“智慧政务”中的文本挖掘应用

摘要

随着时代的发展，智慧政务也渐渐进入大众的视野，网上政务平台凭借其便捷性，普遍性，成本低的优点，成为了目前政府了解民生的主要平台。因此，政务平台中的留言挖掘，文本信息的处理，热点问题的发现，留言的回复,对政府了解民生，汇聚民生问题，解决民生问题有着重大意义。

针对问题一：首先，通过对获取的数据进预处理，将文本标签转化为id，删除无意义的用词，然后利用jieba分词工具实现对留言信息进行分词，文本向量化，预处理完成后，在对数据进行$TF-IDF$值的计算，使用sklearn方法抽取文本的$TF-IDF$的特征值。再采用sklearn中的chi2进行卡方检验用来检验数据的拟合度和关联度，通过尝试不同的模型评估它们的准确率，然后使用朴素贝叶斯分类器测试，建立关于留言内容的一级标签分类模型，F分数（$F-score$）是分类问题的一个重要衡量指标，最后对模型进行评估。

针对问题二：明白热度指标的定义后，首先进行文本去词与文本分词的文本预处理，再利用python词云图绘制出热度指标词云图，得到词频统计的结果，然后再计算文本相似度，再将这些处理好文本统计，取出前五得到热点问题表。

针对问题三：对于广大民众的留言回复之后，对留言的回复进行相应的评价，根据留言回复的充实性，及时性，相关性，可解释性，完整性五个方面，利用余弦相似度算法、层次分析法和基于理想点逼近法（TOPSIS）评估评价等级进行留言回复的评价，利用层次分析法得到其五个方面不同的权重，进行评分，可以得出该回复的最终质量,利用基于理想点逼近法（TOPSIS）评估评价等级,将回复评出优秀、合格和不合格。构建一套完整的智慧政务平台留言回复评价系统。

关键字:$TF-IDF$算法，$F-score$, 词频统计, 层次分析法, 基于理想点逼近法.

Application of text mining in "smart government"

Abstract

With the development of the times, smart government has gradually come into the public's vision. With the advantages of convenience, universality and low cost, online government platform has become the main platform for the government to understand people's livelihood. Therefore, the mining of message, the processing of text information, the discovery of hot issues and the reply of message in the government platform are of great significance for the government to understand the people's livelihood, gather people's livelihood problems and solve people's livelihood problems.

To solve the problem 1: firstly, through preprocessing the acquired data, the text label is transformed into ID, and the meaningless words are deleted. Then, the message information is segmented by using the word segmentation tool of Jieba, and the text is vectorized. After preprocessing, the TF-IDF value of the data is calculated, and the feature value of TF-IDF of the text is extracted by using the sklear method. Chi2 in sklearn is used to test the fit and correlation of data. Different models are used to evaluate their accuracy. Then naive Bayes classifier is used to test and build a first level label classification model about message content. F- score (F- score) is an important measure of classification problem. Finally, the model is evaluated.

For problem 2: after the definition of heat index is understood, the text preprocessing of text word removal and text segmentation is carried out first, and then the word cloud chart of heat index is drawn by using Python word cloud chart to get the results of word frequency statistics, then the text similarity is calculated, and then the text statistics is processed well, and the first five hot issues are obtained.

In response to question 3: after the response of the general public's message, the response of the message will be evaluated accordingly. According to the five aspects of the message response, namely, the substantiality, timeliness, relevance, interpretability and integrity, the cosine similarity algorithm, AHP and TOPSIS are used to evaluate the evaluation level of the message response, The final quality of the reply can be obtained by using the analytic hierarchy process (AHP) to get the different weights of the five aspects, and the evaluation grade can be evaluated based on TOPSIS, and the reply will be evaluated as excellent, qualified and unqualified. Build a complete set of intelligent government platform message reply evaluation system.

Keywords:TF-IDF algorithm,F-score,word frequency statistics,AHP,TOPSIS.

目录

1. 引言

1.1 背景

随着时代的发展，微信、微博等问政平台的大量出现，以及大数据，人工智能，云计算等技术的不断进步成熟，广大民众可以在相关政务平台留言反馈自己的问题，寻求解决方法，围绕着这些问题，智慧政务的文本挖掘显得极其重要，它可以更加便捷，高效的解决政府对于民生，民意的了解和留言回复。

1.2 目标

智慧政务平台作为政府了解民意、汇聚民智、凝聚民气的重要渠道，本次数据挖掘针对群众问政留言记录的文本数据。

首先对文本进行基本的及其预处理、中文分词、停用词过滤后，文本向量化，进行TF－IDF值计算，再通过建立包括朴素贝叶斯分类器，机器学习模型的选择与评估，深度学习等方式，实现对留言文本的标签规划、热点词频统计，词云图显示等以及进一步挖掘并分析答复的多方面判断，通过使用对数量化文本长度判断回复充实性，通过量化回复与留言的时间差判断回复及时性通过余弦相似度算法得到回复相关性的判断，通过层次分析法得到多方面的权重，然后对答复进行各方面的评价。利用基于理想点逼近法（TOPSIS）评估评价等级。以期望有善提升政府的管理水平和施政效率。

2. 分析方法与过程

2.1 总体流程

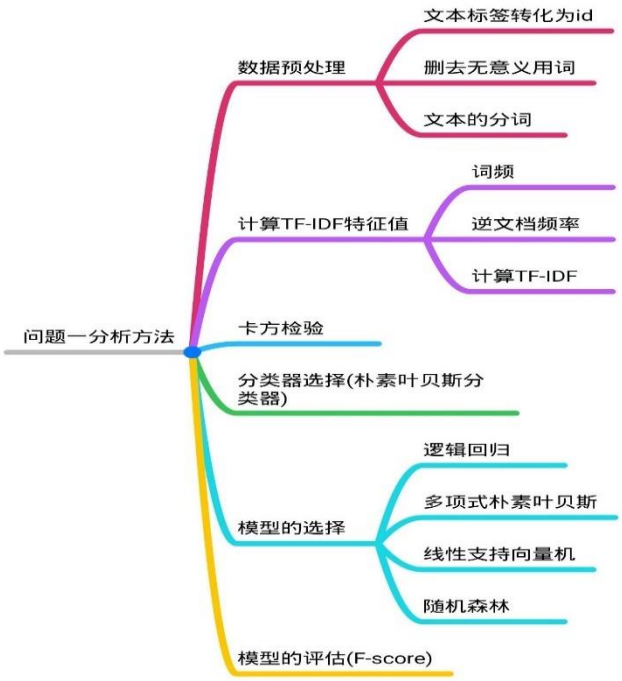

图 1 问题一的分析方法

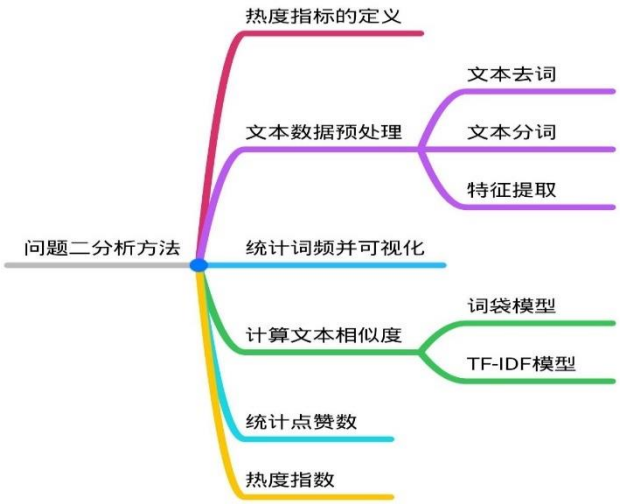

图 2 问题二的分析方法

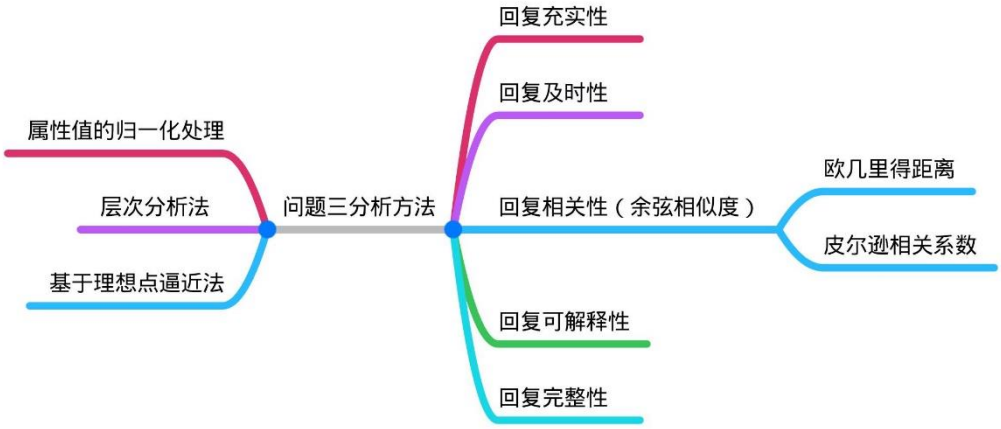

图 3 问题三的分析过程

2.2 具体步骤

2.2.1 问题一

2.2.1.1 文本数据

问题一使用的实验数据为题目所给的文本评论数据，在这些数据中本文要筛选出留言主题内容和对应的一级分类，每条留言根据其内容分别对应不同的留言主题。这些留言主题都是来自于部分群众的留言意见所归纳整合，用群众的留言意见来进行一级分类的构建虽然会更加简便以及计算量小，但是模型的准确度不高，所以我们选择对于留言的内容构建一级分类模型。

由附件一可知总共有 15 个一级分类，包括城乡建设、党务政务、国土资源、环境保护、纪检监察、交通运输、经济管理、科技与信息产业、民政、农村农业、商贸旅游、卫生计生、政法、教育文体、劳动和社会保障。并且分别对应了不同的二级，三级标签。由附件二可知，给出的文本数据里面的 9211 条留言对应了 7 个一级标签。

2.2.1.2 文本评论预处理

在取到留言内容和对应的一级分类后，本文首先要进行文本数据的预处理。在留言数据里存在着大量价值极低的条目，如果将这些留言数据也纳入到分词中，得

到的结果会影响后面的分析以及模型的建立，得出的结果质量也是会存在问题的。那么在利用这些文本评论数据之前必须要进行文本预处理，把这些大量的没有价值的词语给剔除掉。

本文运用python3.6 对这些文本数据预处理，主要由 3 个部分组成：将一级分类（文本标签）转化成id格式，删去留言主题无意义用词以及文本的分词。

1.文本标签转化成id

文本标签转化成id顾名思义是由文本标签转化成数字。因为在机器学习中，机器是根据词向量来进行学习的，所以本文必须将文本转化成向量或者数字的形式来更好的构建模型以及后续模型的训练。

由于这是一个基于文本的多分类问题所以本文不能简单的将标签转化成 0 和 1，由于这里有 7 个分类，这 7 个分类，分别是城乡建设，劳动社会保障，教体文体，商贸旅游，环境保护，卫生计生和交通运输。所以在这里本文将文本标签转化为 0-6 以便下一步的处理，经过处理后，可以将 9211 条留言划入相应的标签，这样方便下面对数据进行处理。处理后对应的各一级对应的id和数量如下表：

表 1 一级标签对应的id

一级标签	id	数量
城乡建设	0	2009
劳动和社会保障	1	1969
教育文体	2	1589
商贸旅游	3	1215
环境保护	4	938
卫生计生	5	877
交通运输	6	613

2.删去无意义用词

由于群众留言文本数量较多，难免会出现一些无意义的用词，所以需要删去这部分无意义用词的处理。在群众留言中，大多数都有带有地名以及标点符号的出现，例如“A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市A市，尽快整改这个极不文明的路段。”，在此句中出现了大量标点符号以及无意义用词，所以为了保证后续模型的建立的准确性，要删去这些无意义用词，同时在中文中还有着停用词的出现，例如“在”“啊”“呢”等出现频率极高但是又毫无意义的词汇，对系统分析和预测文本

不但没有任何帮助，反而还会增加运算的复杂度和系统的开销，所以本文在使用这些数据前要将文本处理干净。

这里可以定义一个删除除了汉字之外所有符号的函数，首先将这些群众留言主题转化成字符串，然后定义一个正则表达式来筛选出文本中除了汉字之外的所有内容，最后我们导入停用词表，将这些无意义且出现频率高的停用词过滤掉。将一些无意义的词语清除后，得到的文本会更加精简，方便下一步将文本数据进行分词。

3.文本的分词

分词就是将连续的字符序列按照一定的规范重新排列组合而成此序列的过程。在英文分词中，以空格作为一个自然分界符，但在中文中，字、句、段都可以通过明显的分界符来简单划界，但是唯有词没有一个形式上的分界符，所以相比于英文的分词，中文的分词就要复杂得多。

而分词结果的准确性又对于后续的模型构建有着不可忽视的影响。如果分词效果不佳，后续算法也无法达到理想的效果，不同分词的选择将直接影响词语在文本中的重要性，从而影响特征的选择。

中文分词算法主要由基于字符串匹配分词方法，基于理解分词方法和基于统计分词方法 3 部分构成。

(1)基于词典分词算法

基于词典的分词算法也叫做字符匹配分词算法。该算法是按照一定的策略将待匹配的字符和一个已建立好的“充分大的”词典中的词进行匹配，若找到某个词条，则说明匹配成功，识别了该词。常见的基于词典的分词算法有：正向最大匹配法、逆向最大匹配法和双向匹配分词法。

(2)基于统计的机器学习算法

近年来,基于统计的分词方法成为中文分词技术的主流。该算法的依据是:在汉语上下文中,虽然没有任何的词边界,但相邻字之间联合出现的频率越高,则其越有可能形成一个词。因此,该算法首先对语料中的字符因此,该算法首先对语料中的字符串进行全切分,然后对所有可能相邻共现的字组合进行频率统计,计算它们的互现信息。这样便将语言问题转化成了统计问题,继而建立反映相邻字互信度的概率模型,从而完成新词识别和切分。这种算法仅需对语料中每个词的频率进行计算,而不依赖于大规模的机器词典,因此也称为无词典分词方法。该方法常用的统计量主要有:信息熵、互信息、t-测试差等;相关的分词模型包括:N元文法模型（N－gram）、最大熵模型（MaximumEntrop，ME）、隐马尔可夫模型(HiddenMarkovModel，HMM)以及条件随机模型(ConditionalRandomFields,CRFs)等。

(3)基于理解的分词方法

该方法也称为基于人工智能的分词算法。它在一定程度上是基于“先理解后分

词”的技术路线设计的，旨在解决词典分词方法缺少全局信息，而统计分词方法缺少句子结构信息的问题，在理论上是一种最理想的分词方法。其基本思想是：在分词时模拟人脑对于语言的理解方式，根据句法、语义以及构词特点来进行分析，以达到词汇识别的效果。基于理解的分词系统一般分为分词系统、句法语义系统和总控系统三个部分，在总控系统的统一协调下，分词系统模拟人类的语言理解能力和思考过程，从句法语义系统中获得相关信息，并将其应用到词切分的歧义消除过程中常见的分词方法有专家系统分词法（Exper System）和神经网络分词法（BPNeural Network）两种。

而现在常用的分词器都是采用机器学习算法和词典相结合，一方面能提高分词的准确率，另一方面可以改善领域适应性。

在本问题中本文采用的python中优秀的第三方分词工具：jieba分词工具，jieba分词工具基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能的词情况构成有向无环图，采用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。使用jieba分词后的群众留言部分示例图如下：

	clean_review	cut_review
0	A3区大道西行便道未管所路口至加油站路段人行道包括路灯杆被圈西湖建筑集团燕子山安置房项目施工...	A3 区 大道 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑...
1	位于书院路主干道的在水一方大厦一楼至四楼人为拆除水电等设施后烂尾多年用护栏围着不但占用人行道...	位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电 设施 烂尾 多年 护栏...
2	尊敬的领导A1区苑小区位于A1区火炬路小区物业A市程明物业管理有限公司未经小区业主同意利用业...	尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 市程明 物业管理 有限 公...
3	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉...
4	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 自来水 龙头 水 霉...
5	我在2015年购买了盛世耀凯小区17栋3楼4楼两层共计2千平方一直以来我们按时足额缴纳物业费...	2015 年 购买 盛世 耀凯 小区 17 栋 楼 楼 两层 共计 平方 足额 缴纳 物业费...
6	由于西地省地区常年阴冷潮湿的气候加之近年气候逐渐更加恶劣地处月亮岛片区近年规划有楚江集中供暖...	西地省 地区 常年 阴冷 潮湿 气候 近年 气候 恶劣 地处 月亮 岛 片区 近年 规划 楚...
7	尊敬的胡书记您好家住A市A3区桐梓坡西路可可小城的居民长期以来经常停水小区业主业委会多次找物...	尊敬 胡书记 您好 家住 市 A3 区 桐梓 坡 西路 可可 小城 居民 停水 小区业主 业...
8	我们是梅家田社区辖区内的小区居民我们每年都依法依规向小区物业公司交纳了城市垃圾处理费我也认为...	梅家田 社区 辖区 小区 居民 依法 依规 小区 物业公司 交纳 城市 垃圾处理 费 环卫局...
9	尊敬的A市政府领导您们好我是A市A3区魏家坡巷的业主多年以来我们小区的脏乱差多次向社区反映都...	尊敬 市政府 领导 您们好 市 A3 区 魏家坡巷 业主 多年 小区 脏乱差 社区 得不到 ...

图 4 部分留言分词

2.2.1.3 计算 TF-IDF 特征值

在对文本进行预处理后，接下来就要计算与处理过后的文本的TF－IDF特征值，TF－IDF是一种用于信息检索与数据挖掘的常用加权技术。TF意思是词频，IDF意思是逆文本频率指数。TF－IDF是在单词计数的基础上，降低了常用高频词的权重，增加罕见词的权重。因为罕见词更能表达文章的主题思想，比如在一篇文章中出现了“中国”和“卷积神经网络”两个词，那么后者将更能体现文章的主题思想，而

前者是常见的高频词,它不能表达文章的主题思想。所以“卷积神经网络”的TF－IDF值要高于“中国”的TF－IDF值。这里本文使用机器学习中的sklearn方法来抽取文本的TF－IDF的特征值，并且使用了参数$ngram_range = (1,2)$,这表示文中除了抽取评论中的每个词语外,还要抽取每个词相邻的词并组成一个“词语对”。这样就扩展了特征集的数量,有了丰富的特征集才有可能提高本文分类文本的准确度。

TF－IDF算法具体公式如下：

1.计算词频TF

词频（TF）＝某个词在文章中出现的次数　　（2-1）

由于文章有长短之分，所以为了便于比较，将词频进行标准化。

$$\mathrm{TF} = \frac{\text{某个词在文章中出现的次数}}{\text{文章的总词数}} \quad (2\text{-}2)$$

2.计算逆文档频率IDF：它是用来是用于衡量关键词权重的指数，该关键词IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。

$$\mathrm{IDF} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}+1}\right) \quad (2\text{-}3)$$

注意：分母之所以要加一，是为了确保分母不为零。

3.计算TF－IDF：TF－IDF在本质上应该是TF×IDF

TF－IDF＝词频（TF）×逆文档频率（IDF）　　（2-4）

可以看出，TF－IDF与一个词在文档中出现的次数成正比，与该词在整个语言中出现的次数成反比。所以根据词语的TF－IDF值可以提取出各个一级分类所对应的特征。

2.2.1.4 卡方检验

卡方检验是一种统计学的工具，用来检验数据的拟合度和关联度。卡方检验在留言的处理中要进行特征选择，首先要去除重复的相似特征词汇，要保留其特有的词汇。比如，“医院，医生，医保”这三个词汇差别不大，它们这个特征是属于“卫生计生”这个标签。词汇间的差别越小，理论值与观察值的差别越小。它们之间的关联度最大。

卡方检验要得出每个分类中相关程度最大的词汇，自由度指的是不受限制的变量的个数。首先，根据预测得到期望频数，再得到实际频数，记期望频数为A，实际频数为B，又检验公式：

$$X^2 = \sum \frac{(B-A)^2}{A} \quad (2\text{-}5)$$

计算出来后，可以根据X^2的值与自由度的值查表得出P的值。总而言之，P值越小，观察值与理论值偏离程度越大。P越大，观察值与理论值偏离程度越小。两个变量间越没有差别，它俩具有很强的“相关性”，即它们之间的关联度最大。

在本问题中本文采用sklearn中chi2方法来进行检验，找出每个分类中关联度最大的两个词语和两个词语对。部分实例如下图：

表 2 一级分类对应关联度最大的词语

关联度 一级分类	关联度最大的两个词语	关联度最大的两个词语对
交通运输	（快递） （出租车）	（的士司机） （出租车司机）
劳动和社会保障	（退休） （社保）	（劳动关系） （退休人员）
卫生计生	（医生） （医院）	（社会抚养费） （乡村医生）
商贸旅游	（传销） （电梯）	（小区传销） （传销组织）
城乡建设	（小区） （业主）	（住房公积金） （公积金贷款）
教育文体	（学生） （学校）	（教育局领导） （培训机构）
环境保护	（环保局） （污染）	（周边居民） （环保局领导）

2.2.1.5 分类器的选择

为了训练监督学习的分类器，本文将文本转化成了TF－IDF值，有了词向量之后就可以开始训练我们的分类器，在分类器训练完成后，就可以对没有见过的文本来进行预测。

在这里本文先选择朴素贝叶斯分类器来进行测试，朴素贝叶斯分类器最适合用于基于词频的高维数据分类器，最典型的应用如垃圾邮件分类器等,准确率可以高达 95%以上。当模型训练完成后，我们可以自定义一个预测函数，让函数预测一

些文本的分类。

朴素贝叶斯算法：

1.贝叶斯理论：贝叶斯理论是在一个已发生事件的概率下，计算另一个事件的发生的概率。贝叶斯理论在概率论中可以表示为：

$$P(A|B)=\frac{P(B|A)P(A)}{P(B)} \tag{2-6}$$

记A，B为两个事件，P(A|B)表示在事件B发生的情况下事件A发生的概率，属于后验概率。P(A)表示随机事件A出现的概率，属于先验概率。P(B|A)表示在事件A发生的情况下事件B发生的概率，属于后验概率。P(B)表示随机事件B出现的概率，属于先验概率。

这里将贝叶斯理论运用到数据集中，这里Y是类变量，X是特征向量，得到：

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)} \tag{2-7}$$

2.朴素假设：对贝叶斯理论进行假设，假设事件A与事件B独立即：

$$P(A|B)=P(A)P(B) \tag{2-8}$$

则这里假设每个特征之间都是相互独立的，可以得出：

$$P(Y|X_1,\cdots,X_n)=\frac{P(X_1|Y)P(X_2|Y)\cdots P(Y)}{P(X_1)P(X_2)\cdots P(X_n)} \tag{2-9}$$

3.建立分类模型：这里运用类变量Y的所有可能的值计算概率，同时给出概率的最大的结果。数学表达式为：

$$y=\arg\max P(Y)\prod_{i-1}^{n}P(X_n|Y) \tag{2-10}$$

这里P(Y)代表类概率，$P(X_n|Y)$代表条件概率。

4.离散属性与连续属性：

离散属性：如果条件概率P(Y|X)中如果Y_n为离散值属性，则公式为：

$$P(Y_n|X)=\frac{|D_{X,Y_n}|}{|D_X|} \tag{2-11}$$

$|D_X|$表示在集合D中第 X 类样本的集合中的元素，$|D_{X,Y_n}|$表示是D_X中第n个属性值上取值为Y_n的样本集合的元素。

连续属性：如果条件概率P(Y|X)中如果Y_n为连续值属性，就需要用到概率密度函数，假设$P(Y_n|X)$满足正态分布，其公式为：

$$P(Y_n|X)=\frac{1}{\sqrt{2\pi}\sigma_{X,n}}\exp\{-\frac{(Y_n-\mu_{X,n})^2}{2\sigma_{X,n}^2}\} \tag{2-12}$$

2.2.1.6 模型的选择

在对进行卡方检验后，我们找到了一级分类对应的词语和词语对并且得到了他们的特征值接下来我们尝试不同的机器学习模型并且评估它们的准确率，我们将使用如下四种模型：

1. Logistic Rergession（逻辑回归）

逻辑回归是一种分类学习方法。通过对历史数据的表现来对未来的结果发生预测。例如，我们可以将购买的概率设置为因变量，将用户的特征属性设置为自变量，根据特征属性来判断购买的概率。在本问题中，我们将一级分类标签设置为因变量，将用户的留言设置为自变量，通过已知数据的学习来判断模型的准确率。

2. Mutinomial NaiveBayes（多项式朴素贝叶斯）

朴素贝叶斯算法是一组基于贝叶斯定理的监督学习算法，其“朴素”假设是：给定类别变量的每一对特征之间条件独立。多项式朴素贝叶斯算法是文本分类中使用的经典算法之一。通过特征数量和特征类别出现的概率来预估样本的应该出现的次数。

3. Linear Support Vector Machine（线性支持向量机）

支持向量机是一套由监督学习的方法集，经常用于分类、回归和异常点检测。它的基本思想时在特征空间中寻找间隔最大的分离超平面使数据得到高效的分类。在本问题中训练数据近似线性可分，可以引入松弛变量，通过软间隔最大化，来学习一个线性分类器。

4. Random Forest（随机森林）

随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别输出类别的众数决定。针对本问题的分类问题，随机森林中的每颗决策树通过学习后会预测最新数据属于那个分类。最终，哪一个分类被选择的最多，就预测这个最新数据属于哪个分类。

在介绍完以上 4 种模型后，本文采用sklearn中的 4 种模型来对分好类的数据进行学习，并且将这 4 种模型的准确率可视化如下：

表 3 各模型的准确率

模型	准确率
Linear Support Vector Machine（线性支持向量机）	0.872638
Logistic Rergession（逻辑回归）	0.817372
Mutinomial NaiveBayes（多项式朴素贝叶斯）	0.651466
Random Forest（随机森林）	0.408035

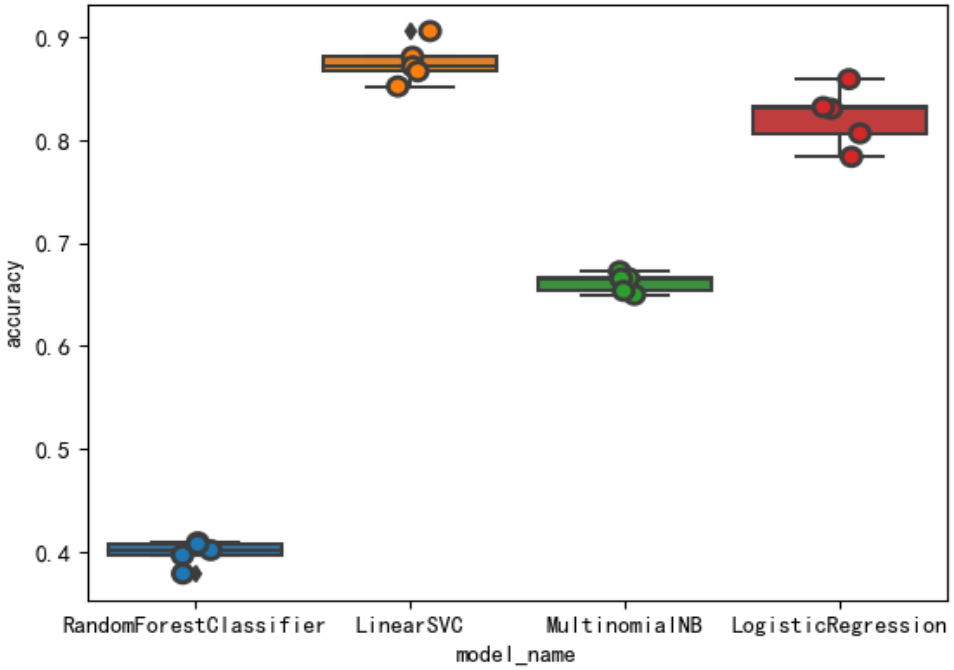

图 5 模型准确率

从以上箱体图上可以看出随机森林分类器的准确率是最低的，因为随机森林属于集成分类器(有若干个子分类器组合而成)，一般来说集成分类器不适合处理高维数据(如文本数据),因为文本数据有太多的特征值,使得集成分类器难以应付,另外三个分类器的平均准确率都在 80%以上。其中线性支持向量机的准确率最高。其次是逻辑回归的朴素贝叶斯。

2.2.1.7 模型的评估

在进行模型的比较后，可以看出效果最好的就是支持向量机模型，下面就针对平均准确率最高的LinearSVC（支持向量机）模型，来查看这个模型预测的结果。本文将数据的百分之八十作为训练集样本，百分之二十作为测试集样本，通过混淆矩阵来查看预测值和实际值的差距。

混淆矩阵也称误差矩阵，是表示精度评价的一种标准格式，用n行n列的矩阵形式来表示。具体评价指标有总体精度、制图精度、用户精度等，这些精度指标从不同的侧面反映了图像分类的精度。通过代码将混淆矩阵可视化如下：

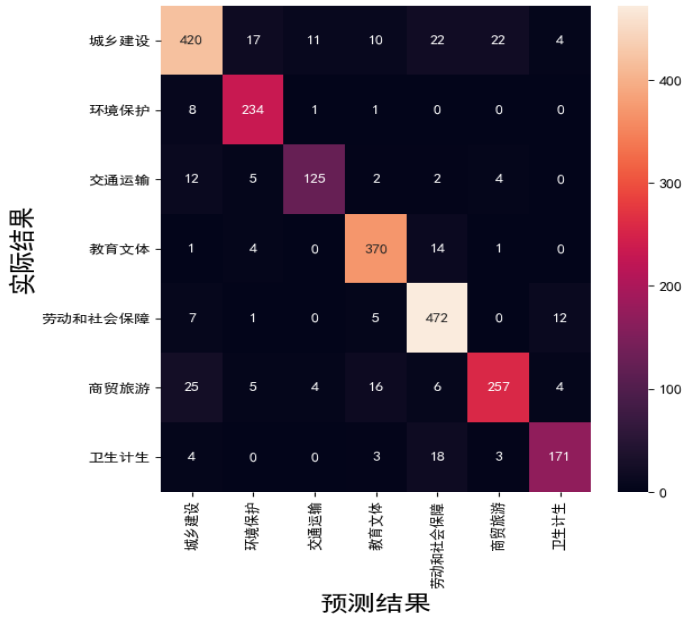

图 6 一级分类的混淆矩阵

从上图可以看出混淆矩阵主对角线上的元素就是预测正确的数量，除了主对角线以外都是预测错误的数量。从上图的混淆矩阵来看“交通运输”和“商贸旅游”的预测数据较差，而“环境保护”和“教育问题”的预测数据较好。究其原因可能是“商贸旅游”在卡方检验中有关联的词语“传销”、“电梯”等词语并不是太符合“商贸旅游”的特征导致预测结果略微有些偏差，而“环境保护”所关联的词语对“环保局”、“污染”等词语，也比较符合我们主观上的感受，所以预测结果较好。

在多分类模型一般不使用准确率(accuracy)来评估模型的质量,因为accuracy不能反应出每一个分类的准确性,因为当训练数据不平衡(有的类数据很多,有的类数据很少)时，accuracy不能反映出模型的实际预测精度,这时候我们就需要借助于F1分数指标来评估模型。

F1分数（F1 − score）是分类问题的一个重要衡量指标。一些多分类问题常常用F1 − score来作为最终评测方法。它是精准率和召回率的调和平均数，最大值为1,最小值为0。

$$F1 = 2\times\frac{precision\cdot recall}{precision+recall} \qquad (2\text{-}13)$$

1. 首先定义以下几个概念：

TP（True Positive）：预测答案正确

FP（False Positive）：错将其他类预测为本类

FN（False Negative）：本类标签预测为其他类标

2. 通过第一步的统计值计算每个类别下的precision和recall

精准度/查准率(precision)：指被分类器判定正例中的正样本的比重

$$\text{precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2-14}$$

召回率/查全率(recall)：指的是被预测为正例的占总的正例的比重

$$\text{recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{2-15}$$

准确率(accuracy)：代表分类器对整个样本判断正确的比重

$$\text{accuracy}=\frac{\text{TP+TN}}{\text{TP+TN+FP+FN}} \tag{2-16}$$

下面我们利用python来查看各个类的F1分数

表 4 各分类的F1分数

一级标签	**precision**	**recall**	**F1 − score**	**support**
城乡建设	0.82	0.95	0.88	402
环境保护	0.97	0.93	0.95	188
交通运输	0.95	0.71	0.82	122
教育文体	0.95	0.93	0.94	318
劳动和社会保障	0.90	0.94	0.92	394
商贸旅游	0.89	0.83	0.86	243
卫生计生	0.95	0.83	0.89	175
macroavg	0.92	0.88	0.90	1842
weightedavg	0.91	0.90	0.90	1842

其中在上表中macroavg代表的是所有F1值的算术平均值，可以代表整体样本的正确率，weightedavg是对每一类的F1值进行加权平均，权重为各个类别在样本正确值中所占的比例，同样也可以大致代表整体样本的正确率。support指每个分类中样本的数量。

从以上F1分数上看,"环境保护"类的F1分数最大为 0.95(只有 10 个预测错误)，“交通运输”类F1分数较差只有 0.82，究其原因可能是因为“商贸旅游”分类的词语对不能反映出其特征值，还有可能是“商贸旅游”用来学习特征值太少导致学习得不够充分。

2.2.2 问题 2

2.2.2.1 热度指标的定义

随着新媒体技术的发展，网络逐渐成为我国民众沟通情感、表达意见以及释放情绪的重要工具。当网民对某一事件的议论发展到一定热度，并外化到公共网络中，就形成了网络舆情，网络舆情是一种注意力召集令。新媒体中热度的定义也同样来源于舆情。所以在本问题中，可以将群众留言内容做一个归类，相似留言最多的作为一个权重指标，同时在网络平台中，点赞数也是一个重要的量化指标，所以本文将点赞数也作为一个重要的权重指标，通过对相似留言的查找以及点赞数的多少来综合确定留言的热度。

这里根据附件三给出的留言的点赞数，和将留言的内容做一级标签后的分类，可以知晓留言中一些相关内容的频率，本文可以把这两点作为留言热度指标定义的标准，从而得出留言热度的排名，任何汇总，分析。

2.2.2.2 文本预处理

1.文本数据：

本问题中采用的文本数据同样来自问题中所给的数据，从所给的文本数据中提取群众留言主题作为热度指标参考的依据，因为对于群众的留言内容来讲，留言主题不仅可以概括群众的留言内容，而且留言主题准确度更好同时也大大减少了对于计算机的计算量。

2.文本去词：

(1)文本去重：得到留言数据后，首先要进行文本的去重，就是将文本中重复的部分去除，这样处理可以让得到的数据方便下一步文本的挖掘。数据的预处理对文本进行去重是因为可能会出现一个人就一个问题多次进行留言，这样会导致后面对热点问题的挖掘出现偏差，比如：一个用户因为对于小区的周围噪音不满，他对此问题进行留言，这样得到的信息就是一个人的留言问题，这种信息没有太大的意义，对于热度问题的定义是再留言中许多人都反应了这个问题，一个人多次相同的留言不利于对文本的后期处理。

(2)文本清洗：得到数据后，文本中存在一些无用的词和标点，首先去除留言文本中的标点符号，将文本的主干显示出来。之后需要对一些停用词和一些无意义的词进行清洗，这样可以消除它对后来的文本挖掘的不良影响，停用词它是一些没有实际意义的词，但出现频率极高。例如：“的”，“了”，“啊”等，有些时候，留言文本中会出现一些意思一样的文本，比如：“希望政府尽快处理此类问题”，“希

望问题能得到解决”等，这些词和短句的出现，会让从留言文本中提取到的特征词都是一些无实际意义的词，这样对于热点事件的挖掘有极大的负面影响。这里需要获取停用词库和低频词库，停用词库我们可以直接在搜索引擎上搜索“停用词库”，能找到很多停用词库。可以在此基础上，自己添加一些留言文本中的常见停用词，低频词库，可以使用Counter等库获取所有句子中所有词的词频，通过筛选词频获得低频词库。获取停用词库和低频词库后，将词库中的词语删除。去除不必要的标签这一块在实际工作中需要灵活的使用，例如使用re库对文本做正则删除、替换，利用json库去解析json数据，又或者使用规则对文本进行相应的处理。

在本文体中去词方法与问题一大致相同，在此不再过多赘述。

3. 文本分词

在文本去词后同样要进行文本的分词处理，本文采用了python中的jieba库来进行分词，但唯一不同的是本文在本问题中需要加入用户自定义的词典，因为在群众留言这一文本中，有一些特定的文本被jieba工具分开，例如“A市”等地名分词后得到“A”“市”，经济学院分词后得到“经济”“学院”等等这些特殊的地名，首先构建未分词文件、已分词文件两个文件夹，将未分词文件夹按类目定义文件名，各个类目的文件夹下可放置多个需要分词的文件。为了保证分词后结果的准确性，本文加入自定义词典来辨别这些特殊的词汇以便后续工作的开展。去词并且分词后的部分结果如下图：

	clean_review	cut_review
0	A3区一米阳光婚纱艺术摄影是否合法纳税了	A3区 一米阳光 婚纱 艺术摄影 合法 纳税
1	咨询A6区道路命名规划初步成果公示和城乡门牌问题	咨询 A6 道路 命名 规划 初步 成果 公示 城乡 门牌
2	反映A7县春华镇金鼎村水泥路自来水到户的问题	A7 县 春华 镇金鼎村 水泥路 自来水 到户
3	A2区黄兴路步行街大古道巷住户卫生间粪便外排	A2区 黄兴路 步行街 古道 巷 住户 卫生间 粪便 外排
4	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	A市 A3区 中海 国际 社区 三期 四期 空地 夜间 施工 噪音 扰民
5	A3区麓泉社区单方面改变麓谷明珠小区6栋架空层使用性质	A3区 麓 泉 社区 单方面 改变 麓 谷 明珠 栋 架空层 性质
6	A2区富绿新村房产的性质是什么	A2区 富绿 新村 房产 性质
7	对A市地铁违规用工问题的质疑	A市 地铁 违规 用工 质疑
8	A市6路公交车随意变道通行	A市 路 公交车 随意 变道 通行
9	A3区保利麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2点施工扰民	A3区 保利 麓 谷林语 桐梓 坡路 与麓 松路 交汇处 地铁 凌晨 点 施工 扰民

图 7 部分群众留言的分词效果

4. 特征提取

通常会采用TF－IDF、Word2Ver等方式实现对文本特征的提取。此文本选取的是TF－IDF进行文本特征的提取。

TF意思是词频，即在留言中关键词出现的频率。记N为该关键词在留言文本中出现的次数，M为该留言文本中总体的词数。TF（词频)为两者之比 ，IDF意思是

逆文本频率指数。它是用来是用于衡量关键词权重的指数，该关键词的IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。记C为留言的总文档数，C_W为包含了该关键词的文档数。IDF（逆文本频率指数）就是两者之比去对数，注意：分母之所以要加一，是为了确保分母不为零。$TF-IDF$的计算：$TF-IDF$在本质上应该是$TF \times IDF$，所以$TF-IDF=TF \times IDF$。运用这个公式可以得出其权重。

运用$TF-IDF$进行特征提取，优点为简单快速，结果比较符合实际，缺点为单纯考虑词频，忽略了词与词的位置信息以及词与词之间的相互关系。

这里对于$TF-IDF$的算法，只进行简短的介绍，具体参见问题一中的$TF-IDF$的计算。

2.2.2.3 统计词频并可视化

在进行数据预处理操作后，本文可以统计词频来看出初步观察热度较高的几个地方或者事件是什么，然后再用python中的词云图来直观的看出热度较高的几个地方。词云图中字越大代表出现的频次越高。群众留言的词云图和频次表如下：

图 8 留言主题的词云图

表 5 留言主题的词频统计

Words	Counts
A 市	1878
A7	682
A3 区	437
扰民	278
A2 区	264
A1	209
投诉	202
噪音	198

从上图的词云图和词频统计表中我们可以看出，热点事件发生在A市、A7、A3、A2、A1区的事件较多，关于扰民、噪音、房屋、施工等问题的热点事件较多，对此我们可以进行一个初步的判断热点的地点以及事件。

2.2.2.4 计算文本相似度

本文虽然已经初步得出了热度较高的几个事件和地点，但是没有具体的得出所发生的事件，所以接下来要找出文本中相似的事件，计算文本的相似度，将相似事件进行归类。

相似度是数学上的概念，自然语言肯定无法完成，所有要把文本转化为向量。两个向量计算相似度就很简单了，欧式距离、余弦相似度等等各种方法都可以计算文本的相似度，那么如何才能将自然语言转化成向量是我们需要解决的关键问题。

1.词袋模型

最简单的表示方法是词袋模型。把一篇文本想象成一个个词构成的，所有词放入一个袋子里，没有先后顺序、没有语义。

例如：

“我喜欢看电影和打篮球。”

“他喜欢玩游戏和打篮球。”

由于在中文中文字是没有间隔的所以先要进行分词本问题也是使用结巴分词，分词后得到的词语是：

“我”“喜欢”“看电影”“和”“打篮球”

“我”“喜欢”“玩游戏”“和”“踢足球”

这两个句子，可以构建出一个词典， key为上文出现过的词， value为这个词的索引序号

{"我":1,"喜欢":2,"看电影":3,"和":4,"打篮球":5,"玩游戏":6“踢足球”:7}
那么，上面两个句子用词袋模型表示成向量就是：

[1,1,1,1,1,0,0]

[1,1,0,1,0,1,1]

这样我们就完成了对于自然语言转化成向量。

2. TF－IDF模型

在问题一中对于TF－IDF特征值有了一定的了解，在这里就不过多赘述TF值和IDF值的含义。

词袋模型简单易懂，但是存在问题。中文文本里最常见的词是“的”、“是”、“有”这样的没有实际含义的词。一篇关于足球的中文文本，“的”出现的数量肯定多于“足球”。所以，要对文本中出现的词赋予权重。

回过头看词袋模型，只考虑了文本的词频，而TF－IDF模型则包含了词的权重，更加准确。文本向量与词袋模型中的维数相同，只是每个词的对应分量值换成了该词的TF－IDF值。

在这里本文使用词袋模型以及TF－IDF特征值算法模型来对群众留言的主题进行向量化，运用dow2box来构建词袋模型，但并不是简单的根据词出现的频率得出向量，通过TF－IDF来计算出每个词对应的TF－IDF值，最后得出每一条群众留言主题所对应的向量，然后根据每条留言的向量来找出相似的留言将其归类。

2.2.2.5 统计点赞数

在网络中，人们对于自己者关心的事件会表达出认同感。在很多情况下，点赞数是人们认同感的数字体现，比如在微博或者其他社交平台上，点赞数往往和事件热度本身有着直接的关系。所以我们在这里也将点赞数作为热度事件的一个重要量化指标。下面统计点赞数最高的 10 条留言主题：

表 6 点赞数前 10 的留言主题

留言主题	点赞数
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	2097
反映 A 市金毛湾配套入学的问题	1762
请书记关注 A 市 A4 区 58 车贷案	821
严惩 A 市 58 车贷特大集资诈骗案保护伞	790
承办 A 市 58 车贷案警官应跟进关注留言	733
A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到	669
A 市富绿物业丽发新城强行断业主家水	242

建议西地省尽快外迁京港澳高速城区段至远郊	80
请问 A 市为什么要把和包支付作为任务而不让市场正当竞争？	78
关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	78

从上表中可以看出群众对于车贷的放映非常强烈，在点赞数前 10 的留言中就有 3 条留言与车贷有关，同时小区物业与学生入学问题也是市民们关注的热点问题，点赞数同样居高不下。下面就文本相似度和点赞数两方面综合得出热度排名前 5 的事件。

2.2.2.6 热度指数

先前定义了热度的指标由相似留言的多少以及留言的点赞数两部分权重构成，相似留言的多少表明了群众对于这个问题反映的迫切性同时也间接证明了此问题需要解决的必要性，所以权重占 0.8，点赞数确实一定程度上反映了群众对于此问题发生的关注度，但由于点赞数不排除有的人只是好奇或者其他因素来点赞，所以点赞数有一定的偶然成分，故权重占 0.2。热度指数最高位 100，最低为 0。在统计好相似留言后取相似留言前 10 同点赞数前 10 加权得到最终热度排名前 5 的事件。

表 7 热度排名

问题 id	相似留言占比（0.8）	点赞数（0.2）	热度指数
1	0.313	0.102	90.26
2	0.273	0.113	85.33
3	0.174	0.204	71.33
4	0.16	0.192	66.13
5	0.08	0.389	60.26

2.2.3 问题 3

对于做好“网络问政”相关工作，留言回复质量是一项必不可少的评判过程。群众的留言与政府相关工作人员的高质量回复就好像使群众与政府之间建立了沟通的桥梁，使群众内心的问题得到了良好的反馈。这里，根据附件四包含的 2817 条留言的回复和回复时间，本文从留言回复的充实性，及时性，相关性，可解释性，完整性五个方面来评估留言的回复的质量。从而建立一套完整的留言回复评价系

统。

2.2.3.1 回复的充实性

工作人员的回复内容与回复的文本长度有直接的关系，简短内容的回复信息量一般不够，评分应该较低；同时，较长文本的回复评分也不应该过高。总的来说，文本长度是一项评判标准，但它并不是非常具有代表性的，所以我们考虑使用对数函数来量化回复的文本长度与评分的关系，建立“回复内容是否充实”的评价项F：

$$F = \frac{1}{N}\sum_{i=1}^{N} \log_m L_i \qquad (2\text{-}17)$$

其中，L_i为针对第i个问题回复的文本长度，m为常数。

下图为计算结果（全部结果详见数据附件）：

表 8 回复充实性的结果（部分）

回复充实性
8.703903573
8.113742166
8.344295908
8.076815597
7.118941073
7.721099189
7.7807354922
9.169925001
8.87036472
7.665335917

2.2.3.2 回复的及时性

在群众回复与工作人员留言的交流中，回复是否及时往往是很重要的。本文在评判这项标准时，分别以一个星期、两个星期、一个月、两个月、三个月、半年、一年等时间为间隔长度，去量化它们之间的及时性，然后根据每条留言回复的时间，量化后，评价其留言回复的及时性。

根据附件四，每条留言的回复对有其对应的留言时间和留言的回复时间，这里将留言的两个时间提取出来，而后建立一个留言回复时间的标准，比如留言 2549 的留言时间是 2019/4/259:32:09，回复时间是 2019/5/1014:56:53，通过对留言回

复时间的比较分析发现，留言的回复大多数不超过一个月，这里需要将留言回复的时间标准改为一个星期，两个星期，三个星期，四个星期，一个月以上，把它们作为留言回复的时间间隔长度，这里设置一个星期到两个星期内回复最快，三个到四个星期内回复，其回复速度一般，一个月以上为回复速度最慢，不及时。对每条留言的回复时间和留言时间进行相减，得到其留言回复的时间长度，然后将这些时间长度与标准时间间隔来相互比较，就可以得到每条留言的回复是否及时。

下图为计算结果（全部结果详见数据附件）：

表 9 回复及时性计算结果（部分）

回复及时性
0.9875
0.9875
0.9875
0.9875
0.986666667
0.974166667
0.965833333
0.975833333
0.986666667
0.986666667

2.2.3.3 回复的相关性

工作人员的回复内容与留言的相关性是一项非常重要的检测标准。我们要根据留言的内容和留言的回复内容来判断留言回复的相似性，我们选择余弦相似度来评估回复与留言的相似性。

2.2.3.3.1 余弦相似度的优势

余弦相似度衡量的是维度间取值方向的一致性，注重维度之间的差异。这里用几种常见的计算相似度算法进行比较：

1.欧几里得距离:

欧几里得度量（Euclindean metric）（也称欧氏距离）是一个通常采用的距离定义，指在 m 维空间中两个点之间的真实距离，或者向量的自然长度（即该点到原

点的距离）。在二维和三维空间中的欧氏距离就是两点之间的实际距离。显然，它更注重数值间的差异，而我们在评估文本相关性时，更应该注重的是回复的内容是否与留言内容关键词匹配，而不是关键词之间存在的数值差异性。因为一条留言与一条回复它所包含的不仅仅只是一些具有代表性的关键词，它们还由许多其他的词组成，所以如果更注重关键词的数值差异则会对评估文本的相关性有较大的误差。

2. 皮尔逊相关系数：

Pearson相关系数是用协方差除以两个变量的标准差得到的，虽然协方差能反映两个随机变量的相关程度（协方差大于 0 的时候表示两者正相关，小于 0 的时候表示两者负相关），但其数值上受量纲的影响很大，不能简单地从协方差的数值大小给出变量相关程度的判断。为了消除这种量纲的影响，于是就有了相关系数的概念。

当两个变量的方差都不为零时，相关系数才有意义，相关系数的取值范围为[-1,1]。《数据挖掘导论》中给了一个很形象的图来说明相关度大小与相关系数之间的联系：

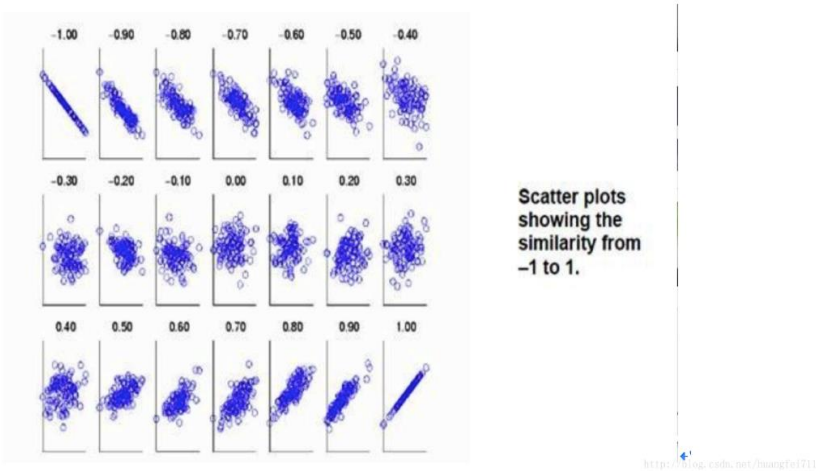

图 9 相关度大小与相关系数的关系

由上图可以总结，当相关系数为 1 时，成为完全正相关；当相关系数为-1 时，成为完全负相关；相关系数的绝对值越大，相关性越强；相关系数越接近于 0，相关度越弱。

且皮尔逊相关系数的约束条件：

(1)两个变量间有线性关系

(2)变量是连续变量

(3)变量均符合正态分布,且二元分布也符合正态分布

(4)两变量独立

因此它们不太适用于比较文本分词进行关键词匹配这类相关的算法，而余弦相似性度可以更好地避免以上出现的问题。

2.2.3.3.2 余弦相似度算法的思路与过程

余弦距离，也称为余弦相似度，是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小的度量。

余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，这就叫"余弦相似性"。

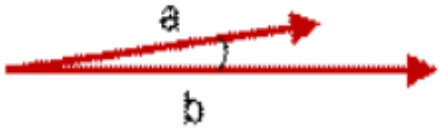

图 10 向量 a,b 的夹角

上图两个向量a,b的夹角很小可以说a向量和b向量有很高的的相似性，极端情况下，a和b向量完全重合。如下图：

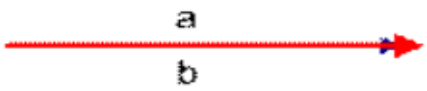

图 11 向量 a,b 的夹角

如上图：可以认为a和b向量是相等的，也即a，b向量代表的文本是完全相似的，或者说是相等的。如果a和b向量夹角较大，或者反方向。如下图

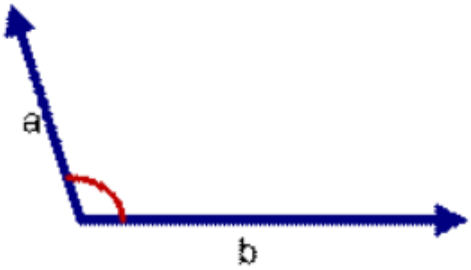

图 12 向量 a, b 的夹角

如上图:两个向量a, b的夹角很大可以说a向量和b向量有很低的的相似性，或者说a和b向量代表的文本基本不相似。那么是否可以用两个向量的夹角大小的函数值来计算个体的相似度呢？

向量空间余弦相似度理论就是基于上述来计算个体相似度的一种方法。下面做详细的推理过程分析。

想到余弦公式，最基本计算方法就是初中的最简单的计算公式，计算夹角的余弦定值公式为：

$$\cos\theta = \frac{a}{c} \qquad (2\text{-}18)$$

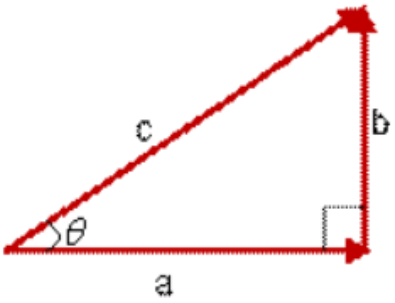

图 13 直角三角形

但是这个是只适用于直角三角形的，而在非直角三角形中，余弦定理的公式是：

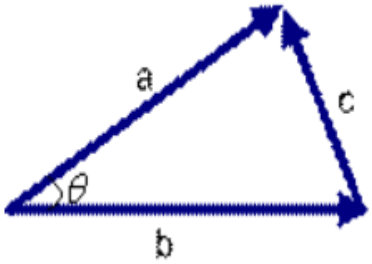

图 14 三角形

三角形中边a和b的夹角的余弦计算公式为：

$$\cos(\theta) = \frac{a^2+b^2+c^2}{2ab} \tag{2-19}$$

在向量表示的三角形中，假设a向量是 (x_1, y_1)，b向量是(x_2, y_2)，那么可以将余弦定理改写成下面的形式：

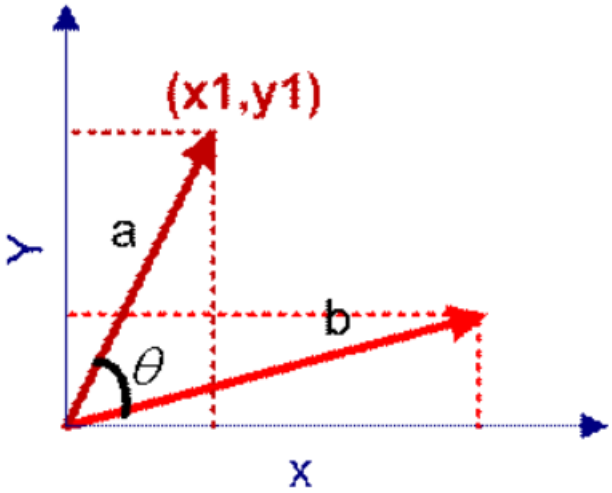

图 15 建立的直角坐标系

向量a和向量b的夹角的余弦计算如下：

$$\cos(\theta) = \frac{a \cdot b}{|a| \cdot |b|}$$

$$= \frac{(x_1,y_1)\cdot(x_2,y_2)}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}} = \frac{x_1x_2+y_1y_2}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}} \tag{2-20}$$

扩展，如果向量a和b不是二维而是n维，上述余弦的计算法仍然正确。则a与b

的夹角的余弦等于：

$$\cos(\theta) = \frac{a \cdot b}{|a| \cdot |b|} = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} {x_i}^2} \times \sqrt{\sum_{i=1}^{n} {y_i}^2}} \quad (2\text{-}21)$$

余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，夹角等于 0，即两个向量相等，这就叫"余弦相似性"。

另外：余弦距离使用两个向量夹角的余弦值作为衡量两个个体间差异的大小。相比欧氏距离，余弦距离更加注重两个向量在方向上的差异。

【下面举一个例子，来说明余弦计算文本相似度】

举一个例子来说明，用上述理论计算文本的相似性。为了简单起见，先从句子着手。

句子A：这只皮靴号码大了。那只号码合适

句子B：这只皮靴号码不小，那只更合适

怎样计算上面两句话的相似程度？

基本思路是：如果这两句话的用词越相似，它们的内容就应该越相似。因此，可以从词频入手，计算它们的相似程度。

第一步，分词。

句子A：这只/皮靴/号码/大了。那只/号码/合适。

句子B：这只/皮靴/号码/不/小，那只/更/合适。

第二步，列出所有的词。

这只，皮靴，号码，大了。那只，合适，不，小，很

第三步，计算词频。

句子A：这只 1，皮靴 1，号码 2，大了 1。那只 1，合适 1，不 0，小 0，更 0

句子B：这只 1，皮靴 1，号码 1，大了 0。那只 1，合适 1，不 1，小 1，更 1

第四步，写出词频向量。

句子A：(1，1，2，1，1，1，0，0，0)

句子B：(1，1，1，0，1，1，1，1，1)

到这里，问题就变成了如何计算这两个向量的相似程度。可以把它们想象成空间中的两条线段，都是从原点（[0,0,...]）出发，指向不同的方向。两条线段之间形成一个夹角，如果夹角为 0 度，意味着方向相同、线段重合,这是表示两个向量代表的文本完全相等；如果夹角为 90 度，意味着形成直角，方向完全不相似；如果夹角为 180 度，意味着方向正好相反。因此，我们可以通过夹角的大小，来判断向量的相似程度。夹角越小，就代表越相似。

使用以下公式：

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} {x_i}^2} \times \sqrt{\sum_{i=1}^{n} {y_i}^2}} \quad (2\text{-}22)$$

计算两个句子向量

句子A：(1，1，2，1，1，1，0，0，0)

句子B：(1，1，1，0，1，1，1，1，1)的向量余弦值来确定两个句子的相似度。

计算过程如下：

$$\cos(\theta) = \frac{1\times1+1\times1+2\times1+1\times0+1\times1+1\times1+0\times1+0\times1+0\times1}{\sqrt{1^2+1^2+2^2+1^2+1^2+1^2+0^2+0^2+0^2} \times \sqrt{1^2+1^2+1^2+0^2+1^2+1^2+1^2+1^2+1^2}}$$

$$= \frac{6}{\sqrt{7} \times \sqrt{8}}$$

$$= 0.81$$

图 16 余弦相似度计算结果

计算结果中夹角的余弦值为 0.81 非常接近于 1，所以，上面的句子A和句子B是基本相似的。

由此，就得到了文本相似度计算的处理流程是:

(1)找出两篇文章的关键词。

(2)每篇文章各取出若干个关键词，合并成一个集合，计算每篇文章对于这个集合中的词的词频。

(3)生成两篇文章各自的词频向。

(4)计算两个向量的余弦相似度，值越大就表示越相似。

因为留言和回复的关键词不是像以上这样单一，且在留言和回复时往往会出现很多与问题相关性不大的其他词组成这个句子。所以不可以简单地如以上实例进行对比。进行以下几个步骤：

本文对留言的文本进行数据预处理，包括停用词过滤，分词等一些操作，提取到留言中问题的关键词。

本文将在回复中去检测这些关键词出现的次数。

将步骤 1、2 进行余弦相似度处理

【以提供的文本数据举例】

这是文本数据中的某留言（全部结果详见数据附件）：

2019 年 4 月以来，位于A市A2 区桂花坪街道的A2 区公安分局宿舍区（景蓉华苑)出现了一番乱象，该小区的物业公司美顺物业扬言要退出小区，因为小区水电

改造造成物业公司的高昂水电费收取不了(原水电在小区买，水 4.23 一吨，电 0.64 一度）所以要通过征收小区停车费增加收入，小区业委会不知处于何种理由对该物业公司一再挽留，而对业主提出的新应聘的物业公司却以交 20 万保证金，不能提高收费的苛刻条件拒之门外，业委会在未召开全体业主大会的情况下，制定了一高昂收费方案要各业主投票，而投票不采用投票箱只制定表格要物业公司人员这一利害关系机构负责组织，对投票业主隐私权没有任何保护，还对投反对票的业主以领导做工作等方式要求改变为同意票，这种投票何来公平公正公开，面对公安干警采用这种方式投票合法性在哪？

下面是有关关键词的提取：

2019 位于 A市 A2 桂花 坪 街道A2 公安分局 宿舍区 景蓉华苑 乱象 物业公司 美顺 物业 扬言 退出 水电 改造 物业公司 高昂 水电费 收取 原 水电 买水 423 一吨 电 064 一度 征收 停车费 增加收入 业委会 不知 处于 何种 理由 物业公司 一再 挽留 业主 提出 新 应聘 物业公司 以交 20 万 保证金 提高 收费 苛刻 条件 拒之门外 业委会 未 业主大会 情况 制定 高昂 收费 方案 业主 投票 投票 采用 投票箱 制定 表格 物业公司 人员 这一 利害关系 机构 负责 组织 投票 业主 隐私权 保护 投 反对票 业主 领导 做 工作 方式 改变 同意 票 投票 何来 公平 公正 公开 面对 公安干警 采用 方式 投票 合法性

很显然，这个留言集中于“景荣华苑”、“物业”、“水电”、“停车费”、“业委会”“高额收费”、“投票”、“公平性”等关键词。我们将在回复中去搜索这些关键词。

以下是这段留言的相关回复：

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”的调查核实情况向该网友答复如下：您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“A2 区景蓉花苑物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”，景蓉华苑业委会于 2019 年 4 月 10 日至 4 月 27 日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真疏理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5 月 5 日下午,辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。再次感谢您对我区工作的理解和关心。

本文可以看到文本中对上文提到的那几种关键词均有反馈，接下来，本文利用

余弦相似度算法进行评估。

在得到相似度后进行一定的量化，则可以作为文本相关性的一个评判标准。

算法流程如下：

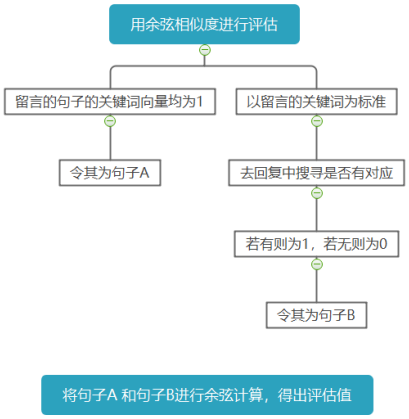

图 17 余弦相似度算法

下图为计算结果（全部结果详见数据附件）：

表 10 回复相关性计算结果（部分）

回复相关性
0.5
0.105263158
0.363636364
0.285714286
0.473049917
0.2
0.2
0.2
0.105263158
0.363636364

由于在回复的文本中不可避免地会出现一些和留言无关但是不可缺少的词语，如一些连接词等，所以我们在此基础上将每一条留言的相关度都乘以 2，以得到一个更好的标准。

2.2.3.4 回复的可解释性

在留言的回复中，观察到一条高质量的留言往往有以下特征：

(1)具有很强的条理性

(2)对于留言中的相关问题，都给予了相应的文件或政策进行回复。因此增强了留言的可解释性。

将在回复中搜索类似于“根据《xxx》文件规定”“按照xxx政策”“第一，xxx；第二，xxx；第三，xxx”等句段，用来评估此项标准。这里可以对留言回复文本中此类词的提取，然后对留言回复中此类词进行统计，定制一个回复可解释性的标准。

下图为计算结果（全部结果详见数据附件）：

表 11 回复可解释性计算结果

回复可解释性
0.870390357
0.811374217
0.834429591
0.80768156
0.711894107
0.772109919
0.780735492
0.9169925
0.887036472
0.766533592

2.2.3.5 回复的完整性

一条留言的回复是否完整于其对于留言中的相关问题是否回复/解决完毕具有联系。因此，我们将根据留言的始、末,进行分析。根据观察，发现一条高质量的留言，在留言的回复中，如果担心民众提出的问题没有完全回答，留言的末尾会给出进一步解决的途径，如“如还有疑问，请到当地相关部门咨询”“如还有疑问，请致电xxx /请继续咨询xxx”并给出相应的电话号码，或者投诉电话。因此，本文将会在留言的回复中搜索类似句段进行有效评估。可以对留言回复中是否有此类词语的出现，对留言回复的完整性做评价。

下图为计算结果（全部结果详见数据附件）：

表 12 回复完整性计算结果（部分）

回复完整性
0.65792768
0.60852219
0.625822193
0.60576117
0.53392058
0.579082439
0.585551619
0.687744375
0.665277354
0.574900194

2.2.3.6 属性值的归一化处理

考虑到本文在评估留言的充实性、及时性、相关性、完整性、可解释性时，数据的属性值并不是统一的---有的属性值越大越好，如充实性、相关性等，这称为效益型；有的属性值越小越好，如及时性，这称为成本型。所以本文在数据处理时要将这些数据的属性进行统一。

在这里，为了方便计算，将所有指标属性统一为效益性指标（即极大型指标）。

$$\frac{\max(ai)-ai}{\max(ai)-\min(ai)} \qquad (2\text{-}23)$$

2.2.3.7 层次分析法

1.建立回复评论的层次结构模型

将决策问题分解成三个层次，最上层为目标层D，即评价答复意见质量；最下层为对象层O，即针对这个对象进行评估；中间层为准则层C，包括充实性C_1，及时性C_2,相关性C_3，完整性C_4，可解释性C_5（如下图所示），确定对象层对每个准则的权重，然后二者综合得到对象对目标的权重。

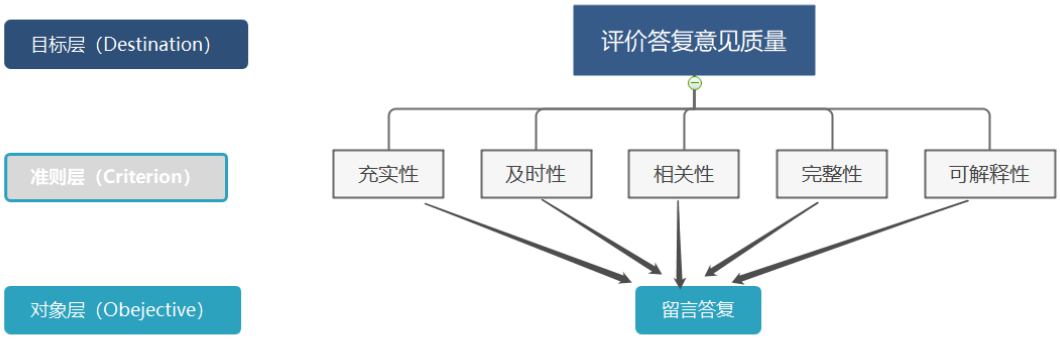

图 18 层次分析图

2. 建立成对比较矩阵

构造判断模型$D-C$，将准则层中五个元素充实性C_1，及时性C_2，相关性C_3，完整性C_4，可解释性C_5两两比较，得成对比较矩阵：

$$D=\begin{bmatrix}1 & 1/2 & 1/4 & 1/2 & 1/3\\ 2 & 1 & 1/2 & 1 & 2/3\\ 4 & 2 & 1 & 2 & 4\\ 2 & 1 & 1/2 & 1 & 2/3\\ 3 & 3/2 & 1/4 & 3/2 & 1\end{bmatrix} \quad (2\text{-}24)$$

3. 成对比较矩阵和权向量的一致性比较

一致性矩阵要具有以下的代数性质：矩阵D的秩为 1，唯一非零特征根为n，任一列向量都是对应于特征根n的特征向量，仅相差一个常数因子，D的归一化特征向量可作为权向量。对于成对比较矩阵需要进行检验，它存在一定的标准，就是取决于矩阵的一致性，给定的一致性矩阵的定义，条件严格，需要放弱条件，不追求强制一致性，相似就行。那么观察矩阵分析知道矩阵D并不完全满足一致性矩阵的要求，$a_{ij}\cdot a_{jk}=a_{ik},i,j,k=1,2,3\cdots n$, 对于不一致但在准许的一定范围内的成对比较矩阵D，建议用对应的最大特征根和最大特征根对应的特征向量来比较，这里我们可以通过MATLAB求出其最大特征根和其对应的特征向量。

4. 一致性指标和一致性检验

上一步我们求出了D对应的最大特征根d和其特征向量作为的权向量w, 对于n阶的成对比较矩阵，其最大特征根$c\geq n$，并且一致性矩阵的的充要条件是$c=n$。由此可以用c和n的数值大小来衡量D的一致性。由公式

$$CI=\frac{c-n}{n-1} \quad (2\text{-}25)$$

可知当$CI=0$时D是一致性矩阵，CI越大，D越不一致。

表 13 随机一致性指标 RI 的数值

n	3	4	5	6	7	8	9	10
RI	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

定义一致性比率$CR=\frac{CI}{RI}$当 $CR<0.1$ 时通过一致性检验。由公式$CI=\frac{c-n}{n-1}$，于是根据$CR=\frac{CI}{RI}$，用MATLAB计算得到CR=0.0337<0.1，通过一致性检验。

5.计算权重

分别运用①算术平均法②几何平均法③特征值法计算权重。

以往的论文利用层次分析法解决问题时，都是采用其中某一种方法求权重，而不同的计算方法可能会导致结果有所偏差。为了保证结果的稳健性，本文采用了三种方法分别求出了权重，再根据得到的权重矩阵计算各方案的得分，并进行排序和综合分析，这样避免了采用单一方法所产生的偏差，得出的结论更全面更有效。

下面，运用MATLAB来计算权重。

```
%%%%%%%%%%%%%方法1：  算术平均法求权重%%%%%%%%%%%%%
Sum_A = sum(A);
SUM_A = repmat(Sum_A,n,1);
Stand_A = A ./ SUM_A;
disp('算术平均法求权重的结果为：');
disp(sum(Stand_A,2)./n)
%%%%%%%%%%%%%方法2：  几何平均法求权重%%%%%%%%%%%%%
Prduct_A = prod(A,2);
Prduct_n_A = Prduct_A .^ (1/n);
disp('几何平均法求权重的结果为：');
disp(Prduct_n_A ./ sum(Prduct_n_A))
%%%%%%%%%%%%%方法3：  特征值法求权重%%%%%%%%%%%%%
[V,D] = eig(A);
Max_eig = max(max(D));
[r,c]=find(D == Max_eig , 1);
disp('特征值法求权重的结果为：');
disp( V(:,c) ./ sum(V(:,c)) )
```

图 19 权重代码

得出算术平均法求权重的结果为：

0.0800

0.1600

0.4000

0.1600

0.2000

几何平均法求权重的结果为：

0.0807

0.1614

0.4021

0.1614

0.1944

特征值法求权重的结果为：

0.0783

0.1566

0.4139

0.1566

0.1947

根据观察，三种方法计算出的权重相差很小，在这里我们可以选择特征值法作最后的评估。即，充实性的权重为0.0783；及时性的权重为0.1566；相关性的权重为0.4139；完整性的权重为0.1566；可解释性的权重为0.1947。分别用这五个方面的权重乘以相对应的得分就可以评估出这条答复的最终质量。

2.2.3.8 基于理想点逼近法（TOPSIS）评估评价等级

TOPSIS是一种用于多目标决策的方法，通过检测评价对象与最优解、最差解的距离来排序。若评价对象最靠近最优解同时又最远离最差解为最好；否则为最差。考虑TOPSIS法具有原理简单并对实验样本要求不高，并且可生成一个明确的评价等级，因此选TOPSIS来做为评价回复的等级方法。

1.构建初始评价指标矩阵

在对2816条留言的5个评判指标进行统计分析后，我们给出了留言的评价矩阵Z：

$$Z = \begin{pmatrix} Z_{1.1} & \cdots & Z_{1.5} \\ \vdots & \ddots & \vdots \\ Z_{2816.1} & \cdots & Z_{2816.5} \end{pmatrix} \tag{2-26}$$

上文中已对各指标统一化。

2.构建加权评价指标矩阵

在上一节中，本文使用了层次分析法对各个指标的权重进行了计算，最终得到了综合权重向量$W = [W^1 \quad W^2 \cdots W_5]$，并根据该向量得到了综合权重矩阵W：

$$W = \begin{pmatrix} W_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & W_5 \end{pmatrix} \tag{2-27}$$

在得到五个指标的特征权重矩阵后，即可计算得到加权评价指标矩阵F：

$$F = Z \times W \tag{2-28}$$

$$F = \begin{pmatrix} W_1 Z_{11} & \cdots & W_1 Z_{15} \\ \vdots & \ddots & \vdots \\ W_5 Z_{2816.1} & \cdots & W_5 Z_{2816.5} \end{pmatrix} \tag{2-29}$$

其中，$f_{i,j} = W_j \cdot Z_{i,j}, i = 1,2 \cdots 120, j = 1,2 \cdots 5$

下图为部分计算结果（全部结果请详见数据附件）：

表 14 矩阵计算结果（部分）

相关性	充实性	及时性	可解释性	完整性
0.4139	0.054521	0.154643	0.169465	0.102227
0.087137	0.050824	0.154643	0.157975	0.095296
0.301018	0.050824	0.154643	0.162463	0.098004
0.236514	0.050593	0.154643	0.157256	0.094862
0.391591	0.044593	0.154212	0.138606	0.083612
0.16556	0.048365	0.152555	0.15033	0.090684
0.16556	0.048905	0.151125	0.152009	0.091697
0.16556	0.05744	0.152816	0.178538	0.107701
0.087137	0.055564	0.154512	0.172706	0.104182
0.301018	0.048016	0.154512	0.149244	0.090029

同时计算得出：2816 条数据的均值为：0.650773；方差为 0.02844。符合预期情况。

3.确定正负理想解

对于加权评价指标矩阵F，找到第j列的最优解为F_j*，最劣解为$F_j\hat{}$，最终得到正理想解和负理想解。在加权评价指标矩阵中，元素越大，说明留言的回复质量越高；元素越小，说明留言的回复质量越低。因此，最大值为最优解，最小值为最劣解。

$$F_j = \sum_{i=1}^{5} F_{i,j} \qquad (2\text{-}30)$$

下图为部分计算结果（全部结果请详见数据附件）：

表 15 五种指标总和（部分）

相加后得
0.894756
0.545874
0.768397
0.693868
0.812914
0.607494
0.609421
0.662055
0.574101
0.742819

最优为 1.265194，最劣为 0.172225。

4.根据正负理想解确定评价等级

根据观察需要设置一定的阈值对数据进行分类，分别分为不合格回复，合格回复，良好回复和优秀回复。

下图为最终数据指标分数的统计图形：

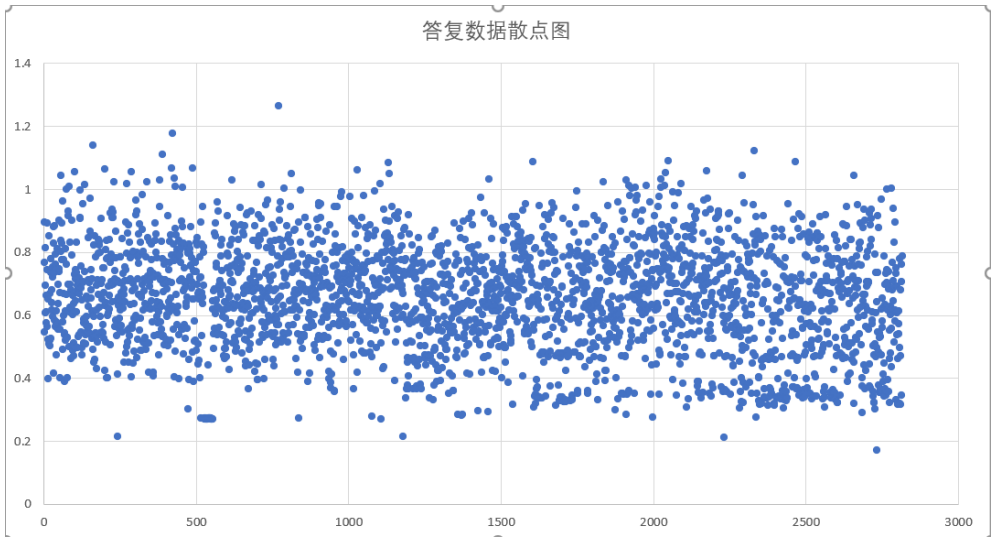

图 20 答复数据散点图

最终得出数据值处于 0.9 及 0.9 以上为优秀答复，数据值处于 0.9-0.7 为良好答复，数据值处于 0.7-0.4 为合格答复，数据值处于 0.4 以下为不合格答复。

下图为数据等级的数量扇形图：

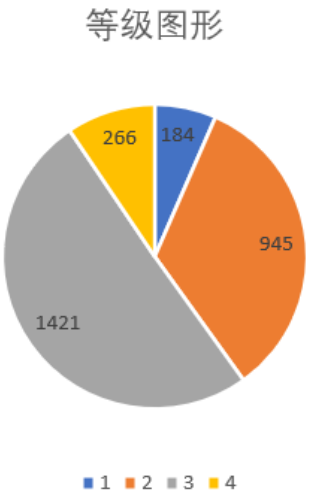

图 21 等级对应数量扇形图

这样就根据一系列的留言回复方法得到了一套相对完整的留言评估方法，对于附件四中给出的两千多条留言回复进行了相关的分类。其中优秀答复为 184 条，良好答复为 945 条，合格答复为 1421 条，不合格答复为 266 条。

2.2.3.9 结果分析

根据结果可以看出，一条高质量的答复往往具有特征：

(1)对于留言所提出的疑问大多给予了充分详细的回答。

(2)答复具有很强的条理性且有理有据，让留言的市民一目了然.

(3)答复距离市民留言的时间往往很近，对于留言中的问题有及时的反馈等。

而一条低质量的留言往往对于所提问的问题并没有给予回答、无条理性、时间间隔长，造成了市民心中的疑问并未很好地解决。

下图分别是抽取了答复中不合格、合格、良好、优秀的样本：

你好。请向当地民政部门询问。2017年8月18日

图 22 不合格样本

网友“UU00877”您好！您的留言已收悉。现将有关情况回复如下：经查，A4区政府已责成区市政局牵头，区城乡建设局、区规划分局配合进行具体选址，招标（邀标）进行方案设计等，尽快启动万国城小区人行天桥建设并投入使用，方便出行，感谢您对我们工作的支持、理解与监督！2019年1月8日

图 23 合格样本

网友“UU0081878”您好！您的留言已收悉。现将有关情况回复如下：全装修政策文件中全装修内容为套内装修部分，不含电梯厅、大厅、走道等公共部分（该公共部分装修已计入毛坯价格），且应计入毛坯部分的普通入户门等不得计入全装修价格，我委在《A市住房和城乡建设委员会关于新建商品住宅全装修价格核算规定的通知》（[政府发文]〔2018〕53号）中已进行了明确规定。A市目前全装修政策实施时，全装修总价格是由所有装修实物累加价格和税费、利润等构成，该全装修总价中并未将公共装修部分内容（如电梯间瓷砖、墙漆等）计入。在实际操作中，预售时官方测量了预售许可面积（参照建筑面积），作为毛坯销售的基准。为与毛坯同步，全装修《实施细则》和《价格核算规定》中也将全装修总价按预售许可面积进行折算。该折算不会影响全装修总价是实物累加而来的逻辑。全装修《实施细则》中“限房价、竞地价”等刚性需求项目，全装修价格不得超过2500元/平方米，是指“全装修总价/预售许可面积”不得超过2500元/平方米，金晖优步花园项目在销售现场都有全装修公示，业主应知晓是全装房销售，后期企业针对市场需求对自身产品进行更改，属于市场行为。如您对第三方造价咨询机构全装修核算价格存在异议，须提供第三方造价咨询机构未客观公正核算价格相关证明，向市住建部门工程造价管理机构申请复核，感谢您对我们工作的支持、理解与监督！2018年12月17日

图 24 良好样本

UU0网友“UU008835”　　您好！来信收悉。现回复如下：　　经镇综治办调查核实，信访人黄尊富，系我镇乌川湖村楼梯坡组村民。2010年—2012年期间，浏醴高速公路修建途经我镇乌川湖村，并在该村楼梯坡组设计有一涵洞，该涵洞能安全通过小轿车、农业用车等车型。黄尊富的房屋在楼梯坡涵洞边，距离高速主线有200多米远，不符合拆迁条件。自涵洞开工建设以来，黄尊富以及其父黄刚明以楼梯坡涵洞设计位置较低，涵洞出口接线坡度较疑，造成其本人和家人出行不便为由，多次向江背镇浏醴高速公路指挥部反映，要求解决该问题，并要求进行拆迁。　　江背镇党委、政府高度重视此事，从2011年至今，多次协调浏醴高速公路指挥部采取措施保障黄尊富一家的安全出行，并给以一定经济补偿，具体情况如下：　　1、在浏醴高速建设期间，积极解决黄尊富一家的出行问题并给以补偿，由浏醴高速八标段项目部承担原材料、人工等所有开支，由黄尊富之父黄刚明按照其家人出行方便的需求，现场指挥施工，对涵洞出口到其家门口的坡道泥路进行水泥硬化，有效减小了涵洞建设对黄尊富一家生产生活的影响，已于2012年10月完工。同时该户户主黄刚明与指挥部签订了因涵洞设计较低导致出行不便的一次性补偿协议，协议明确要求黄刚明一家领取补偿款后，不得再就涵洞出口坡度较陡造成出行不便的问题提出其它要求。2012年1月20日黄刚明领取补偿款21200元，　　2、黄尊富否认第一次协议效力，并以自己名义签订合约再次拿到补偿，黄尊富认为2012年1月20日签订协议时他不在家，是户主黄刚明签订的，否认第一次协议的效力。2012年12月4日，指挥部再次组织乌川湖村、楼梯坡组与黄尊富协商，并与其签订了2万元的一次性经济补偿协议，黄尊富承诺领取资金后，不再就此事向镇政府提出其他任何的补偿要求。　　3、江背镇政府积极召开信访答复会。在该次补偿后，黄尊富仍不满意，每隔一个月到江背镇政府上访，并多次向伺政西地省、县长信箱、“12345”等平台提出拆迁的无理诉求，2014年5月22日，江背镇党委、政府组织镇派出所、国土所、经贸办、综治办、浏醴高速公路拆迁指挥部等部门负责人和黄尊富本人召开了有关信访答复会，明确答复黄尊富，因其房屋不在浏醴高速拆迁红线范围内，要求拆迁房屋的诉求没有相关的政策依据，也没有现实操作拆迁的可能，由于高速公路建设造成的客观影响已得到妥善解决，如果仍然不能认同，必须合理合法的到相关部门反映情况。　　4、2015年12月17日，江背镇综治、信访、派出所、原浏醴高速指挥部成员代表、黄尊富参加了最近一次的信访答复会。会上，我镇对黄尊富进行了政策宣讲与解释，黄尊富不理解、不满意，并通过打电话、发短信等方式多次骚扰我镇工作人员的正常办公与休息，我镇明确告知信访人：一是由于浏醴高速建设对其家庭出行造成的不便，指挥部已两次给予其家庭补助共计4万元，信访人也在协议上签字认可；二是根据信访人家庭实际情况，不符合拆迁的标准，不在拆迁范围之内；三是征地拆迁的主体是浏醴高速建设指挥部，不是江背镇人民政府；四是信访人反映诉求必须依法依规进行，不得以发短信威胁或以打电话骚扰等非正常方式干扰、影响国家机关或国家工作人员的正常工作、生活，在本次答复会议对其训诫的基础上，如若继续违反，将承担相应的法律责任。2016年1月7日

图 25 优秀样本

3. 结论

本文主要由以下几点来构建一套完整的智慧政务系统：

首先，对文本进行预处理后，将文本标签转化为id和文本特征值提取，将多条留言信息分别对应 7 个一级标签转化为 0 到 6，对每个标签留言数量进行统计，然后使用到TF − IDF算法计算TF − IDF值，卡方检验得到一级标签中意思相近的词语对，经过对比与选择后使用朴素贝叶斯分类器进行基于词频的高维数据分析，最后，运用F − score来进一步的评估模型。

其次，再进行热度指标的定义，数据预处理，统计词频并绘制词云图，可以得出热点事件主要发生在A市、A7、A3、A2、A1 区，关于扰民、噪音、房屋、施工等问题的热点事件较多，然后使用词袋模型和TF − IDF模型计算文本的相似度，找出文本中相似的事件。

最后，在评估答复意见时，本文首先从 5 个指标（充实性、及时性、相关性、可解释性、完整性）综合评价答复质量，并进行了属性值的归一化处理，后使用层次分析法为 5 个指标赋权值，构建了 2816 条留言答复的评价矩阵，并使用TOPSIS算法，对答复进行了标记，同时对答复的评分进行分析，判断答复的质量，给出优秀、良好、合格、不合格的等级，最终构建出一套评价方案。

由此建立一套完整的智慧政务系统，更加便捷的，高效的解决问题。

参考文献

[1] 桑亮.基于 BP 神经网络的装备失败效率预测研究[J].四川兵工报,2014(02).
[2] 张旭.一个基于词典和统计的中文分词算法[D].成都:电子科技大学,2006.
[3] 杨涛.中文信息处理中的自动分词方法研究[J].现代交际,2019(07):93-95.
[4] 任雅仙.中国重大突发事件中的网络舆论研究[D].四川:四川大学,2010.
[5] 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.
[6] 杨开平,李明奇,覃思义.基于网络回复的律师评价方法[J].计算机科学,2018,45(09):237-242.
[7] 张继勋,韩冬梅.网络互动平台沟通中管理层回复的及时性、明确性与投资者投资决策——一项实验证据[J].管理评论,2015,27(10):70-83.
[8] 王玥,郑磊.中国政务微信研究:特性、内容与互动[J].电子政务,2014(01):66-77.
[9] 李传军,李怀阳.公民网络问政与政府回应机制的建构[J].电子政务,2017(01):69-76.
[10] 费尔德曼.文本挖掘(英文版)[M].人民邮电出版社,2019(08).
[11] 任崇广.面向海量数据处理领域的云计算及其关键技术研究[D].南京理工大学,2013.
[12] 马轶婷,高洁.电子政务信息服务质量评价模型研究综述[J].电子政务，2014 年 12 期.
[13] 牛力.政务信息资源“云服务”整合模式研究[J].情报杂志,2013 年 01 期.
[14] 胡吉明.智慧政务研究与应用进展[J].智慧城市评论,2017 年 02 期.
[15] 李其玲.基于双循环组织学习理论的政务社交媒体文本挖掘研究[D]:[硕士学位论文].华中科技大学,2018 年.
[16] 常永华，李春玲.电子政务信息服务模型研究——基于演化博弈的分析[J].情报理论与实践,2011 年 09 期.