摘要

近年来，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。

针对问题一：群众留言分类。对数据进行预处理和清洗，采用贝叶斯、TF-IDF、TextCNN 等模型进行比较，选取最优模型，并给出分类结果。同时，为了优化分类情况，在分词等部分进行一定的人工干预，扩展修改停用词表、同义词表等，将预测的误差降到最低。以此来将群众留言分派至相应的职能部门处理，提高工作效率。

针对问题二：热点问题挖掘：结合 LDA 模型和实际数据意义给出热点问题的评价标准，从三个维度来考核问题热度，并进行分类分主题两步走的热点探析，找到热点问题，并且输出对应的热点留言。

针对问题三：答复意见评价：本文从解决度、可解释性、完整性、相关性、及时性五个维度综合评价，构成完整有效的评价体系，在问题是否解决的前提下，研究了不同类别下相关性、完整性、可解释性和及时性的关联和关系，将不同维度赋予权重并具体实践。良好的评价系统是保证反馈渠道价值的重要保障，切实符合市民的心理需求。

关键字：朴素贝叶斯；TF-IDF 模型；LDA 模型；答复意见评价；

Abstract

In recent years, the network political platform has gradually become an important channel for the government to understand public opinion, gather people's wisdom, rally the people's spirit, all kinds of social and public opinion-related text data volume continues to climb, to the past mainly rely on manual to carry out the message

division and hot-spot finishing of the relevant departments of the work has brought great challenges. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government system based on natural language processing technology has become a new trend of innovation and development of social governance, which has a great role in promoting the level of government management and the efficiency of governance.

For question one: mass message classification. The data are pre-processed and cleaned, compared with Bayesian, TF-IDF, TextCNN and other models, selectthe optimal model, and give the classification results. At the same time, in order to optimize the classification, some manual intervention is carried out in part of the part, such as part-word, extending the modification of the de-performing word list, synonym list, etc., to minimize the error of prediction. In order to assign the mass message to the corresponding functional departments to deal with, improve work efficiency.

For question two: hot issue mining: combined with the LDA model and the actual data significance to give the evaluation criteria of hot issues, from three dimensions to assess the problem heat, and classify the topic two steps of hot spot analysis, find hot issues, and output the corresponding hot spot message.

For question three: Reply comment valuation: This paper from the solution, interpretability, integrity, relevance, timeliness of the five dimensions of comprehensive evaluation, constitute a complete and effective evaluation system, under the premise of whether the problem is resolved, the correlation, integrity, interpretation and timeliness under the premise of studying the correlation, integrity, interpretation and timeliness of the correlation and relationship, the different dimensions to weight and concrete practice. A good evaluation system is an important guarantee to ensure the value of feedback channels, and it is in line with the psychological needs of the public.

Keywords: Plain Bayes; TF-IDF model; LDA model; Reply comment valuation

1 问题重述

1.1 研究背景

近年来，为了更好地倾听群众心声，发挥广大市民的监督作用，市长信箱、阳光热线等公开平台逐渐走进人们的生活，随着平台的使用率不断提升，各类社情民意的反馈信息逐渐增多，如果仅仅靠人工来处理、整理并解决各类复杂繁琐或重复的问题，对相关部门的工作是极具有挑战性的。

从目前的平台来看，实际效果的实施方面存在几点问题，一是重复问题较多，可能同一个问题被反复提及，未得到解决或是解决了 A，B 并没有看到，从而造成了工作的重复；二是问题分类不明显，市民在提问时并不会在意问题属于的类别是什么，这也对解决处理的工作人员带来的效率考验，良好的问题分类，能够使得更有针对性的获得帮助，从而提高解决问题的效率与质量；三是，仅靠人工在海量的留言中找到值得重点关注的问题是不切实际的，这样可能会导致平台成为摆设，小问题变成大漏洞。

因此为了更好地利用市民反映问题的平台，真正做到倾听民声、汇聚民智、凝聚民智，并需要依靠大数据与人工智能等技术的发展，基于自然语言处理的智慧政务系统是发展的新趋势与目标，它能够依靠人工智能等技术更有效的提升政府的管理水平与处理政务的效率。本文也将借助于大数据与人工智能等的综合应用，探究解决问题的可行性建议。

1.2 研究问题

结合题设要求，本文主要需要解决三大类问题：

问题一：群众留言分类。依靠人工进行分类存在耗时长、效率低等问题，因此本文在数据预处理的基础上，采用贝叶斯、TF-IDF、TEXTCNN 等多种模型进行比较，选取最优的分类模型，并给出分类结果。同时，为了优化分类情况，本文在分词等部分进行一定的人工干预，扩展修改停用词表、同义词表等，以便在实际应用中得到更好的结果。

问题二：热点问题挖掘：本文结合实际数据意义给出热点问题的评价标准，从三个维度来考核问题热度，并进行分类分主题两步走的热点探析，找到热点问题，并且输出对应的热点留言。

问题三：答复意见评价：良好的评价系统是保证反馈渠道价值的重要保障。本文从解决度、可解释性、完整性、相关性、及时性五个维度综合评价，构成完整有效的评价体系，切实符合市民的心理需求。

1.3 本文结构

本文主要由六个章节组成，这六章的结构安排如下：

第一章是介绍背景知识的部分，先介绍了问题的研究背景，总结了自然语言文本分类的发展和应用，最后提出了研究的问题和总结了本文的主要工作；

第二章针对群众留言分类，分别介绍了朴素贝叶斯理论、TF-IDF 模型和 TextCNN 模型，文本分类过程中预处理及算法性能验证方式，总结了针对朴素贝叶斯算法特征条件独立性假设所作出的特征加权的算法，与传统的 TF-IDF 和 TextCNN 模型相比，在分类准确度上有了明显提升。

第三章针对热点问题挖掘，在每一类下进行主题提取，将每一主题下的留言按照时间排序，以人气、共性、历史性三个角度来综合评分，最后提取出前五的热点问题以及对应的留言分布。

第四章介绍了传统的文本相似度的算法-余弦相似度，以及根据已有文献主观的从五个维度对答复意见的质量进行了综合评价，为今后的群众留言工作提供了有效意见。

第五章是最后的总结部分，总结全篇内容，对模型进行了详细评价，并对本文未来的研究重点和有待完善的部分做出了改善。

2 问题一：群众留言分类

2.1 问题分析

本题的目标是对群众留言分类，依靠人工进行分类存在耗时长、效率低等问题，因此本文在数据预处理的基础上，采用贝叶斯、TF-IDF、TextCNN 等多种模型进行比较，选取最优的分类模型，并给出分类结果。同时，为了优化分类情况，本文在分词等部分进行一定的人工干预，扩展修改停用词表、同义词表等，以便在实际应用中得到更好的结果。

本题的主要解题思路如下，具体分析以及实现过程结果会在之后的小节中陈述。

2.2 朴素贝叶斯

2.2.1 贝叶斯分类原理

当我们有样本（包含特征和类别）的时候，我们非常容易通过

$$p(x)p(y|x)=p(y)p(x|y) \tag{2-1}$$

统计得到 p(特征|类别) 即 p(类别)p(特征|类别)=p(特征)p(类别|特征)，有

$$p(\text{类别}|\text{特征})=\frac{p(\text{类别})p(\text{特征}|\text{类别})}{p(\text{特征})} \tag{2-2}$$

独立假设：特征往往是多维的，$p(features|class)=p(f_0,f_1,\cdots f_n|c)$，这里假设 2 维，有

$$p(f_0,f_1|c)=p(f_1|c)p(f_0|c) \tag{2-3}$$

假设特征之间是独立的（朴素贝叶斯的思想）

$$p(f_0,f_1,\cdots f_n|c)=\prod_i^n p(f_i|c) \tag{2-4}$$

贝叶斯分类器：对每个类别计算一个概率 $p(c_i)$，然后再计算所有特征的条件概率 $p(f_j \mid c_i)$，那么分类的时候我们就是依据贝叶斯找一个最可能的类别：

$$p(class_i \mid f_0, f_1, \cdots f_n \mid c) = \frac{p(class_i)}{p(f_0, f_1, \cdots f_n \mid c)} \prod_j^n p(f_j \mid c_i) \quad (2\text{-}5)$$

2.2.2 贝叶斯分类步骤

数据集说明：一条记录为“一段长留言详情”，lable 为“城乡建设”，“交通运输”，“卫生计生”等类别。

（1）分词：把每一条记录进行分词，保存为 content_S，所有的记录保存在 df_content；

（2）划分训练集和测试集；

（3)统计词频:统计所有训练集的每个词出现的次数,即词频,放入 all_words_dict 字典中，对词频进行降序排序，保存为 all_words_list；

（4）停用词表：载入停用词表 stopwords；

（5）文本特征提取：选取 $n=1000$ 个特征词（词频高到低依次选）保存在 feature_words（选取 all_words_list 中（all_words_list 中词频进行降序排序）排除出现在停用词表中的词，排除数字，并且要求词的长度为(1,5)）

（6）每条记录的表示（表示成长度为 1000 的特征词)：依次遍历 feature_words 中每个词，对于每次遍历的词 word，如果在当前需要表示的记录中，则该记录的这一位为 1，否则为 0。即 features = [1 if word in text_words else 0 for word in feature_words]，这里的 text_words 表示一条记录。

（7）训练，预测：使用 MultinomialNB 进行训练，预测。

2.3 TF-IDF 模型

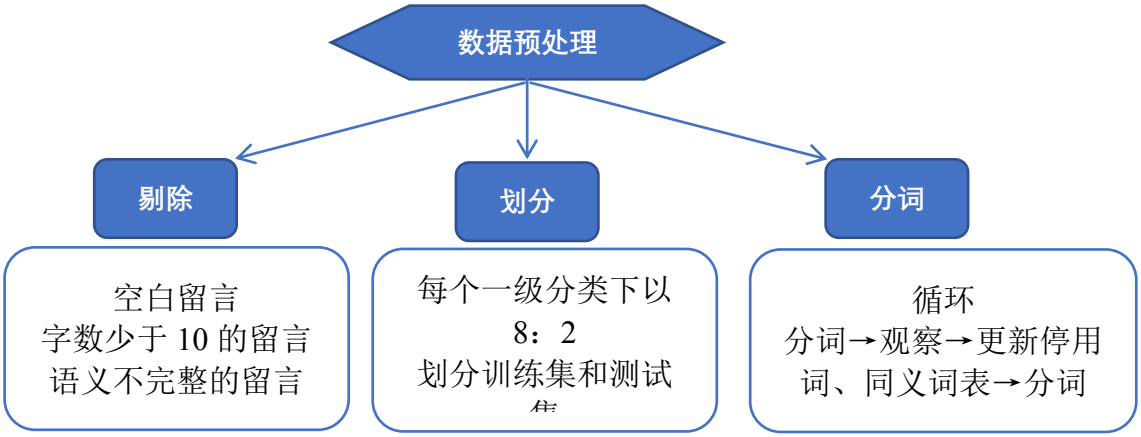
文本处理领域还有一种特征提取方法，叫做 TF-IDF 模型（term frequency－inverse document frequency，词频与逆向文件频率)。TF-IDF 是一种统计方法，用以评估某一字词对于一个文件集或一个语料库的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF 加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。

TF-IDF 的主要思想是，如果某个词或短语在一篇文章中出现的频率 TF（Term Frequency，词频），词频高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。TF-IDF 实际上是：TF * IDF。TF 表示词条在文档 d 中出现的频率。IDF（inverse document frequency，逆向文件频率）的主要思想是：如果包含词条 t 的文档越少，也就是 n 越小，IDF 越大，则说明词条 t 具有很好的类别区分能力。如果某一类文档 C 中包含词条 t 的文档数为 m，而其他类包含 t 的文档总数为 k，显然所有包含 t 的文档数 $n=m+k$，当 m 大的时候，n 也大，按照 IDF 公式得到的 IDF 的值会小，就说明该词条 t 类别区分能力不强。但是实际上，如果一个词条在一个类的文档中频繁出现，则说明该词条能够很好代表这个类的文本的特征，这样的词条应该给它们赋予较高的权重，并选来作为该类文本的特征词以区别与其他类文档。

2.4 TextCNN

2.4.1 数据预处理

对于大量的留言数据，可能存在质量较差的留言，例如，留言内容空白、留言内容信息不完整等，对于这一类留言直接做剔除处理。其次以 8：2 的比例随即划分训练集与测试集，并且由于这里关系到 15 个一级标签，所以这里的切割是在每一个标准下以 8：2 划分，详情代码见 first.py。随后，进行分词并优化分词结果。我们认为对于目前的人工智能而言，在前期阶段需要一定的人工干预来优化结果，因为停用词表和同义词表虽然范围广，但是针对性偏弱，因此不断分词查看结果，添加新的停用词与同义词表，从而得以优化。这个过程虽在前期较为繁琐，但当停用词表完善后，后期实际应用则会提高分类的准确率。



2.4.2 文本分类

1、前期准备

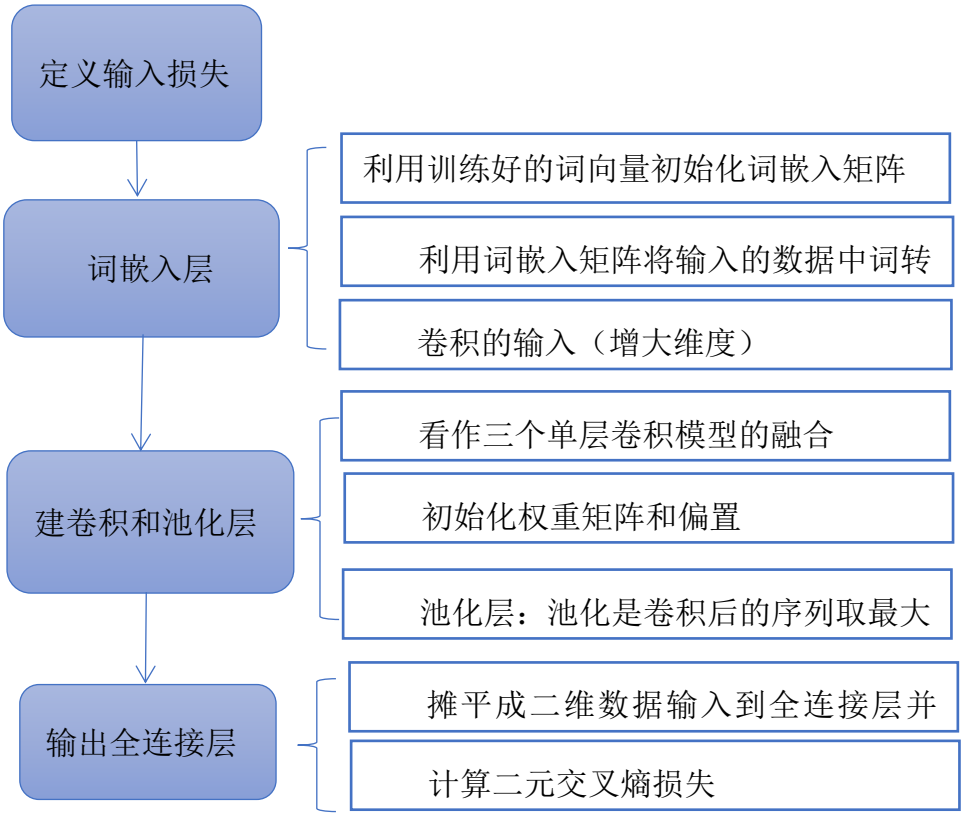
对划分后的训练集做简单处理，替代不可识别或多余的符号，保证文本的连续性与可读性，随后将留言详情提取出来分词后作为字典储存备用，即为训练好的词向量 word2vec，在这个过程中，将文本语言数值化，从而使得后续的处理中更简便。

读取数据集后，将标签与留言分词结果转化成索引储存，并且生成词向量和词汇-索引映射字典，该字典可是用于全部数据集。随后进行上文所提到的分词去停用词的循环处理，并且考虑到大量低频词的实际影响较低，去除低频词（这里定义词频小于 10 的为低频词）使得最后得到较好的分词结果。

分词后所得到的词汇对应从前期训练好的词向量中取出相关值，并且输出不存在于词向量中的词，观察是否存在一些关键词被排除，从而拓展词向量包。从而完成前期的参数设定与数据生成。

2、训练模型

TextCNN 具有网络结构的优势，引入训练好的词向量后能够得到不错的效果，也由于其网络结构比较简单，使得它具有较小的参数数量与计算量，从而能够在较短时间内完成模型的训练。



3、在构建模型的过程中，各类性能指标对应如下，具体计算公式设置详见代码：

表 2-1 变量的说明

变量	含义
Mean	计算列表中元素的平均值
Accuracy	计算二类和多类的准确率
binary_precision	计算二类的精确率
binary_recall	计算二类的召回率
binary_f_beta	计算二类的 f-beta 值
multi_precision	计算多类的精确率
multi_recall	计算多类的召回率
multi_f_beta	计算多类的 f-beta 值
get_binary_metrics	二分类的性能指标集合
get_multi_metrics	多分类的性能指标集合

2.5 结果展示

定义优化函数后，传入学习速率参数，从而计算梯度，应用到变量下生成训练器，随后初始化变量进行模型训练，最终经过词训练后，得到最终 TextCNN 模型性能如下：

表 2-2 优化模型性能情况表

指标	情况
Step	918
Loss	1.51
Acc	46.875%
Recall	0.18
Precision	14.34%
F-beta	0.156

表 2-3 不同模型的具体情况

	TRUE
朴素贝叶斯	"0.8257279"
TF-IDF 模型	"0.790525"
TextCNN	"0.46875"

这里整体看“朴素贝叶斯”的命中率最佳，然后利用其进行预测。调参之后也能达到较好的效果，从运行速度和预测正确率看朴素贝叶斯的结果更好。下面

将看不同类别下对准确率的影响。

利用训练好的模型对测试集进行预测，测试集实际情况与预测结果如下：

表 2-4 预测结果对比

类别	实际	预测结果
城乡建设	402	764
党务政务	0	0
国土资源	0	0
环境保护	188	45
纪检监察	0	0
交通运输	123	51
教育问题	318	275
经济管理	0	0
科技与信息产业	0	0
劳动与社会保障	394	491
民政	0	0
农村农业	0	0
商贸旅游	243	141
卫生计生	176	77
政法	0	0

可以看出，原本缺乏的列别无法进行分类，在城乡建设和劳动与社会保障方面过饱和，而环境保护、交通运输、教育问题、商贸旅游、卫生计生则都存在一定的误判。综合数据我们推测，主要是由两方面原因导致，一是数据集本身类别之间的样本量差异较大，部分类别样本数偏少，特征提取不明显，二是，分类之间的关键词较为接近，容易产生误判。

2.6 问题一结论

在当今数据大爆炸时代，每天所产生的文本量数以亿计，急需整理分类，然而传统的数据分类的文本处理方式过于烦琐，在浩瀚的数据流中迅速，高效，精确地找到需求信息极其困难。怎么有效地区分鉴别杂乱的信息，怎么迅速地满足用户的需求，都面临着困难。为了解决信息无序的问题，文本的自动分类技术自然成了处理和组织大量信息的一个重要技术。因此众多文本分类方法应运而生，朴素贝叶斯也是其中一种。朴素贝叶斯作为数据的十大算法之一，由于其易于构造和解释，并具有良好的性能，因此被广泛用于解决分类和排序问题。本文研究基于朴素贝叶斯算法的中文文本分类改进算法。

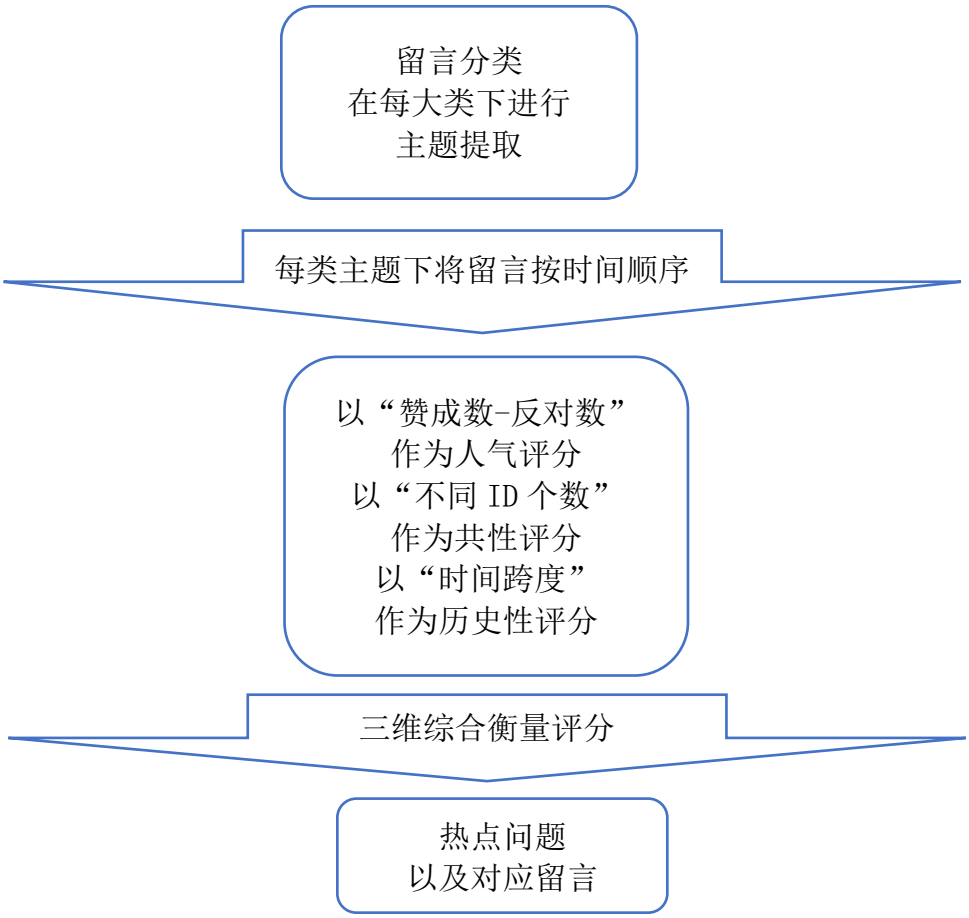
在本文中，我们首先介绍了现有的朴素贝叶斯分类方法、TF-IDF 模型和 TextCNN 模型。然后，我们通过改进 TF-IDF 加权方法，该方法通过对训练数据

的特征加权频率进行深度计算来估计朴素贝叶斯的条件概率。实验结果表明，与之前方法相比，我们的改进 TF-IDF 加权方法很少会降低模型的质量，而且在很多情况下，可以显著提高模型的质量。最后，我们对朴素贝叶斯中文文本分类器进行了改进 TF-IDF 加权，并取得了显著结果。

3 问题二：热点问题挖掘

3.1 问题分析

根据题目要求，将在某一时间段内重复出现的问题被称为“热点问题”。结合考虑实际问题，反复提出的问题可能存在两种情况，一是多名用户同时反映，二是一名用户反复放映，认为第一类的热度是高于第二类，其次问题本身的影响范围也是值得考虑的问题，个人问题还是集体问题在热度评分上应该有所差异，在此定义下，结合本题数据结构本文热点问题评选方法如下：



3.2 问题求解

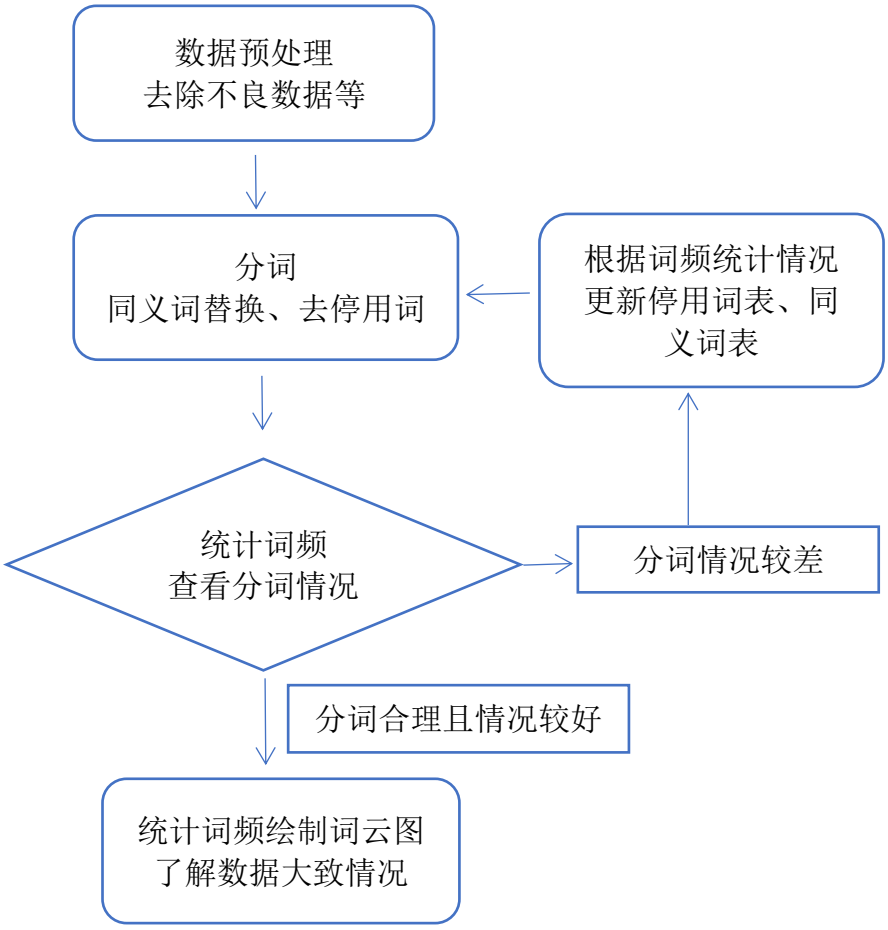
1、数据预处理和去除无效留言

考虑到实际问题中数据可能存在的问题，首先留言数据进行预处理，一是去除空留言，在留言主题或留言详情中出现空值的记录通过直接过滤法去除，二是去除无效虚假留言，考虑到实际情况，将留言详情字数小于 10 的留言看做无效留言一并处理；三是去除矛盾留言，即对比留言主题和留言详情关键词相似度，对于明显不相符的留言作为无效留言去除。

2、优化分词结果

将所给的文本内容转化为可进行挖掘的结构化信息，并且为了之后实际模型中更好的效果，本文针对分词结果进行优化。由于 jieba 包能够在扫描基于前缀词典的词图，并将所有可能组成词的情况形成有向无环图，从而找到最右的切分组和，选取 jieba 包并借助 HMM 模型的汉字成词能力更好的实现中文分词。

同时，为了得到更有效的词汇，本文进行人工干预，不断扩充修改停用词表以及同义词转化题库，以便实际应用中得到更好的效果。



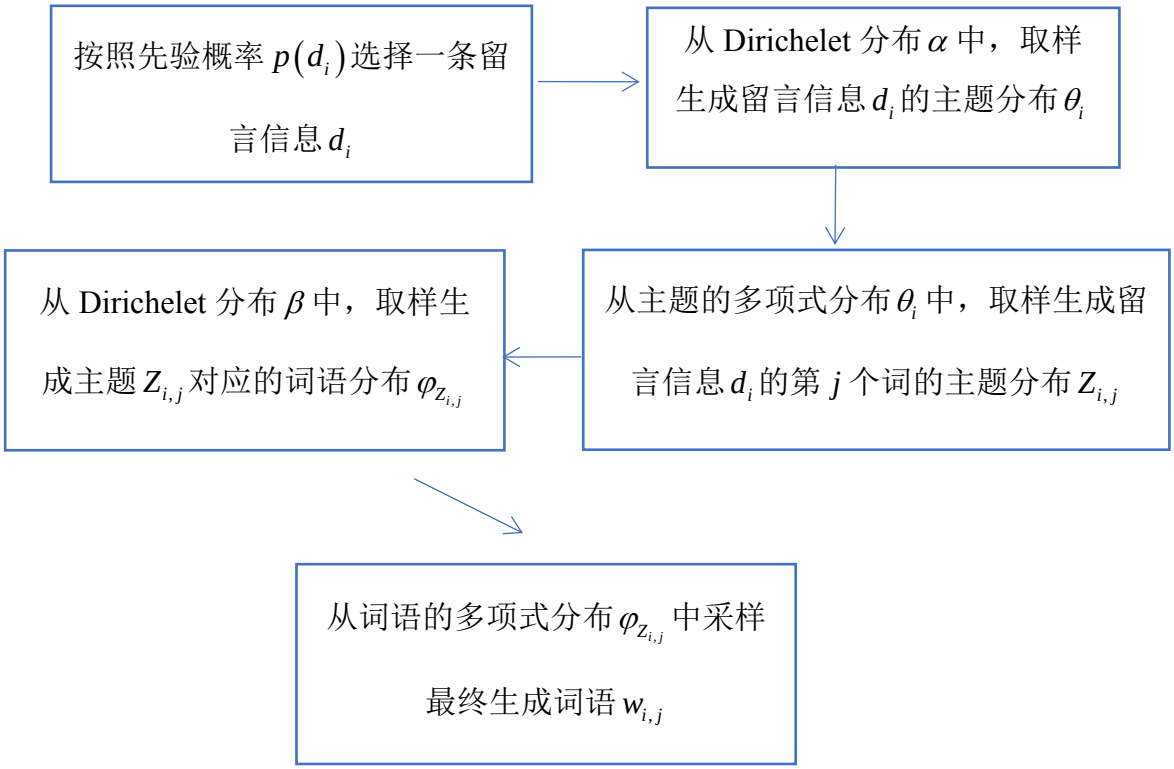
我们将问题二看做在分类的基础上进一步信息的提取，因此我们借助第一问训练好的模型先将问题 2 的数据信息进行分类，同时通过相似度与二三级标签进行拟合，考虑到数据类别的情况，在这里主要选取的是包括“城建建设”“交通

运输”“教育文体”等在内的七大类别进行分析，得到大致分类的结果如下：

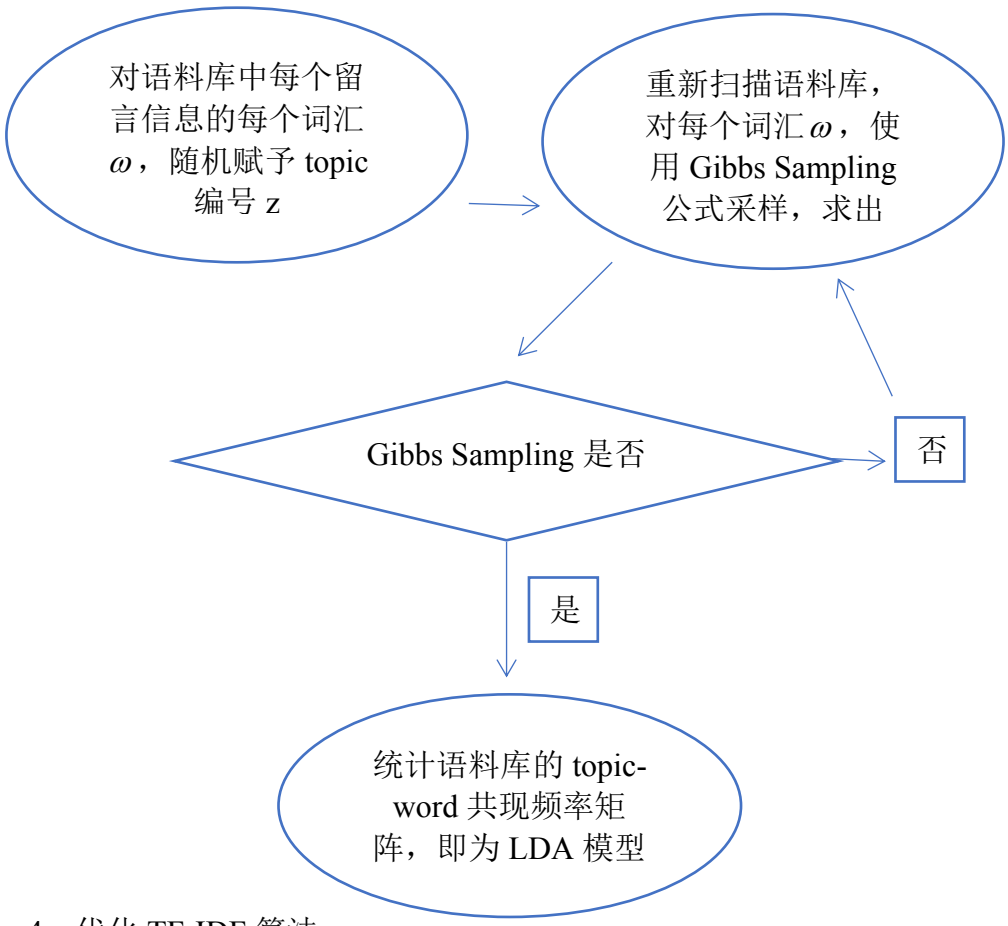
图 3-1 留言类别统计图

3、LDA 主题分类

由于本题中并不能够留言具体包括了多少主题，因此本文采取 LDA 模型来推测文档的主题分布，LDA 在 PLSA 的基础上加入了两个 Dirichlet 先验，其生成文档的过程如下：



在 LDA 建模的过程中，主要目标一是确定主题与词语相关参数，二是针对给定的文本进行主题分布，模型的理论训练过程如下：



4、优化 TF-IDF 算法

在这个寻找关键词的过程中，融入 TF-IDF 算法进行语料训练构建向量。

（1）计算词频（TF 权重）

词频=某个词在文本中出现的次数　　（3-1）

在这一过程中，为了充分利用好一级分类的作用，适当将二级、三级标签中所含有的词比重加大，由于留言存在个人表达习惯倾向，差异较大，因此进行标准化处理。

词频=该词在文本中出现次数÷改文本中出现次数最多的词出现次数（3-2）

（2）计算 IDF 权重

借助于二级、三级标签等，将留言与所属分类的环境语料相配合，当 IDF 值越大，即更具有代表性。

逆文档频率=log（语料库文本总数÷（包含该词的文本数-1））（3-3）

（3）计算 TF-IDF 值

TF-IDF=词频×逆文档频率　　（3-4）

选取部分数据进行模拟，发现 TF-IDF 的值与在留言中出现的次数存在正比

关系，TF-IDF 的值一定程度上可以刻画词的重要性，为 LDA 模型的关键词提供有力支持。

5、确认 LDA 主题数

正如上文所提到的，对于形形色色的留言信息，事先并无法知道分为多少主题更为合适，因此通过计算 perplecity（困惑度）来确认，计算公式为：

$$perplexity = e^{\frac{-\sum \log(p(w))}{N}} \tag{3-5}$$

其中，N 是测试集中词的总数，$p(w)$代表测试集中每个词的概率，在 LDA 模型中，其计算公式为：

$$p(w) = \sum z p(z \mid d) * p(w \mid z) \tag{3-6}$$

其中，z，d 代表训练过的主题和测试集的文档。

简单来说，对于每一条留言详情，LDA 模型判断它是属于某个主题的不确定性有多少，显然，当设定的主题越多，Perplexity 越小，但同时也容易出现过饱和的情况，因此构建循环，使 LDA 模型自我训练找到最合适的主题数，并通过绘制 Perplexity vs num of topics 曲线直观得到结论，在这里我们将数据按每个类别分类后进行主题提取，以环境保护、交通运输为例。

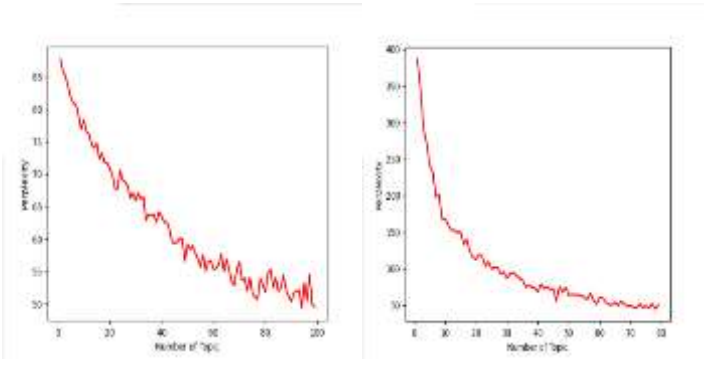


图 3-2 环境保护、交通运输困惑度曲线图

由图可见，对于环境保护类别选取 79 为主题数即可，而交通运输类别选取 46 个主题数最为合适。八种类别下主题数确定对比图如下：

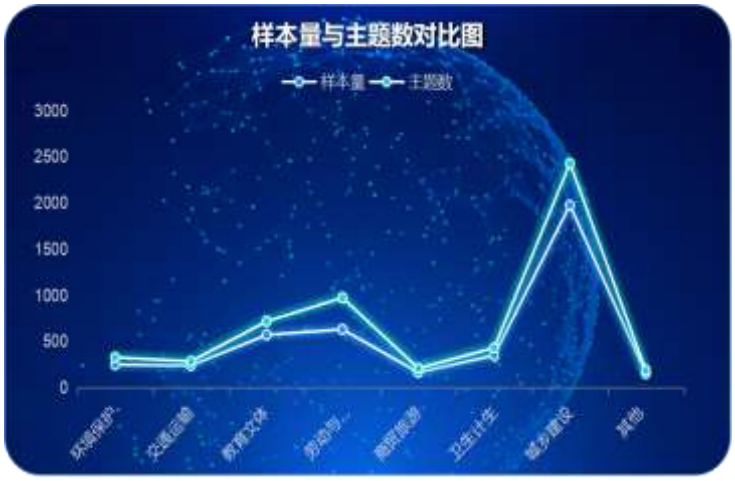


图 3-3 样本量与主题对比图

在每个类别下，进行 LDA 建模，并输出每个主题下前五个关键词，选取每个类别下数量较多的主题情况结合词频统计，进行进一步分析。

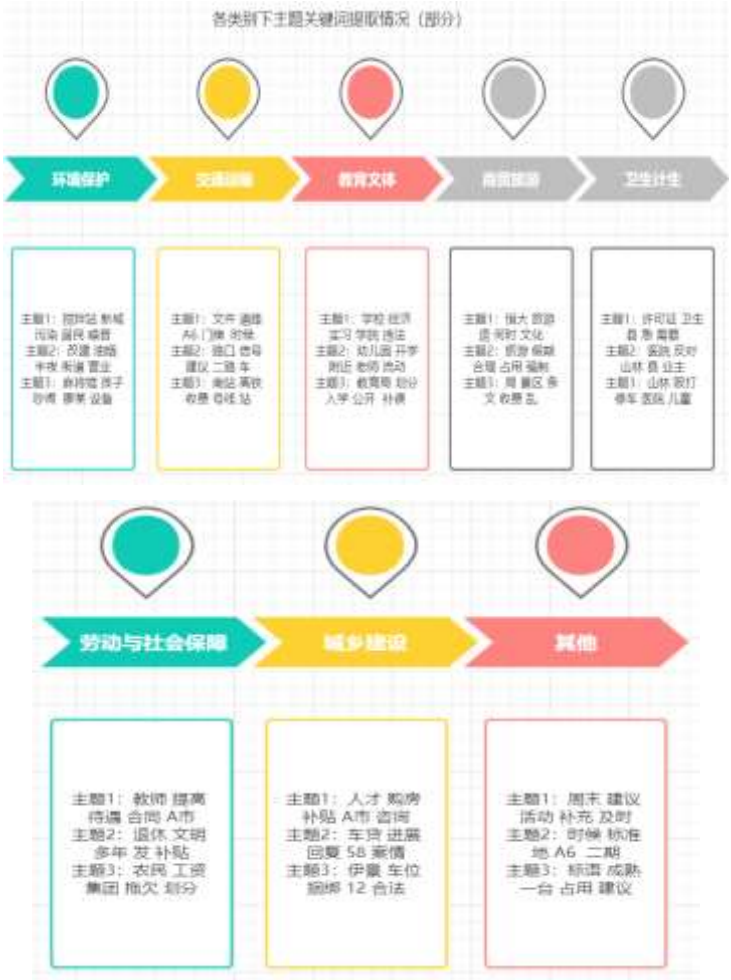


图 3-4 不同类下主题关键词提取

在每个类别下任意选取几个主题展示如上，可以看出虽仍存在一些指向性不明确的词，但大体分类效果还是比较好的，能够大致推测出留言相关内容，从而

可以作为关键词去寻找匹配留言。

6、探寻热点问题

将LDA建模中所提取出的关键词分别建立字典库，并且结合从而词频统计，为词赋予不同分值，从而与分词后的每条留言进行匹配，从而找到被提最多的问题，输出对应的留言情况。将对应的热点问题相关留言按时间排序，综合考虑上文所提到的三个维度进行评价计分，热度的评分思路如下：

图 3-5 热度评分流程图

其中历史评分代表的是问题影响时间长度，即未得到解决的遗留问题，分为三档，3 个月以内记作 1 分，6 个月以内记作 2 分，超过六个月记作 3 分。最终找到在本文所构建的评分方法下，5 个热点问题为：

表 3-1 热点问题详情表

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	4780	2019/1/11 至 2019/09/04	A 市/58 车贷受害人	A 市 58 车贷案件进展情况
2	2	210	2019/07/03 至 2020/01/16	丽发新城/居民	丽发新城附近的搅拌站，噪音污染、环境污染严重
3	3	84	2018/11/15 至 2019/9/27	A 市/待租房购房人群	A 市人才租房购房补贴问题
4	4	57	2019/7/21 至 2019/04/17	劳动东路魅力之城小区/居民	魅力之城小区一楼商业门面，噪音扰民严重，环境污染
5	5	23	2019/07/30 至 2019/08/07	伊景园滨河苑/购房者	A 市伊景园滨河苑捆绑销售车位

并给出五个热点问题三维分数的分布情况（由于数值差异较大，在作图中，统一对数化处理）。

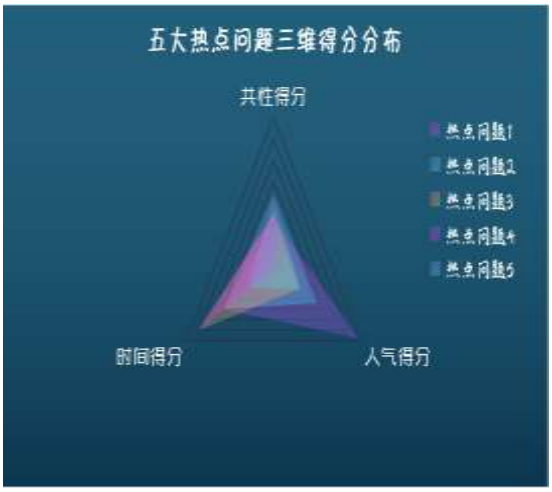


图 3-6 三维得分分布图

并将结果输出到热点问题表，对应具体留言信息按时间排序输出至热点问题明细表，详情可见附件。

4 问题三：答复意见评价

4.1 问题分析

在本题中，目标是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。因此本部分将分为如下几步进行构建。首先根据已有的参考文献，选取五个角度（相关性、及时性、完整性、可解释性、解决度）进行评价，使其尽可能全面准确地反映答复意见；其次用余弦相似度来计算答复意见和留言详情之间的相似度，以防答非所问；最后群众在网上的留言应及时回复，至少给群众指明反映问题的方向，对答复意见的及时性进行了详细的分析，让相关部门工作更加贴近群众。

4.2 文本相似度

在生活中，信息检索、文档复制检测等领域都应用到“文本相似度”。它的应用度更广，除了文字的匹配，还可以是图片，音频等，因为实质都是在计算机中都是以二进制的方式存在的。 相似度，实质就是计算个体间相程度。什么是个体？对于语句，个体就是语句，对于图片，个体就是图片。本文采用余弦相似度计算不同文本之间的相似度。

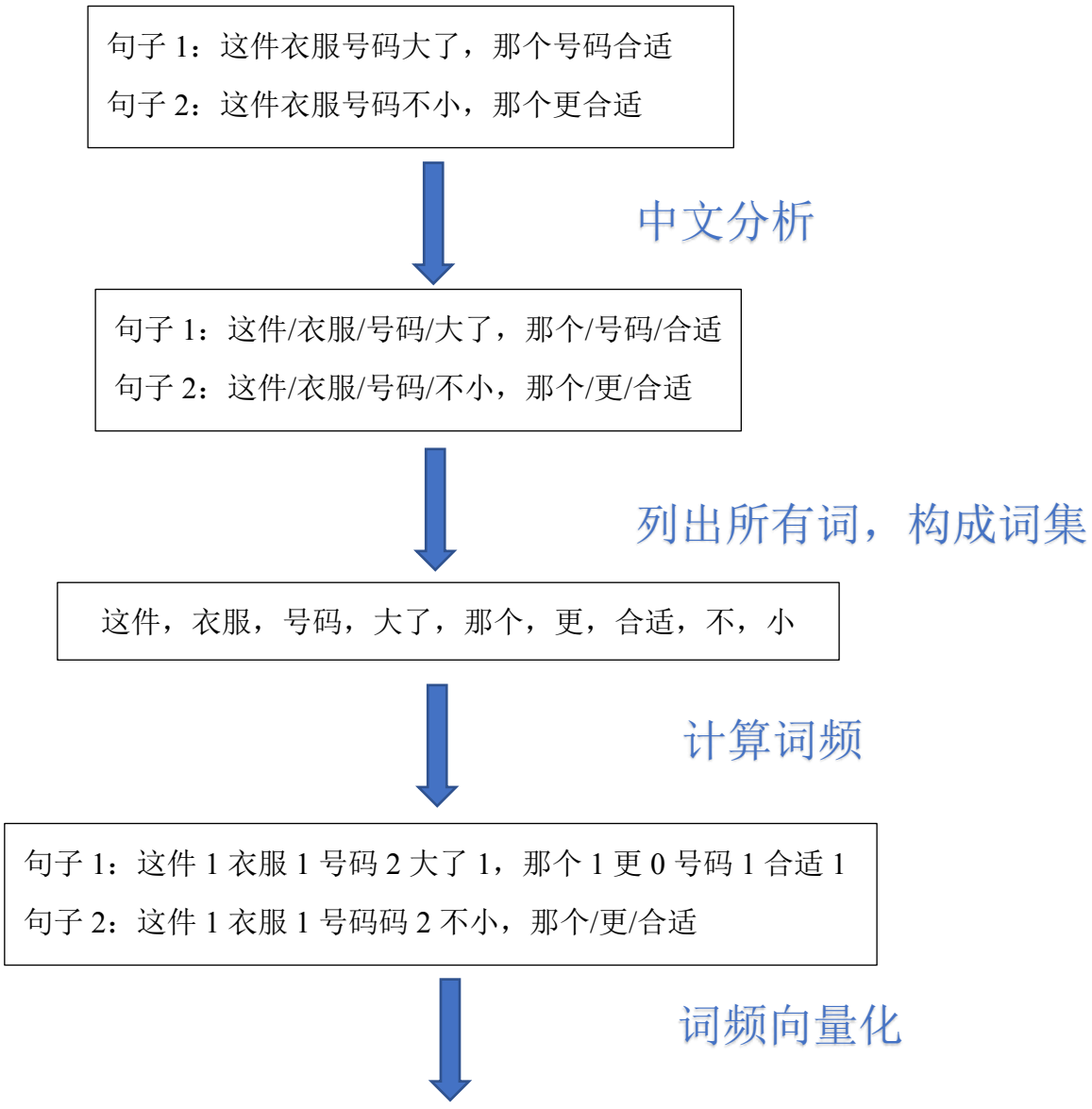
余弦相似度就是通过一个向量空间中两个向量夹角的余弦值作为衡量两个个体之间差异的大小。把 1 设为相同，0 设为不同，那么相似度的值就是在 0~1

之间，所有的事物的相似度范围都应该是 0~1，如果不是 0~1 的话，就不是我们应该研究的事了，那是神经学家和生物学家的事了。余弦相似度的特点是余弦值接近 1，夹角趋于 0，表明两个向量越相似。 即三角形越扁平，证明两个个体间的距离越小，相似度越大；反之，相似度越小。

由图 4-1 可知，两个句子的相似度计算的步骤是：

通过中文分词，把完整的句子根据分词算法分为独立的词集合；求出两个词集合的并集（词包）；计算各自词集的词频并把词频向量化；带入向量计算模型就可以求出文本相似度。注意，词包确定之后，词的顺序是不能修改的，不然会影响到向量的变化。

以上是对两个句子做相似度计算，如果是对两篇文章做相似度计算，步骤如下：找出各自文章的关键词并合成一个词集合；求出两个词集合的并集；计算各自词集的词频并把词频向量化；带入向量计算模型就可以求出文本相似度。



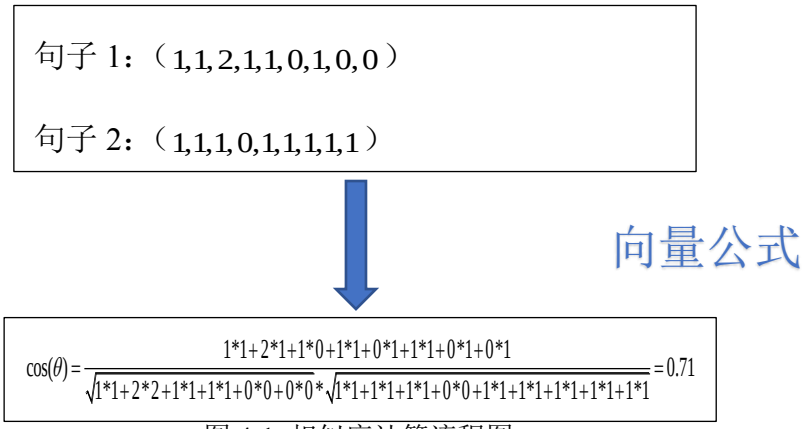


图 4-1 相似度计算流程图

4.3 问题求解

4.3.1 描述性统计分析

采用了比赛中附件 4 的数据集，该数据集一共包括了 2816 条数据，采取了七个变量，变量说明如表 4-1 所示：

表 4-1 数据集变量说明

变量名	解释	详细说明	取值范围
分类	留言类型	共 8 个水平	城乡建设、卫生计生等类型
解决度	留言的否解决	定性变量	0 表示未解决；1 表示已解决
可解释性	答复意见中是否有法律文件的解释	定性变量	0 表示有；1 表示没有
完整性	答复意见的标准化回复	定性变量	0 表示有；1 表示没有
相关性	答复意见和留言之间的联系		0-1
及时性	答复时间和留言时间的差	定性变量	High；Medium；Low

（1）根据每一个特征的数值情况，可以将不少特征因子化，方便后期做不同类别的差异分析。

（2）针对答复意见可解释性和完整性，各单位对网络留言提出的问题，能够一次性解决和回复的，要一次性解决到位，不得要求网民再通过其它方式进行咨询或诉求；对作出承诺解决的，解决后要及时予以回复； 对情况不清或难以回复的，要说明情况，予以解释。在回复语言表述上，要确保文字规范精准，把好政策关、法律关、文字关，避免出现政策、法律适用不当等问题。

（3）针对相关性，先余弦相似度计算了答复意见和留言详情的文本相似度，看二者是否有相关关系，因为有的工作人员会敷衍了事，会大大降低留言者对组织的信任也不利于实际问题的解决，其次对数值进行了归一化处理，将其映射到了 0-1 之间。

4.3.2 答复意见评价模型的构建

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、及时性、解决度五个角度对答复意见的质量给出一套评价方案，逐个分析并得到每个类别综合评价分数来判断答复意见的质量和效率。同时，为了更好的分析不同类别问题的解决度和相关性等方面，延用问题一的分类标准以及社会上热点问题，主观将文本分为 8 类，并把城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游和卫生计生分别记为 1~7。

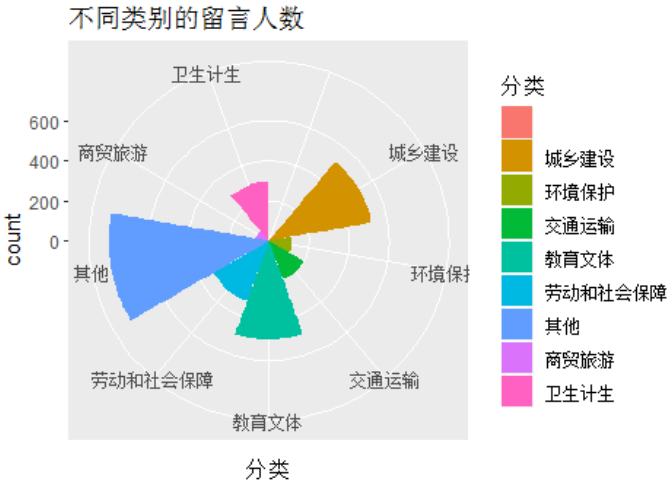
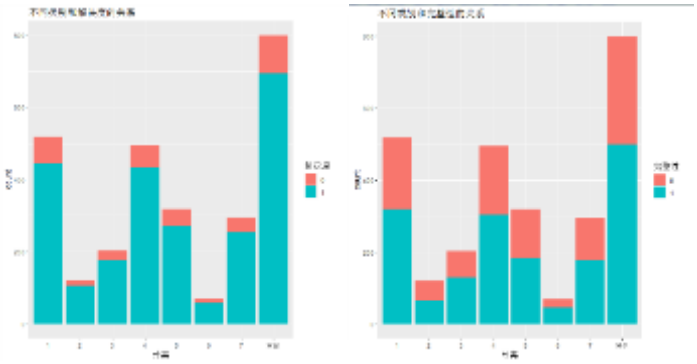


图 4-2 不同类别的留言人数

首先观察不同类别的留言人数。教育文体、城乡建设和其他类别占据人数排行榜前三，其余类别的人数分布较为均匀。



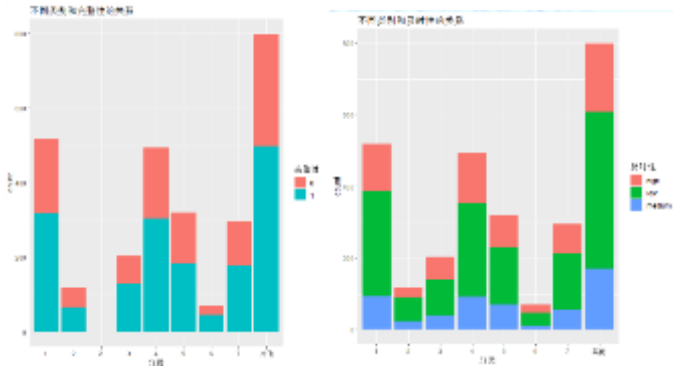


图 4-3 不同类别的完整性、可解释性、及时性和解决度

（1）现在社会越来越“高度重视留言者”的要求，对留言采取分级负责的办法进行分类处理和分别反馈，并建立了协调落实的工作机制。对民意的重视不能停留在口头上，而要认真吸纳其提出的好的建议，切实解决他们反映的一些实际问题；

（2）社会是一个大集体遇到的问题形形色色，不同的问题有不同的处理方式，不是所有的问题都能用法律形式解释，当然在相关人员在答复留言时候也不可能给出详细的法律解释；

（3）之前相关部门就做出了关于规范网上留言办理工作的通知，为进一步提高网上留言办理工作的质量和效率，做到件件有回音，事事查清楚，实现留言办理工作常态化、规范化、制度化；

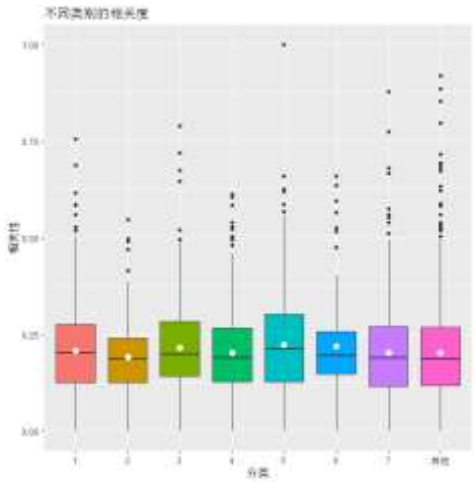


图 4-4 不同类别的相关程度

首先对之前相关系数图中，相关性较高的满意度进行探究，从箱状图来看，黑线和白点分别代表不同类相关程度的中位数和均值，可以看出不同类的相关程度的分布大体相当，相较而言，卫生计生类的留言意见和答复意见之间的相关性

偏低，其他的几个类别均值和中位数都没有明显差别。

然后对于变量相关性、完整性、可解释性、及时性我们分别用核密度曲线和柱形图来探查其与解决度关系。

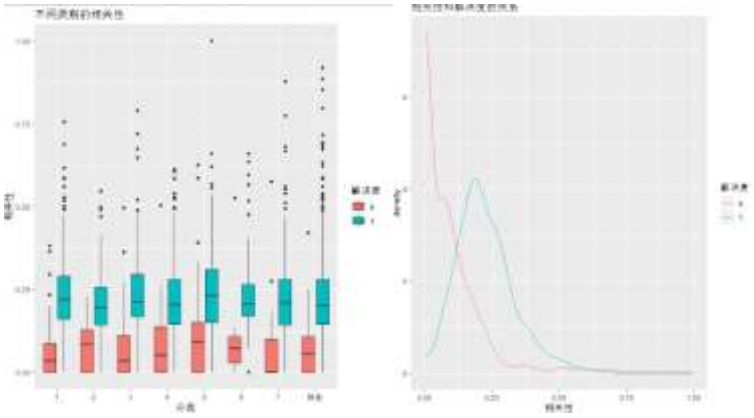


图 4-5 不同类别相关性和解决度

内容越相关，说明问题能更好的解决；绩效评估 0.5 和 0.9 左右的解决度比较高。

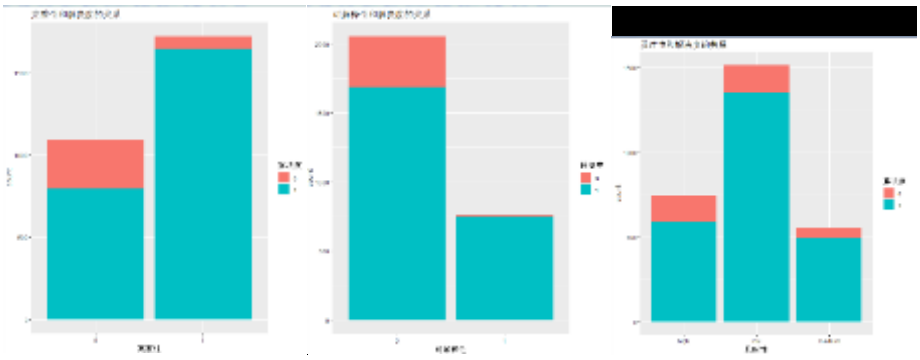


图 4-6 及时性、完整性、可解释性和解决度的关系图

（1）留言详情越完整，工作人员给出的答复意见就越全面，解决度越高，因为只有明白留言者的需求才能更好的解决问题，在此后分析问题未解决的原因过程中，留言信息的不全面是问题难以解决的重要因素。

（2）留言内容是否有法律等相关内容的解释不影响解决度，二者并没有太大的关系。

（3）理论上答复意见越及时，问题解决的越快，反之从上图发现结果却与之相反，说明答复不仅要速度还要质量，效率才是最重要的，因为从答复意见来看回复的及时并不能保证其内涵。

最后通过扇形图来观察留言内容未得到解决和准确答复的原因，表示如下：

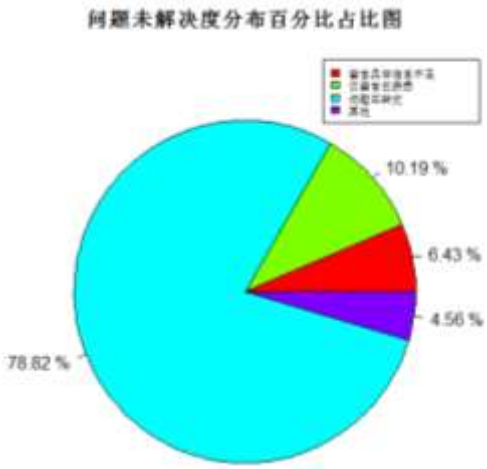


图 4-7 问题未解决的百分比图

从图 4-7 可以看出，在问题未解决度百分比占比图中，占据最大比例的是因为留言者提供的信息和描述问题不清晰，让工作人员无法和其以及相关单位取得联系，进而不能为其提供意见和解决措施，这也说明了在今后的工作生活当中，无论是提出建议还是给予投诉，都应该先清晰明了的表达自己的需求，双方才能更好的交流。

4.3.3 答复及时性

首先根据留言时间的数据，分解成留言日期和留言时刻，由此分析日颗粒度的忙闲时段和小时颗粒度的忙闲时段。

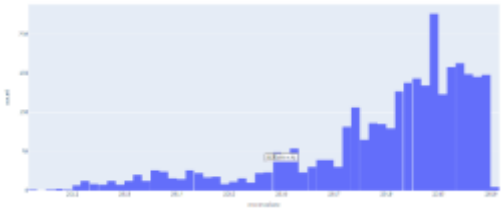


图 4-8 留言热线的时间段

数据的时间范围是 2018.10.2-2019.2.15。留言数量每年呈明显性变化，周末少工作日多。2012 年前数据样本少，后面几天的数据相比其他周也略少，可能系统升级或者坏了吧！最近四年留言的数据急剧上升，因为如今各种社交媒体越来越受到大家和政府的关注，为群众和政府沟通交流搭建了平台，也成为了网络问政新利器，为公民求助、投诉提供了便捷的渠道，这样既调动了群众参与民主管理的积极性，也扩大了政府、社区等社会管理的职能，发现问题才能更好的解决问题。

然后，看分钟颗粒度的留言数量以便更直观的看出留言的热线时间段。

图 4-9 留言热线时间段

留言的上午热门时间段是 10:00-11:00 左右，下午的热门时间段是 3:30-4:30 左右。

接着，查看留言的处理时间即看看一个问题大概需要多长时间解决，将留言时间减去答复时间，得出处理用时，以 10 分钟为 1 个单位。

图 4-10 留言处理时间

可以看到前台工单暂存处理时间会比较快，1 个小时内就能解决掉大半。剩下比较显眼的是绿色选手，时间分布比较平均，让我们仔细看看其他朋友哪个最慢。

最后，对群众的留言详情进行一波分词分析。

图 4-11 留言详情的词频图

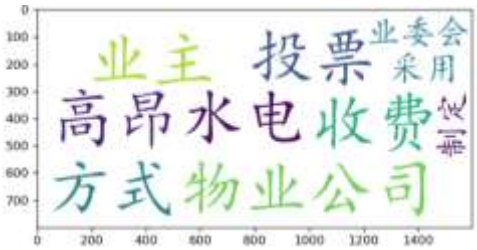


图 4-12 留言详情的词云图

从图 4-11 和图 4-12 中，可得到一下几点信息：

现在，随着大环境下经济条件的改善，相对于过去，生活水平逐渐提高，却发现生活成本却越来越高了。这一切，都体现在衣食住行上。空气，水，天然气，煤，石油等等，这些资源本身就是有限的，而我们生活又离不开这些。平常用的日常用品，食品，医药的制造原料都离不开这些。随着这些资源的消耗，越用越少，自然而然的，生活成本会增加。譬如鱼离开水，就会死亡。水越少，鱼也会越少，要活着，就必须去和更多的鱼，竞争有限的水源。如此下去，相应的不满和矛盾也增加，这就需要各行各业人员携手一方能合理提出意见另一方及时有水平给出方案才能大事化小小事化无，降低摩擦。

4.4 综合评价结果

本小结从五个维度分析了答复意见的质量，为了更好的分析不同类别的质量，将解决度、及时性、相关性、可解释性和完整性赋予不同的权重，即 0.5、0.2、0.1、0.1、0.1 计算出 8 类的综合评价分数。

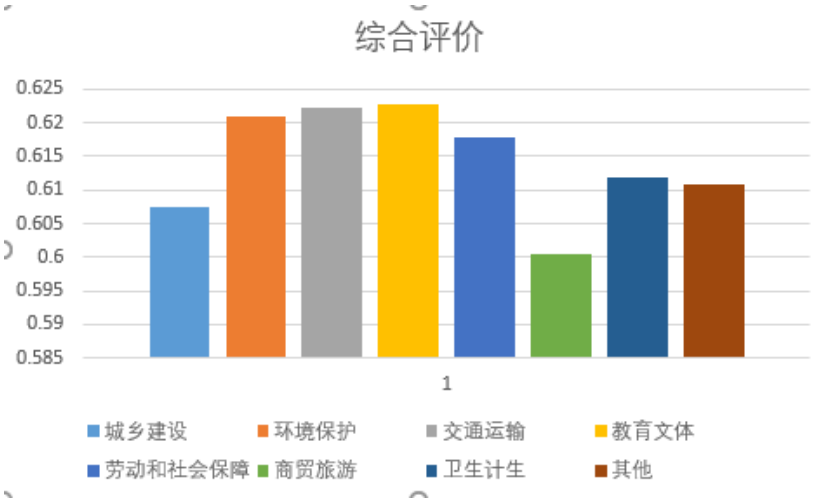


图 4-13 综合评价图

（1）网络是个相当开放的平台，留言群众来自不同的群体，在那里表达诉求，大多是发自内心的真话、直话，有时也免不了有不中听的话，对此，倾听者既要引导，也要有耐性和宽阔的胸襟，善于设身处地体谅对方，把群众的事当作自己的事，把群众的难当做自己的难，真诚以待，听进耳中，装入心头。

（2）要在回应上多多努力。对网民合理的诉求和提出的问题，除了及时回答外，能采纳的应采纳，并尽可能体现到相关政策中去；该办又能办的事，就应及时办理或者限时办好，不能搪塞推诿。要树立网民诉求无小事的观念，真心诚意纳谏

言，实实在在解难题，扎扎实实办实事。对于重要的网络舆情，则要以强烈的政治责任感，敏锐把握，积极因应，做到发现得早、处理得好、化解得了，真正让网民满意，使工作受益。

5 总结与展望

5.1 模型评价

本文利用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，在经过数据清理、分词、修正后，依据要求利用自然语言处理和文本挖掘的方法对群众留言分类、热点问题挖掘和答复意见的评价三个问题进行建模和求解，并得到良好的结果。对提升政府的管理水平和施政效率具有极大的推动作用。

针对问题一，本文通过群众留言和一定的划分体系对留言进行分类，将群众留言分派至相应的职能部门处理。在数据处理时先对数据进行预处理和清洗，采用贝叶斯、TF-IDF、TEXTCNN 等多种模型进行比较，选取最优模型，并给出分类结果。同时，为了优化分类情况，在分词等部分进行一定的人工干预，扩展修改停用词表、同义词表等， 将预测的误差降到最低。

针对问题二，要将在某一时间段内重复出现的问题提取出来，结合考虑实际问题，利用 LDA 模型中结合词频统计，为词赋予不同分值，与分词后的每条留言进行匹配，从而找到被提最多的问题，输出对应的留言情况。同时将对应的热点问题相关留言以三个维度进行评价计分最终得到结果。

针对问题三，本文首先选取了相关性、及时性等在内的 5 个角度，经过余弦相似度、描述性分析等，对原指标的解释能力达到了好的高度。随后在问题是否解决的前提下，分别研究了不同类别下相关性、完整性、可解释性和及时性的关联和关系，每个维度都其有各自的优缺点最后将不同维度赋予不同权重得到综合评价分数，在具体实践后发现，教育文体、交通运输和环境保护问题还是群众关注为止最多的方面，具有较高的有效性和代表性。

5.2 模型改进

本文在整个建模的过程中，在以下几个环节仍存在改进空间，一是在数据的预处理上由于为了数据的精确度经过多轮循环，可能在面对更大的数据集时会

需消耗较高的时间成本，二是对于群众留言分类，虽在分类方法上尝试多种方法对比的形式，但“朴素贝叶斯+TF-TDF”的大体框架已经有了较广泛的应用，创新度有待进一步提高。三是在答复意见评价模型的构建上，根据已有参考文献，设立的指标带有主观色彩，事实数据和理论的严谨上有待突破。

参考文献

[1] 姜天宇，王苏，徐伟．基于朴素贝叶斯的中文文本分类[J] ．电脑知识与技术，2019，15(23) ：253-254+263．

[2] 许甜华，吴明礼．一种基于 TF-IDF 的朴素贝叶斯算法改进[J] ．计算机技术与发展，2020，30(02) ：75-79．

[3] 邓志远．基于自然语言处理的电信系统热点问题的提取[J] ．信息技术与信息化，2020（1） ：31-33．

[4] 刘春磊，武佳琪，檀亚宁.基于 TextCNN 的用户评论情感极性判别[J] ．电子世界，2019(03) ：48+50．

[5] 康雁，杨其越，王沛尧．基于主题相似性聚类的自适应文本分类[J].计算机工程，2020，46(03) ：93-98．

[6] 刘春磊，梁瑞斯，邸元浩．基于 TFIDF 和梯度提升决策树的短文本分类研究[J] ．科技风，2019(24) ：231．

[7] 王仲远，程健鹏，王海勋，文继荣．短文本理解研究[J] ．计算机研究与发展，2016，53(02) ：262-269．

[8] 李颢，张吉皓．基于文本挖掘技术的客服投诉工单自动分类探讨[J] ．移动通信，2017，41(23) ：66-72．