

文本信息挖掘在“智慧政务”中的应用

摘要：随着信息时代的发展，了解民意的方式变得多样化，伴随而至的是数据量的大幅增长，传统的留言处理已经难以负荷，因此，如何提高政务部门工作效率并减轻工作人员负担已成为当下社会需要考虑的问题。本文通过自然语言处理对留言内容及答复意见进行文本挖掘，针对问题一、二、三给出了恰当的解决方案。

针对问题一，本文首先将附件 2 中的文本数据进行查空、分词等预处理。由于分类数据不平衡，文章采用了近义词替换和回译两种方法进行文本数据增强，然后对数据进行词向量转换，通过 7 个常用的机械学习模型对结果进行预测，从中挑选出 3 个表现最佳的模型 (Random forest, XGBoost, Logistic regression)，结合投票法进行模型融合，最终得到留言内容的一级标签分类模型，并使用 F-Score 评价模型，计算结果 F1 值为 0.926。

针对问题二，本文首先使用问题一中的模型对附件 3 中的文本数据进行预分类，以此提高后续聚类的准确率并减少运算量，接着使用 K-Means 算法对同属一个类型的留言内容进行聚类，其中，采用轮廓系数寻找每一类的最优聚类值 k ，然后将每一类事件合并，采用熵值法对事件点评数(点赞数和反对数)以及留言数量进行处理，求出其权重并用于定义热度指数并输出评价结果，然后通过正则匹配法提取出事件的地点与人群，最后按照要求的格式将结果保存为文件。

针对问题三，本文通过相关性 (Correlation)、完整性 (Completeness)、及时性 (Timeliness) 以及可解释性 (Interpretability) 四个指标基于层次分析法与模糊综合评价法建立 CCTI 评价体系，用于评价答复意见的质量，首先，以余弦相似度来计算答复意见的内容与问题的相关性；通过正则匹配法并根据“文本长度”、“文明用语”、“答复时间”以及“条理清晰”这一整套规范，来评价答复意见的完整性；计算答复与留言问题的时间距离来评价答复意见的及时性；以“引用官方条例”、“直接解决”和“间接解决”三个部分来评价答复意见的可解释性。最终，CCTI 评价体系会对每一条答复意见的质量进行评价并点评。

关键词：文本增强 模型融合 K-Means 算法 熵值法 层次分析法

Application of Text Information Mining in "Smart Government Affairs"

Abstract: With the development of the information age, the ways to understand public opinion have become diversified, and with the substantial increase in the amount of data, traditional message processing has been difficult to load. Therefore, how to improve the efficiency of government departments and reduce the burden on staff has become Issues that society needs to consider now. This article uses natural language processing to carry out text mining on the content of the message and the answers, and gives appropriate solutions to the problems 1, 2, and 3.

For the first problem, this article first preprocesses the text data in Annex 2 for gap checking and word segmentation. Due to the imbalance of categorical data, the article uses two methods of synonym replacement and back translation for text data enhancement, and then performs word vector conversion on the data, predicts the results through 7 commonly used mechanical learning models, and selects the three with the best performance. A good model (Random forest, XGBoost, Logistic regression), combined with the voting method for model fusion, and finally obtained a first-level label classification model of the message content, and using the F-Score evaluation model, the calculation result F1 value is 0.926.

For problem two, this paper first uses the model in problem one to pre-classify the text data in Annex 3, in order to improve the accuracy of subsequent clustering and reduce the amount of calculation, and then use the K-Means algorithm to the same type of message content Perform clustering, in which the contour coefficient is used to find the optimal clustering value k of each class, and then each type of event is merged, and the number of comments (likes and objections) and the number of comments are processed by the entropy , Find its weight and use it to define the heat index and output the evaluation result, then extract the location and crowd of the event through the regular matching method, and finally save the result as a file in the required format.

In response to question three, this article establishes a CCTI evaluation system based on the analytic hierarchy process and the fuzzy comprehensive evaluation method through the four indicators of Correlation, Completeness, Timeliness and interpretability for evaluation. For the quality of the response, first of all, the cosine similarity is used to calculate the relevance of the content of the response and the question; through the regular matching method and according to the "text length", "civilized language", "response time" and "clear structure" A complete set of specifications to evaluate the completeness of the reply opinion; calculate the time distance between the reply and the message question to evaluate the timeliness of the reply opinion; evaluate the reply opinion with three parts: "quoting official regulations", "direct solution" and "indirect solution" Interpretability. Ultimately, the CCTI evaluation system will evaluate and comment on the quality of each response.

Keywords: text enhancement, model fusion , K-Means algorithm entropy method , analytic hierarchy process

目录

一、引言	4
二、问题分析	5
2.1 问题一分析	5
2.2 问题二分析	5
2.3 问题三分析	6
三、问题一	7
3.1 处理流程图	7
3.2 数据预处理	7
3.3 文本数据增强	9
3.4 文本特征提取	10
3.5 模型介绍	11
3.6 模型训练	12
3.7 模型融合与结果	13
四、问题二	14
4.1 处理流程图	14
4.2 数据预处理	14
4.3 预分类	15
4.3 降维	15
4.4 K-means 聚类	16
4.5 数据提取与合并	18
4.6 热度指数	18
五、问题三	20
5.1 处理流程图	20
5.2 CCTI 评价体系构建	20
5.2 数据处理	22
5.3 层次分析法	24
六、总结	27
参考文献	28

一、引言

随着信息通讯技术的高速发展，政府正面临着政务电子化、信息化、网络化的压力，传统政务工作是按业务、管理职责设定，要求各个部门各司其职，存在较严重的信息壁垒，使得数据孤立的存在于不同的“烟囱”中。值得注意的是，传统政务中遵循政务边际成本递增法则，即社会化任务越重，管理范围越大，相应的管理成本也就越高。

在“互联网+政务服务”的引领下，中国政务信息化的步伐明显加快。国务院总理曾表示要改变“信息孤岛”以及“数据烟囱”^[1]，促进政府信息共享，提高政府的效能，方便企业和群众办事。一系列新技术的发展令传统政务正在朝着“智慧政务”的方向推进，帮助政府解决信息孤岛、数据烟囱和碎片化应用等问题。

“智慧政务”的核心功能之一是管理基础信息资源，通过信息管理，建立各街道、社区基础信息资源台账，收集群众的信息，提高信息采集的及时性和准确性。基础信息资源的管理需要建立基础数据库并对基础业务进行数据采集、更新以及网络化，政府各部门必须按其职责权限去采集、使用和管理数据。

本文的任务是从政务平台收集的数据中挖掘有价值的信息，通过数据分析达到预测、划分标签、归类留言、筛选热点以及建立一套答复意见评价方案的目的。

二、问题分析

2.1 问题一分析

针对附件 2 的留言信息建立内容分类的一级标签分类模型，本文首先对留言内容进行数据预处理，数据预处理包括了空值处理、标签编码以及中文分词与去停用词的过程，通过数据预处理可以更好地建立模型并进行分析，其中，空值处理是指检测出文本数据中包含的“不知道”、“不存在”或者“无意义”的值并进行相应的处理，提高后续运算的准确率，防止代码运行报错；标签编码指的是将一级标签中的 7 大类转为由 0 至 6 的数字表示，由此可简化运算；中文分词与去停用词的意思是将留言的句子切分为词语，并去除无关的词语和字以此提高模型训练的准确率。由于各类标签的数据不平衡，部分标签数据较少，因此需要做文本增强。为了使模型更加多样性，本文使用两种文本增强方法，第一种使近义词替换，采用的知乎问题训练出来的词向量模型，使用此预训练词向量进行近义词查询，对句子中能替换的词语进行替换。第二种方法是回译，本文采用百度翻译的 API 进行回译处理，对每一个数据进行中译英再英译中处理，以此获取不一样的数据。文本增强后数据相对平衡，然后采用常用的 7 种机械学习模型进行，采用 5 折交叉检验得出结果，然后挑选出来前三个准确率最高的模型，最后使用投票法对这三个模型进行模型融合。

2.2 问题二分析

这个问题主要是聚类问题，由于数据较大，所以本文先采用第一题的模型进行留言预分类的方法，这样做可以提高后续聚类的准确率，因为问题一的分类问题是有监督学习，相对无监督学习能优更好的效果，而且先分类再聚类也能减少训练用时。但是聚类后的数据维度仍然很大，所以本文使用 PCA 方法对数据进行降维，保留 95% 的数据，使训练速度提升几倍。然后使用 K-means 算法对每一类进行聚类，由于 K-means 聚类算法是无监督学习，需要定义 k 的值，所以本文使用轮廓系数找出最优聚类数，然后进行聚类。聚类后对数据进行合并，并使用熵值法对点评数(点赞数和反对数)以及留言数量求权重，并得出热度指数公式。最后通过正则匹配法提取出事件的地点与人群，最后按照要求的格式将结果保存为文件

2.3 问题三分析

针对附件 4 提供的留言及答复内容制定一套用于评价答复内容的方案，本文将以相关性（Correlation）、完整性（Completeness）、及时性（Timeliness）以及可解释性（Interpretability）四个描述性指标建立 CCTI 评价体系，对答复内容进行全面的评价。首先，相关性指的是答复内容与相应问题的相关程度，文章将会计算留言问题与答复内容之间的相似度，根据二者相似度的大小来评价答复内容与相对应的问题是否相关。第二，完整性将由“文本长度”、“文明用语”、“答复时间”以及“条理清晰”四个标准来定义，目的是用于规范答复内容，“文本长度”指的是答复内容的字数，本文将对附件 4 中答复内容的字数进行统计，并选定一个字数范围，若文本长度在该范围中则视为达到标准，反之则未达到标准，即会对答复内容的完整性造成负面影响；“文明用语”，顾名思义，即答复内容中用到了诸如“您好”、“谢谢”等文明用语，从中反映出相关部门的礼貌与素质，本文将检索答复内容中是否存在文明用语来评价其是否达到标准；“答复时间”指的是答复内容中结尾处有答复时间，体现出相关部门的规范性和严谨性，文章将通过定位答复内容中的结尾处并检测该处是否包含时间来决定其是否达到标准；“条理清晰”指的是答复内容的结构性，使留言用户能够更加直观感受到答复内容的完整性，例如答复内容中带有“首先”、“第一”等字眼可令内容更加有层次性，本文将检索答复内容中的相关字眼以评价其是否达到标准。第三，及时性指的是答复是否及时，即答复时间与留言时间的差距，表现出相关部门对问题的回复效率，为了使计算结果尽可能反映出真实状况，文章将仅计算政务人员工作日之间的时间差，即排除双休和节假日的情况，根据计算结果对答复的及时性进行评价。第四，可解释性指的是答复意见中的相关解释，将由“引用官方条例”、“直接解决”和“间接解决”三个部分组成，该三个部分都能体现出答复内容的可解释性，但每个部分产生的可解释性大小却不尽相同，“引用官方条例”指答复内容中引用了法律法规或相关部门出台的文件或政策，使答复内容有理可依，有据可循；“直接解决”指的是直接帮助留言用户解决问题的回复；“间接解决”指的是提供可以解决问题相关部门的联系方式，或者承诺和相关部门反映问题的回复，本文将为这三个部分分别建立词袋，通过检索答复内容寻找词袋中的词语来评价答复内容的可解释性。最后，通过对相关性、完整性、及时性以及可解释性四个描述性指标的评价，建立一个基于层次分析法和模糊综合评价法的 CCTI 评价体系，综合评价答复意见的质量，针对每一条答复内容的质量给予评价和点评。

三、问题一

3.1 处理流程图

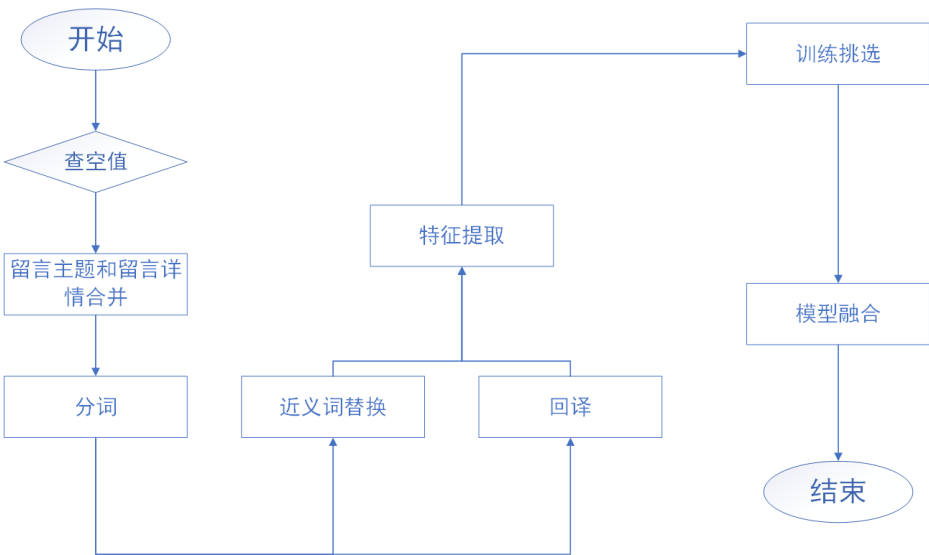


图 1 问题一流程图

3.2 数据预处理

3.2.1 空值处理

空值指的是数据表中出现“不知道”、“不存在”或者“无意义”的值，如果不及时处理空值会造成计算结果出错，更直观的表现的代码在运行时会报错，因此，本文首先需要预先判断数据中是否存在空值并及时对空值进行处理。

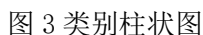
```
In [3]: print("在 留言详情 列中总共有 %d 个空值." % df['留言详情'].isnull().sum())
print("在 留言主题 列中总共有 %d 个空值." % df['留言主题'].isnull().sum())
print("在 一级分类 列中总共有 %d 个空值." % df['一级标签'].isnull().sum())

在 留言详情 列中总共有 0 个空值.
在 留言主题 列中总共有 0 个空值.
在 一级分类 列中总共有 0 个空值.
```

图 2 空值检测

经统计结果显示，留言详情、留言主题以及一级分类这三列数据中皆无空值，因此无需对数据进行空值处理。

本文将一级分类标签中的七大类转换为七个标签并分别对其进行编码,目的是简化后续运算,转换后的结果为:“0”代表“城乡建设”,“1”代表“环境保护”,“2”代表“交通运输”,“3”代表“教育文体”,“4”代表“劳动和社会保障”,“5”代表“商贸旅游”,“6”代表“卫生计生”。其中,属于“城乡建设”的数据最多,共有 2009 条,数据最少的是“交通运输”类,只有 613 条。如图 3,4。

图 4 类别数量

中文分词是指将一个句子分割成一个个单独的词,分词的过程即是把连续的字序列通过某种定义去重新结合成词序列。本文借助 **jieba** 分词的工具包对中文文本进行分词处理, **jieba** 分词包括四种分词模式,第一种是精确模式,意思是尽可能精确地切分句子,此模式适合文本分析;第二种是全模式,指的是搜索句子里能够组成词的词语,运算速度相对较快,但缺点是无法避免出现歧义的问题;第三种是搜索引擎模式,其经过精确模式的处理后,对长词进行二次分割,此模式可以提高结果的召回率,主要针对搜索引擎分词;最后一种是 **paddle** 模式,其基于 **PaddlePaddle** 深度学习框架,对序列标注的网络模型进行训练并分词。为提高分词效率,本文结合去停用词的方法对无意义的词语进行过滤,文章中采用的停用词来自四种常用的中文停用词表,它们分别是中文停用词表、哈工大停用词表、百度停用词表以及四川大学机器智能实验室停用词库^[2]。此外,本文还通过正则匹配法对文本中的数字、英文、标点符号以及带有“省”、“市”、“县”、“区”的关键词进行过滤,目的是减少噪声,其过滤规则为:

0 西湖 建筑 集团 占道 施工 安全隐患 大道 西行 便道 未管 路口 加油站 路段 人行道...
1 在水一方 大厦 人为 烂尾 多年 安全隐患 严重 位于 书院 路 主干道 在水一方 大厦 一...
2 投诉 苑 物业 违规 收 停车 费 尊敬 领导 苑 位于 火炬 路 物业 程明 物业管理 有限...
3 蔡铿 南路 华庭 楼顶 水箱 长年 洗华庭 高层 二次 供水 楼顶 水箱 长年 不洗 现在...
4 华庭 自来水 好大 一股 霉味 华庭 高层 二次 供水 楼顶 水箱 长年 不洗 现在 自来水...

8 / 28

3.3 文本数据增强

文本数据增强是一种扩大数据样本规模的有效方法，通过文本数据增强，可以解决数据不足或者数据不均衡的问题，进而提升模型的准确率。由于文章在标签编码过程中统计出属于“城乡建设”的数据共有 2009 条、“劳动和社会保障”的数据共有 1969 条、“教育文体”的数据共有 1589 条、“商贸旅游”的数据共有 1215 条、“环境保护”的数据共有 938 条、“卫生计生”的数据共有 877 条、“交通运输”的数据共有 613 条，其中数据最多的“城乡建设”类与数据最少的“交通运输”类相差了 1396 条，因此本文发现数据存在不均衡的现象。为了提高模型的适应性和准确率，文章分别使用近义词替换和回译两种方法进行文本数据增强。

3.3.1 近义词替换

近义词替换是指将句子中的词语用近义词替换后生成新的句子的过程，本文使用预训练的词向量进行相近词查询，根据以知乎问答社区为语料库^[3]，Word 和 Ngram 为语境特征，基于 Word2vec 原理训练得到的词向量，找出与之相关性最强的词语并进行替换。如图 6。

```
cn_model.most_similar(positive=['开心'], topn=1)
[('高兴', 0.7564147710800171)]
```

图 6 近义词查找例子

由于“交通运输”类、“卫生计生”类以及“环境保护”类中的数据均与“城乡建设”类的数据相差了一千条以上，因此分别对这三类的数据使用与预训练的词向量相关性最强的的词语进行近义词替换，生成更多的新数据，扩大数据样本的规模。

3.3.2 回译

回译，指的是翻译后的译文用原文中使用的语言进行二次翻译，也被称为“往返翻译”。通过回译可改善文本数据不足的问题，增加数据样本，起到文本数据增强的作用。本文采用百度翻译 API^[4]对文本进行回译，回译的方式为首先将原文中的中文翻译成英文，再将译文中的英文翻译回中文。

由于数据仍存在不均衡的现象，本文随机抽取类别五“商贸旅游”的 600 条数据进行回译并生成 600 条新的“商贸旅游”类数据，同时，对类别二“交通运输”类的 613 条原文数据进行回译，生成 613 条新的“交通运输”类数据。

经过文本数据增强后现在城乡建设”类共有 2009 条数据，“环境保护”类共有 1876 条数据，“交通运输”类共有 1812 条数据，“教育文体”类共有 1589 条数据，“劳动和社会保障”共有 1969 条数据，“商贸旅游”共有 1815 条数据，“卫生计生”共有 1754 条数据如图 7。

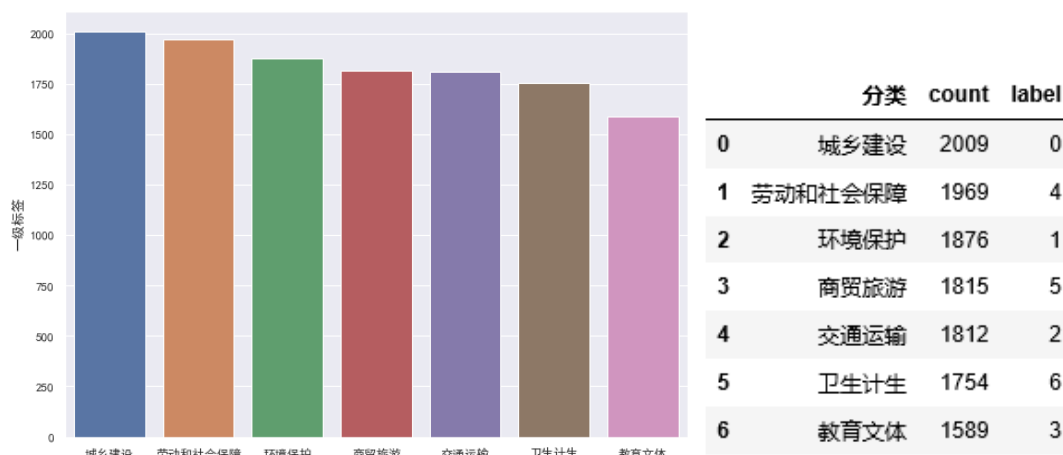


图 7 文本增强后各类的数量

3.4 文本特征提取

文本特征提取的意思是将长度不一的文本内容经过处理简化为长度固定的数值，即提取出词向量作为文本的基本结构。该过程可降低向量的空间维度但又不破坏文本内容的核心意思，提高了文本挖掘的计算速度和处理效率。其中，词向量指的是将词数字化，即转化为向量，较好的词向量可以实现语义相近的词在词向量空间里聚集在一起，为后续的文本分类和文本聚类等操作提供了便利。

首先，由于已对文本内容进行中文分词以及去停用词，现将分词后的词进行向量化处理，接着，本文采用 Scikit-learn^[5]（简称 SKlearn）中的 CountVectorizer 函数进行词向量训练，该方法被称为词袋法，此方法根据每个单词出现的频率组成一个特征矩阵，每一行表示一个训练文本的词频统计结果。需要注意的是，本文还增加了一个过滤条件，对于在文档中词频超过 80% 的关键词进行删除，理由是该类词过于平凡，无法代表文本的核心内容，同时，对于在文档中出现次数小于 2 的关键词也一并删除，理由是该类词过于独特，会使模型训练结果过于主观。

文本特征提取的参数配置：

```
CountVectorizer(analyzer='word', binary=False, decode_error='strict',
dtype=<class 'numpy.int64'>, encoding='utf-8', input='content',
lowercase=True, max_df=0.8, max_features=None, min_df=2,
ngram_range=(1, 1), preprocessor=None, stop_words=None,
strip_accents=None, token_pattern='(?u)\\b[^\\d\\W]\\w+\\b',
tokenizer=None, vocabulary=None)
```

3.5 模型介绍

(1) Logistic regression

Logistic Regression Classifier^[6] 逻辑回归的主要思想就是用最大似然概率方法构建出方程，为最大化方程，然后利用牛顿梯度上升求解方程参数。

(2) XGBoost

XGBoost 是一个优化的分布式梯度增强库^[7]，它是对梯度提升算法的改良，在求解损失函数极值过程中运用了牛顿法，把损失函数用泰勒公式展开到二阶，且在损失函数中放入正则化项。

(3) Random forest

Random forest 又叫随机森林，该算法是通过训练多个决策树，而在构建决策树过程中可进行任何减枝动作，接着生成模型，然后结合多个决策树的分类结果进行投票，以此来实现分类。

(4) SVM

支持向量机是一类可监督学习方式对数据进行二分类的广义线性分类器，

(5) Decision tree

决策树是一个预测模型；他代表的是对象属性与对象值之间的一种映射关系。树中每个节点表示某个对象，而每个分叉路径则代表的某个可能的属性值，而每个叶结点则对应从根节点到该叶节点所经历的路径所表示的对象的值。

(6) Naive Bayes

贝叶斯分类器是一类算法的总称，因为这类算法的基础都是贝叶斯定理。而本文用到的是朴素贝叶斯分类器 MultinomialNB

(7) KNN

KNN 算法又称为 k 近邻分类(k-nearest neighbor classification)算法^[8]。它是从训练集里找出和新数据最接近的 k 条记录，然后根据这些记录的主要分类来决定新数据的类别。在 KNN 算法中，通过计算对象之间的距离来作为各个对象之间的非相似指标，避免了对象之间的匹配问题，一般计算距离使用欧式距离或者曼哈顿距离。同时 KNN 算法通过根据 K 个对象里占优的类别进行决策，而不是单一的对象类别决策

(8) VotingClassifier

VotingClassifier 是一个结合多个概念上不同的机器学习分类器，使用多数投票或平均预测概率的方法让这些不同的分类器对同一组数据来预测类标签，最终通过投票的方式来选出模型公认的最佳结果。这样的集成分类器适用于一组结果同样表现良好的模型，以便于平衡它们各自的弱点。

3.6 模型训练

本文首先选择 7 种常用的机器学习分类器对数据进行分类：Logistic regression, XGBoost, Random forest, SVM, Decision tree, Naive Bayes, KNN, 从数据集中随机选取 80%的数据作为训练集, 20%的数据作为测试集, 验证方法采用 5 折交叉检验, 最后得到各分类器的准确率, 其中准确率最高的前三个分类器为 Logistic regression, XGBoost, Random forest, 分别为 0.9172、0.9010 以及 0.8759。

表 1 各分类器的结果

分类器	准确率
Logistic	0.917243
Xgboost	0.901
Randomforest	0.87591
SVM	0.8516
DecisionTree	0.78857
Bayes	0.789
KNN	0.565

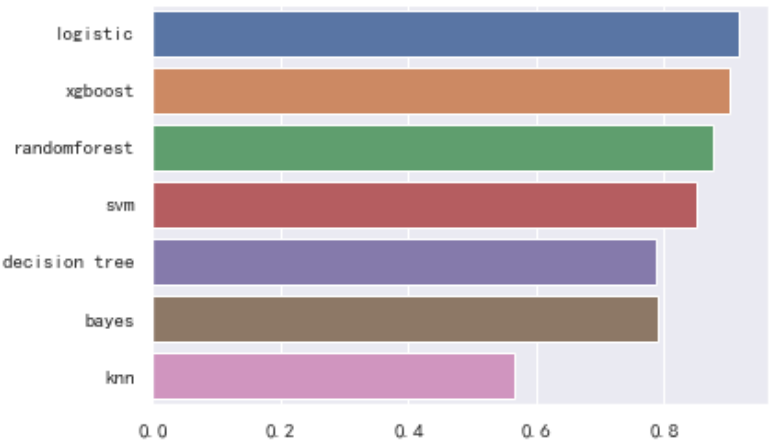


图 8 各分类器结果

3.7 模型融合与结果

根据各个分类器训练的结果，本文选出准确率最高的前三个分类器分别是 Logistic regression, XGBoost, Random forest 来进行模型融合，模型融合的方法采用投票法。通过 SKlearn 中的 VotingClassifier 的 Soft Votting 方法进行模型融合，此方法不是直接地按少数服从多数的原理决定结果，而是增加了权重的功能，即可以为不同的分类器设定不同的权重，区分每个分类器的重要程度。根据模型融合后的最终结果得出 f1_macro 的值为 0.9267。模型融合对 7 个标签的分类结果如图 8。

	precision	recall	f1-score	support
class 0	0.87	0.88	0.88	416
class 1	0.97	0.94	0.96	395
class 2	0.93	0.92	0.93	365
class 3	0.95	0.94	0.95	309
class 4	0.93	0.94	0.94	374
class 5	0.89	0.92	0.90	354
class 6	0.96	0.93	0.95	352
accuracy			0.93	2565
macro avg	0.93	0.93	0.93	2565
weighted avg	0.93	0.93	0.93	2565

图 8 各分类表现

四、问题二

4.1 处理流程图

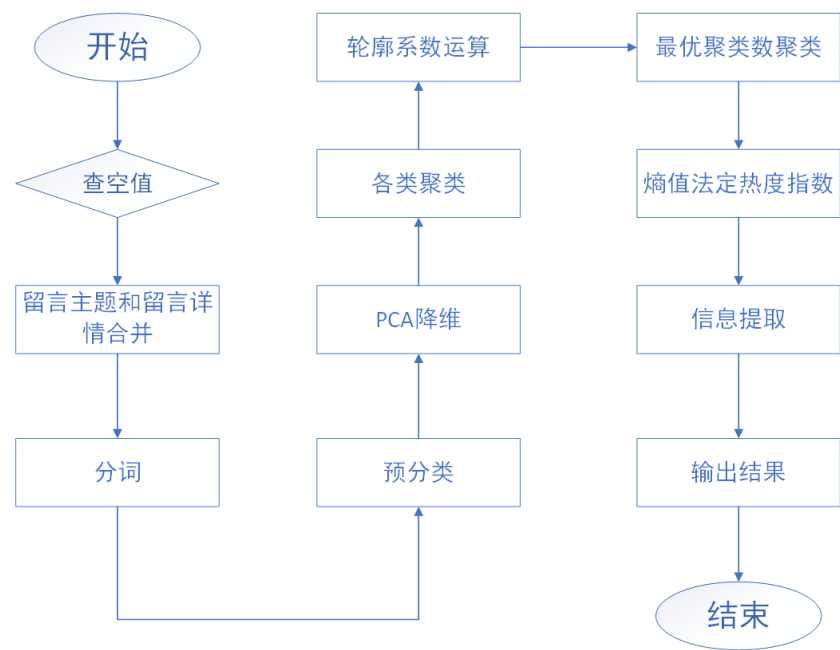


图 9 问题二流程图

4.2 数据预处理

首先，本文对附件 3 的数据进行空值处理，经检测，附件 3 中的数据并无空值，因此无需对其处理，接着，将留言主题和留言详情结合，即将两类句子拼凑在一起，同时考虑留言主题和留言详情的情况，提高结果的准确性，最后，采用问题一所述的中文分词与去停用词方法对文本内容进行分词并利用文本特征提取的方法对文本内容向量化，提高文本挖掘效率。

4.3 预分类

为了提升后续聚类的准确性且减少运算时间，本文先使用问题一里面用到的模型，即最终集成的融合模型，对预处理过的数据进行分类，分类结果为：标签 0 “城乡建设” 共有 2264 条，标签 1 “环境保护” 共有 309 条，标签 3 “教育文体” 共有 380 条，标签 4 “劳动和社会保障” 共有 341 条，标签 5 “商贸旅游” 共有 600 条，标签 6 “卫生计生” 共有 98 条。

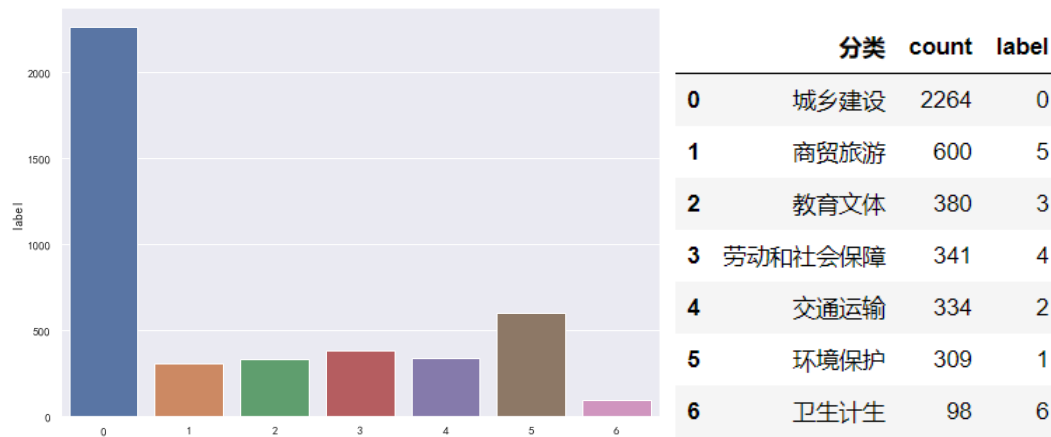


图 10 分类数量

4.3 降维

4.3.1 PCA 算法原理

PCA 算法又称主成分分析法，其原理是假设需将 $X=\{x_1, x_2, x_3, \dots, x_n\}$ 降维至 k 维，需要先去中心化，计算协方差矩阵 $\frac{1}{n}XX^T$ ，接着，用特征值分解方法求出协方差矩阵 $\frac{1}{n}XX^T$ 的特征值与特征向量，然后对特征值依照从大到小的顺序进行排序，选择其中最大的 k 个特征值，再将它们所对应的 k 个特征向量分别作为行向量组成特征向量矩阵 P ，最后，将数据转换为以 k 个特征向量构建的新空间中，即 $Y=PX$ 。

4.3.2 处理过程

由于数据量较大会导致运算时间较长，为避免出现这种情况，本文采用 PCA 降维的方法以降低向量的维度空间，减少数据量并设定保留 95%的数据，最终降维后结果如表 2。

表 2 降维前后比较

类别	降维前数据维度	降维后数据维度
卫生计生	(98, 3366)	(98, 75)
环境保护	(334, 6228)	(334, 286)
交通运输	(334, 6228)	(334, 286)
劳动和社会保障	(341, 8252)	(341, 280)
教育问题	(380, 9161)	(380, 302)
商务旅游	(600, 12958)	(600, 487)
城乡建设	(2264, 26214)	(2264, 1714)

4.4 K-means 聚类

4.4.1 K-means 算法原理

K-means 算法又称 K-平均或 K-均值算法^[9]，是一种常用的聚类算法。该算法的主要思想是通过迭代的过程把数据集分成不同的类别，将评价聚类性能的准则函数达到最优。其原理是先将数据分为 K 组，然后随机选取 K 个数据点来作为初始的聚类中心，接着，计算每个数据点到各个聚类中心之间的距离，再把每个数据点分配给距离它最近的聚类中心，以聚类中心和分配完毕的数据点共同构成一个聚类，因此，每分配一个样本，聚类中心就会根据现有的数据点进行重新计算，这一过程将会不断重复直至达到期望的结果为止。

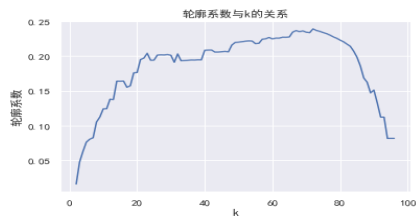
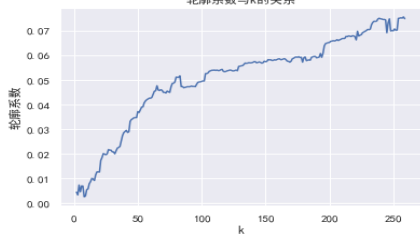
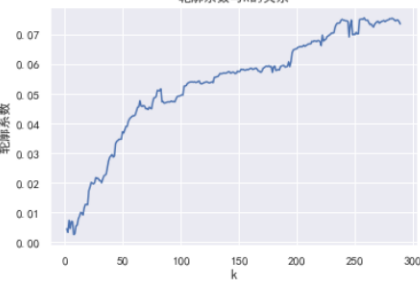
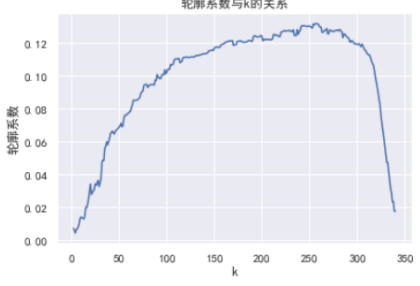
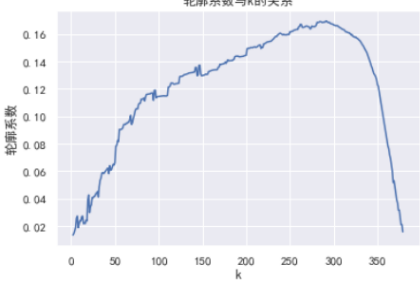
4.4.2 轮廓系数

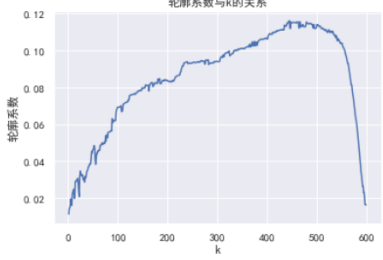
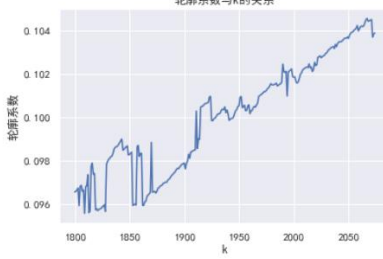
轮廓系数结合了聚类的凝聚度和分离度，其可用于评估聚类的效果，该系数处于-1 至 1 之间，系数越大，表示聚类效果越好，反之，则聚类效果越差。轮廓系数公式为：

$$s = \frac{a - b}{\max(a, b)}$$

其中 b 为该点与本类中其他点之间的平均距离，a 为该点与非本类点之间的平均距离，若 s 值越接近 1，则说明聚类的表现越佳。聚类结果如下表 3。

表 3 轮廓系数与聚类数关系

类别	最优聚类数	轮廓系数与聚类数的关系图
卫生计生	72	
环境保护	258	
交通运输	258	
劳动和社会保障	257	
教育文体	291	

商务旅游	445	
城乡建设	2067	

4.5 数据提取与合并

数据提取指的是将文本内容中的地点与人群提取出来，为了提取地点，本文采用了正则匹配法搜索与地点有关的关键词，公式为：

```
re.search( r' ([a-zA-Z]+[0-9] | [a-zA-Z]) + ([市] | [县] | [区] | [镇]) ', line, re.M | re.I)
```

对于人群的提取，文章根据词袋法自定义了有关人群的词袋并进行匹配，人群词袋为：['学生', '老师', '医生', '业主', '住户', '工人', '司机', '村民', '家长', '居民']

数据合并的意思是将聚类后的留言内容以类为单位计算该类的留言数量以及点赞数和反对数，方便后续建立热度指数。

4.6 热度指数

4.6.1 热度指数

热度指数反映了留言所述事件的受关注度，本文通过熵值法以及归一化的方法对留言数量及点评数(点赞数和反对数)进行了数据处理，定义热度指数公式，根据公式计算事件的热度指数。

$$Hot(\text{热度指数}) = D * Wd + L * Wl$$

其中 D 为点评数归一化的结果，L 为留言数量归一化的结果，Wd 为点评数得权重，Wl 为留言数量得权重

4.6.2 熵值法

熵值法是一种客观赋权法,其根据各个数据指标值所提供的信息的大小来确定指标权重,对某个指标,如果数据之间的差距越大,则该指标在综合评价中所起的作用越大;如果某项指标的指标值全部相等,则该指标在综合评价中不起作用。

本文定义一个新指标,点评数,即为点赞数与反对数之和,而同属某一事件的留言数量为类留言数量,由于聚类后的类留言数量与点评数的重要程度不同,本文通过熵值法对两者设定权重,达到突出重要性的目的,最终得到权重为:

表 4 指标与权重

指标	权重
点评数	0.625499
类留言数量	0.374501

4.6.3 归一化方法

由于点评数与类留言数量在数值大小方面差距较大,会对最后的计算结果造成较大误差,本文运用归一化的方法使点评数以及类留言数量的数值范围均处于 0 至 1 之间。

点评数的归一化处理公式为:
$$\frac{(\text{点评数量大于 1 的个数} - \text{点评数占总排名})}{\text{点评数量大于 1 的个数}}$$

类留言数量的归一化处理公式为:
$$\frac{(\text{当前类留言数量} - \text{类留言数量})}{(\text{最大类留言数量} - \text{最小类留言数量})}$$

根据热度指数公式,本文最终确定了热度指数最高的前五个事件,并对热点问题表进行部分展示如图 11。

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1	0.754635	2019-03-09至2019-07-23	西地省市民	对西地省聚利网的联名控诉书
2	2	0.750661	2019-08-12至2019-08-26	A4区市民	建议A市经开区泉星公园项目规划进一步优化
3	3	0.732286	2019-04-17至2019-09-06	A市市民	反映A市地铁3号线松雅湖站点附近地下通道问题
4	4	0.729996	2019-07-08至2019-08-16	A市市民	反对在A7县诺亚山林小区门口设立医院
5	5	0.714005	2019-01-03至2019-12-27	A市市民	“民生三问”A市住建委

图 11 热点问题表部分展示

以及对热点问题留言明细表进行部分展示如图 12。

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	207791	A0001061	西地省聚利人普	2019/3/9 1	胡书记:	15	0
1	227371	A0001192	对西地省聚利网	2019/4/11	对西地省聚	16	0
1	227676	A0001192	A市聚利网诈骗	2019/7/23	我是西地省	1	0
1	229554	A0002544	西地省聚利人普	2019/4/14	胡书记:您	0	0
1	253735	A0002544	西地省聚利人普	2019/4/14	您好!请	1	0
2	238692	A0008034	建议A市经开区	2019/8/12	目前A市经	16	0
2	273741	A0008034	建议进一步优化	2019/8/13	目前A市经	0	0
2	278281	A0003269	建议A市经开区	2019/8/22	目前A市经	0	0
2	278545	A0003684	给A市经开区泉	2019/8/26	目前A市经	13	0

图 12 热点问题留言明细表进行部分展示

五、问题三

5.1 处理流程图

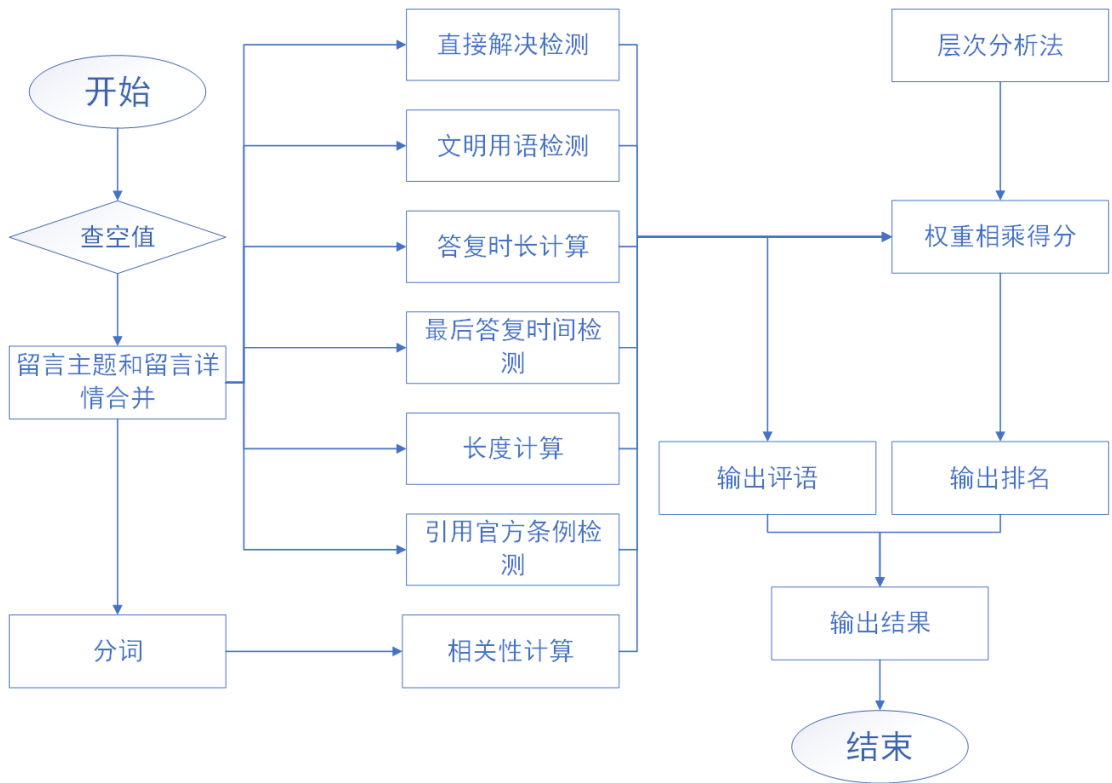


图 13 问题三流程

5.2 CCTI 评价体系构建

本文通过相关性 (Correlation)、完整性 (Completeness)、及时性 (Timeliness) 以及可解释性 (Interpretability) 四个指标基于层次分析法与模糊综合评价法建立 CCTI 评价体系，对答复意见的质量进行评价及点评。

5.1.1 相关性

相关性指的是答复内容与留言问题之间的相关程度，即两者应属于同一话题，而非答非所问，本文用余弦相似度来计算两者之间的相似度，根据相似度的大小来评价答复内容与对应问题的相关性。

5.1.2 完整性

本文根据“文本长度”、“文明用语”、“答复时间”以及“条理清晰”四个标准来定义完整性，目的是用于规范答复内容，文章采取正则匹配法搜索答复意见中的关键词，并从这四个标准对每一条答复内容的完整性进行评价。

(1) 文本长度

文本长度即答复内容的字数，字数是平台用户对答复意见最直观的感受，字数的长度与信息量正相关，因此，本文统计除标点符号之外的答复内容字数，以达到某一字符数范围视作达到文本长度标准，反之则未达到标准。

(2) 文明用语

文明用语是指答复内容中带有如“您好”、“谢谢”等礼貌用词，可以反映出政务部门工作人员的礼貌与素质，文章通过搜索答复意见中是否含有文明用语的关键词来评价句子是否达到标准。

(3) 答复时间

答复时间指的是答复句子结尾处带有的答复时间，从中体现出政务部门的规范性和严谨性，本文通过定位句子的结尾处并搜索该附近是否包含时间来决定其是否达到标准。

(4) 条理清晰

条理清晰的意思是答复段落的结构是否完整，可使留言用户能够更加直观感受到答复内容的完整性，文章采取检索答复句子中例如“首先”、“第一”等关键词来评价答复意见是否达到标准。

5.1.3 及时性

及时性是根据计算答复时间与留言时间的差来判断答复是否及时，反映出相关部门对问题的回复效率，及时性的计算排除双休和节假日，减小结果的误差，根据计算结果对答复的及时性进行评价。

5.1.4 可解释性

可解释性根据“引用官方条例”、“直接解决”和“间接解决”三个部分判断答复内容的解释强度，通过关键词搜索分别得出结果。

(1) 引用官方条例

引用官方条例是指答复内容中提到了一条或多条法律法规或相关部门制定的政策，增强答复内容的可靠性，使答复意见更有说服力，提高答复意见的可解释性。

(2) 直接解决

直接解决指的是直接帮助留言者解决问题的答复，答复内容不一定会引用官方条例，但是会通过工作人员的工作经验知识或调查核实直接回复于留言者，此答复意见的可解释性也会相对较高。

(3) 间接解决

间接解决的意思是指非直接解决问题，即没有出现直接解决的信号。

5.2 数据处理

5.2.1 数据合并与分词

本文针对附件 4 中的文本数据进行数据处理, 首先将留言主题和留言详情进行数据合并, 将其拼凑在一起, 然后对留言主题和留言详情以及答复意见的文本内容采用中文分词和去停用词的分词方法, 将句子切分为词并去掉句子中的标点符号、数字、英文及“省”、“市”、“县”、“区”, 为后续计算相关性等作准备, 提高运算的效率。

5.2.2 相关性

本文通过余弦相似度来计算留言主题和留言详情合并后的内容与答复意见之间的相关性, 余弦相似度的公式为:

$$\cos(\theta) = \frac{\sum(a \cdot b)}{\sqrt{\sum a^2} \cdot \sqrt{\sum b^2}}$$

a,b 分别为 a 句子和 b 句子计算句子词频后得出的向量

5.2.3 长度

本文针对留言主题和留言详情合并后的内容以及答复意见的长度进行统计, 长度是指排除了标点符号、数字、英文及“省”、“市”、“县”、“区”之后的留言内容和答复意见的长度。

5.2.4 最后答复时间

最后答复时间的意思在于判断答复意见里是否在句子的末尾留下了答复时间, 若有, 则为 1, 反之, 则为 0。

5.2.5 条理清晰

本文针对答复意见中是否出现类似“首先”、“第一”等词语判断答复意见是否条理清晰, 若出现关键词, 则为 1, 反之, 则为 0。

5.2.6 文明用语

本文针对答复意见中是否带有文明用语来判断其是否符合文明用语的标准, 若带有关键词, 则为 1, 反之, 则为 0。

5.2.7 直接解决

本文针对答复内容是否包含“回复如下”或“经过调查”等关键词来判断答复内容是否属于直接解决, 若属于, 则为 1, 反之, 则为 0, 即属于间接解决。

5.2.8 引用官方条例

判断文章中是否带有书名号等来判断答复内容是否有引用官方条例, 如有, 则为 1, 反之, 则为 0。

5.2.9 答复用时

本文计算答复时间距留言时间的天数, 计算结果已排除非工作日(双休及节假

日)。
答复用时=回答时间-提问时间-非工作日(节假日+双休)

5.2.10 归一化处理

为了将不同大小的数量级数据转变为互相可以进行运算或比较的数据，本文采取归一化的方式处理相关性、长度、最后答复时间、条理清晰、文明用语、直接解决、引用官方条例以及答复用时。其中，相关性采用偏大型梯形型隶属函数进行归一化处理，隶属函数的公式为：

$$Col(相关性) = \begin{cases} 0 & , x \leq 0.04 \\ \frac{x - 0.04}{0.446 - 0.04} & , 0.04 < x < 0.446 \\ 1 & , x \geq 0.446 \end{cases}$$

相关性采用偏大型梯形型隶属函数进行归一化，0.04 为总体数据 10%的位置数，0.446 为整体位置的 90%位置数

$$len(长度) = \begin{cases} 0 & , x \leq 36 \\ \frac{x - 36}{562 - 36} & , 36 < x < 562 \\ 1 & , x \geq 562 \end{cases}$$

其中 36 为总体数据 10%的位置数，562 为整体位置的 90%位置数

需要注意的是，由于答复用时是越小越好，因此采用偏小型梯形型隶属函数进行归一化，隶属函数的公式为：

$$time(答复时间) = \begin{cases} 1 & , x \leq 1 \\ \frac{27 - x}{27 - 1} & , 1 < x \leq 27 \\ 0 & , x > 27 \end{cases}$$

其中 1 为总体数据 10%的位置数，27 为整体位置的 90%位置数

最终，数据处理的结果如图 14。

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	答复意见分词结果	留言分词结果	相关性	长度	最后答复时间	条理清晰	文明用语	直接解决	引用官方条例	答复用时
0	2549	A00045581	2019/4/25 9:32:09	2019/4/25 9:32:09 2019年4月以来，位于A市A2区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核...	2019/5/10 14:56:53	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核...	景蓉花苑物业管理问题位于桂花坪街道公安分局宿舍景蓉	0.742602	0.686312	1	1	1	1	1	0.692308

图 14 处理结果展示

5.3 层次分析法

5.3.1 层次分析的概念

层次分析（简称：**AHP**）是说把一个复杂的多目标决策问题看成一个系统^[10]，先将目标分解为多个目标或准则，从而进行分解为多指标、条件或约束的若干层次，通过指定性模糊量化方法算出层次单排序和总排序。该方法较适用于具有分层交错评价指标的目标系统。

5.3.2 层次分析基本原理

层次分析法根据问题的性质和需达到的总目标，把问题分解成不同的组成因素，并按照因素间的相互关联影响以及隶属关系将因素按不同的层次进行聚集组合，形成一个多层次的分析结构模型，从而将问题归结为最底层相对于高层的相对重要权值的确定或相对优劣次序的排定。

下表是两两指标对比得打分表，根据重要性进行赋值。

表 5 打分表

因素 i 比因素 j	量化值
同等重要	1
稍微重要	3
较强重要	5
强烈重要	7
极端重要	9
两相邻判断的中间值	2, 4, 6, 8

根据重要性构建判断矩阵

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

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，需要进行一致性检验，所谓一致性检验是指判断矩阵允许有一定不一致的范围，但需要明确得是在范围内。一致性的检验是通过计算一致性比例 **CR** 来进行

$$\text{CR 公式: } CR = \frac{CI}{RI}$$

$$\text{其中 } CI = \frac{\lambda - n}{n - 1} \quad RI = \frac{CI1 + CI2 + \cdots + CIn}{n}$$

注：**n** 为 **n** 阶一致阵的唯一非零特征根为 **n**，**CI** 的值由判断矩阵计算获得，**RI** 的值查表获得。**CI**=0，有完全的一致性；**CI** 接近于 0，有满意的一致性；**CI** 越大，不一致越严重。

5.3.2 构建评价公式

层次结构图如图 15。

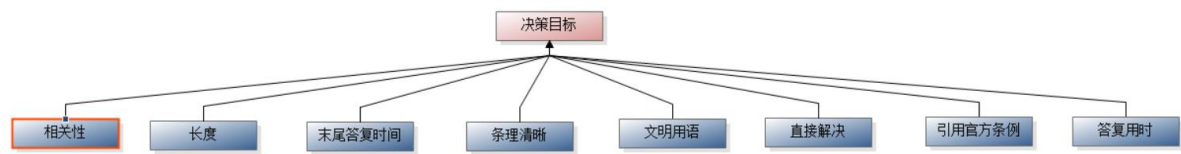


图 15 层次结构图

判断矩阵处理结果如图 16。

	相关性	长度	末尾答复时间	条理清晰	文明用语	直接解决	引用官方条例	答复用时
相关性		3	8	3	2	1/3	1/5	2
长度			5	4	3	1/4	1/7	2
末尾答复时间				1/3	1/2	1/7	1/8	1/6
条理清晰					2	1/5	1/7	1/4
文明用语						1/5	1/8	1/4
直接解决							1/2	2
引用官方条例								7
答复用时								

图 16 判断矩阵图

5.3.3 一致性检验

由于 $CR=0.0753<0.1$ ，即表示通过。计算结果如表 6。

表 6 权重结果

因子	权重
相关性	0.1211
直接解决	0.2073
答复用时	0.0977
引用官方条例	0.3774
长度	0.0933
文明用语	0.0361
条例清晰	0.0450
末尾答复时间	0.0221

评分公式如下：

$$w = \text{相关性} * 0.1211 + \text{直接解决} * 0.2073 + \text{答复用时} * 0.0977 + \text{引用官方条例} * 0.3774 + \text{长度} * 0.0933 + \text{文明用语} * 0.0361 + \text{条例清晰} * 0.045 + \text{末尾答复时间} * 0.0221$$

评价体系如表 7。

表 7 评分体系

评分	评价
0.8-1	极好
0.6-0.8	较好
0.4-0.6	一般
0.2-0.4	较差
0-0.2	极差

为了提升答复质量，本文加入留言评价，让答复者知道自己不足所在,可有知道如何改善并进步，最后自动化输出评分和对应的评价，输出结果在第三题附件/处理结果.xls。

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	评分	评价								
74486	UU00863	举报F9市	2019/3/17	F9市长塘村	您的留言	2019/4/26	0	极差,文本长度较小,缺失最后答复时间,条理不清晰,缺少使用礼貌词,间接解决问题,答复用时较短								
74555	UU008839	建议F市对	2019/1/22	现在农村	您的留言	2019/3/8	0	极差,文本长度较小,缺失最后答复时间,条理不清晰,缺少使用礼貌词,间接解决问题,答复用时较短								
75000	UU008153	请求恢复F	2018/3/25	F7县	您的留言	2018/6/29	0	极差,文本长度较小,缺失最后答复时间,条理不清晰,缺少使用礼貌词,间接解决问题,答复用时较短								
74851	UU008665	咨询F2区	2018/7/12	我叫李晨,	您的留言	2018/10/3	0.000156	极差,文本长度较小,缺失最后答复时间,条理不清晰,缺少使用礼貌词,间接解决问题,答复用时较短								
93090	UU008165	咨询J11市	2018/7/10	J11市政府	您的留言	2018/9/20	0.001064	极差,文本长度较小,缺失最后答复时间,条理不清晰,缺少使用礼貌词,间接解决问题,答复用时较短								

图 17 输出结果部分展示

六、总结

本文的主要目的是利用文本挖掘与数学建模技术建立关于留言内容的标签分类评价模型，之后对归类后的留言建立了热度评价模型，然后对留言的答复建议建立了评价方案并且进行实现。

对于问题一，本文对附件二的留言主题、留言详情进行分词，并将词向量化，把得到的词向量放入多个机械学习分类器模型进行训练，最后利用模型融合训练出最优的标签分类模型。这次使用的是机械学习的模型，以后可以根据实际情况尝试使用深度学习的模型；对于问题二，本文对附件三分类后使用无监督学习的K-means 算法进行聚类，利用熵值法和模糊综合评价法来确定热度指标，未来可以尝试使用多种无监督学习的算法进行对比；对于第三题，本文使用层次分析法对答复意见进行评价，不足的是结果可能存在主观性判断，未来可以研究如何更客观地进行评定。

最后，C 试题要解决的是“智慧政务”会面临的问题，即从群众中收集到的文本数据量不断在增长，若单纯依靠人工对留言进行划分以及对热点事件进行整理，不仅导致工作效率降低，而且会加重相关部门的工作量。因此通过解决 C 题的问题，不仅可有效推进“智慧政务”的步伐，文章中提到的方法也可尝试用于医院里病例的分类、划分出医院里受理最多的病种以及对医生给病人的回复做出质量评价等问题。

参考文献

- [1] www.wdzt.com.[EB/OL]. <https://www.wdzt.com/news/hydongtai/29400.html>.-.
- [2] 中文常用停用词表（哈工大停用词表、百度停用词表等 <https://github.com/goto456/stopwords>
- [3] 100+ Chinese Word Vectors 上百种预训练中文词向量 <https://github.com/Embedding/Chinese-Word-Vectors>
- [4] 百度翻译 api. <https://api.fanyi.baidu.com/>
- [5] SKlearn 介绍 <https://scikit-learn.org/stable/>
- [6] 维基百科 <https://zh.wikipedia.org/wiki/%E9%82%8F%E8%BC%AF%E8%BF%B4%E6%AD%B8>
- [7] XGBoost 介绍 <https://xgboost.readthedocs.io/en/latest/>
- [8] KNN 算法 <https://www.cnblogs.com/ybjourney/p/4702562.html>
- [9] K-均值聚类（K-means）算法 <https://www.cnblogs.com/ybjourney/p/4702562.html>
- [10] 层次分析法(AHP) <https://zhuanlan.zhihu.com/p/39993228>