基于神经网络、主题模型和层次分析的留言文本分析

摘要

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，在互联网的各类社情民意文本数据量与日俱增的大背景下，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。与此同时，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。本文将基于数据挖掘技术对分类标签、留言以及部门答复内容等文本数据信息的挖掘和分析。

在本次数据挖掘过程中，我们首先对获取得到的评论数据利用python工具进行数据预处理、分词以及停用词过滤操作，实现了对留言数据的优化，并提升了其可建模度。

接着，采用多种方法来进行数据挖掘模型的构建，为后面的留言分析构建分析的基础。为此，我们先利用深度学习的办法，通过构建神经网络模型对留言文本进行一级标签分类，采用5折交叉验证法对模型进行验证，该分类模型得分F_1=90.34%。其次，利用LDA主题模型的思想，结合统计学的角度实现留言主题归类模型的构建，并给出热度指数的定义公式，得到各类热点问题。再有，构建TF-IDF模型与正则表达式，量化答复的相关性、完整性、可解释性这三个指标，通过层次分析法评价留言答复质量。

最后，运用构造出来的多种数据挖掘模型的结果，对这些留言数据进行多方面多角度的文本分析，以提取其中隐藏信息。神经网络模型被用以进行词语之间的相似性分析；LDA主题模型则提取出了从统计学角度上的不同类型留言的热点，以了解群众最关心的问题。层次分析法模型从专家获得三个指标的相对重要程度，一定程度上得到相关部门答复质量的评价方案。

关键词： 文本分析；Python；神经网络；5折交叉验证；LDA；TF-IDF模型

Abstract

In recent years, with the online questioning platforms such as WeChat, Weibo, mayor's mailbox, and sunshine hotline, it has gradually become an important channel for the government to understand public opinion, gather people's wisdom, and gather people's popularity. In the past, it has brought great challenges to the work of relevant departments that used to manually divide messages and sort hot spots. 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 has become a new trend in the development of social governance innovation, which is extremely important for improving the government's management level and governance efficiency. This article will be based on data mining technology to mine and analyze text data such as classification labels, messages and department replies.

In this data mining process, we first used the python tool for data preprocessing, word segmentation, and stopword filtering operations on the obtained comment data to optimize the message data and improve its modelability.

Then, a variety of methods are used to construct the data mining model, and the basis for the analysis of the subsequent message analysis is constructed. To this end, we first use the method of deep learning to classify the message text by constructing a neural network model, and use five cross-validation methods to verify the model. The classification model score F_1=90.34%. Secondly, the idea of LDA topic model is combined with statistical perspective to realize the construction of message topic classification model, and the definition formula of heat index is given to get various hot issues. Furthermore, the TF-IDF Model and regular expressions are constructed to quantify the three indicators of relevance, completeness, and interpretability of responses, and the quality of message responses is evaluated by AHP.

Finally, using the results of the constructed multiple data mining models, a multi-faceted and multi-angle text analysis is performed on these message data to extract hidden information. The neural network model is used to analyze the similarity between words; the LDA topic model extracts the hotspots of different types of messages from a statistical perspective to understand the issues that the people are most concerned about. The analytic hierarchy process model obtains the relative importance of the three indicators from the experts, and to a certain extent, the evaluation plan of the response quality of the relevant departments.

Key Words: text analysis; Python; neural network; 5-fold cross-validation; Dirichlet Allocation(LDA); TF-IDF Model

目录

1 挖掘目标

本次建模针对互联网公开来源的群众问政留言记录的文本评论数据，在对文本进行基本的机器预处理、中文分词、停用词过滤后，通过建立LDA主题模型和TF-IDF模型等多种数据挖掘模型，实现对文本群众留言信息分类评价、热点问题挖掘以及给出答复意见的质量评价方案。

2 分析方法与过程

2.1 总体流程

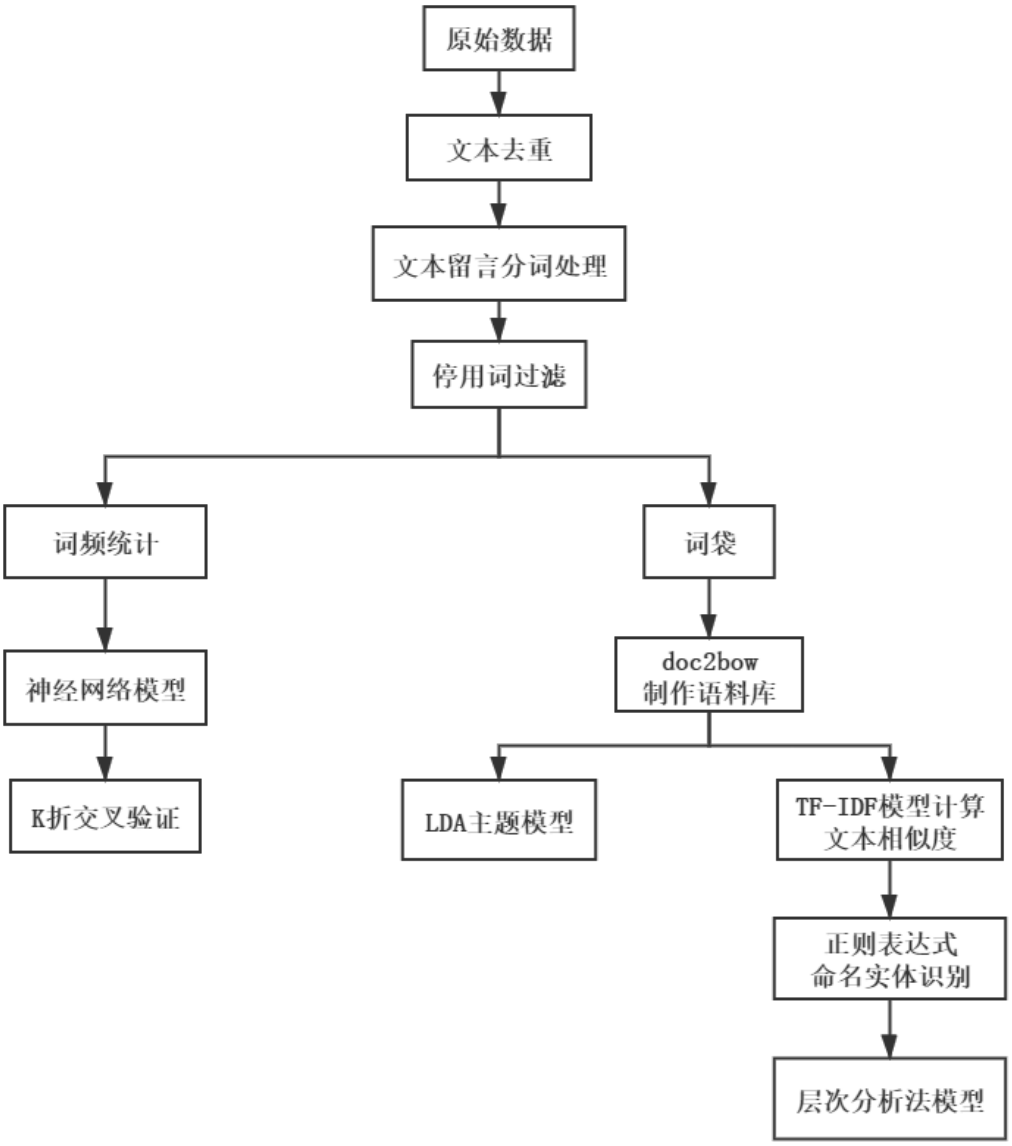


图 1: 总体流程

本论文的分析流程可大致分为以下四步：

(1) 读取分析用的原始数据（文本留言主题、留言详情等）；

(2) 数据预处理：文本去重、留言分词处理、过滤停用词等；

(3) 文本留言及回复经过处理后，运用多种方法对数据进行多方面分析；

(4) 获取文本留言数据中有价值的内容，建立模型并检验模型的合理性。

2.2 具体步骤

2.2.1 数据介绍

本文使用的实验数据为三级分类标签、互联网公开来源的群众问政留言记录以及相关部门对部分群众留言的答复意见，均来自附件所给数据。

2.2.2 数据预处理

2.2.2.1 文本去重

对于获得的文本数据后，我们首先要进行文本留言数据的预处理。考虑到文本数据里可能存在重复的条目，如果将这些留言数据也引入进行分词、词频统计等，则必然会对分析造成很大的影响，得到的结果的质量也必然存在问题。那么在利用文本数据之前必须先进性文本去重处理，把一些重复的文本去除。

文本去重，顾名思义，就是去除文本留言数据中重复的部分。无论获得什么样的文本留言数据，首先要进行的预处理应当都是文本去重。使用Pycharm软件进行去重，结果显示并无重复文本数据。

2.2.2.2 文本留言分词

中文的语义中，只有字、句和段落能够通过明显的分界符进行简单的划界，而对于“词”和“词组”来说，它们的边界模糊，没有一个形式上的分界符。因此，进行中文文本挖掘时，首先应对文本分词，即把连续的字序列按照一定的规范重新组合成词序列的过程。

分词结果的准确性对后续文本挖掘算法有着不可忽视的影响，如果分词效果不佳，会导致后续算法达不到理想效果。例如在特征选择的过程中，不同的分词效果，将直接影响词语在文本中的重要性，从而影响特征的选择。

本文采用python结巴分词库（jieba），对附件的excel文档中的留言主题、留言详情、留言回复等数据进行中文分词处理。jieba库优点在于支持繁体分词、支持自定义词典等功能。

本文使用jieba库实现了中文文本分词。部分示例结果如下：

- A市西湖建筑集团占道施工有安全隐患

  'A'，'市'，'西湖'，'建筑'，'集团'，'占'，'道'，'施工'，'有'，'安全隐患'

- A3区梅溪湖壹号御湾业主用水难

  'A3'，'区梅'，'溪湖'，'壹号'，'御湾'，'业主'，'用水'，'难'

- 地铁5号线施工导致A市锦楚国际星城小区三期一个月停电10来次

  '地铁'，'5'，'号线'，'施工'，'导致'，'A'，'市锦楚'，'国际'，'星城'，'小区'，'三期'，'一个月'，'停电'

- 请加快A6区时代倾城小区小区管理监督力度

  '请'，'加快'，'A6'，'区'，'时代'，'倾城'，'小区'，'小区'，'管理'，'监督'，'力度'

- 长株潭城铁的运行时间有点不合理

  '长株'，'潭'，'城铁'，'的'，'运行'，'时间'，'有点'，'不合理'

2.2.2.3 停用词过滤

经过中文分词处理这一步骤，将初始的留言文本处理成为词的集合，即为$d=(\omega_1;\omega_2;\ldots;\omega_N)$，其中$N$为文本$d$中出现词语的个数。但是文本中含有对文本含义表达无意义的词语，应进行删除，以消除它们对文本挖掘工作的不良影响，此类词称为停用词。停用词的两个特征为：一是及其普遍、出现频率高；二是包含信息量低，对文本标识无意义。例如中文的“了”、“的”、“地”、“啊”等，标点符号“—”、“？”、“。”、“、”等，在特征选取的过程中，停用词的介入可能会造成选出的特征几乎都是停用词，从而影响结果的分析。但是在停用词的去除中，应注意要保留其中的否定词，可以对停用词表进行人工筛选相结合的方式，对停用词进行处理。

文中采用基于停用词表与自定义词典的文本停用词过滤方式，将分词结果和停用词表与自定义词典中的词语进行匹配，若匹配成功，则进行删除处理。示例结果示例如下：

- A市—C市往返上班的群众呼声

  'A'， '市'，'C'，'市'，'往返'，'上班'，'群众'，'呼声'

2.2.3 词频统计

本文随机选取80%数据作为训练集，20%数据作为测试集。利用文本分类中文档的表示方法，提出了词频统计方法，实现了关键词的高效匹配。

文本表示成计算机可以处理的结构化信息的过程在文本分类的整个过程中称为文本预处理，现在最为广泛被采用的使用方法是向量空间模型表示。即用一个向量来表示一个文本的信息，使得文本成为特征空间中的一个点。在向量空间模型中文本集合形成一个矩阵，也就是特征空间中点的集合。词频矩阵就是应用向量空间模型表示文本的一种形式，其表示方法如下表所示：

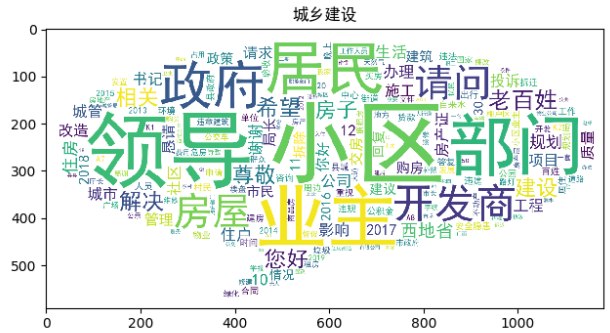
表 1: 词频矩阵

	$Word_1$	$Word_2$	$Word_3$	$\cdots$	$Word_n$
$File_1$	a_{11}	a_{21}	a_{31}	$\cdots$	a_{n1}
$File_2$	a_{12}	a_{22}	a_{32}	$\cdots$	a_{n2}
$File_3$	a_{13}	a_{23}	a_{33}	$\cdots$	a_{n3}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$File_m$	a_{1m}	a_{2m}	a_{3m}	$\cdots$	a_{nm}

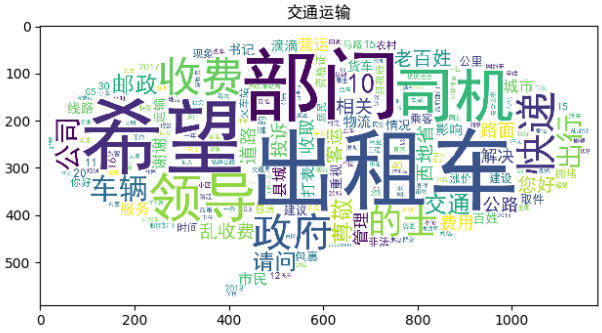
在词频矩阵中，$Word_i$是向量空间模型中的特征向量，a_{ij}是项的权重，通常是指第i个词在第j篇文本中出现的频率。

$CountVectorizer$是属于常见的特征数值计算类，是一个文本特征提取重要方法。对于每一个训练文本，它只考虑每种词汇在该训练文本中出现的频率。$CountVectorizer$会将文本中的词语转换为词频矩阵，它通过$fit_transform$函数计算各个词语出现的次数。通过$get_feature_names()$可获得所有文本的关键词，通过$toarray()$可看到词频矩阵的结果。注意到在训练集和测试集上提取的$feature$维度不同，让两个$CountVectorizer$共享$vocabulary$。

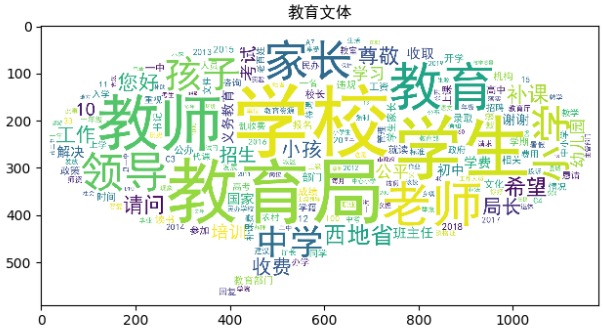
“词云”，指的是对文本留言中出现频率较高的关键词予以视觉上的突出，形成“关键词云层”或“关键词渲染”，从而过滤掉大量的文本信息，使读者只要一眼扫过文本就可以领略文本的主旨。使用$python$中的$WordCloud$绘制词云图，各一级标签分类的词云图如下所示：



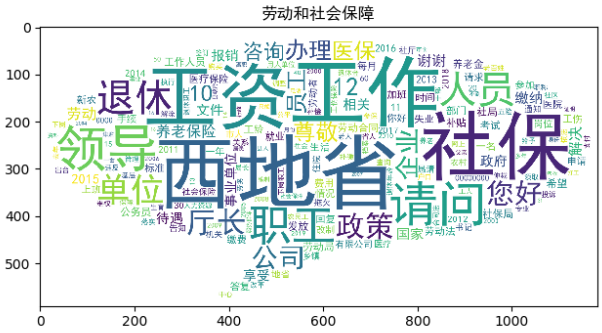
(a) 城乡建设　　(b) 环境保护



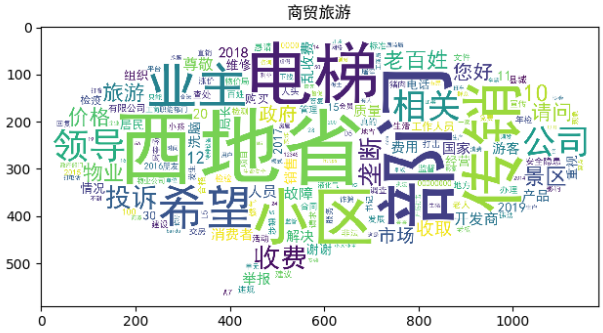
(c) 交通运输



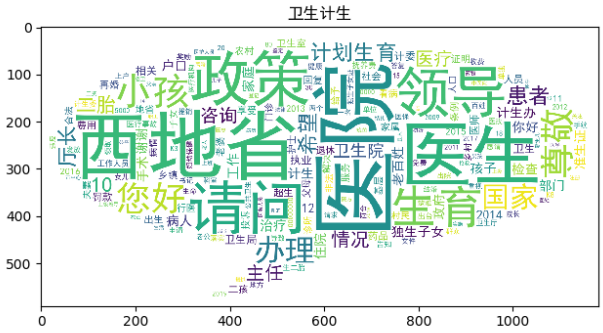
(d) 教育文体



(e) 劳动和社会保障



(f) 商贸旅游



(g) 卫生计生

图 2: 词云图

2.2.4 神经网络模型

神经网络和深度学习目前提供了针对图像识别，语音识别和自然语言处理领域诸多问

题的最佳解决方案。传统的编程方法中，我们告诉计算机如何去做，将大问题划分为许多小问题，精确地定义了计算机很容易执行的任务。而神经网络不需要我们告诉计算机如何处理问题，而是通过从观测数据中学习，计算出他自己的解决方案。

假设我们有一个网络：

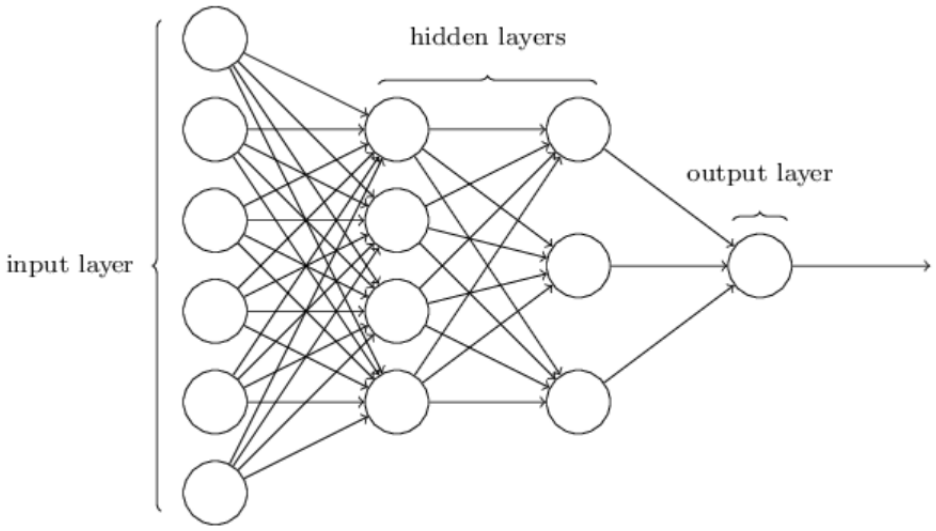


图 3: 神经网络的架构

最左边的一层称为输入层，位于这一层的神经元称为输入神经元。最右边的输出层包含了输出神经元。中间的层被称为隐藏层，因为这些神经元既不是输出也不是输入。隐藏层意味着既不是输入也不是输出。上图的神经网络中只包含了两个隐藏层，有些网络有许多隐藏层。

使用神经网络最重要的就是明确输入输出是什么，输入需要量化成数学表示的向量或矩阵，便于神经网络计算。对于文本而言，最传统的表示方式是$One-hot$表示，即建立一个N维的词典，然后文本中出现某个单词，就用1表示，从而表示这个特征。但这种表示方式不能明确词语之间的相似性，另外还导致特征空间过大，不便于计算。现在出现了各种$embedded$的方法，就是把主要特征映射到某一特征空间，用向量表示，相似词语在该空间中距离较近，能够明确词语之间的相似性。常见的表示工具有谷歌的$word2vec$工具。

得到文本的特征表示后，使用神经网络进行文本分类步骤：

(1) 针对文本需要划分的输出类，进行相关特征提取；

(2) 针对提取的特征（可能是词语、词性、语言模型等），将其数学表示、向量化；

(3) 把所有特征的向量表示拼接；

(4) 将拼接得到的特征表示输入到神经网络，使用已标记的训练数据集合进行神经网络参数训练。

$sklearn$是机器学习中经常用到的一个库，库中含有大量的训练数据集以及大部分经典机器学习算法的封装，我们直接在$python$中导入库中需要使用的文件即可。$neural_network$是$sklearn$库中的一个分文件，用于神经网络模型的训练。

2.2.5 K折交叉验证

Train/test split 是将原始数据集划分为训练集/测试集，避免了为了追求高准确率而在训练集上产生过拟合，从而使得模型在样本外的数据上预测准确率高。但是，划分出训练集/测试集的不同会使得模型的准确率产生明显的变化。为了消除这一变化因素，我们可以创建一系列训练集/测试集，计算模型在每个测试集上的准确率，然后计算平均值。这就是K折交叉验证（$K-foldcross-validation$）的本质。

$K-foldcross-validation$的步骤：

(1) 将原始数据集划分为相等的K部分（“折”）；

(2) 将第1部分作为测试集，其余作为训练集；

(3) 训练模型，计算模型在测试集上的准确率；

(4) 每次用不同的部分作为测试集，重复步骤(2)和(3)K次

(5) 将平均准确率作为最终的模型准确率。

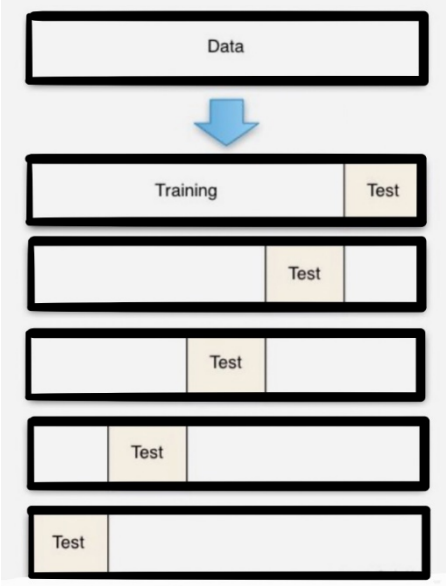


图 4: 5折交叉验证

本文分类模型构建思路是：运用了多个模型（包括：朴素贝叶斯、多项式贝叶斯、伯努力贝叶斯、随机森林、KNN、神经网络模型等），用同一份训练集和测试集分别计算出每个模型的F_1，然后每个模型都用5折交叉验证来验证每个模型算出来的结果是否误差大。

5折交叉验证给出的结果是把数据分割成5份数据集，每份轮流当测试集，然后算出每次的F_1，然后取F_1的平均值。

2.2.6 词袋模型

词袋模型（Bag of Words,简称BoW），所谓词袋模型是一种用机器学习算法对文本进行建模时表示文本数据的方法。词袋模型假设我们不考虑文本中词与词之间的上下文关系，仅仅只考虑所有词的权重。而权重与词在文本中出现的频率有关。与词袋模型非常类似的一个模型是词集模型(Set of Words,简SoW)，和词袋模型唯一的不同是它仅仅考虑词是否在文本中出现，而不考虑词频。也就是一个词在文本在文本中出现1次和多次特征处理是一样的。在大多数时候，我们使用词袋模型。词袋模型被广泛应用在文件分类，词出现的频率可以用来当作训练分类器的特征。

还有很多不同的文本表示方法，比如TF-IDF、LSA、LSI、LDA、HDP、NMF、Word2vec等。但是，不同的应用场景需要不同的文本特征，选择适当方法的原则是：垃圾入，垃圾出。

使用词袋模型将多个文档转换为向量，每个文档由一个向量表示，其中向量元素“i”表示第i个单词出现在文档中的次数。仅通过它们的（整型）id来表征词汇是有利的，问题和ID之间的映射称为字典（dictionary）。

词袋模型具体思想是：分类一个留言主题，假设留言内容是一堆字并且随机倒在一堆袋子的其中一个袋子里，之后使用贝氏机率去决定哪个袋子是较有可能的。

载入中文数据以及对应的包，corpora是构造词典的, similarities求相似性可以用得到。将词语进行分词，并进行存储。寻找整篇语料的词典、所有词，corpora.Dictionary，由doc2bow变为词袋。

2.2.7 语料库制作

在这里，我们通过gensim.corpora.dictionary.Dictionary这个类为处理过的语料库中出现的每个词汇分配一个独一无二的整数ID。这会扫描整个文本，统计所有的词汇计数和词汇相关数据。最后，我们看到在处理的语料库中有不同的词汇，这意味着每个文档将由这些不同词汇的数字表示。可以查看每个词汇与其对应ID之间的映射关系。函数doc2bow()只是计算每个不同词汇的出现次数，将词汇转换为整数词汇id，并将结果作为一个词袋（bag-of-words）— 一个稀疏向量返回，形式为$(word_id1, word_count1)$，$(word_id2, word_count2)$，$(word_id3, word_count3)$……

在token_id中，“A市”对应的为0，“地铁”为1，...，“违规”为11。因而，新文档“A市地铁违规的施工（A市地铁违规的施工）”将被转换为[(2, 1), (10, 1), (11, 1)]。”A市”、“地铁”、“违规” 出现在词典中并出现一次。因此，它们在稀疏向量中分别变为(2, 1), (10, 1), (11, 1)。而“的”等词汇在字典中不存在，因此不会出现在稀疏向量中。词汇计数为0的词汇不会出现在稀疏向量中，并且稀疏向量中将永远不会出现像(3,0)这样的元素。

与scikit-learn相比，doc2bow()与在CountVectorizer上调用transform()有类似的作用。doc2bow()也可以像fit_transform()那样运作。

2.2.8 LDA主题模型

基于语义留言分析进行初步数据感知后，我们从统计学习的角度，对主题特征词出现的频率进行量化表示。本文运用LDA主题模型，用以挖掘留言中更多信息。主题模型在机器学习和自然语言处理等领域是用来在一系列文档中发现抽象主题的一种统计模型。直观上来说，传统判断两个文档相似性的方法是通过查看两个文档共同出现的单词的多少，如TF、TF-IDF等，这种方法没有考虑到文字背后的语义关联，可能在两个文档共同出现的单词很少甚至没有，但两个文档是相似的，因此在判断文档相似性时，应进行语义挖掘，而语义挖掘的有效工具即为主题模型。

如果一篇文档有多个主题，则一些特定的可代表不同主题的词语会反复的出现，此时，运用主题模型，能够发现文中使用词语的规律，并且把规律相似的文本联系在一起，以寻求非结构化的文本集中有用信息。LDA模型作为其中一种主题模型，属于无监督的生成式主题概率模型。

2.2.8.1 LDA主题模型介绍

潜在狄利克雷分配（Latent Dirichlet Allocation, LDA）是由Blei等人在2003年提出的生成式主题模型。生成模型，即认为每一篇文档的每一个词都是通过“一定概率选择了某个主题，并从这个主题中以一定的概率选择了某个词语”。LDA模型也被称为三层贝叶斯概率模型，包含文档（d）、主题（z）、词（w）三层结构，能够有效对文本进行建模，和传统的空间向量模型（VSM）相比，增加了概率的信息。通过LDA主题模型，能够挖掘数据集中的潜在主题，进而分析数据集的集中关注点及其相关特征词。

LDA模型采用词袋模型（Bag Of Word, BOW）每一篇文档是为一个词频向量，从而将文本信息转化为易于建模的数字信息。

定义词表大小为V，一个V维向量$(1,0,0,\ldots,0,0)$表示一个词。有N个词构成的评论记为$d=(w_1,w_2,\ldots,w_N)$。假设某一商品的评论集D由M条留言构成，记为$D=(d_1,d_2,\ldots,d_M)$。M条留言分布着K个主题，记为$z_i=(1,2,\ldots,K)$。记α和β为狄利克雷函数的先验参数，θ为主题在文档中的多项分布的参数，其服从超参数为α的$Dirichlet$先验分布，ϕ为词在主题中的多项分布的参数，其服从超参数β的$Dirichlet$先验分布。LDA模型图示见下图：

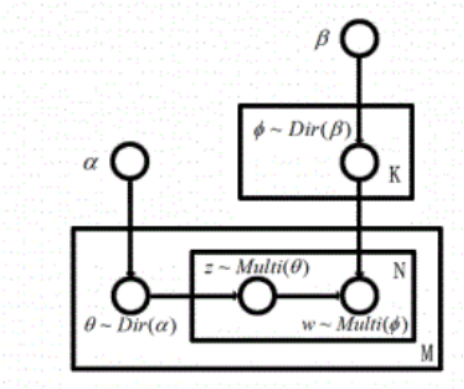


图 5: LDA模型结构示意图

LDA模型假定每条留言由各个主题按一定比例随机混合而成，混合比例服从多项分布，记为：

$$Z|\theta = Multinomial(\theta) \tag{1}$$

而每个主题由词汇表中的各个词语按一定比例混合而成，混合比例也服从多项分布，记为：

$$W|Z, \phi = Multinomial(\phi) \tag{2}$$

在留言d_j条件下生成词w_i的概率表示为：

$$P\left(w_i, d_j\right) = \sum_{s=1}^{K} P\left(w_i|z = s\right) \times P\left(z = s|d_j\right) \tag{3}$$

其中，$P(w_i|z = s)$表示词w_i属于第s个主题的概率，$P(z = s|d_j)$表示第s个主题在留言d_j中的概率。

2.2.8.2 主题模型估计

LDA模型对参数θ、ϕ的近似估计通常使用马尔科夫链蒙特卡洛（Markov Chain Monte Carlo, MCMC）算法中的一种特例Gibbs抽样。利用Gibbs抽样对LDA模型进行参数估计，根据下式：

$$P\left(z_i = s|Z_{-i}, W\right) \propto \left(n_{s,-i} + \beta_i\right) / \left(\sum_{i=1}^{V} n_{s,-i} + \beta_i\right) \times \left(n_{s,-j} + \alpha_s\right) \tag{4}$$

其中，$z_i = s$表示词w_i属于第s个主题的概率，Z_{-i}表示其他所有词的概率，$n_{s,-i}$表示不包含当前文档d_j的被分配到当前主题z_s下的个数。

通过对上式的推导，可以推导到词w_i在主题z_s中的分布参数估计$\phi_{s,j}$，主题z_s在文档d_j中的多项分布的参数估计$\theta_{j,s}$:

$$\phi_{s,j} = \left(n_{s,i} + \beta_i\right) / \left(\sum_{i=1}^{V} n_{s,i} + \beta_i\right) \tag{5}$$

$$\theta_{j,s} = (n_{j,s} + \alpha_s) / \left(\sum_{i=1}^{V} n_{j,s} + \alpha_s \right) \tag{6}$$

其中，$n_{s,i}$表示词w_i在主题z_s中出现的次数，$n_{j,s}$表示文档d_j包含主题z_s的个数。

LDA主题模型在文本聚类、主题挖掘、相似度计算等方面都有广泛的应用，相对于其他主题模型，其引入了狄利克雷先验知识，因此，模型的泛化能力较强，不易出现过拟合现象。其次，它是一种无监督的模式，只需要提供训练文档，它就可以自动训练出各种概率，无需任何人工标注过程，节省大量人力及时间。再者，LDA主题模型可以解决多种指代问题。例如：在留言评论中，根据分词的一般规则，经过分词的语句会将“小区”一词单独分割出来，而“小区”是指“新都小区”，还是其他小区等情况，如果简单地进行词频统计，是无法识别的，从而也无法准确地了解人群反映的情况。运用LDA主题模型，可以求得词汇在主题中的概率分布，进而判断“小区”一次属于哪个主题，并求得属于这一主题的概率和同一主题下的其它特征词，从而解决多种指代问题。

2.2.8.3 运用LDA模型进行主题分析的实现过程

在本文留言内容的研究中，即对留言中的潜在主题进行挖掘，留言的特征词是模型中的可观预测变量。一般来说，每条留言中都存在一个中心思想，即主题。如果某个潜在主题同时是多条留言的主题，则这一潜在主题很可能是整个留言语料集的热门关注点。在这个潜在主题上越高频的特征词将越可能成为热门关注点中的留言词。

本文运用python软件编写LDA主题模型的算法，并采用Gibbs抽样对LDA模型的参数进行近似估计。由上文的模型介绍可知，模型中存在3个可变量需要确定最佳取值，分别是狄利克雷函数的先验参数α和β、主题个数K。本文中将狄利克雷函数的先验参数设置为经验值，分别是$\alpha = 20/K$, $\beta = 0.1$。而主题个数K采用统计语言模型中常用的评价标准肯火毒来选取，即令$K = 20$。

困惑度可以用来度量一个概率分布或概率模型预测样本的好坏程度，低困惑度的概率分布模型或概率模型能更好地预测样本。以下公式计算模型的困惑度：

$$\text{perplexity } (\mathrm{D_{test}}) = P(W|M) = \prod_{m=1}^{m} p\left(w_m|M\right)^{-\frac{1}{N}} = \exp\left\{ \frac{-\sum_{n=1}^{M} \log\left(p\left(w_m\right)\right)}{\sum_{m=1}^{M} N_m} \right\} \tag{7}$$

其中N_m表示第m条留言的词的个数，$p(w_m)$表示词w_m 出现的概率，表示如下：

$$p\left(w_i\right) = \sum_{i \in k} p\left(z_i|m\right) \bullet p\left(w|z_i\right) \tag{8}$$

$p(z_i|m)$表示第m条留言中第i个主题的概率，$p(w|z_i)$在第i个主题下词w的分布概率，计算主题数1到100的困惑度，如图6所示，发现在1到30之间困惑度呈下降趋势，因此再计算主题数1到35的困惑度，如图7所示：

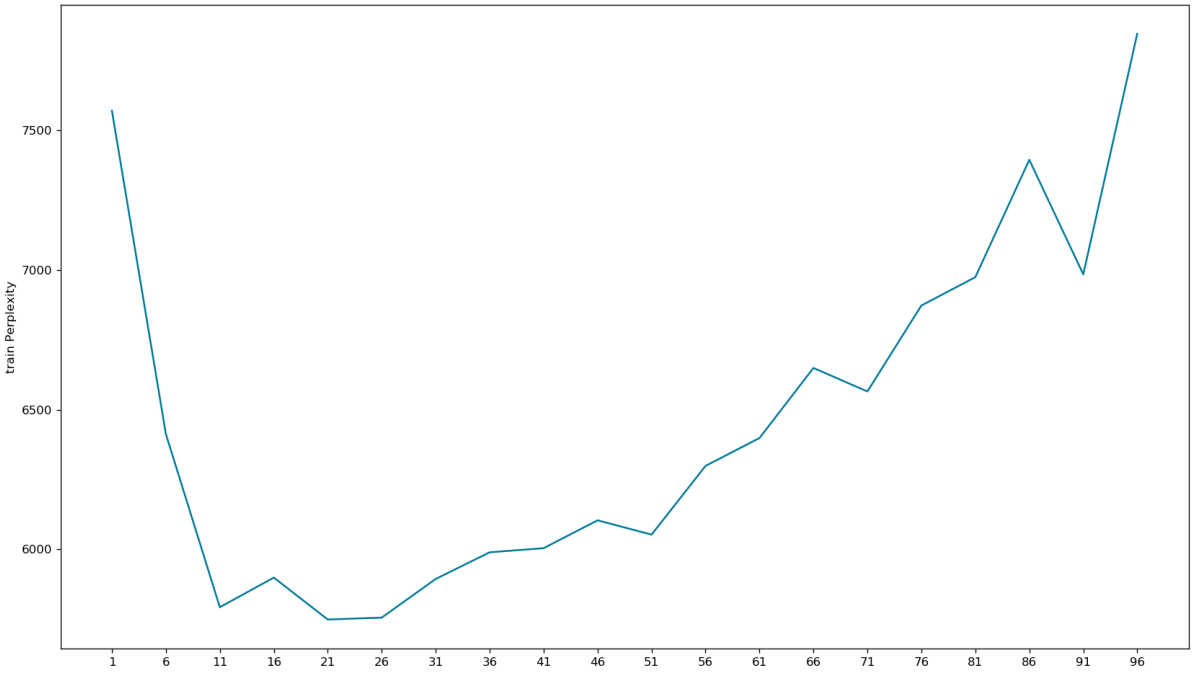


图 6: 主题数1到100的困惑度折线图

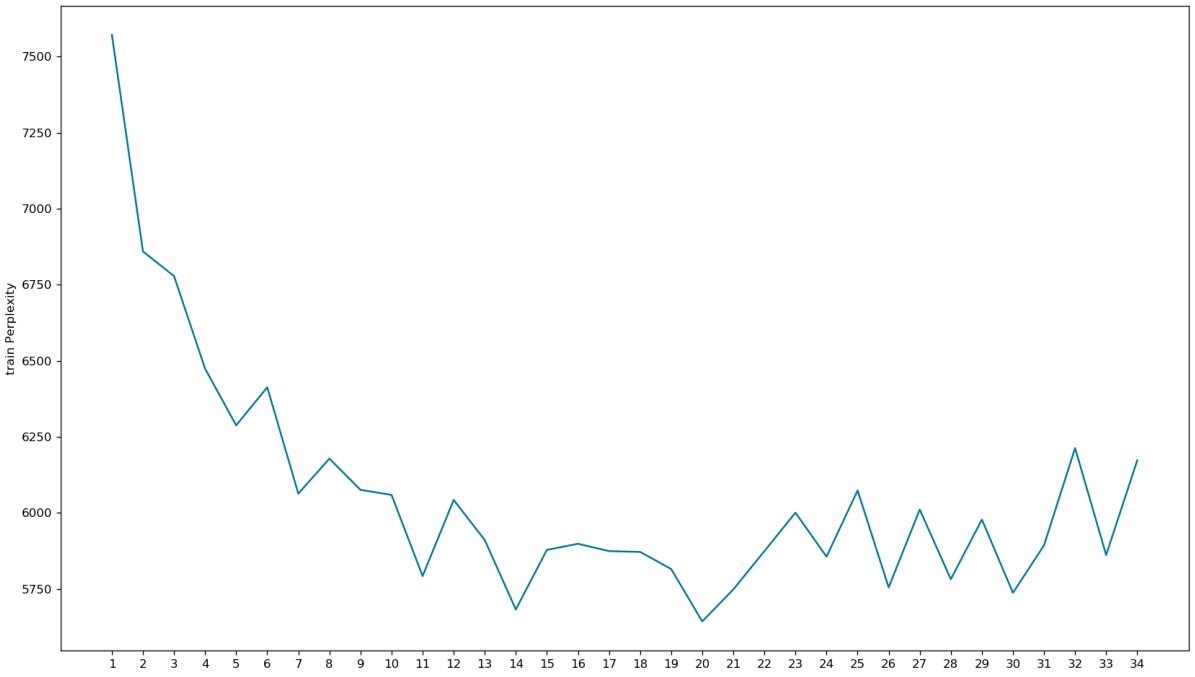


图 7: 主题数1到35的困惑度折线图

主题数为20的时候，困惑度最低。20个主题数目的LDA模型分布如下，其主题基本不重叠：

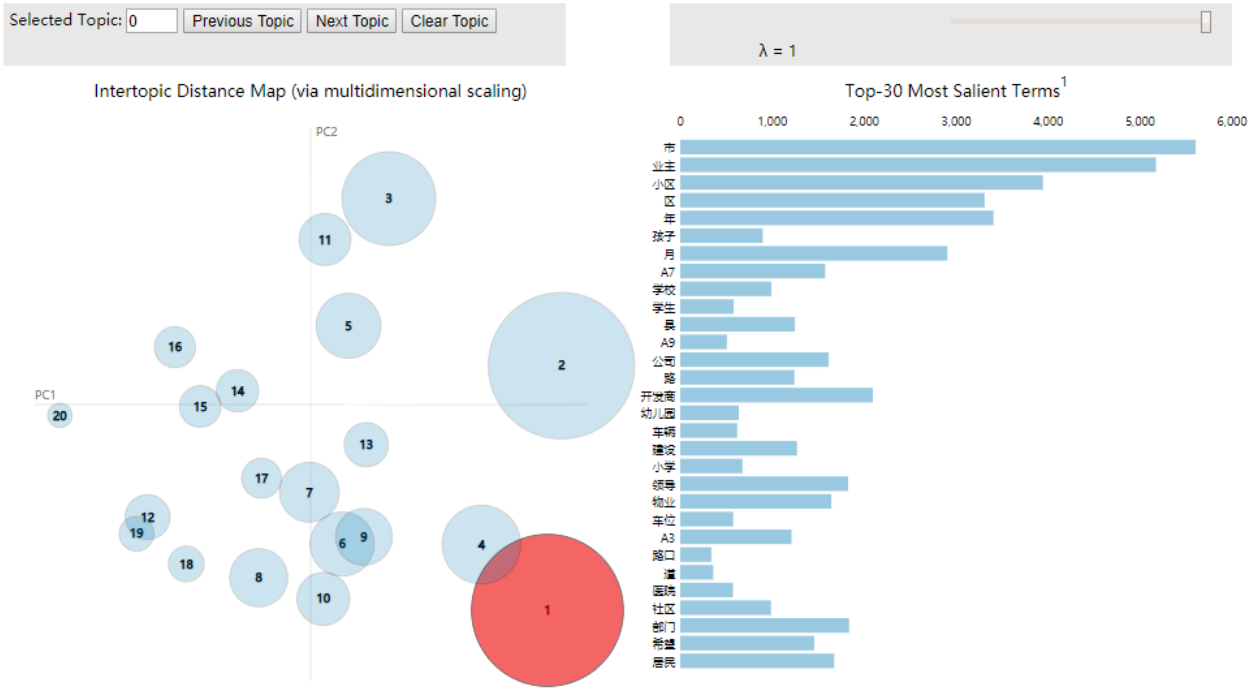


图 8: 各主题容量分布图

2.2.9 训练LDA模型得到主题特征词权重

每一个主题都可以得到特征项的特征权重表达式。特征项即在文档中留言可以用$W(t_1, t_2, t_3, \ldots, t_n)$来表示，其中每一个为特征项，而特征权重则起区分各特征项在留言文档中的作用大小。

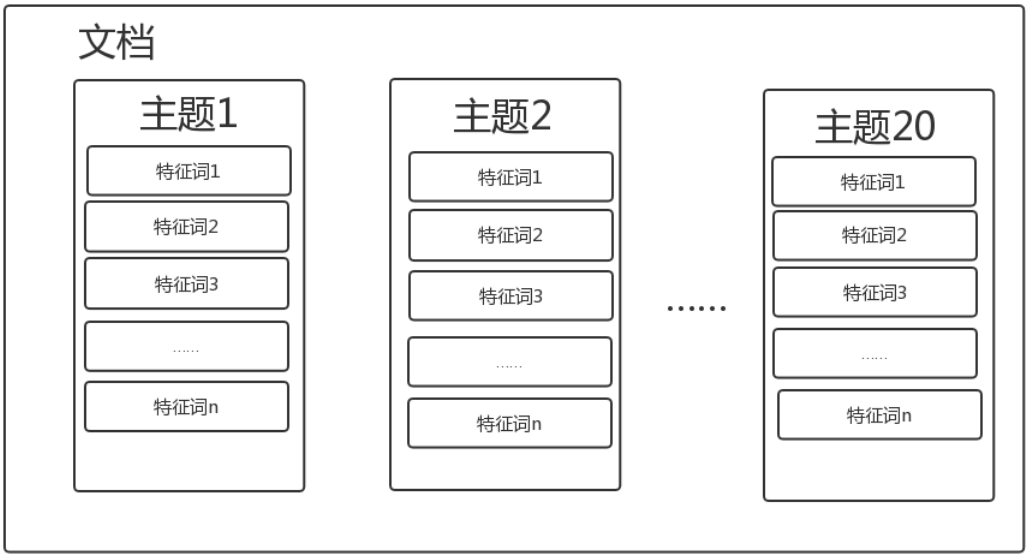


图 9: LDA三层结构

对于已经得到所有文档的主题分布矩阵主题与所有主题的特征词矩阵，在主题-词中选择主题前10个特征词作为主题描述的关键词，结果如下，其中一行表示一个主题：

```
0.032*"市" + 0.024*"A9" + 0.012*"问政" + 0.011*"站" + 0.011*"建设" + 0.009*"公园" + 0.008*"垃圾" + 0.008*"禧" + 0.008*"分拣" + 0.007*"控告人"
0.014*"A7" + 0.013*"市" + 0.010*"路" + 0.010*"县" + 0.006*"建议" + 0.006*"希望" + 0.005*"领导" + 0.005*"小区" + 0.004*"规划" + 0.004*"A3"
0.017*"县" + 0.014*"A7" + 0.009*"业主" + 0.007*"九龙湾" + 0.007*"解决" + 0.006*"年" + 0.006*"小区" + 0.006*"领导" + 0.006*"国泰" + 0.005*"月"
0.011*"车辆" + 0.008*"路" + 0.007*"市" + 0.006*"商业广场" + 0.005*"区" + 0.005*"K4" + 0.005*"路口" + 0.005*"大道" + 0.005*"匝道" + 0.005*"东盈"
0.016*"直行" + 0.013*"左" + 0.011*"路口" + 0.010*"护窗" + 0.008*"道" + 0.008*"拐" + 0.007*"市" + 0.006*"区" + 0.005*"车辆" + 0.005*"泥头车"
0.019*"业主" + 0.014*"年" + 0.013*"市" + 0.011*"开发商" + 0.011*"月" + 0.008*"公司" + 0.007*"物业" + 0.006*"政府" + 0.006*"日" + 0.005*"区"
0.014*"业主" + 0.011*"医院" + 0.009*"车位" + 0.008*"区" + 0.008*"市" + 0.005*"领导" + 0.005*"销售" + 0.005*"街道" + 0.005*"小区" + 0.004*"部门"
0.012*"出借" + 0.010*"月" + 0.010*"立案" + 0.008*"平台" + 0.008*"市" + 0.008*"资金" + 0.007*"公司" + 0.006*"区" + 0.006*"派出所" + 0.006*"xiang"
0.026*"市" + 0.008*"城市" + 0.007*"加快" + 0.007*"国家" + 0.007*"发展" + 0.007*"希望" + 0.006*"建设" + 0.006*"中心" + 0.005*"市政府" + 0.005*"请"
0.024*"业主" + 0.010*"幼儿园" + 0.008*"开发商" + 0.007*"小区" + 0.007*"年" + 0.007*"月" + 0.006*"驾校" + 0.006*"物业" + 0.006*"栋" + 0.005*"市"
0.016*"市" + 0.011*"月" + 0.010*"年" + 0.008*"区" + 0.007*"号" + 0.007*"征收" + 0.006*"窗口" + 0.006*"电话" + 0.006*"村民" + 0.005*"记录"
0.014*"小区" + 0.013*"业主" + 0.008*"幼儿园" + 0.008*"国际" + 0.007*"区" + 0.006*"开发商" + 0.005*"规划" + 0.005*"时代" + 0.004*"A3" + 0.004*"部门"
0.024*"学生" + 0.017*"学校" + 0.011*"市" + 0.011*"学院" + 0.011*"孩子" + 0.010*"实习" + 0.010*"家长" + 0.009*"中学" + 0.008*"训练" + 0.007*"组织"
0.024*"小区" + 0.013*"业主" + 0.011*"区" + 0.011*"居民" + 0.008*"市" + 0.008*"影响" + 0.007*"部门" + 0.006*"生活" + 0.005*"噪音" + 0.005*"请"
0.023*"孩子" + 0.015*"市" + 0.012*"小学" + 0.008*"年" + 0.008*"教育" + 0.007*"区" + 0.007*"月" + 0.007*"学校" + 0.006*"老师" + 0.005*"业主"
0.007*"罚款" + 0.007*"市" + 0.007*"公司" + 0.006*"诈骗" + 0.006*"云顶" + 0.006*"资金" + 0.006*"西地省" + 0.006*"传销" + 0.006*"街区" + 0.005*"富达"
0.012*"年" + 0.009*"交易" + 0.008*"社保" + 0.007*"市" + 0.006*"月" + 0.005*"早教" + 0.005*"巴辛" + 0.005*"区" + 0.004*"项目" + 0.004*"西地省"
0.013*"区" + 0.011*"市" + 0.008*"渣土" + 0.008*"自治" + 0.008*"违规" + 0.007*"A3" + 0.007*"天顶" + 0.007*"青山" + 0.006*"鸡" + 0.006*"改商"
0.011*"商品交易" + 0.010*"学校" + 0.009*"学生" + 0.006*"孩子" + 0.006*"市" + 0.005*"工作" + 0.004*"中心" + 0.004*"家长" + 0.004*"希望" + 0.004*"上课"
```

图 10: 各主题关键特征词权重表达式

2.2.10 依附主题主要特征词挖掘多留言问题

2.2.10.1 整合主题留言问题

在主题1中的124条留言，对比每条留言关键词的相似程度，以此划分问题类型，提取文档超过3的部分问题如下表所示：

表 2: 主题1问题发现结果

问题	主题1留言文档描述关键词	留言数
1	A市 洪山 公园 建设 城市	4
2	梅溪 公园 电站 杨美 居民	4
3	A9 垃圾 分拣 政府 禧	4
4	梅溪 公园 居民 殡仪馆 市民	3
5	控告 平安 A3 民警 证人	3

由表可知主题1中“A市A4区洪山公园建设计划”，“A市楚江新区梅溪湖杨美公园改建成变电站”，“A9区禧和路垃圾分拣中心选址问题”，“A市梅溪公园附近非法建立殡仪馆”“A4区检察院工作人员隐匿证据伪造证据”等问题比较受关注，对其他主题进行同样操作。

2.2.10.2 过滤低频、假性高频问题

通过LDA模型将问题进行主题划分后，在对热点问题的挖掘过程中，容易受到其他与高频留言问题有相同特征词的低频留言问题干扰，因此对低频留言问题进行过滤。

在热点问题的初步挖掘中，发现存在留言问题同时拥有几条留言，但是皆由同一个留言用户提出的现象，且“留言主题”、“留言详情”的内容词汇差异不大。如下表所示，与“A7县深业睿城到松雅中学204公交车的两点问题”相关的问题有三条，与“建议A2区暮云街道月塘路增设限时停车位”相关的留言有两条，但是皆由同一个留言用户提出，且其留言得到的反对数、点赞数皆不高，留言间隔时间短。这类数据会造成留言涉及的问题留言次数较多、关注该问题人数较多的假象，我们把它们归总为一条留言，对于仅一两人反复提及且点赞数、反对数基本为0的问题，显然为非热点问题，不具有热点话题考察性质，对这类留言亦可直接过滤。

表 3: 同用户短时间对同问题留言实例

留言编号	留言用户	留言主题	留言时间	反对数	点赞数
211673	A00088861	反映A7县深业睿城到松雅中学204公交车的两点问题	2019/8/21 15:25:43	0	4
247742	A00088861	反映从A7县深业睿城到松雅中学204公交车的两点问题	2019/8/20 11:24:57	0	1
225812	A00088861	反映A7县204公交车的两点问题	2019/8/26 10:17:37	0	1
237799	A00090621	建议A2区暮云街道月塘路增设限时停车位	2020/1/1 18:35:05	0	0
230131	A00090621	建议在A2区暮云街道月塘路增设限时停车位	2019/12/30 16:24:38	0	0

除此之外，经过观察也发现存在问题留言由同一个留言用户提出，且“留言主题”、“留言详情”的内容词汇差异不大，但是“留言时间”差异较大的留言问题。如下表同一用户“A00074795”在2019/1/10与2019/11/18先后提出“外迁京港澳高速城区段至远郊”的问题，两次提出问题时间间隔较长，且获得的点赞数亦不低。对于这类留言，不考虑归于一条留言，仍视为描述同一问题的不同留言。

表 4: 同用户间隔长时间对同问题留言实例

留言编号	留言用户	留言主题	留言时间	反对数	点赞数
284571	A00074795	建议西地省尽快外迁京港澳高速城区段至远郊	2019/1/10 15:01:26	0	80
253369	A00074795	穿A市城而过的京港澳高速（长楚高速）什么时候可以外迁至远郊？	2019/11/18 15:35:11	0	29

2.2.10.3 初步筛选潜在热点问题

事实上，所挖掘的文本数据中有许多问题仅有寥寥几条相关留言，而本次所求为某一段时间的热点问题，故留言仅一到五条的基本可视为非热点问题，对于这类问题我们选择忽略。选取每个主题中，不同留言用户中描述同样问题超过五条的问题。在问题处理中发现，有些问题尽管提及的次数很少，但是反对和点赞总和很高，说明该问题除了留言用户外，还有许多浏览群众也同样对之加以关注，故将这类问题也选取进来。

下面列出15个留言条数多或者反对与点赞总和较高的问题：

表 5: 潜在热点问题表

问题	时间范围	地点/人群	问题描述	留言条数	反对与点赞数总和
1	2019/2/23至2019/8/23	A市丽发新城小区侧面建设混泥土搅拌站	搅拌站日夜工作，粉尘和噪音污染严重	55	53
2	2019/7/7至2019/8/31	A市伊景园滨河苑	A市伊景园滨河苑捆绑销售车位	40	21
3	2018/11/15至2019/10/29	A市人才	A市人才租房购房补贴问题	18	28
4	2019/3/4至2019/12/4	A市地铁	A市地铁扫码乘车存在各种不方便现象	13	30
5	2019/1/14至2019/7/8	西地省A市58车贷	58车贷A4区立案已近半年毫无进展	12	2377
6	2017/7/21至2019/10/24	A市A5区万科魅力之城商铺	万科魅力之城商铺无排烟管道，小区内到处油烟味	12	18
7	2019/3/28至2019/11/29	A市温斯顿英语培训机构	温斯顿英语培训强设霸王条款	12	8
8	2019/5/31至2019/8/5	A市A3区天顶街道青山新村社区青青家园小吃店	青青家园违规住改商存在安全隐患	9	1
9	2019/2/20至2019/11/15	A市三一大道	A市三一大道改造问题	8	69
10	2019/5/5至2019/9/19	A市五矿万境K9县	A市五矿万境K9县交房后存在诸多问题	7	2106
11	2019/8/23至2019/9/6	A市A4区绿地海外滩小区	A市绿地海外滩二期与长懿高铁问题	7	691
12	2019/3/26至2019/4/12	A市A6区月亮岛路	关于A6区月亮岛路沿线架设110kv高压电线杆的投诉	7	173
13	2019/3/9至2019/5/18	西地省惠普利聚人投资有限公司	西地省惠普利聚人投资有限公司涉嫌诈骗巨额资金	7	33
14	2019/2/23至2019/8/23	A市A2区余易贷平台	A2区余易贷平台涉嫌诈骗	7	22
15	2019/4/11	A市金毛湾	A市金毛湾配套入学的问题	1	1762

2.2.11 TF-IDF模型计算文本相似度

TF-IDF算法是用于信息检索和数据挖掘的常用加权算法。TF-IDF是一种统计方法，用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。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。除了TF-IDF以外，因特网上的搜索引擎还会使用基于链接分析的评级方法，以确定文件在搜寻结果中出现的顺序。

TF-IDF模型的主要思想是:如果词w_i在一篇文档w_i中出现的频率高，并且在其他文档中很少出现，则认为词w_i具有很好的区分能力，适合用来把留言w_i和其他留言区分开来。

TF-IDF是最基础的文本相似度计算方法。TF(Term Frequency)指一篇文档中单词出现的频率，IDF(Inverse Document Frequency)指语料库中出现某个词的文档数，取对数。

$$TF = \frac{\text{词在文档中出现的次数}}{\text{文档中所有词的个数}} \tag{9}$$

$$IDF = \log\left(\frac{\text{语料库的文档总数}}{\text{语料库中出现苯单词的不同文档个数}}\right) \tag{10}$$

TF原理：某个词在一篇文档中出现的频率越多则对这篇文章越重要；

IDF原理：该词在越多的文章中出现，则说明它对文章没有很强的区分度，在文档中所占的权重也就越小，一般采用取词频的逆。还要考虑一个现象，一些通用词出现的次数可能是低频词的几十倍上百倍，如果只是简单的取逆处理，通用词的权重会变动非常小，稀缺词的权重就显得的过大了。为了平衡通用词与稀缺词的权重关系，又对逆采用取对数运算。

TF-IDF则是由TF和IDF相乘得到，如下：

$$TF - IDF = TF \times IDF \tag{11}$$

TF-IDF原理：如果一个词，在该留言中出现次数越多，而在其他留言中出现很少时，则该词汇越能够表示该留言的信息。

TF-IDF算法的优势在于算法简单，并且对文章的所有元素进行了综合考量。

2.2.12 正则表达式命名实体识别

2.2.10.1 命名实体

文本中有一些描述实体的词汇。比如人名、地名、法律名、公司名、专业术语等，称为命名实体。具有以下共性:

(1) 数量无穷。比如宇宙中的恒星命名、新生儿的命名不断出现新组合。

(2) 构词灵活。比如中国工商银行，既可以称为工商银行，也可以简称工行。

(3) 类别模糊。有一些地名本身就是机构名，比如“国家博物馆”。

2.2.10.2 正则表达式匹配、识别命名实体

构造正则表达式的方法和创建数学表达式的方法一样。也就是用多种元字符与运算符可以将小的表达式结合在一起来创建更大的表达式。正则表达式的组件可以是单个的字符、字符集合、字符范围、字符间的选择或者所有这些组件的任意组合。

正则表达式是由普通字符（例如字符 a 到 z）以及特殊字符（称为“元字符”）组成的文字模式。模式描述在搜索文本时要匹配的一个或多个字符串。正则表达式作为一个模板，将某个字符模式与所搜索的字符串进行匹配。

在识别命名实体法律法规中，使用最小匹配的正则表达式可以写为：[《](.*?)[》]，来匹配出符合要求的字符，识别法律法规。

2.2.13 层次分析法

层次分析法（AHP）是将要决策的问题及其有关因素分解成目标、准则、方案等层次，进而进行定性和定量分析的决策方法。它的特征是合理地将定性与定量决策结合起来，按照思维、心理的规律把决策过程细致化（层次化、数量化）。层次分析法广泛地应用到处理复杂的决策问题，而决策是基于该方法计算出的权重，所以也常用来确定指标的权重。

2.2.13.1 构建层次模型

建立层次结构模型，将问题条理化、层次化。该模型主要分为三层：

(1) 最高层（目标层）——只有一个元素：决策目标；

(2) 中间层（准则层）——考虑的因素，决策的准则、子准则；

(3) 最底层（方案层）——决策时的备选方案、措施。

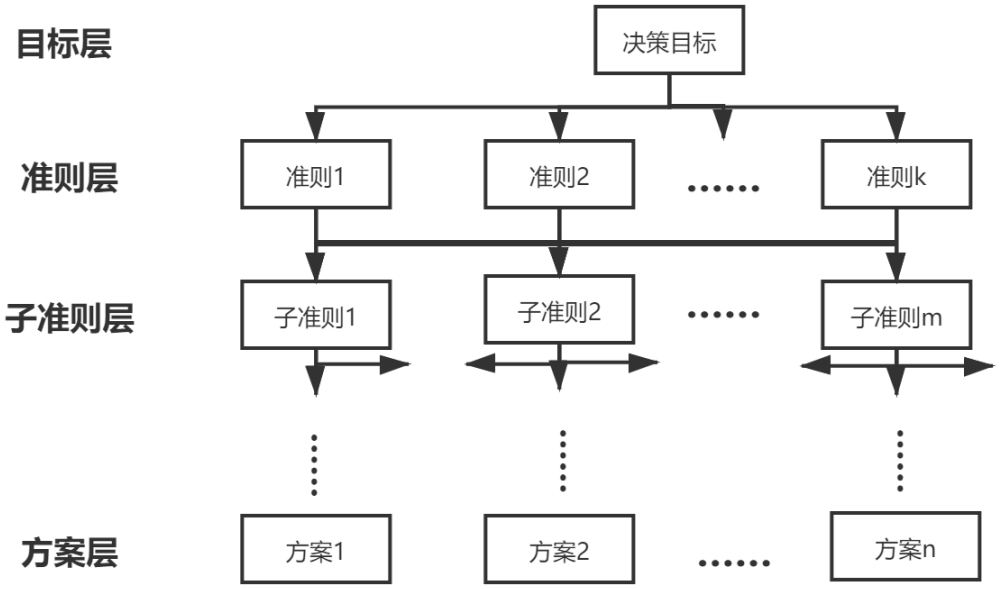


图 11: 层次分析法结构

2.2.13.2 构造判断矩阵（成对比较矩阵）

构造好层次模型后，针对某一层来讲，在比较第i个元素与第j个元素相对于上一层某个因素的重要性时，使用数量化的相对重要度a_{ij}来表示，假设共有 n个元素参与比较，则矩阵表示为：

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} = (a_{ij})_{n\times n} \tag{12}$$

称为判断矩阵（或成对比较矩阵）。

Saaty根据绝大多数人认知事物的心理习惯，建议用$1 \sim 9$及其倒数作为标度来确定a_{ij}的值。如下表所示：

表 6: 评价标度

i比j的重要程度	相等	一般	重要
a_{ij}	1	3	5

其中，2，4分别介于1，3，5对应的重要程度之间。显然，A中的元素满足：

i) $a_{ij} > 0$

ii) $a_{ji} = \frac{1}{a_{ij}}$

iii) $a_{ii} = 1$

称为正互反矩阵。

2.2.13.3 判断矩阵的一致性检验与层次单排序

在实际操作中，由于客观事物的复杂性以及人们对事物判断比较时的模糊性，很难构造出完全一致的判断矩阵。因此，Satty在构造层次分析法时，提出了一致性检验，所谓一致性检验是指判断矩阵允许有一定不一致的范围。

一致性检验步骤如下：

(1) 计算判断矩阵A的最大特征值λ_{max}；

(2) 求出一致性指标（Consistencey Index）：

$$CI = \frac{\lambda_{\max} - n}{n-1} \tag{13}$$

$CI = 0$ 表示完全一致，CI 越大越不一致；

(3) 用随机模拟取平均的方法，求相应的平均随机一致性指标RI，或者直接用Satty模拟1000次得到的RI表。

表 7: RI表

矩阵阶数	3	4	5	6	7	8	9	10	11	12	13
RI	0.52	0.89	1.12	1.26	1.36	1.41	1.46	0.49	0.52	1.54	1.56

(4) 计算一致性比率：

$$CR = \frac{CI}{RI} \tag{14}$$

(5) 判断，当$CR < 0.1$时，认为判断矩阵A有满意的一致性；若$CR \geq 0.1$，应考虑修正判断矩阵A。

通常用特征根法从判断矩阵导出，单一准则下元素相对排序权重。特征根法的基本思想是，当正互反矩阵$(a_{ij})_{n\times n}$为一致性矩阵时，对应于判断矩阵的最大特征根λ_{max}的特征向量，经归一化后（使向量中各元素之和等于1）即为排序权向量w，w的元素为同一层次因素对于上一层次某因素相对重要性的排序权值，这一过程称为层次单排序。

2.2.13.4 计算各元素对目标层的合成权重（层次总排序）

为了实现层次分析法的最终目的，需要从上而下逐层进行各层元素对目标合成权重的计算。

设已计算出第$k-1$层n_{k-1}个元素相对于目标的合成权重为：

$$w^{(k-1)}=\left(w_1^{k-1},w_2^{k-1},\ldots,w_{n_{k-1}}^{k-1}\right) \tag{15}$$

再设第k层的n_k个元素关于第$k-1$层第j个元素$(j=1,2,\ldots,n_{k-1})$的单一准则排序权重向量为：

$$u_j^k=\left(u_{1j}^{(k)},u_{2j}^{(k)},\ldots,u_{n_kj}^{(k)}\right) \tag{16}$$

上式对k层的n_k个元素是完全的，若某些元素不受第$k-1$层第j个元素支配，相应位置用0补充，于是得到$n_k\times n_{k-1}$阶矩阵：

$$U^{(k)}=\begin{pmatrix} u_{11}^{(k)} & u_{12}^{(k)} & \cdots & u_{1n_{k-1}}^{(k)} \\ u_{21}^{(k)} & u_{22}^{(k)} & \cdots & u_{2n_{k-1}}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ u_{n_k1}^{(k)} & u_{n_k2}^{(k)} & \cdots & u_{n_kn_{k-1}}^{(k)} \end{pmatrix} \tag{17}$$

从而可以得到第k层的n_k个元素关于目标层的合成权重向量：

$$w^{(k)}=U^{(k)}w^{(k-1)} \tag{18}$$

按递归展开的

$$w^{(k)}=U^{(k)}U^{(k-1)}\ldots U^{(3)}w^{(2)} \tag{19}$$

写成分量形式为

$$w_{ij}^{(k)}=\sum_{j=1}^{n_{k-1}}u_{ij}^{(k)}w_{ij}^{(k-1)},\quad i=1,\cdots,n_k \tag{20}$$

各层元素对目标层的合成排序权重向量是否可以满意接受，与单一准则下的排序问题一样，需要进行综合一致性检验：

设k层的综合指标分别为一致性指标$CI^{(k)}$，随机一致性指标$RI^{(k)}$，一致性比率$CR^{(k)}$。再设以第$k-1$层上第j层元素为准则的一致性指标为$CI_j^{(k)}$，平均一致性指标为$RI_j^{(k)}(j=1,2,\ldots,n_{k-1})$，则

$$CI^{(k)}=\left(CI_1^{(k)},\cdots,CI_{n_{k-1}}^{(k)}\right)w^{(k-1)}=\sum_{j=1}^{n_{k-1}}w_j^{(k-1)}CI_{ij}^{(k)} \tag{21}$$

$$RI^{(k)}=\left(RI_1^{(k)},\cdots,RI_{n_{k-1}}^{(k)}\right)w^{(k-1)}=\sum_{j=1}^{n_{k-1}}w_j^{(k-1)}RI_j^{(k)}. \tag{22}$$

进而可计算综合一致性比率：

$$CR^{(k)}=\frac{CI^{(k)}}{RI^{(k)}} \tag{23}$$

当$CR^{(k)}<0.1$时，则认为层次结构在第k层以上的判断具有整体满意的一致性。

2.3 结果分析

2.3.1 建立分类模型，使用 F-Score 对模型评价

我们首先采取了多个模型（包括：朴素贝叶斯、多项式贝叶斯、伯努力贝叶斯、随机森林、KNN、神经网络模型等），用同一份训练集和测试集分别计算出每个模型的F_1得分，然后每个模型都用5折交叉验证来验证每个模型算出来的结果是否误差大。5折交叉验证给出的结果是把数据分割成5份数据集，每份轮流当测试集，然后算出每次的F_1，然后取F_1的平均值。通常使用 F-Score 对分类方法进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{24}$$

其中P_i为第i类的查准率，R_i为第i类的查全率。

各种模型测试结果如下所示：

表 8: 各模型比较表

模型	F_1	5折交叉验证的F_1的平均值
朴素贝叶斯	73.28%	67.23%
多项式贝叶斯	83.43%	78.97%
伯努力贝叶斯	83.92%	80.02%
随机森林	79.37%	78.89%
KNN	84.84%	78.79%
神经网络模型	90.34%	86.54%

在多条留言集的测试下，可以看出神经网络模型为最佳模型，F-Score最大，$F_1 = 90.34\%$，而且通过5折交叉验证。神经模型在词向量的基础上，通过特征提取，挖掘出更深层的语义信息，从而达到较好的分类能力。

2.3.2 挖掘热点问题

由上文的模型介绍可知，模型中存在3个可变量需要确定最佳取值，分别是狄利克雷函数的先验参数α和β、主题个数K。本文中将狄利克雷函数的先验参数设置为经验值，分别是$\alpha = \frac{20}{K}$, $\beta = 0.1$。而主题个数K采用统计语言模型中常用的评价标准困惑度来选取，即令$K = 20$。故，我们选取20个主题进行归类分析，每个主题中选取10个高频关键特征词进行挖掘分析。

热点问题往往拥有以下特点：

- 热点问题的产生往往是影响到群众生活，它会在一段时间内持续，故群众会在这段时间内集中反映。

- 热点问题容易受到其他群众的关注，从而相关留言的点赞数或反对数会相对较高。

所以，热点话题的影响因素有以下三个：

- 相关留言数：与问题相关的留言越多，表示留言的潜在热度越高。
- 相关留言影响力：留言点赞量或留言反对量的大小表示了别人对留言用户提出的问题的认同感或不认同感，数值的大小反映了留言的传播影响力。
- 相关留言时间频率：一段时间内相关的留言数量越多，留言的点赞量、反对量越多，留言的热度越高。

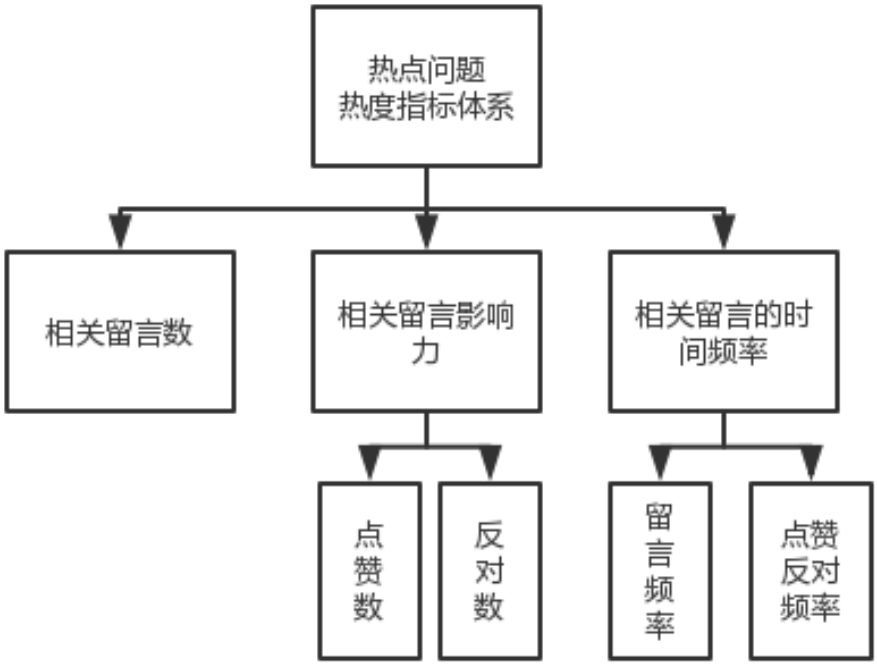


图 12: 热点问题热度指标体系

我们定义了热点问题的评价[4]，如下：

$$Hot\_topic_i = M_i + \log_2(2 + Like_i + Unlike_i) + (M_i + \log_2(2 + Like_i + Unlike_i))/D_i \quad (25)$$

其中Hot_topic_i表示问题的热度指数，M_i表示问题的相关留言数，$Like_i + Unlike_i$表示问题的相关留言点赞量与反对量的总数，D_i表示问题的时间范围，$log_2(2 + Like_i + Unlike_i)$保证了留言的影响力不为0，起调节作用。

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	61.50494733	2019/11/2至2020/1/25	A市丽发新城小区侧面建设混泥土搅拌	搅拌站日夜工作，粉尘和噪音污染严重
2	2	45.33308126	2019/7/7至2019/8/31	A市伊景园滨河苑	A市伊景园滨河苑捆绑销售车位
3	3	23.34880321	2019/1/14至2019/7/8	西地省A市58车贷受害人	58车贷A4区立案半年毫无进展
4	4	22.97271499	2018/11/15至2019/10/29	A市人才	A市人才租房购房补贴问题
5	5	18.17335009	2019/5/5至2019/9/19	A市五矿万境K9县	A市五矿万境K9县交房后存在诸多问题

图 13: 排名前五热点问题

热点问题留言明细表见附件“热点问题留言明细表.xlsx”。

2.3.3 评价方案

针对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。本文对三个角度进行量化处理，其中，使用TF-IDF模型计算文本相似度，从留言和答复的对比得到相关性得分；从模板和答复的对比得到完整性得分。用正则表达式对命名实体进行识别得到可解释性得分。由于三个指标的量纲不一样，因此需要对其进行标准化处理。综合得到如下结果（具体参见附录）：

留言编号	相关性	完整性	可解释性	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	留言	相关性（标准化）	完整性（标准化）	可解释性（标准化）	层次分析法
88069	63.29189	44.81291	120	UU008119	反映J3县	2019/7/1	尊敬的J3	网友：您	2019/8/2	反映J3县	0.653496143	0.603571757	1	0.737764337
9744	77.66399	36.82297	110	UU008132	关于对A市	2016/10/	小区	网友“UU	2016/11/	关于对A市	0.801889736	0.495957675	0.916666667	0.799494803
4331	63.1036	49.40525	100	UU008233	质疑A市新	2018/10/	尊敬的书	网友“UU	2018/11/	质疑A市新	0.651552077	0.665424552	0.833333333	0.699956304
19133	74.954	28.86257	100	UU008720	万明村征	2015/2/7	尊敬的领	网友“UU	2015/8/2	万明村征	0.773908765	0.38874142	0.833333333	0.748918882
133336	64.71087	42.67459	80	UU008104	咨询L6县	2019/5/1	尊敬的领	网友：您	2019/5/1	咨询L6县	0.668147317	0.574771327	0.666666667	0.657985673
9372	73.05285	51.61853	60	UU008573	关于小孩	2017/2/2	本人	网友“UU	2017/3/1	关于小孩	0.754279172	0.69523466	0.5	0.682418979
92030	60.36927	53.78316	70	UU008219	咨询J11市	2018/9/3	现在农合	网友同志	2018/10/	咨询J11市	0.623319705	0.724389364	0.583333333	0.623576794
99145	66.48697	39.87788	70	UU008130	对K10县关	2018/6/7	再次对K1	“UU0081	2018/6/2	对K10县关	0.686485687	0.53710324	0.583333333	0.644198243
12669	67.63062	46.13608	60	UU008154	强烈呼吁	2013/10/	近段	网友“UU	2013/12/	强烈呼吁	0.69829403	0.621393133	0.5	0.63902387
115274	65.24283	38.10119	70	UU008130	对K10县关	2018/6/7	再次对K1	“UU0081	2018/6/2	对K10县关	0.6736398	0.513173541	0.583333333	0.633509454
103903	57.71249	53.97505	60	UU008222	投诉K市香	2019/4/1	香水湾2期	“UU0082	2019/6/2	投诉K市香	0.595888186	0.726973809	0.5	0.58485023
128875	70.59181	35.21014	60	UU008100	L1区幼儿	2018/2/2	尊敬	网友：	2018/2/2	L1区幼儿	0.728868709	0.474234866	0.5	0.643087697
33235	81.37335	61.87183	20	UU008220	咨询B7县	2019/6/1	五菱之光	尊敬的“	2019/6/1	咨询B7县	0.84018929	0.83333328	0.166666667	0.665510477
11819	61.98255	39.8862	60	UU008167	关于对A市	2014/11/	尊敬的易	网友“UU	2014/12/	关于对A市	0.639977042	0.53721535	0.5	0.593060899
93031	80.51092	60.10408	20	UU008497	咨询J11市	2018/12/	请问公务	网友：您	2018/12/	咨询J11市	0.831284645	0.80952392	0.166666667	0.657344806
6325	61.25898	48.75084	50	UU008131	关于A5区	2018/4/1	我们是A市	网友“UU	2018/5/1	关于A5区	0.632506153	0.656610561	0.416666667	0.579282494
27418	72.99966	56.69467	30	UU008409	反映A5区	2018/10/	尊敬的各	“UU0084	2018/11/	反映A5区	0.753729967	0.763603599	0.25	0.624658135
22977	51.49397	58.08377	50	UU008914	投诉A8县	2019/5/1	在这里想	网友“UU	2019/5/2	投诉A8县	0.531681274	0.782312949	0.416666667	0.52822324
136055	63.38663	55.21763	40	UU008432	投诉L5县	2019/7/2	L5县环境	网民朋友	2019/8/1	投诉L5县	0.654474299	0.743709803	0.333333333	0.58087399

图 14: 三个指标得分值

接下来使用层次分析法对回复信息进行评价，采用专家评价法，基于重要程度对三个指标进行重要程度评价，评价准则如下表所示：

表 9: 相对重要程度

	完整性	可解释性	相关性
完整性	1	1/3	1/5
可解释性	3	1	1/3
相关性	5	3	1

故，判断矩阵表示如下：

$$A=\begin{pmatrix}1 & \frac{1}{3} & \frac{1}{5}\\ 3 & 1 & \frac{1}{3}\\ 5 & 3 & 1\end{pmatrix} \tag{26}$$

相应的值如下表所示：

表 10: 层次分析法结果

	行内连乘M_i	开n次方w_i	归一化得到权重w_i
完整性	0.066666667	0.405480133	0.104729434
可解释性	1	1	0.258284994
相关性	15	2.466212074	0.636985572
合计		3.871692207	1

表 11: 一致性检验结果

一致性检验	
λ_{max}	3.038511091
CI	0.019255545
RI	0.033199216

由$CR=\frac{CI}{RI}<0.1$，可认为层次结构的判断具有整体满意的一致性。

故，每条留言答复质量的评价指标定义为三个指标权重分别乘以得分，再计算这三项的和。答复质量评价指标如下所示：

$$\text{答复质量}=0.10\times\text{标准化的完整性得分}+0.26\times\text{标准化的可解释性得分}+0.64\times\text{标准化的相关性得分} \tag{27}$$

3 结论

本文通过对处理过的群众问政留言记录数据利用神经网络、LDA主题模型、TF-IDF模型等方法建立多种数据挖掘模型，得到了具有一定价值的结果，实现了对文本留言数据的分类；基于文本相似度的留言归类，根据留言点赞数、反对数、留言天数挖掘热点问题；以及对相关部门的答复质量评价等在内的更细节的文本信息的挖掘与认识，而这些结果对于微信、微博、市长信箱、阳光热线等网络问政平台都具一定的指导意义，使政府更加方便智能地了解民意、汇聚民智、凝聚民气，在留言群体与相关部门更直接有效沟通，从而促进解决相关问题，以提高人民幸福感，达到为人民谋幸福，为民族谋复兴目的。

但是从我们的建模过程发现其中的一些缺陷与不足，比如在分词过程中没有考虑分词精度，如果分词精度没有达到一定标准，这会对数据预处理以及后面的一些工作造成误差。其次TF-IDF算法提取关键词严重依赖语料库，需要选取质量较高且和所处理文本相符的语料库进行训练。另外，对于IDF来说，它本身是一种试图抑制噪声的加权，本身倾向于文本中频率小的词，这使得TF-IDF算法的精度不高。在计算可解释性得分时提取的法律法规可能有错误。层次分析法模型在专家评分法的过程中赋值会存在主观因素。以上涉及的四个方面的问题也是我们在后期进一步对中文文本数据研究过程中可以继续深入探讨的地方。

参考文献

[1] 董静. 基于主题模型和聚类算法的网络热点话题发现[D].河北大学,2019.

[2] 张国锋. 在文章聚类中话题热度排序的研究与实现[D].东华大学,2019.

[3] 李情情. 基于话题热度的微博推荐算法研究[D].山东师范大学,2016.

[4] 伊秀娟. 基于LDA主题模型的高校新闻话题发现研究[D].北京交通大学,2019.

[5] 陈龙. 新闻热点话题发现及演化分析研究与应用[D].南京理工大学,2017.

[6] 炎士涛. 基于词频统计的文本分类模型研究[D].上海师范大学,2007.

[7] 吴柳,程恺,胡琪. 基于文本挖掘的论坛热点问题时变分析[J].软件,2017,38(04):47-51.

[8] 陈珊珊. 基于LDA模型的文本聚类研究[D].苏州大学,2017.

[9] 曹茂林. 层次分析法确定评价指标权重及Excel计算[J].江苏科技信息,2012(02):39-40.