基于机器学习的“智慧政务”文本挖掘

摘 要

近年来，各种网络问政平台成为了政府了解民意、解决问题的重要渠道，各类民情相关的文本数据量也在不断攀升，随着大数据、人工智能等方面的发展，建立自然语言处理技术来提升政府的管理水平是社会创新发展的新趋势。本文基于机器学习方法建立模型，将大量留言文本进行标签自动分类、挖掘其中的热点问题、设立评价系统自动评价答复意见质量情况。

对于问题一，使用 R 3.6.3 版本中 jiebaR 包对样本的留言主题、详情进行分词，基于 TF-IDF 算法提取 7 个一级标签对应留言主题、详情中 tf-idf 值排名前 10%的词汇分别建立成对应的标签类别词库，基于该词库计算所有样本在各标签类别下的得分权重，最终利用机器学习模型根据不同样本在各标签下的得分权重来判断最终的所属一级标签类别。本文通过三种机器学习模型（随机森林、ANN 人工神经网络、KNN）进行交叉检验、重复训练后，结果基于留言主题计算权重得分的模型中，KNN 表现最优，平均预测准确率为 85.66%，平均 kappa 值为 0.8281,随机森林为 85.57%(0.8269)，ANN 为 85.4%(0.8254)。同时，使用相同方法基于留言详情计算得分权重的模型中，KNN 的预测准确率提升到 88.58%，kappa 值提高到了 0.8631。通过循环迭代最终选定最优 k 值后，基于留言详情的 knn 模型最终全标签类别 F1 得分均值达到了 0.9211055。

对于问题二，通过非监督模型 LDA 主题建模进行主题自动聚类与先前基于留言详情建立的 KNN 标签判别模型来计算得到对应特征词的 TF-IDF 值，通过等权处理得到词汇的最终 TF-IDF 值，先通过贝叶斯平均对词频增量进行修正作为衡量热度评价指标，再结合人工调参词频与词汇 TF-IDF 值阈值来最终确定热点词，进行相关热点词索引进而进一步通过热点词组合提取到热点话题并计算出对应话题热度排名及话题总时间窗口上的热度变化趋势。

对于问题三，通过用户留言与相关部门对相应留言答复意见构建 DTM 矩阵，基于 DTM 矩阵计算用户留言与答复意见的余弦相似度来量化答复相关性，通过对应答复长度与全样本平均答复长度比值来衡量答复完整性，通过基于第一问的最优 KNN 标签判别模型分别对用户留言与对应答复意见进行标签判别，计算出对应属于城乡建设、环境保护、交通运输等七个标签类别的概率。计算用户留言与对应答复意见的各标签类别所属概率的绝对差值和$\lambda(\lambda \in [0,2])$后，选取得分计算曲线$e^{-\lambda}$来度量回复解释度得分。最终通过对相关性度量、完整性度量、解释度度量的量化结果的加权非线性组合来定义最终答复质量评价指数指标，最终实现答复质量自动评价系统。

关键字：文本挖掘；TF-IDF 算法；机器学习；LDA 主题建模；答复质量评价

"Smart Government Affairs" Text Mining Based on Machine Learning

Abstract

In recent years, various online political inquiry platforms have become an important channel for the government to understand public opinion and solve problems. The amount of text data related to various types of public sentiment is also increasing. Technology to improve the government's management level is a new trend of social innovation and development. This paper builds a model based on machine learning methods, automatically categorizes a large number of message texts, mines hot topics, and establishes an evaluation system to automatically evaluate the quality of the answers.

For question one, use the jiebaR package in the R 3.6.3 version to segment the sample message topics and details, and extract 7 first-level tags corresponding to the message topic and details in the top 10% of the vocabulary of the tf-idf value based on the TF-IDF algorithm to build the corresponding tag category thesaurus respectively, calculate the score weights of all samples under each tag category based on the thesaurus, and finally use the machine learning model to determine the final first-level tag category according to the score weights of different samples under each tag . In this paper, after cross-checking and repeated training through three machine learning models (random forest, ANN artificial neural network, KNN), the results of the model that calculates the weight score based on the subject of the message, KNN performs best, and the average prediction accuracy rate is 85.66%. The average kappa value is 0.8281, the random forest is 85.57% (0.8269), and the ANN is 85.4% (0.8254). At the same time, using the same method to calculate the score weight based on the details of the message, the prediction accuracy of KNN is increased to 88.58%, and the kappa value is increased to 0.8631. After the optimal k value is finally selected through loop iteration, the average value of the full label category F1 score of the knn model based on the details of the message reaches 0.9211055.

For the second problem, the topic automatic clustering through the unsupervised model LDA topic modeling and the KNN label discriminant model previously established based on the message details are used to calculate the TF-IDF value of the corresponding feature word, and the final TF-IDF of the vocabulary is obtained by equal weight processing TF-IDF value, the Bayesian average is used to modify the word frequency increment as a measure of heat evaluation, and then combine the frequency of manual parameter adjustment and the threshold of TF-IDF value of vocabulary to finalize hotspot words, and extract related hotspots to further extract hotspot topics through the combination of hotspot words and calculate the hotness ranking of corresponding topics and the trend of hotness change in the total time window.

For question three, construct a DTM matrix based on the user message and the relevant department ’s response to the corresponding message. Calculate the cosine

similarity between the user message and the response opinion based on the DTM matrix to quantify the relevance of the response. To measure the completeness of the responses, the optimal KNN label discrimination model based on the first question is used to label the user's message and the corresponding response opinion respectively, and the probability of corresponding to the seven label categories of urban and rural construction, environmental protection, and transportation is calculated. After calculating the sum of absolute difference λ (λ ∈ [0, 2]) of the probabilities of each tag category corresponding to the user's message and the corresponding reply opinion, the score calculation curve $e^{-\lambda}$ is selected to measure the reply interpretation score. Finally, the final response quality evaluation index is defined by weighted nonlinear combination of the quantized results of correlation measurement, completeness measurement, and interpretation degree measurement, and finally an automatic response quality evaluation system is realized.

Keywords: Text mining; TF-IDF algorithm; Machine learning; LDA topic modeling; Reply quality evaluation

目录

1 挖掘背景与目标

近年来，随着微信、微博等网络问政平台逐渐成为了政府了解民意的重要渠道，各类社情相关的文本数据量不断攀升，给以往依靠人工来进行留言分类和热点整理的部门带来了极大挑战。

网络问政平台大量的文本留言造成了工作效率低下、人工错判率较高等问题，为了解决这一问题，笔者首先利用 jieba 库对文本进行分词、TF-IDF 算法提取词汇并自定义七个类别标签下的主题词库，然后对比随机森林、ANN 人工神经网络、KNN 三个机器学习模型对文本分类的准确率、kappa 值、F1 得分，最终选择最优模型。标签判别最优模型可以很好地辅助解决留言分类的问题，只需提供留言内容模型就能够自动计算该内容在不同类别标签下的得分权重进而自动判断出最终标签类别，可在短时间内处理海量文本数据，大大提高了工作效率。

其次，通过非监督聚类 LDA 主题建模进行自动主题聚类与先前基于留言详情提取的 KNN 一级标签判别模型来计算得到对应特征词的 TF-IDF 值，然后利用贝叶斯平均对词频增量进行修正作为衡量热度评价指标，再结合人工调参词频与词汇 TF-IDF 值来最终确定热点词，进行相关热点词索引进而进一步通过热点词组合提取到热点话题并计算出对应话题热度排名及总时间窗口上的热度变化趋势。这一模型可以解决热点问题的排序，列出某段时间内居民受影响最大、反映最大的热点问题，让热点问题得到及时解决。

最后，度量文本回复意见的相关性、完整性、解释性，基于 DTM 矩阵计算用户留言与答复意见的余弦相似度来量化答复相关性，通过对应答复长度与全样本平均答复长度比值来衡量答复完整性，以用户留言与对应答复意见的各标签类别所属概率的绝对差值和$\lambda(\lambda \in [0,2])$，选取得分计算曲线 $e^{-\lambda}$来度量回复解释度。最终通过对相关性度量、完整性度量、解释度度量的量化结果的加权非线性组合来定义最终答复质量评价指数指标。本文建立这一答复质量评价指标，对答复意见进行评价，给出了一套评价方案，最终实现答复质量自动评价系统。

在利用机器学习、非监督学习等方法建立模型后，将有一套笔者建立的系统，可以直接进行文本分类、热点问题的挖掘、回复意见的评价，解决了部门人工分类效率低、错误率高等问题。

2 分析方法与过程

2.1 总体流程图

总体流程图见图 1。

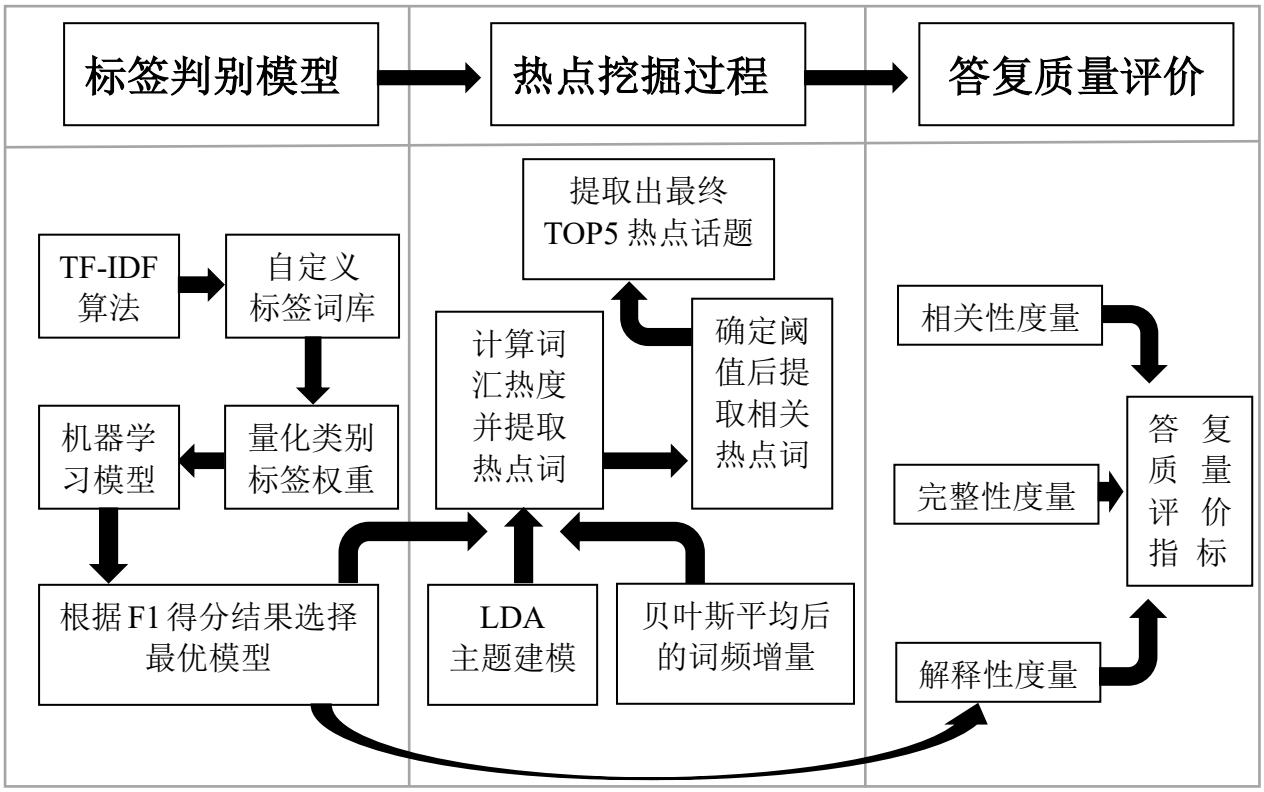


图 1 建模总体流程图

本文主要包括三大建模过程：

（1）标签判别模型：通过对文本清理、去除停用词、设置自定义词库后进行分词，使用 tf-idf 算法得到各标签下词库的 tf-idf 值，通过设定阈值提取各标签下达到预设阈值的高 tf-idf 词汇建立各个标签分类的词库。通过词库量化样本中留言主题、留言详在各标签词库下的权重得分，基于此权重得分进一步结合机器学习模型来建立标签判别模型。

（2）热点挖掘模型：通过标签判别模型与 LDA 主题建模结合贝叶斯平均后的词频增量来建立热度评价指标，通过设定阈值提取全样本中的高热度词汇，并通过重复迭代提取相关热点词组成热点话题，最终索引出组合热度最高的前 5 话题。

（3）答复质量评价模型：通过留言与答复的余弦相似度度量留言相关性；通过独立样本答复字数与全样本平均答复字数度量留言完整性；通过标签判别模

型分别对留言与答复的各标签判定概率的绝对差值和来度量解释性，最终通过对相关性、完整性、解释性结果的加权非线性组合来定义最终答复质量评价指标。

2.1 问题 1 分析方法与过程

2.1.1 问题 1 建模流程图

问题 1 建模流程图见图 2。

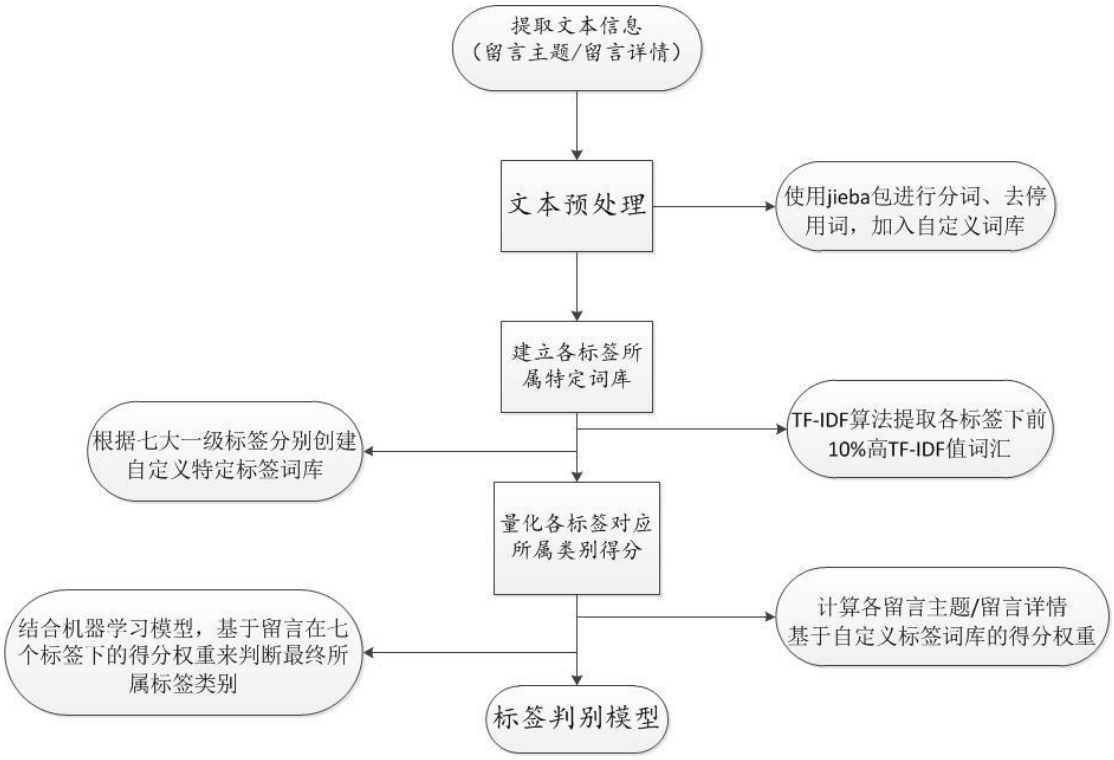


图 2 问题一建模流程图

在处理网络问政平台的群众留言时，为将各类型不同的群众留言分配到各职能部门，会进行一级标签的人工分类，分为城乡建设、环境保护等七个类别，但人工分类有效率低、错误率高且工作量大等问题存在，所以笔者通过机器学习建立高效的标签判别模型处理这一问题，进行留言内容的一级标签自动分类判别，使该问题得到改善。

首先，我们将留言文本使用 *R 3.6.3* 软件进行文本预清洗，去除停用词（中文停用词词库：见附件 *cn_stopwords.txt*），然后使用 *jiebaR* 包进行对样本的留言主题进行分词，jieba 分词基于前缀词典能够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)，通过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。其中对于未登录词，结合 Viterbi 算法采用了基于汉字成词能力的 HMM 模型。对各样本的留言主题进行分词结束后利用 TF-IDF 算法计算出各一级标签背景下词汇的 TF-IDF 值来衡量不同一级标签下不同词汇的相对重要程度以便来提取各一级标签的特征词。我们基于各一级标签样本留言主题中分词后各词汇的 TF-IDF 值排名前 10%的词汇分别建立了七个一级标签的自定义词库，并基于所建立的词库对样本进行类别权重计算。

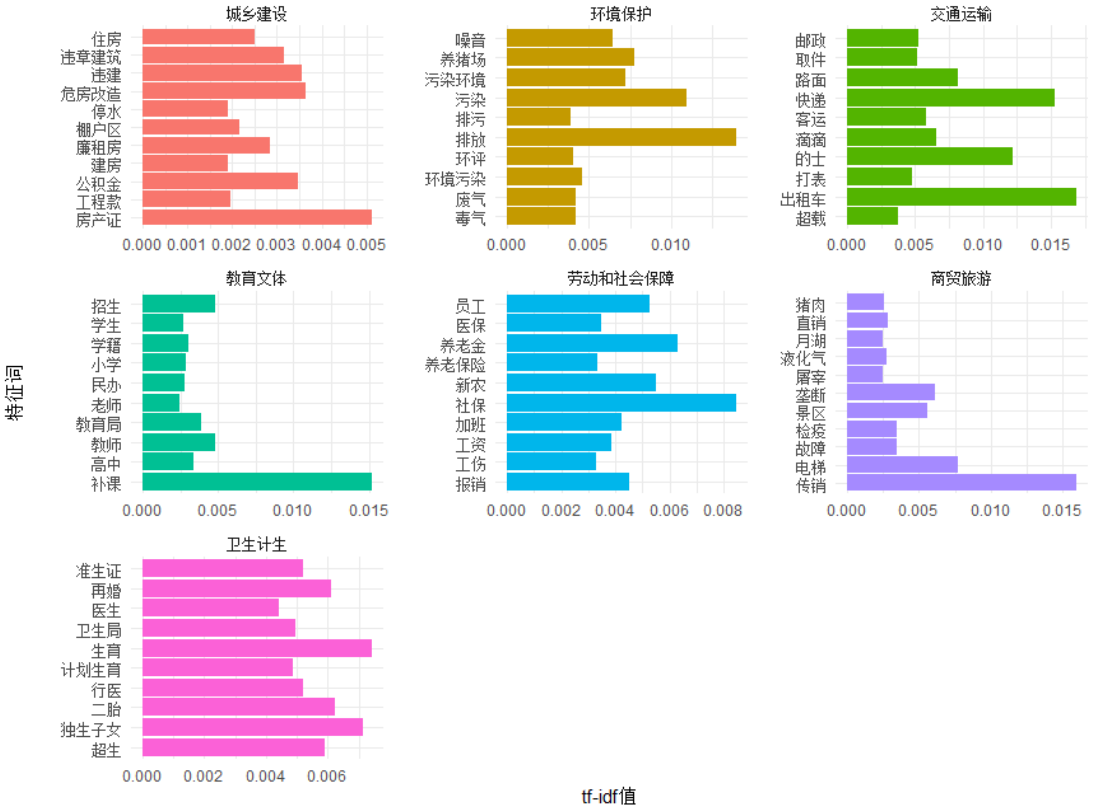


图 3 基于留言主题的各一级标签的排名前 10 的 TF-IDF 词汇

由图 3 可以看出在城乡建设标签下，“房产证”这一词汇 TF-IDF 值最高，同时“住房”、“危房改造”、“违建”、“违章建筑”也具有较高的 TF-IDF 值，表明这类词汇更易出现在城乡建设标签类别中；在环境保护标签下，“排放”、“污染”等词汇 TF-IDF 值在此标签内较高，超过 0.01；在交通运输标签下，“出租车”、“快递”词汇 TF-IDF 值在所有标签中较高，超过 0.015；在教育问题标签下，除了“补课”这一词汇 TF-IDF 值超过了 0.015，其他词汇 TF-IDF 值显著的低于“补课”这一词汇，皆在 0.005 以下；在劳动合同社会保障标签下，“社保”、“养老金”词汇 TF-IDF 值相对更高，都是很受关注的社会保障问题；在商贸旅游标签下，“传销”的 TF-IDF 值超过 0.015，显著超过了其他词汇的 TF-IDF 值，说明在商贸旅游标签

下反映“传销”相关问题的留言内容较多；在卫生计生标签下，TF-IDF 值没有特别突出的词汇，TF-IDF 值皆在 0.006 左右。TF-IDF 值越高说明该词汇越重要，在对应分类标签下出现的频率、比重相对于其他分类标签更高，所以利用 TF-IDF 值排名前 10%的词汇分别建立了七个一级标签的自定义词库进而来量化留言内容在各类标签下的得分权重。

2.1.2 TF-IDF 算法

TF-IDF 算法常用于咨询检索和信息勘探，是一种提取文档关键词汇的统计方法，通常用来评估特定字词某语料库中一份文件或在某文件集的重要程度[1]。字词的重要程度与其在文件中的出现频率呈正比，而与其在语料库中的出现频率成反比。

（1）TF 是词频（term frequency）

词频的含义是指某一特定词在指定文档中出现的频率值，其计算公式为该特定词在文档中出现次数除以所有词在文档中出现次数。

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中 $n_{i,j}$ 是该词在文件中出现的次数，$\sum_k n_{k,j}$ 则是文件中所有词汇出现的次数总和。其中分母的目的在于将词数进行归一化防止偏向长的文档，不管该词语重要与否，同一个词语在长文档里可能会比短文件有更高的词数。

（2）IDF 是逆向文件频率（inverse document frequency）

逆向文件频率用于度量一个词语普遍重要性，其计算公式为总文档数量除以包含该词的文件数目，将所得到的商取对数即为结果。

$$idf_i = \log \frac{|D|}{1+\left|\left\{ j : t_i \in d_j \right\}\right|}$$

其中，$|D|$ 是语料库中的文件总数。$\left|\left\{ j : t_i \in d_j \right\}\right|$ 表示包含词语 t_i 的文件数目（即 $n_{i,j} \neq 0$ 的文件数目）。如果该词语不在语料库中，就会导致分母为零，因此一般情况下使用（$1+\left|\left\{ j : t_i \in d_j \right\}\right|$）。

（3）TF-IDF 是 TF×IDF

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

$$TF-IDF=tf_{ij}\times idf_i$$

本文首先基于各一级标签 TF-IDF 值排名前 100 的词汇分别建立七个自定义词库，并基于所建立的词库对样本进行评分，其计算方法：若样本m中出现的词汇x_i在j_i词库中存在，则对应的类别权重得分+1。

权重评分计算函数关键 R 代码：

```
fun <- function( x, y) x%in% y
getScoreType <- function( x,awords,bwords,cwords,dwords,ewords,fwords,gwords){
  a.weight = sapply(llply( x,fun,awords),sum)
  b.weight = sapply(llply( x,fun,bwords),sum)
  c.weight = sapply(llply( x,fun,cwords),sum)
  d.weight = sapply(llply( x,fun,dwords),sum)
  e.weight = sapply(llply( x,fun,ewords),sum)
  f.weight = sapply(llply( x,fun,fwords),sum)
  g.weight = sapply(llply( x,fun,gwords),sum)
  return(data.frame( a.weight, b.weight, c.weight, d.weight, e.weight, f.weight, g.weight))
}
```

表 1 标签词库及对应权重得分表

词库类型	一级标签	权重得分
城乡建设词库	城乡建设	*a.weight*
环境保护词库	环境保护	*b.weight*
交通运输词库	交通运输	*c.weight*
教育文体词库	教育文体	*d.weight*
劳动和社会保障词库	劳动和社会保障	*e.weight*
商贸旅游词库	商贸旅游	*f.weight*
卫生计生词库	卫生计生	*g.weight*

通过在不同词库类型下的权重得分来判断最终的样本一级标签类型，由于各样本留言主题在各一级标签下的权重得分有时会出现得分较为均衡的情况，仅使用最大的权重得分来判断对应类别缺乏正确性与科学性，因此采用机器学习的统计方法对样本进行训练学习，旨在基于不同一级标签在各标签下的权重得分的分

布情况来判断最终该样本的一级标签类别。

本文使用随机森林、ANN 人工神经网络、KNN 算法三种方法进行对比，通过数据集分割（按 7:3 比例分割成训练集与测试集）后使用三种不同模型进行交叉验证（Cross Validation），最终使用 F-Score 对分类方法进行评价，以此来确定最优模型（自变量为样本留言主题中的词汇在不同一级标签下的对应权重得分，因变量为样本的一级标签类别）。

2.1.3 随机森林模型

随机森林（Random Forest）用于执行回归和分类任务，是一种多功能的机器学习算法。在随机森林中，将生成很多的决策树，当基于某些特征对一个新的对象进行分类判别时，随机森林中的每一棵树都会给出自己的分类选择，并由此进行加权，森林整体的输出结果将会是权数最高的分类选项。

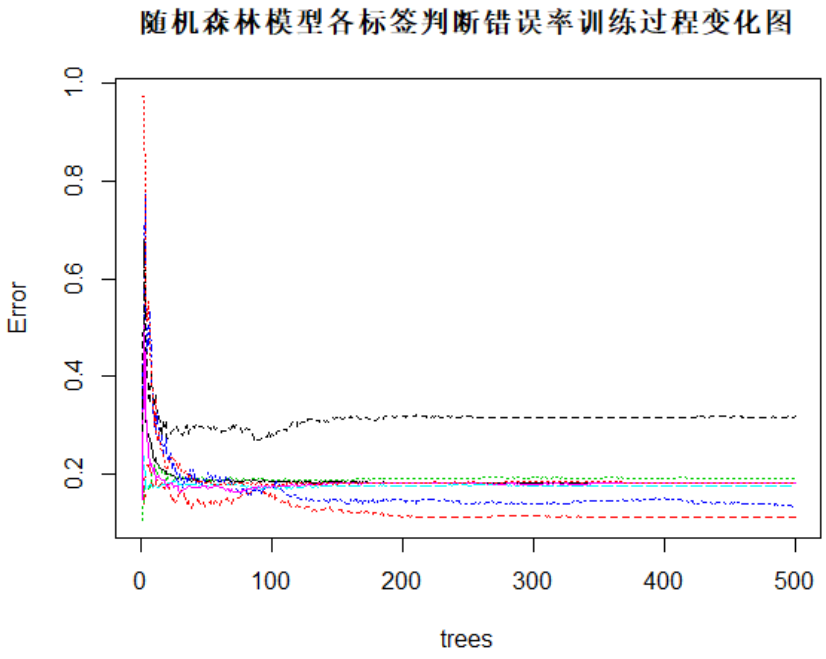


图 4 随机森林模型各标签判断错误率训练过程变化图

由图 4 可以发现通过不断增加树的个数，各一级标签的预测错误率将越来越小直至约迭代到 200 棵树后开始收敛。由于计算量过大，本文仅采用基于留言主题进行各标签类别得分权重计算下且树的个数为 500 的随机森林模型进行进一步研究。

将文本对应七个词库的加权得分作为决策树的特征、给定的一级标签作为决策结果，建立随机森林模型，最终可得到随机森林模型 1 取各一级标签的前 100 个高 TF-IDF 词汇的预测结果混淆矩阵（如下表 2）。

表 2 随机森林模型 1 预测结果混淆矩阵表

预测 / 实际	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	卫生计生	判别错误率
城乡建设	1780	79	32	40	21	51	6	0.1139871
环境保护	157	757	1	8	4	10	1	0.1929638
交通运输	40	6	532	3	5	20	7	0.1321370
教育文体	191	4	6	1305	70	8	5	0.1787288
劳动和社会保障	233	12	11	43	1610	12	48	0.1823261
商贸旅游	292	17	29	26	4	826	21	0.3201646
卫生计生	79	3	3	9	55	10	718	0.1812999

由表 2 可以看出，该模型的 OOB (Out-of-bag) error，即包外误差为 18.26%，且可看出在随机森林模型建立的模型 1 中，一级标签商贸旅游错误率最高，城乡建设的错误率最低。

其中，$OOB\ error = \sum_1^N \frac{Erri_i}{N}_i$，$N$为样本数，对于一个样本$x_i$，没有使用$x_i$训练的数组成的小随机森林来预测$x_i$，其预测误差即在整个随机森林模型中的$OOB\ Error$，即$Erri_i$。通常来说$OOB\ Error$会比交叉验证的误差更大，因为只选用了随机森林中的部分树，没有完整使用模型，限制了模型的发挥。但省去了在交叉验证中需要多次训练的庞大计算量，更加高效。

由于各标签中的高 tf-idf 总数不尽相同，都取定相同个数的词汇数作为阈值忽视了部分标签中词汇总数的影响，以下取各一级标签的 top10%高 TF-IDF 词汇进行随机森林模型 2 的建立。

表 3 随机森林模型 2 预测结果混淆矩阵表

预测 / 实际	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	卫生计生	判别错误率
城乡建设	1776	35	11	37	51	86	13	0.1159781
环境保护	68	809	2	16	15	26	2	0.1375267
交通运输	51	1	521	6	7	23	4	0.1500816
教育文体	44	5	6	1385	91	41	17	0.1283826
劳动和社会保障	88	10	13	61	1689	34	74	0.1422042
商贸旅游	165	8	34	19	22	946	21	0.2213992
卫生计生	17	4	1	13	63	29	750	0.1448119

由表 3 可以看出，模型 2 的 OOB error 为 14.48%，相比于模型 1，模型 2 的准确率更高，说明以比例来取定各标签词汇阈值的方式更加合理有效。最优阈

值可通过循环迭代找到最低错判率的词库作为提取词量的阈值，由于计算量过大，这里不进行进一步探究，仅以 10%设定为默认阈值。

2.1.4 ANN 人工神经网络模型

人工神经网络（Artificial Neural Networks，ANN）是一种模拟人脑思维的计算机模型，人工神经网络由相互连接的神经元组成，该神经元称为节点，节点之间的连接称为边，关联性的强弱体现在边的连接权重上，其中 H 为隐层，I 为输入层，0 为输出层。

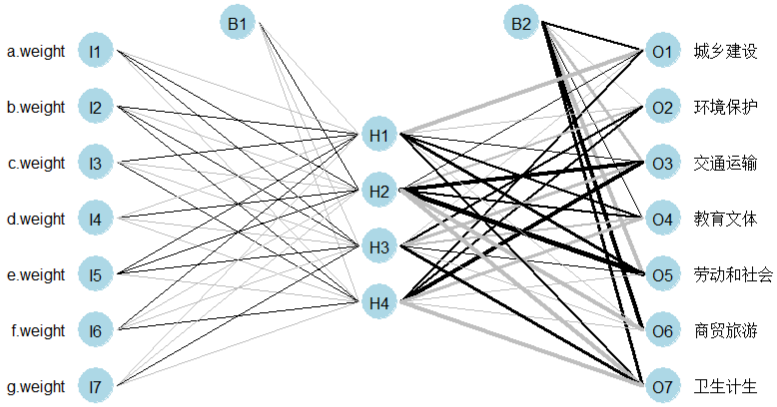


图 5 人工神经网络模型结果可视化

其中输入特征是样本在各类标签词库下的权重得分，B 为偏置参数。在四个隐藏层中通过样本在 7 类标签词库下的得分权重特征进而划分出最终的输出层作为最终的判定结果。

表 4 ANN 人工神经网络模型预测结果混淆矩阵表

预测 / 实际	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	卫生计生	判别错误率
城乡建设	1696	46	38	35	78	90	7	0.147739
环境保护	57	824	1	5	15	15	4	0.10532
交通运输	19	2	517	4	11	25	1	0.107081
教育文体	38	8	5	1427	98	25	11	0.114764
劳动和社会保障	36	16	15	48	1642	12	61	0.102732
商贸旅游	151	37	37	49	46	1025	44	0.262059
卫生计生	12	5	0	21	79	23	749	0.15748

由表 4 可以看出，ANN 人工神经网络模型的全类别平均判别错误率为14.24%，与上文的随机森林模型 2 的预测准确率接近，同样对于商贸旅游类别的判别错误率最高，为 26.2%，ANN 模型在除城乡建设、卫生计生的标签类别判别准确率都优于随机森林模型 2。

2.1.5 KNN 模型

KNN（K Nearest Neighbors）算法基本原理就是当预测一个新的值X的时候，根据它距离最近的K个点是什么类别来判断X属于哪个类别。将样本包含的n个观测数据看成p维特征空间中的点，并根据X_0的K个近邻$(y_1, y_2, \cdots, y_k)$依函数F计算y_0，对于此类多分类的问题来说，函数$F: y_0 = P(y_0 = m|X)$，即类别分别取城乡建设、环境保护、交通运输等七个类别的概率，预测值为最大概率值对应的m。

（1） 距离计算

使用常见的 KNN 算法中的欧氏距离，拓展到多维空间则距离：

$$d_{(x_i,x_j)} = \sqrt{\sum_{l=1}^{n}\left|x_i^{(l)} - x_j^{(l)}\right|^2}$$

其中 $x_i=\left(x_i^{(1)},\ x_i^{(2)},\cdots x_i^{(n)}\right)^T, x_j=\left(x_j^{(1)},\ x_j^{(2)},\cdots x_j^{(n)}\right)^T$

本文中n为 7，当获得某一待测值X属于城乡建设、环境保护、交通运输等七个类别的权重得分后,由此可算出某两观测值间的距离，即可观测出距离待测值X较近的一些点于哪一类别的概率更高，则认为该观测值为该类别。

（2） 分类决策规则

根据给定的距离度量，在训练集中寻找与X最近邻的k个点，涵盖这k个点的x的邻域中根据分类决策规则决定x的类别y:

$$y = \arg\max_{c_j} \sum_{x_i \in N_k(x)} I\left(y_i = c_j\right), i = 1, 2, \cdots, N; j = 1, 2, \cdots, K$$

其中I为指数函数，当$y_i = c_j$时I为1，否则I为0

同时分类函数为：

$$f: R^n \to \{c_1, c_2, \cdots, c_K\}$$

正确分类的概率为：

$$P(Y = f(X))$$

要使正确分类率最高， 就要使 $\sum I(y_i = c_j)$最大，即要最小化经验风险[2]。

（3） K值选择

k值的选择会对 KNN 的结果产生巨大影响。k值的减小模型整体就会变得复杂，易发生过拟合问题。若选择较大的k值，就相当于用较大邻域中的训练实例进行预测。其优点是可以减少学习的估计误差。但缺点是学习的近似误差会增大。k值的增大就意味着整体模型变得更加简单。本文采用交叉验证的方法，将循环代码带入，观测k值从 1 至 60 的 F1 得分，选取全类别平均 F1 得分最高的k值，确定后进行模型的建立。

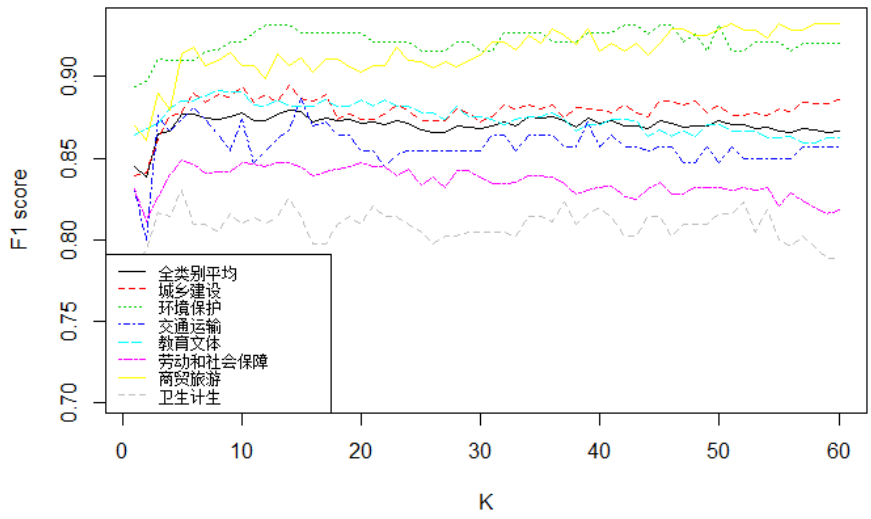


图 6 交叉验证K值 1 至 60 的各标签类别 F1 得分变化图

通过循环迭代、交叉验证K值进一步分析通过K值的选取对各标签 F1 得分的影响，最终选择全类别平均（图中黑色实现）F1 得分最高时的K值作为最优K值。从图 6 中可以发现，绿色点虚线的环境保护标签的 F1 得分相对其他标签的 F1 得分在K值迭代变化过程中始终处于较高的水平，同时卫生计生标签的 F1 得分始终处于最低水平。

交叉验证 K 值 1 至 60 的各标签类别 F1 得分 R 代码：

```
miss <- data.frame('K'=1:60, '全类别平均'=rep(0,60)) # New data frame to store the error value
whichk=c()
for (k in 1:60){
  whichk<-knn(train = data_train2, test = data_test2, cl = label_train2, k=k)
  miss[k,'全类别平均'] <- mean(confusionMatrix(whichk, label_test2)$byClass[,'F1'])
  miss[k,'城乡建设'] <- confusionMatrix(whichk, label_test2)$byClass[1,'F1']
  miss[k,'环境保护'] <- confusionMatrix(whichk, label_test2)$byClass[2,'F1']
  miss[k,'交通运输'] <- confusionMatrix(whichk, label_test2)$byClass[3,'F1']
  miss[k,'教育文体'] <- confusionMatrix(whichk, label_test2)$byClass[4,'F1']
  miss[k,'劳动和社会保障'] <- confusionMatrix(whichk, label_test2)$byClass[5,'F1']
  miss[k,'商贸旅游'] <- confusionMatrix(whichk, label_test2)$byClass[6,'F1']
  miss[k,'卫生计生'] <- confusionMatrix(whichk, label_test2)$byClass[7,'F1']
}
head(miss,15)
which(miss$全类别平均==max(miss$全类别平均)) #最优 k 值
plot(miss$K,miss$全类别平均,xlab="K",ylab="F1 score",ylim=c(min(miss[,-1])-0.08,max(miss[,-1])),type='l') #作图
for(i in 3:9) lines(miss$K,miss[,i],type='l',col=i-1,lty=i-1)
legend("bottomleft",col=1:8,lty=1:8,legend = names(miss[,2:9]),cex=0.7)
```

表 5 不同 K 值下的 F1 得分表（前 15 个 k）

类别 / K	F1 Score							
	全类别平均	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	卫生计生
1	0.8444875	0.8395062	0.8936170	0.8305085	0.8637874	0.8312958	0.8695652	0.7831325
2	0.8386898	0.8407960	0.8969072	0.8000000	0.8684211	0.8109453	0.8606557	0.7931034
3	0.8646409	0.8613861	0.9109948	0.8760331	0.8716216	0.8260870	0.8897959	0.8165680
4	0.8667146	0.8762376	0.9100529	0.8666667	0.8805461	0.8400955	0.8790323	0.8143713
5	0.8772374	0.8780488	0.9100529	0.8739496	0.8851351	0.8487805	0.9142857	0.8304094
6	0.8772166	0.8899756	0.9100529	0.8813559	0.8851351	0.8467153	0.9180328	0.8092486
7	0.8742289	0.8845209	0.9157895	0.8739496	0.8881356	0.8410758	0.9068826	0.8092486
8	0.8739925	0.8894349	0.9166667	0.8644068	0.8911565	0.8418491	0.9098361	0.8045977
9	0.8752505	0.8872549	0.9214660	0.8547009	0.8904110	0.8418491	0.9149798	0.8160920
10	0.8776061	0.8938272	0.9214660	0.8739496	0.8904110	0.8474576	0.9068826	0.8092486
11	0.8726057	0.8839506	0.9270833	0.8474576	0.8827586	0.8461538	0.9068826	0.8139535
12	0.8729106	0.8888889	0.9319372	0.8547009	0.8819444	0.8448687	0.8987854	0.8092486
13	0.8771712	0.8845209	0.9319372	0.8620690	0.8858131	0.8476190	0.9142857	0.8139535
14	0.8794410	0.8948655	0.9319372	0.8672566	0.8819444	0.8476190	0.9068826	0.8255814
15	0.8789709	0.8866995	0.9270833	0.8869565	0.8819444	0.8448687	0.9112903	0.8139535

最终结论可以得出，K 在取值 1-60 的迭代过程中，取 14 时对应 KNN 模型的全类别平均 F1 score 最高，因此选择$k = 14$作为最优k值。

2.1.6 模型对比

通过上述基于主题留言的各标签权重得分计算结合三种机器学习模型随机森林、ANN 人工神经网络和 KNN 算法的测试结果，通过 F-Score、精准率和召回率作为标签分类性能的评估来对比三种模型[3]。

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

其中，P_i为第i类的查准率（Precision，也称精准率），R_i为第i类的查全率（Recall，也称召回率）。

表 6 三种机器学习模型标签判别结果比较

模型 / 分类	随机森林			ANN 人工神经网络网络			KNN		
	F1	*Precision*	*Recall*	*F1*	*Precision*	*Recall*	*F1*	*Precision*	*Recall*
城乡建设	0.8694	0.8439	0.8964	0.8482	0.8522	0.8442	0.8550	0.8446	0.8656
环境保护	0.9103	0.9398	0.8827	0.8864	0.8946	0.8784	0.9213	0.9761	0.8723
交通运输	0.8936	0.9103	0.8776	0.8674	0.8929	0.8433	0.8636	0.8028	0.9344
教育文体	0.9030	0.9248	0.8823	0.8915	0.8852	0.8980	0.8896	0.9194	0.8616
劳动和社会保障	0.8878	0.8914	0.8842	0.8644	0.8972	0.8339	0.8636	0.8592	0.8680
商贸旅游	0.8069	0.7937	0.8205	0.7872	0.7379	0.8436	0.7854	0.7760	0.7950
卫生计生	0.8690	0.8680	0.8700	0.8482	0.8425	0.8540	0.8409	0.8409	0.8409

通过三种模型的平均 F1 得分对比，可以发现随机森林的全标签类别平均 F1 得分为 0.8771 高于 KNN 模型的 0.8599 和 ANN 模型的 0.8562，但随机森林模型在实际应用中往往由于其计算量过大，在设备性能不足的情况下会影响工作效率，且仅通过模型的 F-Score、精准率和召回率来选择最优模型仍有不足，以下进一步通过三种模型进行重复训练、交叉验证后对其模型的预测准确率及 kappa 值来进一步判定最优模型。

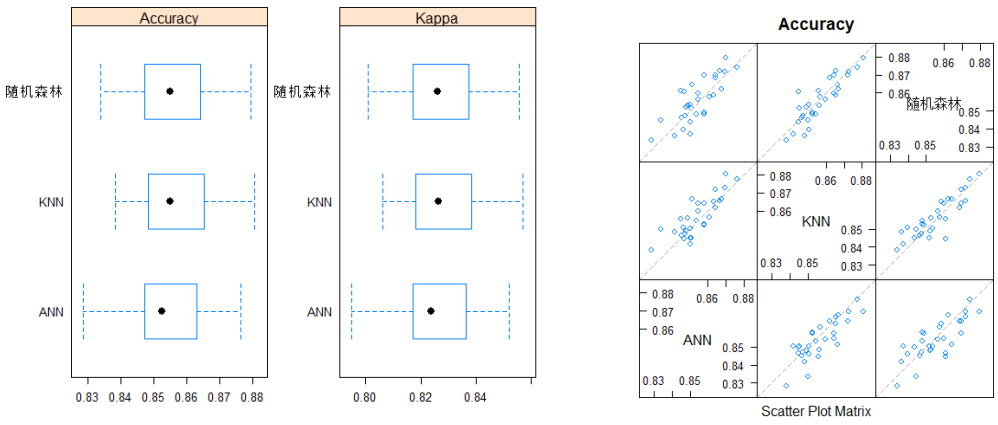


图 7 三种模型重复三次训练后进行 10 折交叉检验的预测结果对比图

从图 7 可以发现，从预测准确率（Accuracy）来看，多次训练模型后进行 10 折交叉验证的结果表明，KNN 的预测平均准确率与预测准确率最大值、最小值都略微优于其他两个模型；从 kappa 一致性系数来看，KNN 的平均、最大、最小 kappa 值都大于其他模型，这说明 KNN 模型明显优于其他两个模型，因此认为 KNN 模型更优。

2.1.7 最优模型

以上模型都是基于留言主题进行样本的标签类别权重得分计算，虽然留言主题涵盖了样本反馈内容的主体，但相对于留言内容详情会缺失很多必要的细节信息，于是下面基于留言详情进行探究。

图 8 基于留言详情的各一级标签的排名前 10 的 TF-IDF 词汇

在上述基于留言主题计算样本各标签得分权重的 KNN 模型中，相对另外的随机森林模型与 ANN 人工神经网络模型在预测准确率和 kappa 值上有着更大的优越性，于是本文基于 KNN 模型将进一步分别对留言主题和留言内容进行建模的以预测准确率和 kappa 值进行比较，得到如下结果图 9。

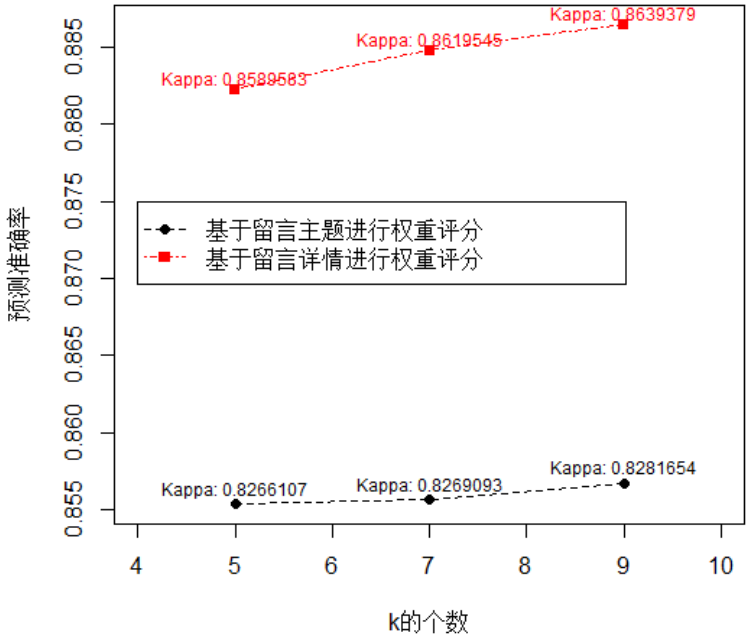


图 9 分别基于留言主题、留言详情进行权重评分的 knn 模型预测准确率与 kappa 值对比图

这里仅通过 k 取值为 5、7、9 的 KNN 模型来对比留言详情与留言主题分别进行权重评分对于模型预测准确率和 kappa 值的影响。从图中可以明显看出基于留言详情进行权重评分的 KNN 模型更优，因此可以认为基于留言详情的权重评分 KNN 模型为最优。其中基于留言内容的 KNN 模型（k=9）各参数值结果如下表 7。

表 7 基于留言详情的权重评分 knn 模型结果参数值（k=9）表

参数 / 标签类别	*Sensitivity*	*Specificity*	*Precision*	*Recall*	*F1*	*Prevalence*	*Detection Rate*	*Balanced Accuracy*
城乡建设	0.8257840	0.9586169	0.8477261	0.8257840	0.8366112	0.21813246	0.18013029	0.8922004
环境保护	0.9285714	0.9892408	0.9072917	0.9285714	0.9178082	0.10184582	0.09457112	0.9589061
交通运输	0.8075041	0.9906944	0.8608696	0.8075041	0.8333333	0.06655809	0.05374593	0.8990993
教育文体	0.8986784	0.9790054	0.8992443	0.8986784	0.8989613	0.17252986	0.15504886	0.9388419
劳动和社会保障	0.9258507	0.9569120	0.8538642	0.9258507	0.8884016	0.21378936	0.19793702	0.9413814
商贸旅游	0.8329218	0.9822389	0.8769497	0.8329218	0.8543689	0.13192182	0.10988056	0.9075804
卫生计生	0.8643101	0.9900396	0.9013080	0.8643101	0.8824214	0.09522258	0.08230185	0.9271749

可以看出，基于留言详情的权重评分且 k 取值为 9 的 knn 模型全类别平均 F1 得分为 0.8731，以下将进一步选取最优 k 值（k=14）且基于留言详情的权重评分的 knn 模型作为最终模型并计算其模型结果。

表 8 基于留言详情的权重评分 knn 模型结果表

模型 / 标签分类	全样本学习模型结果（k=14）	70%训练集 30%测试集 模型结果（k=14）	70%训练集 30%测试集 模型结果（k=9）
城乡建设	0.9155313	0.8762542	0.8747913
环境保护	0.9451124	0.9295775	0.9195804
交通运输	0.9102668	0.8780488	0.8767123
教育文体	0.9312339	0.9030884	0.9020021
劳动和社会保障	0.9192173	0.8834459	0.8837998
商贸旅游	0.9131347	0.8772414	0.8729282
卫生计生	0.9132420	0.8715084	0.8587571
全类别平均 F1 值	0.9211055	0.8884521	0.8840816

从基于留言详情的权重评分且选取最优 k 值的 knn 模型结果可以看出，通过全样本学习的 knn 模型对样本回判的全标签分类平均 F1 得分已经高达了 0.9211，当然这很可能存在一定的过拟合问题，但同时经过数据分割进行交叉验证的 knn 模型的全标签分类平均 F1 得分也达到了 0.8884，其得分已经较为可观了，由此将该模型作为最优模型。

2.2 问题 2 分析方法与过程

网络问政平台在进行一级标签分类后，会将各类型不同的群众留言分配到各职能部门，在各职能部门处理相应的群众留言时，为助相关部门进行有针对性的处理、提升服务效率，本文将某段时间内反映特定地点或特定人群问题的留言进行聚类，利用贝叶斯平均对梯度分数进行修正，以此作为热度评价指标，挖掘出某时段群众集中反映的热点问题。首先对附件 3 基于留言详情使用第一问的最优 KNN 模型进行一级标签的判别后提取 TF-IDF 值。并使用 LDA 模型在未分类一级标签情况下进行归类判别[4]，计算对应的 TF-IDF 值并与基于最优模型判断的一级标签分类进行等权计算，最终得到最后的 TF-IDF 值，结合热度评价指标、TF-IDF 与词频去衡量热点词，基此来提取热点词与热点话题。

对于 Top N 热点话题的检测，该阶段主要分为三个步骤：第一步，首先基于文档关键特征的子话题进行聚类，将抽取的所有关键特征进行去重整合，并建立与文档的对应关系。然后，将关键特征映射到话题空间，建立初始话题。其次，将初始话题通过删去低频词汇，提取高 tf-idf 值词汇以降低后续子话题聚类的复杂度[5]。最后，通过词频与等权处理后的高 tf-idf 值词汇人工调参设定阈值来确

定热点词。

第二步，提取到热点词后，索引出所有包含热点词汇的留言内容，并分别对热点相关话题内容进行词频统计，同样通过词频与等权处理后的高 tf-idf 值词汇人工调参设定阈值来确定相关热点词[6]，找到多个个相关热点词组合出热点话题。

第三步，Top N 热点话题确定。该步骤在重复多次第二步的步骤后，直至提取出完整的且热度得分最高的热点话题。

2.2.1 LDA 主题建模

在信息急速膨胀的网络时代，有监督方法具有较低的可行性，一方面，标注训练集合是一件非常耗时耗力的工作，另一方面，时间变化的同时，文档主题也会发生剧烈的变化，在这种情况下，随时对训练集合进行标注是不现实的。与之相比，无监督方法具有不需要人工标注训练集合的优点，因此更加快捷，但同时也存在无法综合有效的利用多个信息对候选关键词进行排序的缺点，故综合来看有监督学习方法更胜一筹，因此有关于关键词抽取的研究主要集中于无监督方法方面。LDA（Latent Dirichlet Allocation）指的是一种文档主题的生成模型，或者称其为一个三层贝叶斯概率模型，其中包含文档、主题和词三层结构。生成模型的涵义为：一篇文章所有词均通过“以特定概率选择主题，并在此主题中以特定概率选择词语”这样的过程得到。文档到主题与主题到词都服从多项式分布。对于语料库中的每篇文档，LDA 定义了如下生成过程：首先，每一篇文档均从主题分布中抽取一个主题；然后，从被抽到的主题所对应的单词分布中抽取出一个单词；最后，重复上述过程直至遍历文档中的所有单词。

语料库中的每一篇文档与 T（通过反复试验等方法事先给定）个主题的一个多项分布相对应，将该多项分布记为 θ。每个主题又与词汇表中的 V 个单词的一个多项分布相对应，将这个多项分布记为 φ[7]。

其中 LDA 核心公式：

$$P(w|d) = P(w|t)P(t|d)$$

其中 w 为词，d 为文档，t 为主题。以主题为中间层，通过前面两个向量 θ 和 ϕ，分别给出 P(词 | 主题）和 P(主题 | 文档），其学习过程：

（1） 先随机地给 θ_d 、ϕ_t 赋值（对所有的 d 和 t ）。

（2） 针对特定文档 d_s 中的第 i 个单词 w_i，如果令该单词对应的主题为 t_j，则上述公式为：

$$P_j(w_i \mid d_s) = P(w_i \mid t_j)P(t_j \mid d_s)$$

（3） 得到所有的 $P_j(w_i \mid d_s)$，然后可以根据这些概率值的结果为 d_s 中的第 i 个单词 w_i 选择一个主题，此时本文取 $P_j(w_i \mid d_s)$ 概率最大的主题 t_j 。

（4） 如果 d_s 中的第 i 个单词 w_i 在这里选择了一个与原先不同的主题，则会对 θ_d 、ϕ_t 有影响，他们的影响反过来影响 $P(w \mid d)$ 的计算。

（5） 对文档集中所有文档中的所有单词 w 进行一次 $P(w \mid d)$ 的计算，并重新选择主题看成是一次迭代，迭代 n 次后可收敛到 LDA 所需的分类结果。

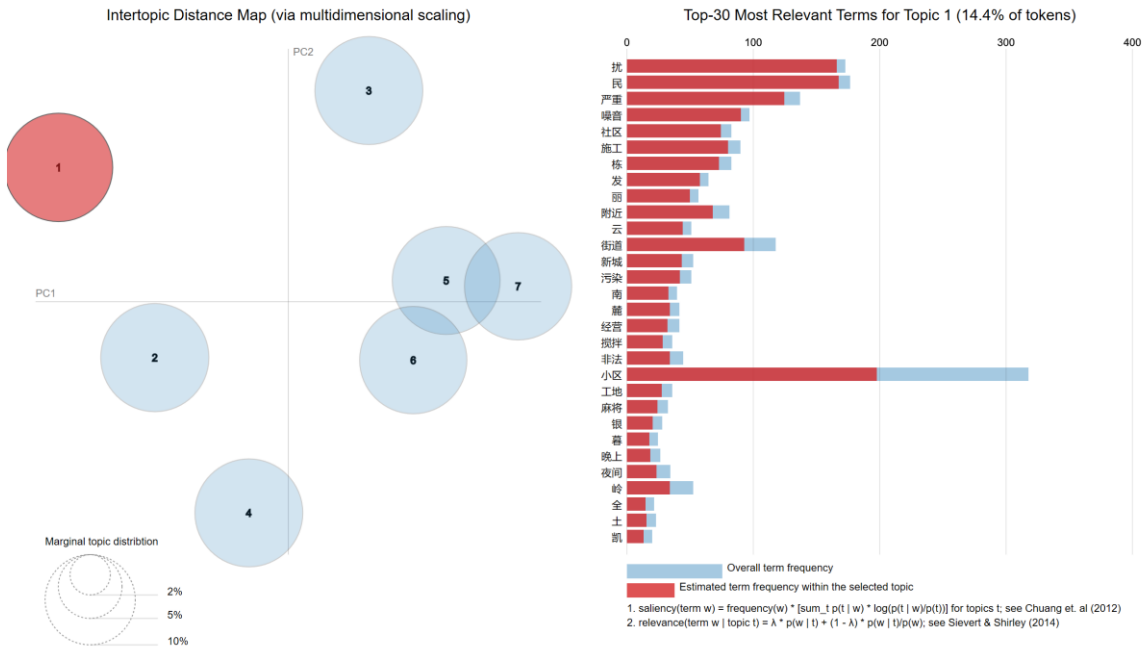


图 10 设定主题个数为 7 后用 LDA 主题建模进行主题生成结果图(λ = 0.1,以 Topic1 为例)

从图 10 可以发现在设定主题个数为 7 后进行 LDA 主题建模生成的主题 1 中提取了前 30 个与该主题最相关的词汇，“扰民”、“噪音”、“施工”等词汇在该主题下词频占比特别高，这接近初始原样本中的“环境保护”标签，一定程度上说明了该模型良好的主题生成效果。

2.2.2 热度评价指标

本文利用修正梯度分数作为热度评价指标，在贝叶斯平均的基础上进行修正，增加该热度评价指标的客观性、科学性。

（1）梯度是词频增量的主要衡量指标。未修正的梯度分数为：

$$S(w_i)=\frac{F(w_i,T_j)}{F(w_i,T_1,T_2,...,T_j)}$$

其中 w_i 表示某个词语，T_j 表示时间窗口，$F(w_i,T_j)$ 表示词语 w_i 在时间窗口 T_j 的出现频次，$S(w_i)$ 表示某词语目前未修正的梯度分数。

（2）对于贝叶斯平均，以用户投票排名为例，若用户投票评分的人很少，则平均分很可能会出现不够客观的情况。这时需要引入外部信息，即假设有一部分人（C 人）投了票，并且都给定了平均分（m 分）。把这类人的评分结果加到已有用户的评分中，再进行平均值计算，则就可对平均分进行修正，在一定程度可增加最终分数的客观性。

贝叶斯平均：

$$x=\frac{C\times m+\sum_{i=1}^{n}x_i}{n+C}$$

其中，C 是修正后的平均词频，m 是修正后的平均分。

（3）利用贝叶斯平均进行梯度分数的修正：

$$\text{修正分数}=\text{平均分}+\frac{\text{词频}}{\text{词频}+\text{平均词频}}\times(\text{梯度分数}-\text{平均分})$$

在热词提取中，我们用梯度分数的平均分作为先验 m，用平均词频作为 C。在热词提取中，词语每出现一次，相当于给词的热度进行了评分。词频少，也就代表了评分的人数少，则评分的不确定性大，需要用平均分来进行修正、平滑。

这里通过修正分数可以有效地将词频较少但词频增量较高的词语进行修正。词频远大于平均词频的词语，很大程度上代表了相关内容留言的人数多，更可能是一个热点词。

2.2.3 热词提取

图 11 留言详情中词频大于平均词频词汇词云

首先将留言详情样本中词频大于平均词频的词汇绘制词云，进行可视化，结果可以发现大部分的留言内容都围绕在“A 市”。由于仅通过词频判断热度词极大程度上会受到部分常用词的干扰，因此下面结合 TF-IDF 算法与词频增量来进一步衡量热点词。将所有标签中词频大于 0.1%总词频且 TF-IDF 值排名前 100 的词汇并提取其中前 15 个词汇定义为基础热点词。

表 9 前 15 个基础热点词表

序号	词汇	词频	标签类别	tf	idf	tf_idf
1	学生	43	教育文体	0.013288010	0.5596158	0.007436180
2	噪音	61	环境保护	0.019383540	0.3364722	0.006522023
3	幼儿园	54	教育文体	0.016687268	0.3364722	0.005614802
4	施工	46	环境保护	0.014617096	0.3364722	0.004918247
5	扰民	93	环境保护	0.029551954	0.1541507	0.004555454
6	麻将馆	40	城乡建设	0.001863586	1.9459101	0.003626370
7	严重	53	环境保护	0.016841436	0.1541507	0.002596119
8	车位	51	城乡建设	0.002376072	0.8472979	0.002013240
9	房屋	66	城乡建设	0.003074916	0.5596158	0.001720772
10	二期	60	城乡建设	0.002795378	0.5596158	0.001564338
11	公园	53	城乡建设	0.002469251	0.5596158	0.001381832
12	地铁	86	城乡建设	0.004006709	0.3364722	0.001348146
13	施工	86	城乡建设	0.004006709	0.3364722	0.001348146
14	质量	51	城乡建设	0.002376072	0.5596158	0.001329687
15	安全隐患	50	城乡建设	0.002329482	0.5596158	0.001303615

从表 9 的 15 个热点词可以看出，“学生”这一词汇在教育文体标签出现的频率较高，说明在教育文体标签热点问题是与学生相关的。而在环境保护标签下“噪音”、“施工”、“扰民”的 TF-IDF 值较高，说明该标签下提出噪音、扰民相关留言的频次较高。在城乡建设标签下，“麻将馆”、“车位”、“房屋”等词汇的 TF-IDF 值较高，说明提出此类问题的留言较多，这些词汇属于热点词汇。且 15 个热点词汇属于城乡建设标签的词汇有 9 个，占比 60%，说明留言用户提出的留言很大一部分在城乡建设标签下，该标签也属于热点主题。以下进一步通过贝叶斯平均词频增量的筛选出前 10 个热点词汇，如图 12。

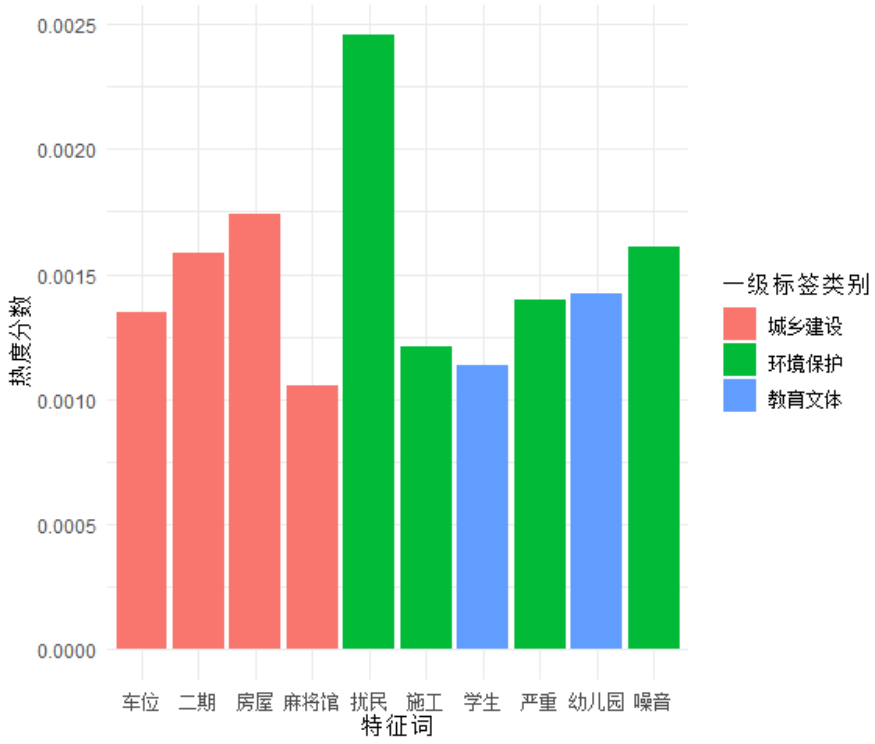


图 12 附件 3 中排名前 10 的热点词

通过词频、词汇 tf-idf 值与贝叶斯平均后的词频增量等权处理后计算得到热点词的热度指数[8]，在附件 3 中排名前 10 的热点词中，环境保护标签下的“扰民”热度分数最高，说明留言用户对“扰民”问题反响很大，属于热点问题，且前十个热度词主要集中在城乡建设和环境保护标签下，说明留言用户对这两个主题的问题更加关注。

以下绘制除部分热点词在时间窗口下的热度曲线：

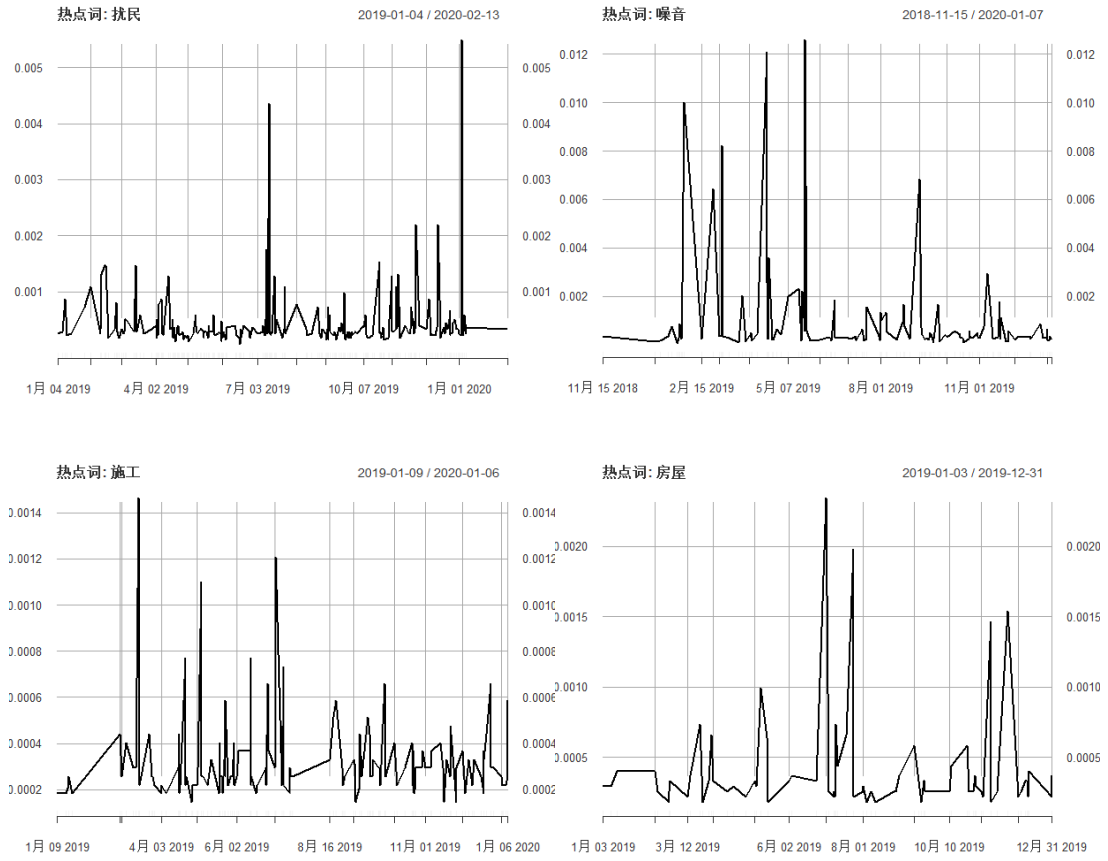


图 13 部分热点词在时间窗口下的热度曲线

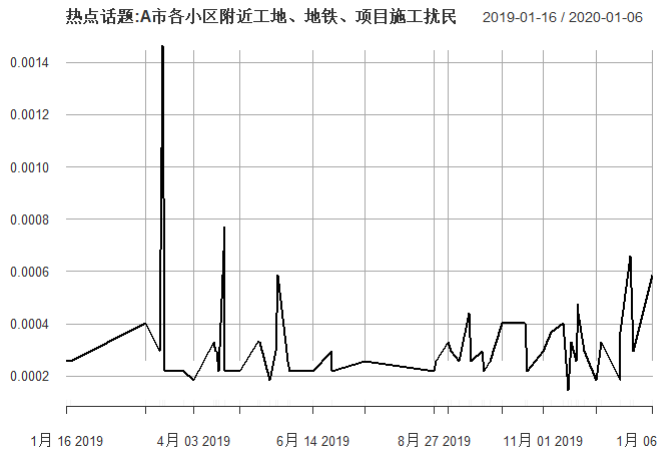
在“扰民”这一热度词的时序图下，可以看出在 2019 年 7 月 3 日和 2020 年 1 月 1 日附近的热度较高，在该时间范围内留言用户提出了较多有关“扰民”的问题，有关部门应及时解决。热点词“施工”和“扰民”有一定的联系，2019 年 4 月 3 日至 6 月 2 日，反映与热点词“施工”有关的留言较多，说明这段时间施工造成的居民困扰较多。热点词“房屋”在 2019 年 8 月和 2019 年 11 月附近热度较高，说明这两段时间范围内房屋有关问题较多。“噪音”这一热词与“扰民”、“施工”也有一定的联系在 2019 年 2 月 15 日至 5 月 7 日热度最高，即认为这段时间施工或其他原因造成噪音对居民影响大，建议有关部门根据热点问题进行解决。

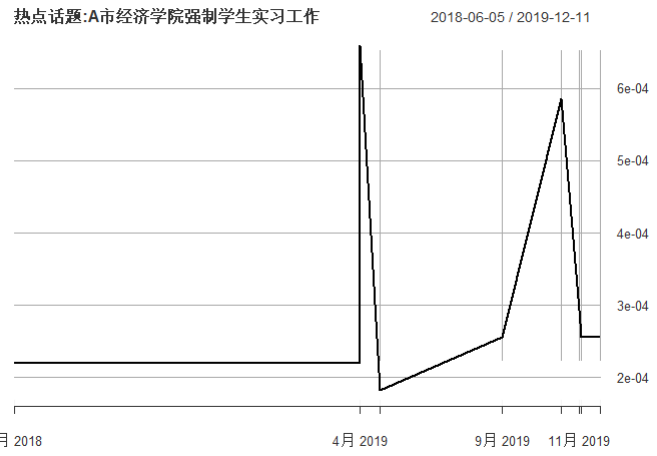
以下进一步对相关热点词进行挖掘、通过前 15 个热点词之间的话题相关度进行进一步归类、挖掘[9],从而提取出热点话题。热点话题中各相关热点词的热度分数进行贝叶斯平均后得到话题热度指数。

表 10 Top5 热点话题汇总表

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
2	1	0.0186	2019/01/16 至 2020/01/06	A 市各小区	附近工地、地铁、项目施工扰民
5	2	0.00293	2018/06/05 至 2019/12/11	A 市经济学院学生	学校强制学生去企业实习
1	3	0.0255	2019/02/21 至 2019/10/05	A7 县	配套幼儿园普惠性问题
4	4	0.0107	2019/01/09 至 2019/12/31	A 市各小区	房屋质量问题
3	5	0.0151	2019/07/07 至 2019/09/20	A 市武广新城伊景园滨河苑	车位捆绑销售

总结得到以上热点问题，热度排名第一的是在 2019/02/21 至 2019/10/05，A7 县的配套幼儿园普惠性问题，这段时间内该问题出现频次较高且热度指数达到 0.0255，其次是 A 市各小区附近工地、地铁、项目施工扰民的问题，时间范围在 2019/01/16 至 2020/01/06，排名最后热点问题为 A 市经济学院学校强制学生去企业实习的问题，时间范围在 2018/06/05 至 2019/12/11，根据热点问题的挖掘分析，有关部门可直接根据挖掘出的热点问题解决目前时间段的重要、热点问题，提高工作效率。





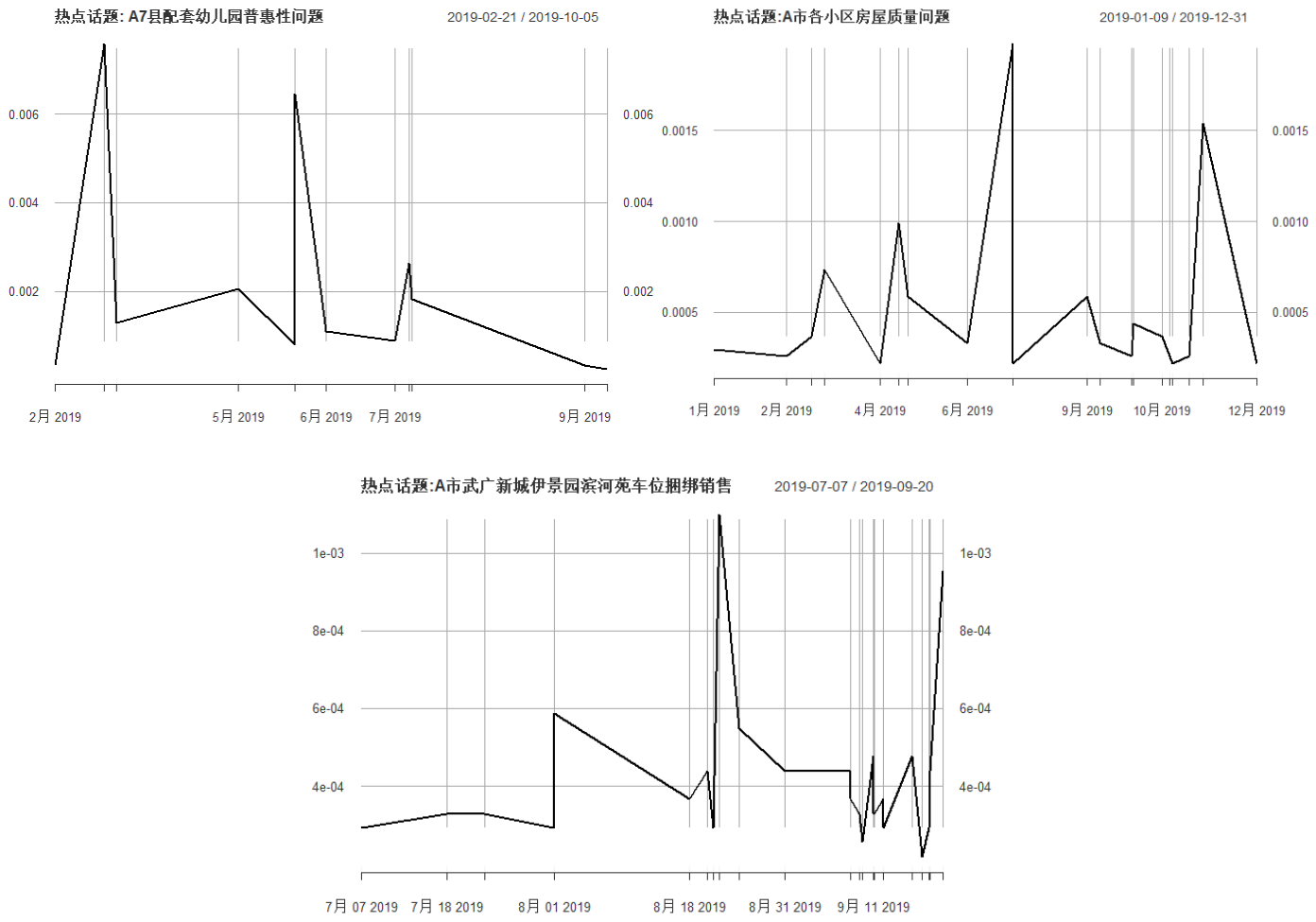


图 14 Top5 热点话题在时间窗口下的热度曲线

从热点话题时序图可以看出这五个热点话题在 2019 年至 2020 年随时间变化的热度，可以看出问题 1 在 2019 年 4 月 3 日热度最高、问题 2 在 2019 年 4 月热度最高、问题 3 在 2019 年 2 月热度最高、问题 4 在 2019 年 7 月热度最高、问题 5 在 2019 年 8 月热度最高，有关部门结合热点问题时间范围可以综合判断某时间点的最热问题，有针对性的解决问题。

2.3 问题 3 分析方法与过程

问题 3 建模流程图见图 15 图 2。

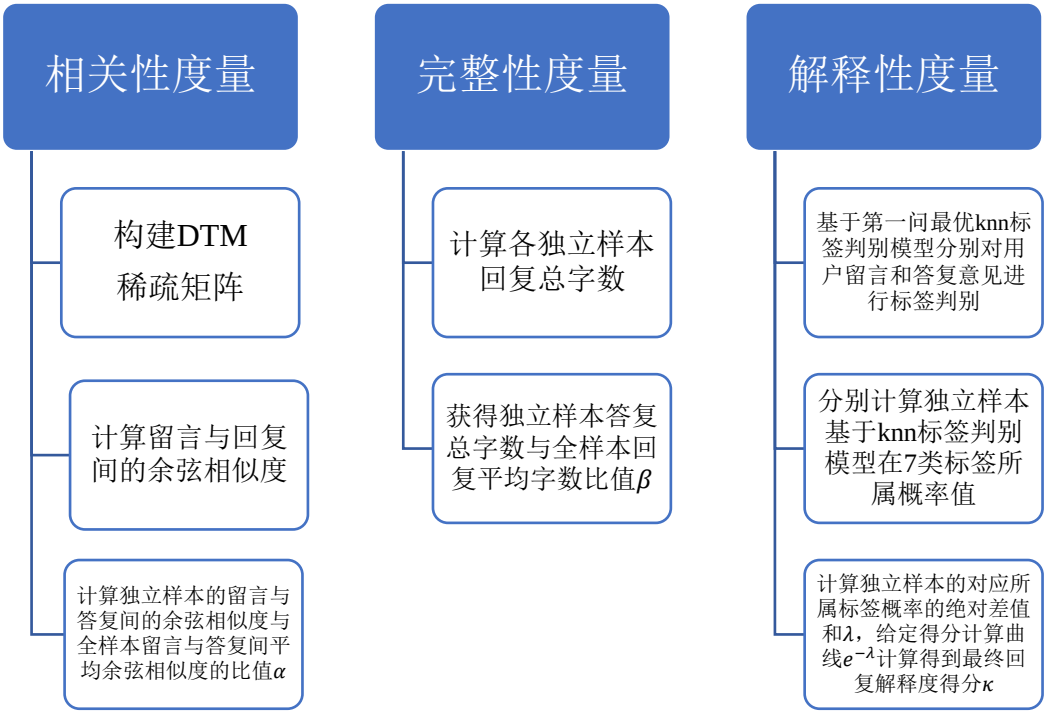


图 15 问题三建模流程图

计算每个答复的得分，答复质量得分公式：

$$\kappa \times (\alpha + \beta)$$

其中κ为解释度得分系数（衡量答复的解释性），α为独立样本的留言与答复间余弦相似度与总体样本留言与答复间平均余弦相似度的比值（衡量答复与留言的相关性），β为独立样本答复总字数全样本回复平均字数比值（衡量答复的完整性）。

2.3.1 相关性度量

（1）构建 DTM 矩阵

首先构建 DTM 矩阵，即文档-词频矩阵（document-term matrix，DTM），将处理后的语料库进行断字处理，生成词频权重矩阵(稀疏矩阵)也叫词汇文档矩阵。本文通过 *R 3.6.3* 版本中 *text2vec* 包构建 DTM 矩阵，过程如下：

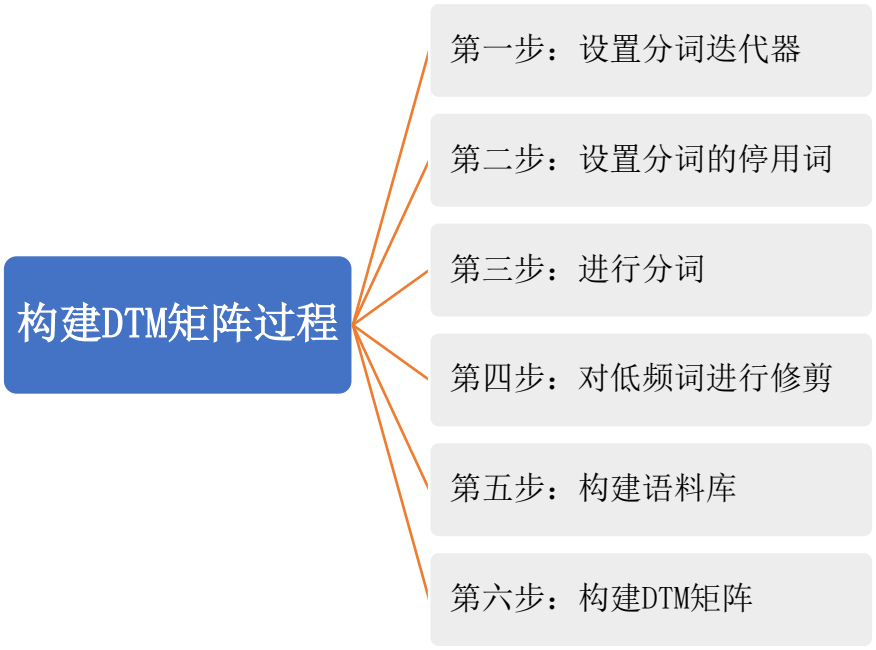


图 16 构建 DTM 矩阵过程图

过程关键代码(*R 3.6.3* 版本下运行)

步骤 1.设置分词迭代器

```
it_train = itoken(as.character(train$topic),       #语料内容
                  tokenizer = tok_fun,
                  ids = train$id,     #  可不设置 Id
                  progressbar = FALSE)
```

步骤 2.设定停用词

```
stop_words<-read.table("cn_stopwords.txt")
```

步骤 3.进行分词、消除停用词

```
vocab = create_vocabulary(it_train, stopwords = stop_words) #分词函数 ： create_vocabulary
```

步骤 4.对低频词的修剪

```
pruned_vocab = prune_vocabulary(vocab,
                                term_count_min = 10,     #删除词频低于 10 个的词汇
                                doc_proportion_max = 0.5,
                                doc_proportion_min = 0.001)
```

步骤 5.设置形成语料文件

```
vectorizer = vocab_vectorizer(pruned_vocab)
```

步骤 6.构建 DTM 矩阵 通过传入分词迭代器和语料文件

```
dtm_train = create_dtm(it_train, vectorizer)
```

（2）余弦相似度计算

基于 DTM 矩阵计算余弦相似度，其中余弦相似度的原理为：以向量空间中两个向量间夹角的余弦值作为度量两个独立个体之间的差异大小，若值越接近 1，则说明夹角的角度越接近 0°，也就是两个向量越相似，即为余弦相似。

其计算公式为：

$$similarity = \cos(\theta) = \frac{\vec{A}\cdot\vec{B}}{\left\|\vec{A}\right\|\left\|\vec{B}\right\|} = \frac{\sum_{i=1}^{n}\vec{A_i}\times\vec{B_i}}{\sqrt{\sum_{i=1}^{n}(\vec{A_i})^2}\times\sqrt{\sum_{i=1}^{n}(\vec{B_i})^2}},$$

对于文本匹配，属性向量A_i和B_i为文档中的词频向量。余弦相似性被看作是一种在比较过程中把文件长度正规化的方法。

其相似性范围为-1 到 1：-1 意味着两个向量指向的方向完全相反，1 表示它们的方向完全相同，0 通常表示它们之间是独立的，而在这之间的值则表示中间的相似性或相异性。

根据计算留言（留言详情）与回复的 DTM 矩阵的余弦相似度稀疏矩阵，求出每个样本的留言与回复余弦相似度之和来作为相似度度量的指标。进而求得系数α作为该样本余弦相似度得分与全样本平均余弦相似度得分的比值，以此来衡量留言与答复的相似性，记作相似度得分α。

其计算公式为：$a_j = \frac{similarity_j}{\sum_{j=1}^{n} similarity_j}$

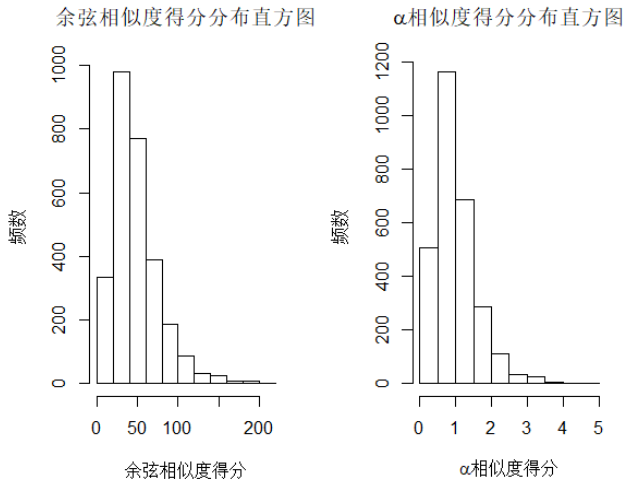


图 17 样本余弦相似度与相似度得分α分布直方图

可以发现在余弦相似度得分中，样本得分方差较大，通过相似度得分α避免了原余弦相似度得分可能带来的异方差性从而影响最终答复质量指标结果，从相似度得分α分布可以发现大部分答复的相似度得分均在 1 左右。

2.3.2 完整性度量

通常来讲，答复的字数数量可以较好的衡量答复的完整性，最终通过与相关性、解释性的非线性加权组合作为答复质量评价指标。

完整性得分的计算公式为：$\beta = \frac{独立样本的答复字符串长度}{总体样本的平均答复字符串长度}$

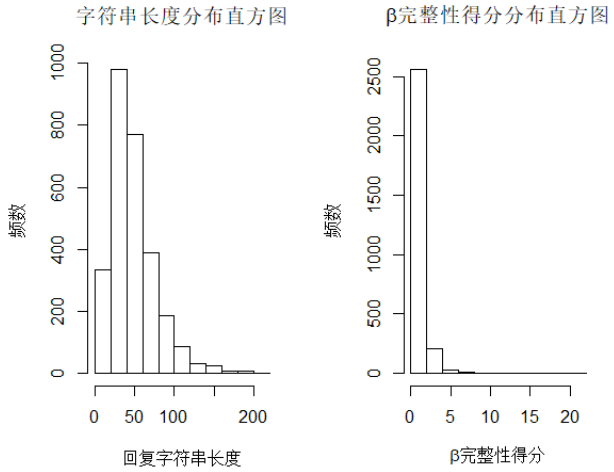


图 18 样本答复字符串长度分布与完整性得分β分布直方图

可以发现绝大部分的样本答复完整性得分均小于 5，同时存在极个别的样本完整性得分达到了 20，这是由于该样本答复字数长度远大于平均值，这种情况在现实生活中也是存在的，这也是一种答复高水平的体现，因此这里不选择去除该异常值。

2.3.3 解释性度量

首先判断对应标签，根据样本的用户留言详情和回复内容分别基于先前的最优 knn 模型进行标签判别，获取在留言、回复内容中对应所属一级标签的概率值。然后使用标签判断模型，分别判断留言与回复的对应标签来衡量回复内容质量。最后计算各标签留言、回复判断概率绝对差值之和。

$$\lambda=\sum_{i=1}^{n}\left|P\left(Y_i=c_i\right)-P\left(Y_j=c_j\right)\right|, i=j=1,2,\ldots,n$$

查看对于回复和留言标签判别的概率值，各概率值差越小说明答复质量越高，解释度越高。因此对应的概率绝对差值和$\lambda(\lambda\in[0,2])$。

表 11 基于标签判别模型分别对留言与答复所属各标签概率及λ结果表（选取 5 个样本）

对应留言回复	留言编号	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	卫生计生	λ值
留言 1	2549	0.6666667	0	0	0.1111111	0	0.2222222	0	0.6666667
答复 1	2549	0.4444444	0	0	0	0	0.5555556	0	
留言 2	2554	0.6666667	0.1111111	0.2222222	0	0	0	0	0.3030303
答复 2	2554	0.8181818	0	0.1818182	0	0	0	0	
留言 3	2555	0.09090909	0	0	0.2727273	0.6363636	0	0	1.232323
答复 3	2555	0	0	0	0.8888889	0.1111111	0	0	
留言 4	2557	0	0	0	0.8181818	0.09090909	0.09090909	0	2
答复 4	2557	1	0	0	0.	0	0	0	
留言 5	2574	1.0	0	0.0	0	0	0	0	0.2
答复 5	2574	0.9	0	0.1	0	0	0	0	

可以发现，在选取的 5 个样本中，基于第一问最优 KNN 标签判别模型分别对留言和答复的标签分类的判别概率结果发现，留言编号为 2574 的答复绝对差值和λ最小，说明该答复非常契合留言用户的提出问题主旨，对问题内容有较高的解释度。

由于λ越小说明答复质量越高，为了将解释度的量化标准转化为指标指数与解释度正比化（使指标指数越高代表解释度越高），这里选取e^{-x}作为得分计算曲线，最终计算得到样本解释度得分κ，则 $\kappa=e^{-\lambda}$。

以下绘制样本解释度得分计算曲线与解释度得分κ分布直方图来进一步分析解释性得分的合理性。

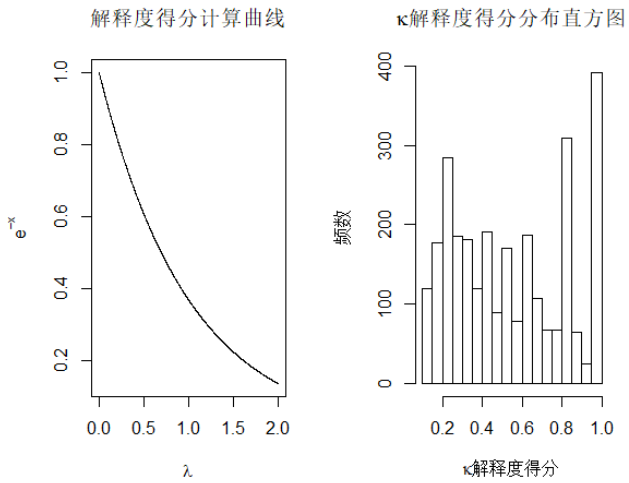


图 19 样本解释度得分计算曲线与解释度得分κ分布直方图

通过得分计算曲线对样本各标签判别概率绝对差值和λ的转换，解释度得分κ范围被限制在了$[e^{-2}, 1]$，κ值越高说明答复解释度越高。由于在实际生活中，对于用户留言的答复的解释度往往是衡量答复质量水平的核心指标，解释度过低，答复拥有再高的完整性或相关性得分也不能成为优秀的答复，因此在最终答复质量评价指标公式中，将κ作为答复相关性得分α与完整性得分β等权相加后的决定系数。

则最终答复质量评价公式：$\textit{回复质量得分} = \kappa \times (\alpha + \beta)$

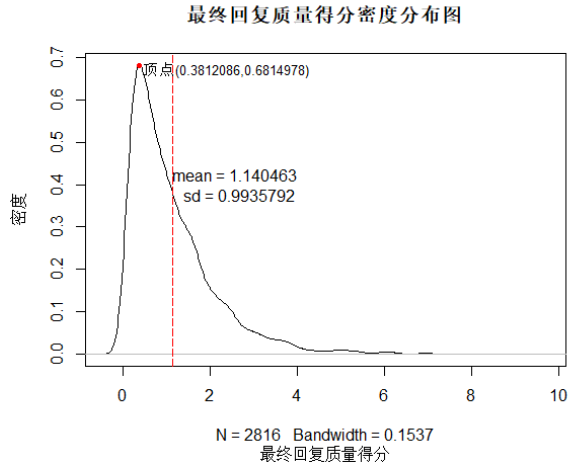


图 20 最终回复质量得分密度分分布图

从图中可知，最终样本中回复质量得分的均值为 1.140，方差为 0.9935，约 68.14%的答复质量得分都在 0.3812 分左右，答复质量得分总体的范围在 [0.024917, 9.289989]，以下分别提取最低、最高、最接近平均得分的回复内容进行对比参考。

表 12 代表性留言内容与对应答复汇总表

回复比较	留言编号	留言用户	留言详情	答复意见	回复质量得分
得分最低回复	6556	UU0081320	″请问领导，农合费用增加了，打狂犬役苗报销比例是多少。盼回音。先谢了！″	″已收悉″	0.024917
得分最高回复	76823	UU0081457	″市长：您好！F5 县东山镇关山村水库采石场肆意妄为破坏生态环境，污染空气，污染水库水源，炮炸声机器声严重扰民，运输石头的车每天络绎不绝，道路被压的坑坑洼洼，每到下雨就泥泞不堪，甚至连水库防洪堤也被压得伤痕累，这样会给全村人民带来安全隐患！请市长为民除害！″	″“UU0081457”您的留言已收悉。关于您反映的问题，已转 F5 县调查处理。我镇于 2018 年 6 月 28 日收到 F 市人民政府办公室《网络舆情交办函》，内容为东山镇一村民请求关闭 F5 县东山镇关山村采石场的投诉件，东山镇人民政府高度重视，并同县环保局一同到实地展开调查了解情况，现回复如下：一、采石场基本情况：**…（省略 2866 字）**	9.289989
得分接近均值回复	16351	UU0081733	″此处违章建筑已经巍峨挺立很多年了，群众不敢举报，领导也无人过问。房子主人特别牛。2018 年 7 月 27 日下午二点，将要真正拆除的消息不径而走，不少群众早已等待在此处违章建筑对面的树林下，目睹那正义的一刻的到来，有的手中还拿着鞭炮以示庆祝。时值下午二点，等待已久的群众激动不已，一个小时过去了，两个小时过去了，动静全无，随后是一场瓢泼大雨，树林中的群众仍不愿离去，他们等待这一天太久了。直到夜色见晚，违章建筑仍然高高挺立在 A7 县榔梨镇陶公庙对面的梨江河边上！与悻悻而散的群众形成了强烈对比。违坚强的主人是陶公庙社区的女能人，他不自己创造了违坚强，还帮他们的朋友策划了违建。地址：榔梨陶公庙社区文化巷户主：粟华爱″	″网友“UU0081733”您好！来信收悉。现回复如下：针对您反映的陶公庙社区梨江河边的违章建筑情况，我街道高度重视，我街道拆违控违办工作人员立即进行了调查，经调查核实，该路段确实存在有一处历史违建行为。该处建筑户主为 2017 年灾后建房对象，在该处建筑隔壁有一处房屋因灾倒损，正在进行灾后重建，该处建筑作为临时过渡用房使用，待该户受灾房屋建成完工后，我街道将督促其对此处房屋进行拆除，逾期不自拆的，街道将联合行政执法局对其进行强制拆除。如果您有新的意见建议，请直接与我们联系，联系电话：0000-00000000。感谢您对我街道工作的支持和监督！2018 年 8 月 31 日″	1.140289

从得分最低回复内容来看：答复者只用“已收悉”三字答复了提问用户，不仅没有正面回复用户的提问且没有给予不作正面答复的解释，因此回复质量得分仅有 0.02 分

从得分最高回复内容来看：答复者的答复内容近 3000 字，且回复内容格式规范，逻辑清晰，条理分明，思路严谨，可见十分关注提问者的问题，回复质量

得分高达 9.28 分。

从得分最接近平均得分的回复内容来看：答复者答复内容较为规范，问题的时间、地点、对象及解决方案都罗列了出来，答复质量较好。

以下为提取代表性留言内容与对应答复的关键 R 代码：

得分最低回复提取：

```
review[which(final.explain.score==min(final.explain.score))]
answer[which(final.explain.score==min(final.explain.score))]
```

得分最高回复提取：

```
review[which(final.explain.score==max(final.explain.score))]
answer[which(final.explain.score==max(final.explain.score))]
```

得分最接近平均值回复提取：

```
mean_review<-abs(final.explain.score-mean(final.explain.score))==min(abs(final.explain.score-mean(final.explain.score)))
review[which(mean_review)]
answer[which(mean_review)]
```

3 结果分析

3.1 问题 1 结果分析

问题 1 主要目的是为解决群众在问政平台留言的分类问题，此题主要分类模型有随机森林、人工神经网络、KNN 模型，利用 F-Score 对分类方法进行评价从而选取一最优模型。首先进行预处理，利用 jiebaR 包对样本的留言主题、详情进行分词、去除停用词等，此步骤是为了去除文本内容中停用词等不重要的词汇，并将重要词汇分开防止停用词等词汇影响机器学习过程，从而无法提取出重要词汇。然后基于 TF-IDF 算法提取 7 个一级标签对应留言主题、详情中 TF-IDF 值排名前 10%的词汇分别建立成对应的标签类别词库，此步骤为建立与七个一级标签对应的自定义词库，自定义词库是根据原本标签内容来设立的，所以准确率较高。在自定义词库中可以看出每个一级标签对应的词库和标签有显著关系，例如在城乡建设标签下，“房产证”这一词汇 TF-IDF 值最高，同时“住房”、“危房

改造”、“违建”、“违章建筑”也具有较高的 TF-IDF 值，这些词汇都与城乡建设有显著关系。最终通过三种机器学习模型（随机森林、ANN 人工神经网络、KNN）进行交叉检验、重复训练后，发现在基于留言主题计算权重得分的模型中，KNN 表现最优，平均预测准确率为 85.66%，平均 kappa 值为 0.8281，随机森林为 85.57%(0.8269)，ANN 为 85.4%(0.8254)。同时，使用相同方法基于留言详情计算得分权重的模型中，KNN 的预测准确率提升到 88.58%，kappa 值提高到了 0.8631。通过循环迭代最终选定最优 k 值后，基于留言详情的 KNN 模型最终全标签类别 F1 得分均值达到了 0.9211055。所以最终选定 KNN 模型为留言分类的最优模型，KNN 算法比较适合于样本较大的情况，且主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属的类别，因此在这种情况下，认为 KNN 模型最适，分类效果最好，最终全标签类别 F1 得分均值达到了 0.9211055。

3.2 问题 2 结果分析

问题 2 主要目的是为了挖掘出某短时间内的热点问题，方便部门进行热点问题的解决和观察。对于该问题，本文建立了一热度评价指标，即利用贝叶斯平均修正梯度分数来作为评价指标，修正后的梯度分数更具客观性。

首先通过 LDA 主题建模进行主题自动聚类与先前基于留言详情建立的 KNN 标签判别模型来计算得到对应特征词的 TF-IDF 值，通过等权处理得到词汇的最终 TF-IDF 值，此方式得到的 TF-IDF 值更全面可靠，再结合人工调参词频与词汇 TF-IDF 值阈值来最终确定热点词，然后进行相关热点词索引进而进一步通过热点词组合提取到热点话题并计算出对应话题热度排名及话题总时间窗口上的热度变化趋势，得到热点问题留言明细表，将前 15 个热点词之间的话题相关度进行进一步归类、挖掘，可以得到热度排名第一的是 A7 县的配套幼儿园普惠性问题，时间范围在 2019/02/21 至 2019/10/05，其热度指数为 0.0255，其次根据热度指数还排出了所有的问题，让部门可以直观的看出某段时间范围内的热点问题，提高了部门工作效率、减少了错误率。

3.3 问题 3 结果分析

问题 3 主要目的是得到一套关于答复意见的评价体系，即给出答复意见的评价得分，本文从相关性、完整性、可解释性三个方面进行度量，综合三个方面给出回复得分公式。

关于相关性的度量，主要是基于 DTM 矩阵计算余弦相似度，然后再将独立样本的留言与答复间余弦相似度与总体样本留言与答复间平均余弦相似度相比作为公式中相关性度量方面，在其他条件不变的情况下，独立样本的留言与回复间的余弦相似度越高，其相关性越强，相关性得分也越高。关于完整性的度量，则是根据独立样本答复总字数和全样本回复平均字数相比作为完整性的度量，在其他条件不变的情况下，独立样本答复总字数越多，说明其完整性越强，则回复完整性得分越高。对于可解释性的度量，主要根据样本的用户留言详情和回复内容分别基于先前的最优 KNN 模型进行标签判别，获取在留言、回复内容中对应所属一级标签的概率值。然后计算留言、回复每个一级标签下的概率绝对值差，再将七个标签的概率绝对值差相加，记该值为λ，若该值越大则表明回复的和留言大相径庭，则认为其解释性弱，其解释性系数κ值为$e^{-\lambda}$，在其他条件不变的情况下，λ越大，k值越小，其解释性越弱，则最终回复质量得分越低。

此套评价方案汇集相关性、完整性、可解释性三个方面，更加全面地将留言和其回复联系起来，构造其评价分数、对回复进行评价，最终得到每个回复的质量得分，回复质量得分越高，说明其回复和留言相关性越高、其回复越完整且对留言详情的解释性越强。

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]. 李航.统计学习方法[M].北京:清华大学出版社,2012.

[3]. 祁小军,兰海翔,卢涵宇,丁蕾锭,薛安琪.贝叶斯、KNN 和 SVM 算法在新闻文本分类中的对比研究[J].电脑知识与技术,2019,15(25):220-222.

[4]. 廖列法,勒孚刚,朱亚兰.LDA 模型在专利文本分类中的应用[J].现代情报,2017,37(03):35-39.

[5]. 张瑞琦.基于关键特征聚类的 Top N 热点话题检测方法研究[D].北京理工大学,2015.

[6]. 张杨子.面向对话系统回复质量的自动评价研究[D].哈尔滨工业大学,2018.

[7]. 张梦笑,王素格,王智强.基于 LDA 特征选择的文本聚类[J].电脑开发与应用,2012,25(01):1-5.

[8]. 刘业政,杜亚楠,姜元春,杜非.基于热度曲线分类建模的微博热门话题预测[J].模式识别与人工智能,2015,28(01):27-34.

[9]. 程克非,邓先均,周科,罗昭,陈旭东.基于微博多维度及综合权值的热点话题检测模型[J].重庆邮电大学学报(自然科学版),2019,31(04):468-475.