例	预测负例
真实正例	TP	FP
真实负例	FN	TN

（1）准确率（Accuracy）

准确率是评判模型整体的识别能力，即正确分类的文档数占总文档数的百分比，公式表示为：

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

（2）查准率（Precision）

查准率反应的是分类模型判定为正的文档中有多少是真正的正例文档，是针对判定结果而言的，公式表示为：

$$P = \frac{TP}{TP + FP}$$

（3）查全率（Recall）

查全率反应的是样本中真正的正例文档有多少被正确判定，是针对原样本而言的，公式表示为：

$$R = \frac{TP}{TP + FN}$$

（4）F1 值（F1-score）

查准率和查全率分别从不同的角度反映出分类算法的性能，二者是互相影响的，一般情况下其中某一个指标较高可能另一个指标较低。然而 F1 指标较好的平衡了查全率和查准率，将二者结合起来进行性能评价。F1 值可表示为公式：

$$F_1 = \frac{2 \times P \times R}{P + R}$$

此外，当我们计算总体的 F1 值时，又分为 micro 和 macro 两种计算方式，前者是微平均，求整个样本集的准确率和查全率，然后求平均；后者是宏平均，假设每个类具有不同权重，先对类别求 F1，然后再求平均。本文采用宏平均，F1 值的表示公式为：

$$F_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{2 P_i R_i}{P_i + R_i}$$

其中 P_i 为第类的查准率，R_i 为第 i 类的查全率。

3.4.4 实验参数设置

对于深度学习模型来说，不同的参数会输出不同的模型，参数设置对实验结果影响很大，下表 3-4 为本文 textCNN 模型的参数值

表 3-4 实验参数表

参数	参数描述	参数值
embeding_dim	嵌入层维度	128
num_filters	卷积核数目	256
kernel_size	卷积核尺寸	3
hidden_dim	全连接层神经元	128
dropout_keep_prob	神经元丢弃率	0.2
learning_rate	学习率	1e-3

3.4.5 对照试验设置

本文主要设置了了多组对比实验,针对留言主题分类设置 SVM,KNN,NB,LSTM Bi-LSTM 模型作为对照实验；针对留言详情分类设置了 Bi-LSTM 模型作为对照实验。实验中使用的符号及其含义如下表 3-5 所示:

表 3-5 实验中使用的符号及其含义

符号名	代表的含义
SVM	支持向量机分类器
NB	朴素贝叶斯分类器
KNN	K 近邻分类器
LSTM	长短时记忆网络模型
Bi-LSTM	双向长短时记忆网络模型
textCNN	卷积神经网络分类模型
percision	准确率
recall	召回率
F1_score	F1 评分

3.5 实验结果分析

3.5.1 各分类模型比较

使用改进的 textCNN 进行分类和 Bi-LSTM 分类的结果如下表 3-6,percision 表示分类的准确率,recall 表示分类的召回率,F1_score 表示 F1 值。

表 3-6 基于留言详情的 Bi-LSTM 分类与 textCNN 分类对结果的影响

类别	percision		recall		F1_score	
	Bi-LSTM	textCNN	Bi-LSTM	textCNN	Bi-LSTM	textCNN
环境保护	0.854	0.943	0.875	0.936	0.864	0.940
卫生卫计	0.786	0.888	0.920	0.907	0.848	0.898
教育文体	0.932	0.940	0.841	0.963	0.885	0.951
交通运输	0.676	0.899	0.705	0.735	0.690	0.809
城乡建设	0.830	0.836	0.811	0.911	0.807	0.871
商贸旅游	0.698	0.856	0.733	0.816	0.715	0.836
劳动和社会保障	0.905	0.934	0.846	0.907	0.875	0.920
总体	0.829	0.899	0.824	0.897	0.825	0.896

如下图 3-5 基于留言详情的 Bi-LSTM 分类与 textCNN 分类结果对比,可以看出 textCNN 分类模型的分类效果优于 Bi-LSTM 分类模型

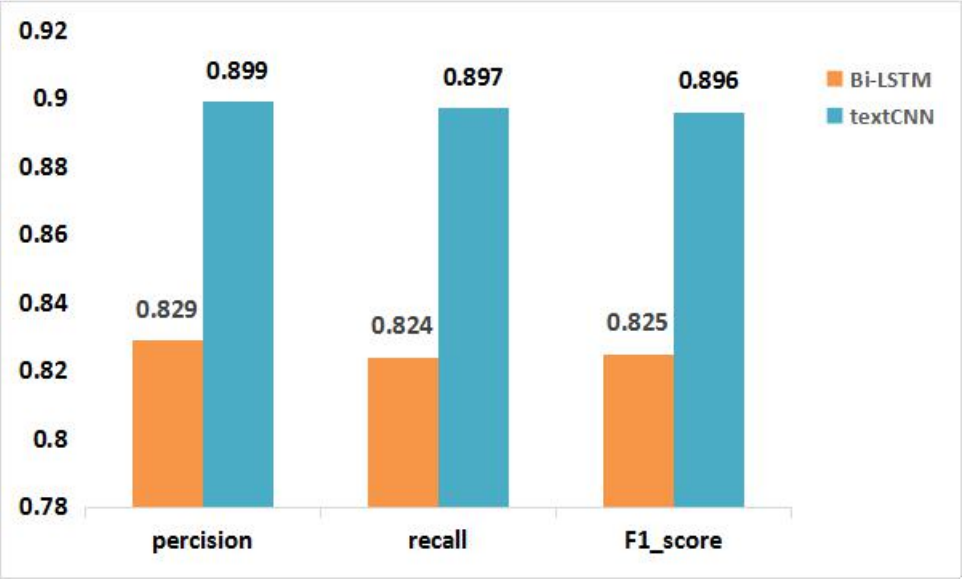

图 3-5 基于留言详情的 Bi-LSTM 分类与 textCNN 分类结果对比

如下图 3-6 是基于留言主题的各种对照分类方法的 F1 值,通过与 SVM,NB,KNN,LSTM,Bi-LSTM 等分类模型对比,可以得出本文选择的方法的分类效果最好,F1_score 为 0.84。

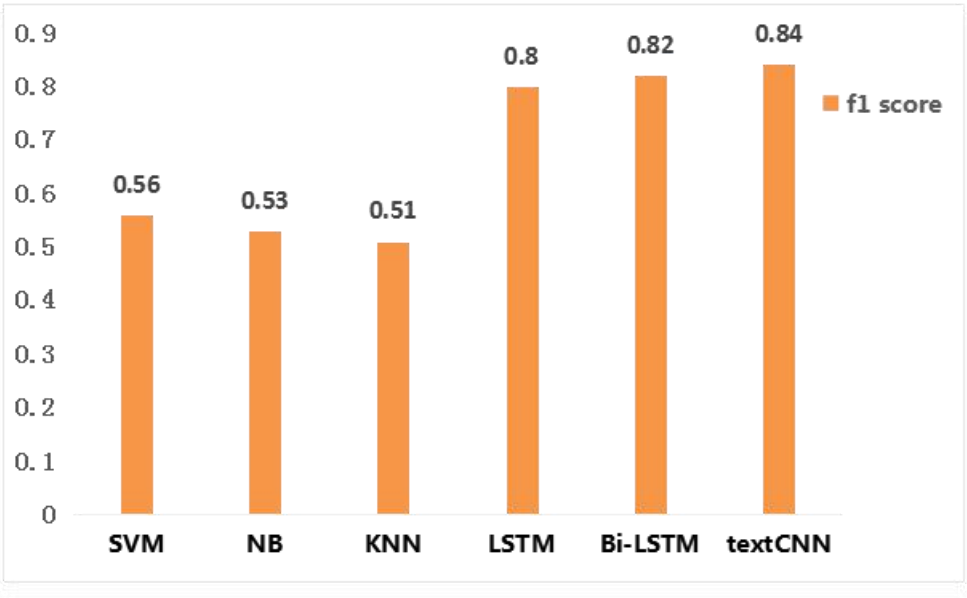

图 3-6 基于留言主题各分类模型对比

3.5.2 主题分类与详情分类对比

前面 3.6.1 我们主要是通过留言主题分类和留言详情分类对各个分类模型进行比较,本小节我们主要通过 Bi-LSTM 模型和 tectCNN 模型对留言主题和留言详情分类比较,分类结果的各指标如下表 3-7,用条形图对比如下图 3-7 可以看出留言详情分类的效果比留言主题分类的效果好。

表 3-7 留言主题与留言详情分类比较

	Flscore	
	Bi-LSTM	TextCNN
留言详情分类	0.825	0.896
留言主题分类	0.832	0.843

图 3-7 留言主题与留言详情分类比较

3.6 本章小结

本章主要介绍了卷积神经网络(textCNN)模型,使用卷积神经网络(textCNN)模型分别对留言主题和留言详情进行文本分类,并与 KNN,NB,SVM,BI-LSTM 等分类方法比较,结果得出基于留言详情分类比基于留言主题分类结果好,各种分类模型中 textCNN 模型的分类效果最好,基于留言详情的 textCNN 分类模型 F1 值为 89.6%。

4.基于文本聚类的话题热度排序

4.1 文本聚类算法概述

近年来,网络评论数据迅速增长,对这些文本数据聚类发现热点问题也成为自然语言处理领域的热点。文本聚类的目标是发现大量文本数据中的内部数据结构并进行划分。文本聚类算法不断改进和发展,现在的文本聚类算法大致有五个类别,分别是基于密度,基于划分,基于层次,基于模型,基于网格的文本聚类方法,本文我们主要使用 K-Means 算法进行文本聚类,K-Means 算法使用非常广泛,在各种聚类方法中聚类效果较好。

本章的主要是使用 K-Means 算法进行文本聚类,首先将文本数据按地区分类,接着按照第一章介绍的方法对数据进行预处理转化为词向量,对词向量使用 TF-IDF 方法进行特征提取,然后进行聚类,定义热度评价指标,给出热点问题排序结果,本章主要实验流程如下图 4-1 所示:

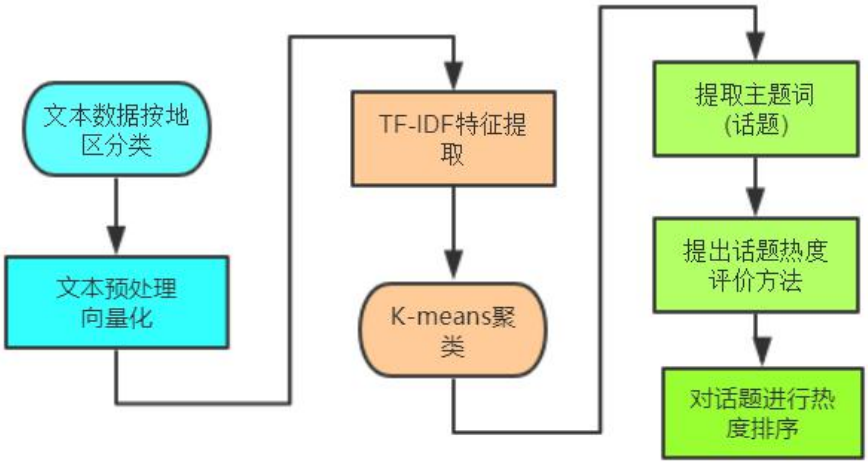

图 4-1 文本聚类、热度排序流程图

4.2 K-Means 聚类模型

K-Means 算法是基于划分方法中最经典的聚类算法，K-Means 算法是围绕若干初始点，对每个簇就近分配数据对象。输入数据,集 $S\{s_1, s_2, \ldots s_n\}$，输出 K 个簇 $C\{c_1, c_2, \ldots c_k\}$,簇的质心集合为 $M\{m_1, m_2 \ldots m_k\}$,质心的计算公式为：

$$M = \frac{1}{n_t} \sum_{s_i \in c_t} s_i$$

其中 c_t 为结果中的一个簇，n_t 是该簇中包含的数据个数。

K-Means 算法的流程如下：

（1）确定聚类的数量 K 并随机确定初始质心的坐标；

（2）计算数据间的相似度，即选择合适的距离公式计算每个数据与各个质心的距离，并将其划分到距离最近的簇中；

（3）在所有数据完成聚类后，更新每个簇的质心坐标并重新计算每个点到质心的距离，将数据点重新聚类到距离最近的簇中；

（4）重复以上步骤直到质心变化值小于设定值，聚类完成。

算法的具体流程图如 4-2 图所示：

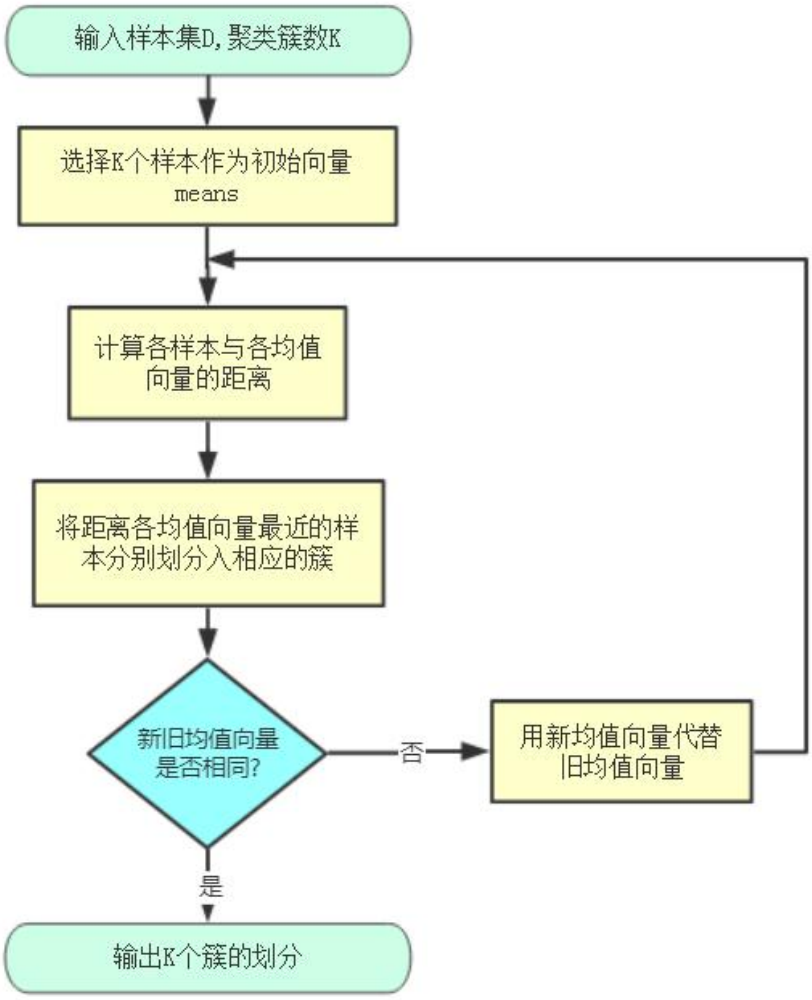

图 4-2K-Means 算法流程图

4.3 热点问题评价方法

话题热度反映了话题本身的重要程度，也反应了公众的参与程度，这是一个综合性指标。本问题中，群众留言是对某个时间所发生的某件事情的即时评论反映，影响群众留言的热度的因素有以下几个方面：话题持续时间、留言数量、点赞数和反对数。

4.3.1 话题热度影响因素及其量化

（1）**话题持续时间**。当留言所反映的问题，持续的时间越长，那么这个问题的热度也就相应的越高。在相同的时间段内，群众对某一问题的反映越多，就说明这一问题的热度越高。

（2）**留言数量**。群众对某个问题留言的条数越多，说明该问题受到群众的反映也就越多，热度也就越大。

（3）**点赞数和反对数**。当群众浏览评论时，会对以有评论给出自己的反馈，

支持或者反对，当某条留言获得的点赞和反对数越多时，那么该条留言的热度越高。

本文采用留言中的文本总数、话题持续时间、点赞和反对总数三个因素作为热度排序依据，并将这些因素量化。

文本总数是指在整个留言文档中，某个留言话题的所有文本数量总和，数学符号记为N（topic）。

话题持续时间是指某个留言话题，在最初发布的日期到最后发布日期的时间长度，数学符号记为T-L(time-long)。若最早发布时间记为T（first），最后发布时间为T(last)，则计算公式为：

$$\mathrm{T-L(time-long)=T(first)+T(last)}$$

点赞数和反对数总和是指所有有关话题的留言下的点赞数和反对数的总和，数学符号记为A-O（agree-oppose)，其中agree表示点赞数，oppose表示反对数，计算公式为：

$$A-O(\mathrm{agree-oppose})=\sum_{i=1}^{N(topic)}(\mathrm{agree+oppose})$$

4.3.2 话题热度排序实现流程

本文排序采用的策略思想是，对话题热度的总评分设置为100分，将三个评价指标按照一定的权重分配相应分数分别为α，β，γ，分别计算每个话题在三个指标上的得分,将三个指标的得分相加记为话题热度评分,最后按照话题热度评分排序。具体计算步骤如下:

Step1:定义三个指标N(topic)、T-L(time-long)、A-O（agree-oppose)对应的权重分数分别为$\alpha=50$，$\beta=30$，$\gamma=20$。

$$100=\alpha+\beta+\gamma$$

Step2:所有话题分别按照三个指标N(topic)、T-L(time-long)、A-O（agree-oppose)大小从低到高排序，计算出第i个话题按指标N(topic)排序结果对应的分位数,记为Z_{iN},第i个话题按指标T-L(time-long)排序结果对应的分位数记为Z_{iT-L},第i个话题按指标A-O(agree-oppose)排序结果对应的分位数记为Z_{iA-O}，

Step3:计算话题在各指标上的得分。分别用α，β，γ乘以相应的排序分位数Z_{iN}，Z_{iT-L}，Z_{iA-O}就可以计算出第i个话题在3个指标上的得分a_i，b_i，c_i，计算公

式如下:

$$a_i = \alpha \times Z_{iN}$$
$$b_i = \beta \times Z_{iT-L}$$
$$c_i = \gamma \times Z_{iAO}$$

Step4:计算各个话题的热度得分。第 i 个话题的热度得分等于其在三个指标上的得分之和:

$$Score_i = a_i + b_i + c_i$$

Step5:对各话题热度得分排序,得出热点话题排序结果

话题热度评估和排序策略流程图如图 4-3 所示:

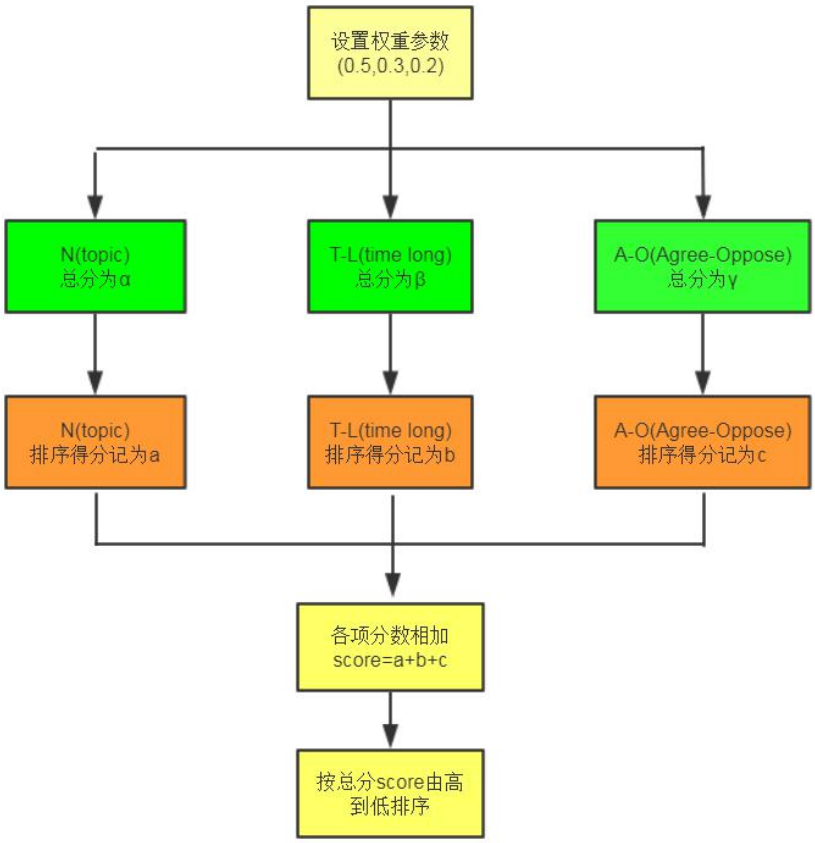

图 4-3 话题热度评估和排序策略流程图

4.4 基于文本聚类结果的话题热度排序

本文对泰迪杯官网方提供的附件三的 4326 条数据，首先对群众留言主题数据通过分词、去噪、停用词过滤等文本预处理方法，然后使用 K-means 方法进行聚类,得到聚类结果一共聚了 531 类,部分聚类结果如图 4-4 所示，其中 label

列代表每一条留言主题所属的组，图中给出了第一组和第二组的聚类结果。

label	反对数	点赞数	留言主题	留言时间	留言用户	留言编号
1	0	2	A5区五矿万境K9县的开发商与施工方建房存在质量问题	2019/5/5 13:52:50	A00094436	20806
1	0	2097	A市A5区汇金路五矿万境K9县存在一系列问题	2019/8/19 11:34:04	A00077171	20863
1	0	1	A市五矿万境K9县存在严重的消防安全隐患	2019/9/12 14:48:07	A00010323	21550
1	0	6	A市五矿万境K9县房子的墙壁又开裂了	2019/6/20 9:30:44	A00099869	23408
1	0	0	A市五矿万境K9县交房后仍存在诸多问题	2019/9/11 15:16:02	A00010531	25265
1	0	0	A市五矿万境K9县房屋出现质量问题	2019/9/19 17:14:49	A00010042	26259
1	0	0	A市五矿万境K9县负一楼面积缩水	2019/9/10 9:10:22	A00061339	27549
2	0	0	反映A市恒大御景天下楼盘二期烟道质量及设计问题	2019/5/9 12:13:39	A00088601	20286
2	0	2	A市时代年华项目被责令限期整改，请问是否真的实施？	2019/7/10 10:46:04	A00011794	23587
2	0	1	反映A市恒大江湾退房退款问题	2019/3/16 18:00:37	A00060834	22625
2	0	1	反映A市禹泰云开一品产权证问题	2019/4/10 19:13:25	A00080564	23746
2	0	0	反映A市鑫华驾校的一些问题	2019/11/27 17:19:28	A00064002	23971
2	0	1	投诉A市红星国际公馆违规交房现象	2019/3/12 10:40:32	A00037852	24428
2	0	9	关于A市金晖优步花园相关问题的反映	2019/1/8 17:07:55	A00080248	28872
2	5	1762	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	A00087522	22329
2	0	0	反映A市合能枫丹丽舍精装修问题及业主子女入读小学事宜	2019/2/27 10:01:21	A00095839	28658
2	0	0	关于A市梅溪湖安置房质量威胁人身安全问题的反映	2019/10/31 19:07:09	A00054159	27053
2	0	3	再次反映A市时代年华项目装修价质不符问题	2019/8/12 16:24:48	A00011794	19748
2	0	0	反映A市金座雅居的一些问题	2019/7/5 16:14:12	A00073189	25937

图 4-4 部分聚类结果

图 4-5 显示了各组中所包含的文本数量，这样可以清楚留言文本的分布情况。

图 4-5 聚类分组留言文本数量

4.4.1 话题的关键词提取

话题是一系列相关的关键词组合。在文本聚类完成后，要根据最终的分类组从中提取关键词，筛选关键词并组成话题来表示分组的核心内容。现将本文已经聚类好的留言文本分组进行关键词提取。聚类结果中，已经确定了所有留言文本共分为 531 个组，因此能够构建出 531 个话题。

本文使用词频-逆文本频率（TF-IDF）算法找出留言主题文本中的关键词。文本的 2。4 章节中对 TF-IDF 进行了描述。将同一分组当作一个单独的文本来处理。由于文本数量以及参与权重计算的词语基数庞大，因此本文从每组文本的权值计算结果中选择权值最大的 20 个词语，当做每组文本的关键词。部分留言主题的关键词如图 4-6 所示：

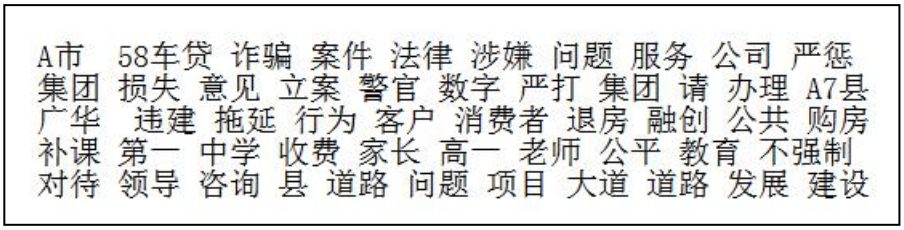

图 4-6 部分文本关键词

在提取出每组的关键字之后，需要对这些关键词进行筛选组成话题。话题通常是对事件精简客观的概括描述。形成每一组的话题，也就是对每组留言进行了关键词问题描述，形成简要的问题概括。

4.4.2 话题的热度排序结果

根据前面 4.3.2 节提出的热度评价方法对前面的各个话题计算热度得分并进行热度排序,给各个热度影响因素分配不同比例，话题最终的排序结果也会不一样。文中 4.3.2 节对三个热度指标分别给予 0.5、0.3、0.2 的权重，综合热度排序的部分热度排序展示如下图 4-7 所示：

热度评分	热度指数	N(topic)	T-L	A-O	组别	反对数	点赞数	留言主题	留言时间	留言用户
93.5133	0.935133	46.75665148	28.05399089	18.70266059	1	0	2	A5区五矿万	2019/5/5	A00094436
93.5133	0.935133	46.75665148	28.05399089	18.70266059	1	0	2097	A市A5区汇	2019/8/19	A00077171
93.5133	0.935133	46.75665148	28.05399089	18.70266059	1	0	1	A市五矿万	2019/9/12	A00010323
93.5133	0.935133	46.75665148	28.05399089	18.70266059	1	0	6	A市五矿万	2019/6/20	A00099869
93.5133	0.935133	46.75665148	28.05399089	18.70266059	1	0	0	A市五矿万	2019/9/11	A00010531
93.5133	0.935133	46.75665148	28.05399089	18.70266059	1	0	0	A市五矿万	2019/9/19	A00010042
93.5133	0.935133	46.75665148	28.05399089	18.70266059	1	0	0	A市五矿万	2019/9/10	A00061339
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	0	0	反映A市恒	2019/5/9	A00088601
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	0	2	A市时代年	2019/7/10	A00011794
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	0	1	反映A市恒	2019/3/16	A00060834
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	0	1	反映A市禹	2019/4/10	A00080564
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	0	0	反映A市鑫	2019/11/2	A00064002
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	0	1	投诉A市红	2019/3/12	A00037852
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	0	9	关于A市金	2019/1/8	A00080248
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	5	1762	反映A市金	2019/4/11	A00087522
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	0	0	反映A市合	2019/2/27	A00095839
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	0	0	关于A市梅	2019/10/3	A00054159
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	0	3	再次反映A	2019/8/12	A00011794
88.15121	0.881512	44.07560254	26.44536152	17.63024102	38	0	0	反映A市金	2019/7/5	A00073189

图 4-7 热度综合性排序结果展示

从结果来看，排名前列的热度排名，是受到公众广泛关注的留言问题，说明本文制定的热度排序策略有实用性和可靠性。

根据题目要求对热度排名前五的组进行话题提取得到相应的问题描述，并按照如下表 4-1:

表 4-1 热点问题表

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	0.935	2019/5/5 至 2019/9/10	A 市五矿万境 K9 县	房屋出现质量问题和安全隐患
2	2	0.881	2019/1/1 至 2019/11/27	A 市住户子女	子女按区入学问题存在困难
3	3	0.751	2019/2/7 至 2019/12/21	A 市网络平台	以 58 车贷为首的特大诈骗案
4	4	0.607	2019/5/21 至 2019/10/23	A9 市车站	车站设立及选址争议

5	5	0.391	2019/1/91 至 2019/7/21	A4 区绿地海外滩小区	小区受到高铁噪音困扰

对热点问题的热度指数、时间范围、地点人群以及问题描述经行整理,对相应热点问题对应的留言信息整理，得到表 4-2“热点问题留言明细表”如下表 4-2 所示：

表 4-2 热点问题留言明细表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
1	208069	A00094436	A5区五矿万境K9县的开发商	2019/5/5 13:52:50	本人是A5区洞井街道汇金路	0	2
1	208636	A00077171	A市A5区汇金路五矿万境K9县	2019/8/19 11:34:04	我是A市A5区汇金路五矿万	0	2097
1	215507	A00010323	A市五矿万境K9县存在严重的	2019/9/12 14:48:07	预交房23栋没有通往负一楼	0	1
1	234086	A00099869	A市五矿万境K9县房子的墙壁	2019/6/20 9:30:44	五矿万境K9县的房子又出问	0	6
1	252650	A00010531	A市五矿万境K9县交房后仍存	2019/9/11 15:16:02	尊敬的相关部门，本人家庭	0	0
1	262599	A00010042	A市五矿万境K9县房屋出现质	2019/9/19 17:14:49	我是西地省A市五矿万境K9县	0	0
1	275491	A00061339	A市五矿万境K9县负一楼面积	2019/9/10 9:10:22	关于五矿万境·K9县负一楼	0	0
2	202862	A00088601	反映A市恒大御景天下楼盘二	2019/5/9 12:13:39	我购买了A市恒大御景天下二	0	0
2	235871	A00011794	A市时代年华项目被责令限期	2019/7/10 10:46:04	2019年7月5日房产报报道了	0	2
2	226254	A00060834	反映A市恒大江湾退房退款问	2019/3/16 18:00:37	胡书记您好：我在A市创业开	0	1
2	237463	A00080564	反映A市禹泰云开一品产权证	2019/4/10 19:13:25	禹泰云开一品小区，2015年	0	1
2	239711	A00064002	反映A市鑫华驾校的一些问题	2019/11/27 17:19:28	举报A市鑫华驾驶员培训有限	0	0
2	244289	A00037852	投诉A市红星国际公馆违规交	2019/3/12 10:40:32	认购的红星实业集团开发楼	0	1
2	288728	A00080248	关于A市金晖优步花园相关问	2019/1/8 17:07:55	关于A市金晖优步花园（金晖	0	9
2	223297	A00087522	反映A市金毛湾配套入学的问	2019/4/11 21:02:44	书记先生：您好！我是梅溪	5	1762
2	286582	A00095839	反映A市合能枫丹丽舍精装修	2019/2/27 10:01:21	合能6、7业主要求参与毛坯	0	0
2	270539	A00054159	关于A市梅溪湖安置房质量威	2019/10/31 19:07:09	关于梅溪湖安置房质量威胁	0	0
2	197483	A00011794	再次反映A市时代年华项目装	2019/8/12 16:24:48	本人是A市A6区时代年华楼盘	0	3
2	259370	A00073189	反映A市金座雅居的一些问题	2019/7/5 16:14:12	1、我们小区位于A4区新河街	0	0

4.5 本章小结

本章介绍了 k-means 聚类模型,定义了影响话题热度的因素,并将各因素指标量化,定义了计算话题热度评分的方法。使用 k-means 对附件三的 4326 条群众留言数据进行聚类分析,定义了热度评价方法,计算各聚类话题的热度评分,并给出话题热度排序结果。

5.答复意见评价方案

针对公众留言，相关部门对留言进行了答复并给予了相关意见，如果相关部门对公众留言给予及时有效的回复，那么公众会更加对留言平台给予信任，愿意在留言平台上诉说问题，同时对部门的公信力度也有所提升。这样一来如何评价答复的质量是一个重要的问题。高效的评估答复意见对日后的平台管理和问题解决都有很大的帮助。评价答复意见可以从以下方面进行度量，分别是相关性、完整性、时效性、可解释性。

5.1 政府留言答复意见评价指标

根据题目所给的数据信息,构建了文本相似度(Cosine Distance),回答中心度(ansCen),语言多样性(lingDiv),时间间隔(ansTime),文本长度(length),情

感支持(emSup)六个指标,从相关性,完整性,回答时效性,可解释性四个角度对政府答复意见进行评价。各评价指标及其测度如下表 5-1 所示:

表 5-1 模型变量与度量指标

变量类别	变量名称	变量	变量测度
相关性	相似度	d	回复与评论的余弦距离
完整性	回答中心度	ansCen	回答文本所有词的权重之和
	语言多样性	lingDiv	不同词语数/回答总词数
回答时效性	时间间隔	ansTime	回答与提问时间之差
可解释性	文本长度	length	回复内容包含字数
	情感支持	emSup	含积极词汇句子数/文本总句子数

5.1.1 相关性

相关性 （correlation）是对答复意见与留言主题的相关性分析，主要从通过相似度来表示,若两者的相似度越高，那么也就说明答复意见与留言主题越相关。文本相似度的计算上更倾向以使用余弦距离。

若两个 n 维向量 A(a1,a2,...an)和 B(b1,b2,...bn)。则余弦距离（Cosine Distance）,是以两个向量之间的夹角余弦值作为判断依据来评估相似度。当夹角较小时，说明向量所指方向较近，可以认为两个向量比较相似。余弦距离计算公式为：

$$d(A,B)=\cos(A,B)=\frac{A\cdot B}{|A|\times|B|}=\frac{\sum_{i=1}^{n}a_i\times b_i}{\sqrt{\sum_{i=1}^{n}a_i^{\,2}\times\sum_{i=1}^{n}b_i^{\,2}}}$$

第 K 条留言与答复的相似度记为d_k

5.1.2 完整性

答复意见是否完整对评价来讲非常重要，对于答复意见的完整性，其主要包括广度和深度两个方面。从广度上看，答复意见的表达内容要全面，对于一个问题的回答要全面，不能缺少任何公众留言中所包含的内容单元。从深度上看，答复意见的信息量要丰富。本文提出两个指标进行完整性度量：回复中心度、语言多样性。

(1)回答中心度

回答中心度（ansCen）是指一条留言回复在所有留言回复中的中心程度,中

心程度越高,表明该解决办法的有用的可能性越大。

对回答中心度进行变量度量,首先对文本预处理计算词向量的 TF-IDF,然后对该回答中所有词 i 的 TF-IDF 值累加,记第 k 条答复的回答中心度为$ansCen_k$，回答中心度计算公式如下，

$$ansCen = \sum_{i=1}^{n} TF_i \times IDF_i$$

(2)语言多样性

在留言答复中,具有良好的书面格式和语言的多样性的回答能够提高回复的可信度。回复的语言风格在很大程度上影响其所回复信息的认同度。语言多样的回答往往比起乏味重复的回答更具有说服力。

对语言多样性（lingDiv）进行度量，即文本中不同词的数量与文本总词数之比。计算回复的语言多样性，其中 n 表示留言中不同词的数量，$\sum_{i=1}^{n} f_i$ 表示回复总词数，记第 k 条答复的语言多样性为:$lingDiv_k$,计算公式如下：

$$lingDiv = n / \sum_{i=1}^{n} f_i$$

5.1.3 回答时效性

回复时效性是指回答与提出问题之间的间隔时间，间隔越小表示其回复的越及时。时效性是留言着做出满意答案判定的指标之一。回答时效性（ansTime）的计算方式为回答与提出问题的时间差的倒数（以天数计算）。第 k 条留言时效性表示记为$ansTime_k$,计算公式为:

$ansTime_k = \frac{1}{t_k}$,其中,$t_k$表示第 k 条留言与答复的时间差

5.1.4 可解释性

在回复中公众主要在意的是是否对提出的问题进行解释,文本提出两个指标进行可解释性度量，分别是：句子的长度和可解释性。

(1)文本长度

文本长度（length）指回复文本中所包含得到字数，显然回复的内容所包含的字数越多，说明回复的更加全面，所涉及的内容也就越多，对于公众留言提到的问题可以解答的方面也就更多。第 k 条答复意见的文本长度为$Length_k$

(2)情感支持

研究发现情感支持（emSup）对用户采纳意见有正向影响。带有正面情感的

信息有助于增强用户的认同。对回复文本进行文本预处理，通过知网发布的正面情感词表进行自动化查询匹配，若句子中包含词表中的情感词，则标注为 1，认为其具有情感倾向；若不包含，则标注为 0，情感支持可以通过回复中含有的正向情感词数量来度量,记第 k 条回复的情感支持记为 $\mathrm{em}Sup_k$

5.2 层次分析法确定各评价指标权重

5.2.1 层次分析法(AHP)介绍

上文中我们根据建立了留言回复的评价指标体系,本节使用层次分析方法的步骤,建立留言回复评价指标的判断矩阵,根据判断矩阵计算特征向量,最大特征根,并且进行一致性检验,最后得出各个指标的权重。

使用层次分析法确定权重的步骤:

步骤 1:分析系统中各因素间的关系，建立层次分析结构模型

在深入探讨实际问题的基础上,将有关的各个因素按照不同属性自上而下地分解成若干层次，同一层的诸因素从属于上一层的因素或对上层因素有影响，同时又支配下一层的因素或受到下层因素的作用。最上层为目标层，通常只有 1 个因素，最下层通常为方案或对象层，中间可以有一个或几个层次，通常为准则或指标层。当准则过多时(譬如多于 9 个)应进一步分解出子准则层。

步骤 2:对同一层次各元素关于上一层次中某一准则的重要性进行两两比较，构造两两比较的判断矩阵；

表 5-2

b_{ij} 的取值	含 义
1	表示两个元素 B_i 和 B_j 相比，同样重要
3	表示 B_i 比 B_j 稍微重要
5	表示 B_i 比 B_j 明显重要
7	表示 B_i 比 B_j 强烈重要
9	表示 B_i 比 B_j 极端重要
2、4、6、8	上述两相邻判断中的值，如 2 为同样重要和稍微重要之间的判断值
1、2、…、9 的倒数	元素 B_i 和 B_j 比较时为 b_{ij}，则 B_j 和 B_i 比较时为 $1/b_{ij}$

步骤 3:由判断矩阵的特征向量 W_i 以及最大特征值 λ_i，将 W_i 平均化即可得到

各影响因素对目标的重要性排序值。

步骤 4:一致性检验

平均随机一致性指标 RI 如表 5-3 所示

表 5-3 平均随机一致性指标 *RI*

n	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.95	1.12	1.24	1.32	1.41	1.45

利用 $CI=\frac{\lambda_i-n}{n-1}$ 求出判断矩阵的一致性指标，再根据 $CR=\frac{CI}{RI}$ 求出判断矩阵的一致性比率，

当 CR=0 时,判断矩阵具有一致性；

当 CR<0.1 时,认为判断矩阵的一致性较好

当 CR>=0.1 时,认为应该对判断矩阵作适当修改

5.2.2 各评价指标权重计算

(1)建立树形评价指标体系

根据上文中建立的留言回复评价指标体系,建立如下的树形层次结构,如图 5-1 所示:

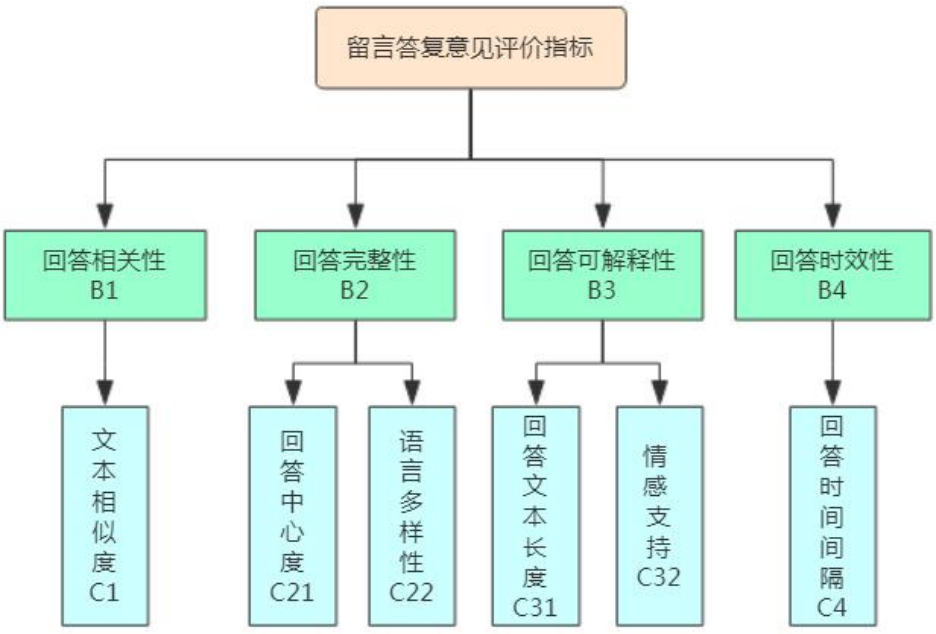

图 5-1 留言答复层次结构图

(2)二阶因子权重向量

根据各指标的重要性,构造两两比较判别矩阵 A-B,如下表 5-4 ,计算特特征向量,最大特征根,并且进行一致性检验:

表 5-4 两比较判别矩阵 A-B

A	B_1	B_2	B_3	B_4	W
B_1	1	5	4	2	0.49
B_2	1/5	1	1/3	1/4	0.07
B_3	1/4	3	1	1/3	0.14
B_4	1/2	4	3	1	0.30

其中 $M_1=B_1*B_2*B_3*B_4$=40, M_2=1/60, M_3=1/4, $M_4=6$

$\overline{W_1}$=2.515/5.146=0.49, $\overline{W_2}$=0.359/5.146=0.07

$\overline{W_3}$=0.707/5.146=0.14, $\overline{W_4}$=1.565/5.146=0.30

$$AW=\begin{bmatrix}1 & 5 & 4 & 2\\ 1/5 & 1 & 1/3 & 1/4\\ 1/4 & 3 & 1 & 1/3\\ 1/2 & 4 & 3 & 1\end{bmatrix}\begin{bmatrix}0.49\\0.07\\0.14\\0.30\end{bmatrix}=\begin{bmatrix}2.00\\0.29\\0.57\\1.25\end{bmatrix}$$

$\lambda_{\max}$=2.0/(4*0.49)+0.29/(4*0.07)+0.57/(4*0.14)+1.25/(4*0.30)=4.111

CI=(4。111-4)/(4-1)=0.037

查表可知,当 n=4 时,RI=0.90

所以 $CR=0.037/0.90=0.041<0.1$,因此判断矩阵 A-B 具有较好的一致性

所以二阶因子的权重 $W=(0.49,0.07,0.14,0.30)^T$

(3)一阶因子的权重向量

由于各二阶因子下的一阶因子较少,这里直接给出权重结果

$$W_{C1}=1, W_{C2}=(0.5,0.5)^T, W_{C3}=(0.5,0.5)^T, W_{C4}=1$$

(4) 经过层次分析法计算之后各指标的权重如下表所示

表 5-5 各指标权重

变量类别(权重)	变量名称	变量	变量权重
相关性(0.49)	相似度	d	0.49
完整性(0.07)	回答中心度	ansCen	0.07*0.5=0.035
	语言多样性	lingDiv	0.07*0.5=0.035
回答时效性(0.30)	时间间隔	ansTime	0.30
可解释性(0.14)	文本长度	length	0.14*0.5=0.07
	情感支持	emSup	0.14*0.5=0.07

5.3 政府答复意见评价结果分析

根据题目所给数据计算处 6 个指标值,并对各指标值进行归一化处理,处理之后根据各个指标的权重就行加权计算第 k 条回复的得分为:

$$S_k = 0.49 \times d_k + 0.07 \times (0.5 \times ansCen_k + 0.5 \times lingDiv_k)$$

$$+ 0.14 \times (length_k \times 0.5 + emSup_k \times 0.5) + ansTime_k \times 0.30$$

(1) 时效性指标统计图

通过对留言与答复时间间隔的倒数进行计算,变量统计图如图 5-2:

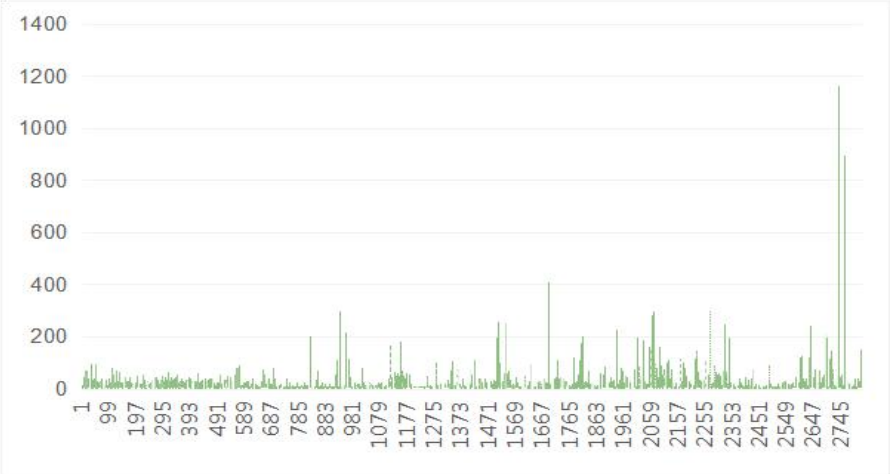

图 5-2 时效性指标统计

图为回复公众留言的时间间隔指标，可以看出对公众所提出的大多数问题，在 20 天内能够得到回复，而有些问题回复的时间超过了一年之久。

(2)答复文本长度统计图

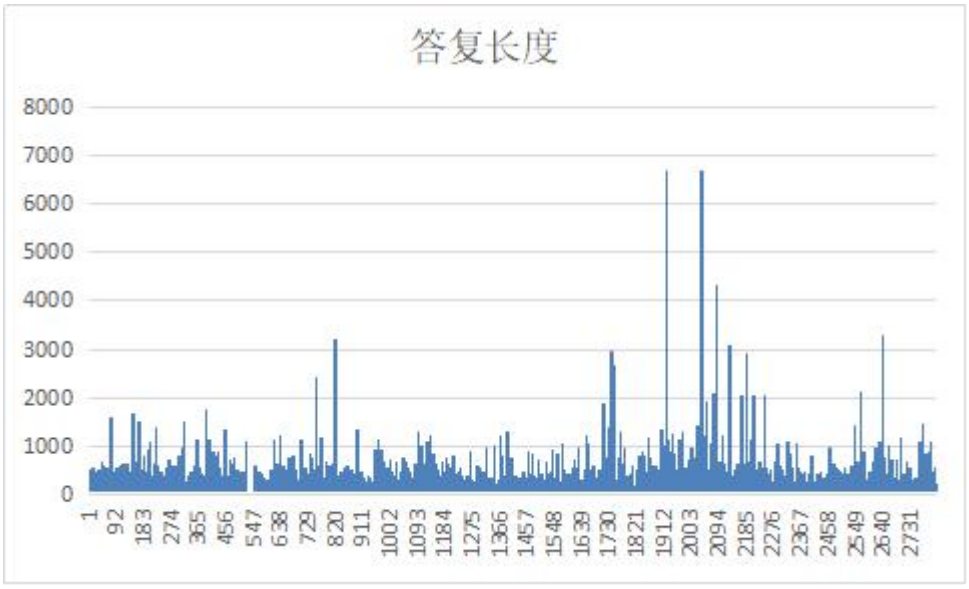

图 5-3 答复文本长度统计图

图为回复公众留言的答复长度，总体来看平台对公众的问题都进行了详细的回答，这也有助于公众对答复系统的信任，更能够通过留言平台解决问题。

最后的留言回复意见评价结果表明,政府的留言回复较为完整,相关性强,可解释性强,具体的政府留言答复意见评分见附件。

6. 总结与展望

本文从多个方面对留言数据进行了分析,首先,介绍了数据预处理工作,讨论了中文分词,去停用词方法,研究了词向量训练 Word2vec 词向量模型,介绍了改进的 TF-IDF 权重模型--CTF-IDF。其次介绍了群众留言文本分类模型,主要介绍了卷积神经网络(textCNN)模型,并与其他的分类模型 KNN,NB,SVM,Bi-LSTM 等进行对比。然后,使用 k-means 聚类方法对留言进行聚类,定义热度评价指标,进行

话题热度排序,整理热点问题;最后介绍了对政府留言答复意见的评价方案,主要通过层次分析法构建评价指标体系进行综合评价。整体来说对文本数据的分析比较全面,这些方法能够极大的提高数据处理的工作效率。

我们认为文中存在的问题就是在问题二中,题目中所给的数据只有留言内容,留言时间,点赞和反对数,没有转发量和评论量等更多热度相关的信息,如果有这些信息,我们的热度评价方法将会更加完善与实用。

参考文献

[1] 杨宇婷.基于分布式表达的微博话题检测与情感分类研究[D].东北林业大学,2016.
[2] 王宁.Word2vec 训练优化的影响因素研究[D].苏州大学,2018.
[3] 施聪莺,徐朝军,杨晓江.TFIDF 算法研究综述[J].计算机应用,2009,第 C1 期
[4] 陈珂,梁斌,柯文德,许波,曾国超.基于多通道卷积神经网络的中文微博情感分析[J].计算机研究与发展,2018,第 55 卷,第 5 期.
[5] 冯超,梁循,李亚平,周小平,李晓菲.基于词向量的跨领域中文情感词典构建方法[J].数据采集与处理,2017,第 32 卷,第 3 期.
[6] 刘华.基于关键短语的文本分类研究 [J].中文信息学报,2007,21(4):34－41.
[7] 徐安滢,吉宗诚,王斌.基于用户回答顺序的社区问答答案质量预测研究[J].中文信息学报,2017,第 31 卷,第 2 期.
[8] 朱晓峰,陈楚楚,尹婵娟.基于微博舆情监测的 K-Means 算法改进研究[J].情报理论与实践,2014,第 1 期.
[9] 王小林,杨林,王东.基于知网的新词语相似度算法研究[J].情报科学,2015,第 33 卷,第 2 期.
[10] 郑梦欣.问答型虚拟社区用户知识共享行为影响因素研究[D].黑龙江大学,2017