基于文本挖掘的群众问政处理

摘要：

近年来，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

针对问题一，基于 TF-IDF 算法方法对留言信息提取关键词，利用监督学习算法中的朴素贝叶斯分类建立了关于留言内容的一级标签分类模型，使用 F-Score 对分类模型进行评价，得到每一类的 F 均大于 0.9，整体的 F 的是 0.9026，说明分类模型的质量很好。

针对问题二，基于分词处理、Word2Vec 模型与 Doc2Vec 模型训练得到句子向量和文本向量，利用 K-means 聚类方法对留言归类，得出排名前 5 的热点问题分别为土地规划、追缴社保、房屋拆迁、捆绑车位销售、设立公交站，热点问题表、以及对应的留言信息表见附件；构建留言条数、点赞数和反对数三个热度评价指标，计算影响力得分，得分高低依次为房屋拆迁、追缴社保、设立公交站、捆绑车位销售。

针对问题三，构建相关性、完整性、可解释性、时效性、客观性、友好性这六个指标，利用熵权法求得指标权重，量化评价进行数值挖掘进而达到对答复意见评价的分析；根据词袋模型、余弦相似度、已发布的措施文件、相关度、客观分词、文本情感分析方面解释上述指标，最后对答复意见的质量给出一套评价方案并进行实现，答复意见的质量得分与定性评价见附件答复质量评价表。

关键词：热点问题挖掘；朴素朴素贝叶斯；word2vec；Doc2Vec；K-means；文本情感分析

Abstract

In recent years, with the development of big data, cloud computing, artificial inte lligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoti ng the management level and efficiency of the government.

In view of question 1 , a method based on TF - IDF algorithm for extracting me ssage information keywords, supervised learning algorithm of naive bayesian classi fication level 1 label on the message content classification model is established, usi ng the F - Score to evaluate classification model, each kind of F were greater than 0 .9, the integral of F is 0.9026, the quality of the classification model is very good.

In view of question 2, based on word processing, Word2Vec model with Doc2Ve c vector and text vector model training sentence, using the K means clustering meth od of message classification, it is concluded that the top 5 issues of land planning re spectively, recovered, social security, housing demolition, sets up the bus station, b ound to car sales, hot issue list, as well as the corresponding message information s ee attachment table;Three heat evaluation indexes of message number, thumb up nu mber and objection number were constructed, and influence score was calculated. T he scores were in order of house demolition, social security recovery, bus station es tablishment, and parking space sales.

In view of question 3, six indexes of relevance, completeness, interpretability, ti meliness, objectivity and friendliness are constructed. The entropy weight method i s used to obtain the index weight, and the quantitative evaluation is carried out to ca rry out numerical mining, so as to achieve the analysis of the evaluation of response opinions.The above indicators are explained according to the word bag model, cosi ne similarity, published measure documents, relevance, objective word segmentatio n, and text emotion analysis. Finally, a set of evaluation scheme is given and imple mented for the quality of reply comments. The quality score and qualitative evaluat ion of reply comments are shown in the attached reply quality evaluation form.

Keywords:hot issues mining;Naive bayes;Word2vec;Doc2Vec;K - means;Textual affe ctive analysis

目录

1、背景

近年来，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

智慧政府充分利用物联网、云计算、大数据分析、移动互联网等新一代信息技术，以用户创新、大众创新、开放创新、共同创新为特征，强调作为平台的政府架构，并以此为基础实现政府、市场、社会多方协同的公共价值塑造，实现政府管理与公共服务的精细化、智能化、社会化。实现政府和公民的双向互动

政府优化自身服务功能的有效途径之一是推进智慧政务建设工作进程，这是促进当前经济社会长期、稳定、可持续发展的重要举措。而微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。

2、挖掘目标及流程

1、建立关于留言内容的一级标签分类模型，对群众留言进行分类，以便在处理网络问政平台的群众留言时，解决人工依据检验处理问题的低效率方式。

2、对热点问题进行挖掘，将某一时段内反映特定地点或特定人群问题的留言进行归类，定义热度评价指标，并给出评价结果，给出排名前 5 的热点问题，以及相应热点问题对应的留言信息。

3、答复意见的评价，针对相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案并实现。

本文的整体思路如下：如图 1：

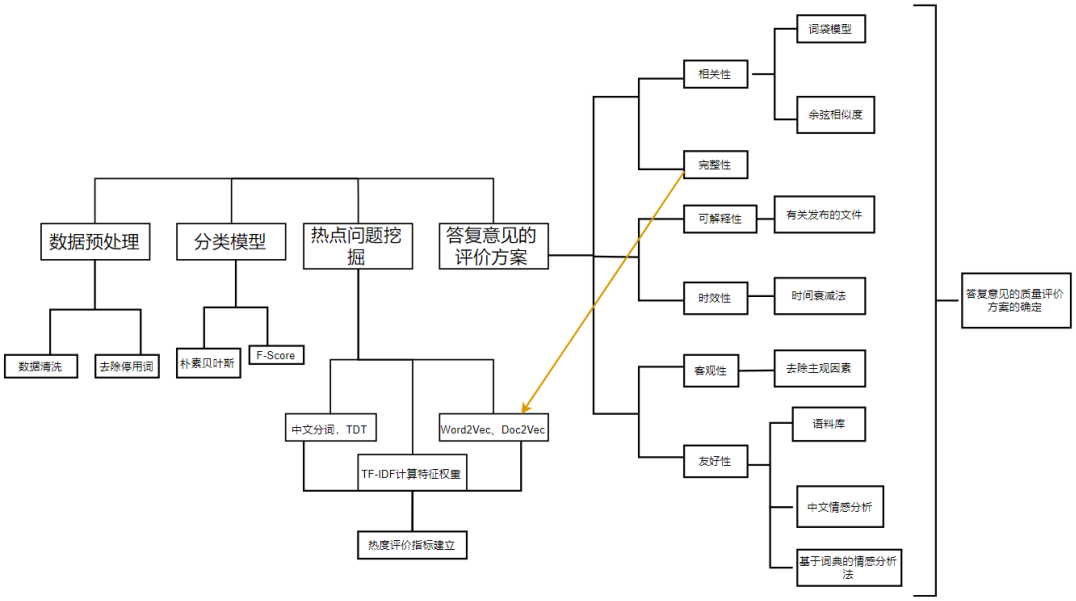

图 1 整体流程图

3、问题分析

3.1 关于留言分类的分析

首先本文对题目所给数据进行预处理，其次基于 TF-IDF 算法提取关键词[1]。

TF-IDF 算法提取关键词是一种简单有效的提取关键词的方法。其思想主要在于预先统计在语料中出现的所有词的词频，针对要提取关键词的文章或句子的每个词计算出 tf 值，再计算出 idf 值，相乘即为 tf-idf 值，tf-idf 值越大表示作为关键词的优先级越高。

TextRank 这种算法主要基于 pageRank 算法。其设计之初使用与 Google 的网页排名的，PageRank 通过互联网中的超链接关系来确定一个网页的排名，其公式是通过一种有向图和投票的思想来设计的。

展示提取的城乡建设一类评论的关键词词云如图 2，全部文件见附件 Keywords_3.csv。

图 2 关键词词云图

在提取关键词以后利用监督学习算法中的朴素朴素贝叶斯进行建模，最后用 F-Score 对模型进行检验，检验模型的优劣并进行改进。

3.2 热点问题的文本挖掘分析

首先对留言进行中文分词处理，用 TF-IDF 算法计算特征权重，word2vec 模型和 Doc2Vec 模型训练得到词向量和短文向量，再通过文本聚类找出话题，构建热度评价指标体系，结合用户行为对留言热度的影响提出衡量话题热度的热点话题影响力，最后利用影响力计算公式计算热点话题得分并进行排名。

3.3 答复意见评价的分析

针对答复意见，本文从相关性、完整性、可解释性、时效性、客观性、友好性六个指标进行分析评价，构建指标体系，熵权法计算权重，量化评价进行数据挖掘进而达到对答复意见评价的分析。

4、数据预处理

为了提高数据挖掘的质量首先对数据进行预处理。

首先查看原数据，截取部分情况如图 3：

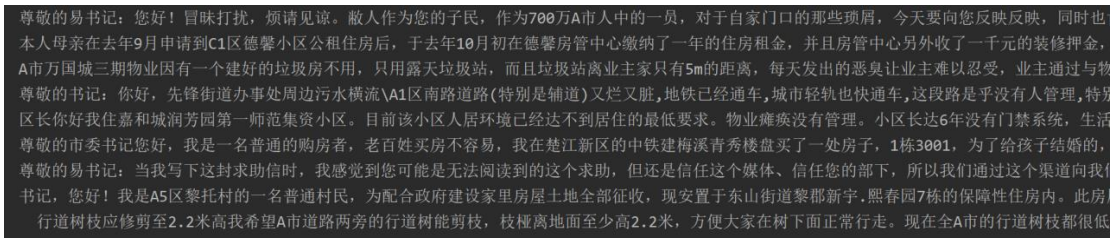

图 3 原始数据情况图

对原始数据进行数据清洗，去除不需要的标点符号，以及处理停用词，得到处理以后的数据如图 4：

图 4 数据处理以后的情况图

在进行正式的数据挖掘之前使用，大大提高了数据挖掘模式的质量，降低实际挖掘所需要的时间。

5、建立关于留言内容的一级标签分类模型

5.1 朴素贝叶斯算法思想

监督学习算法中的朴素贝叶斯分类过程，首先有一批已知分类的数据集，对每个输入对象提取特征，根据输入对象的特征属性和输入对象的所属分类，计算分类与特征属性之间的概率关系，以此来实现样本的训练。当对新的输入对象进行预测所属分类时，提取新输入对象的特征，根据训练好的概率，判断输入对象属于每个分类的概率。

5.2 模型的建立

设 $x=\{a_1,a_2,...,a_m\}$ 为一个待分类项，而每个 a 为 x 的一个特征属性，有类别集合 $C=\{y_1,y_2,...,y_n\}$，计算 $P(y_1|x),P(y_2|x),...,P(y_n|x)$。如果 $P(P(y_k|x))=\max\{P(y_1|x),P(y_2|x),...,P(y_n|x)\}$，则 $x\in y_k$。

要实现对输入对象进行分类，首要要能够对输入对象提取用于分类的特征。在留言信息中特征就是单词，我们就是根据单词来进行类别识别。

将所有数据按照 29：1 的比例进行划分，29 份做训练集，1 份做测试集，具体过程如图 5：

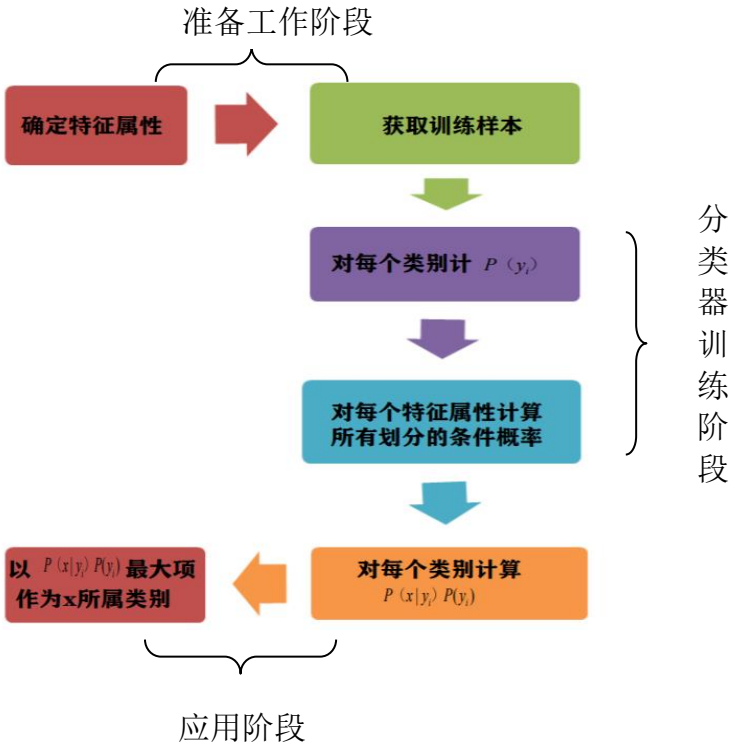

图 5 模型建立过程图

统计得到各类别下各个特征的条件概率估计。即：

$$P(a_1|y_1),P(a_1|y_1),...,P(a_1|y_1);P(a_1|y_2),P(a_2|y_2),...,P(a_m|y_2);\\ ...;P(a_1|y_n),P(a_2|y_n),...,P(a_m|y_n) \quad (1)$$

假设各个特征属性是条件独立的，根据朴素贝叶斯定理有：

$$P(y_i|x)=\frac{P(x|y_i)P(y_i)}{P(x)} \quad (2)$$

因为分母对于所有类别为常数，因此我们将分子最大化处理，且各特征属性是条件独立的，所以：

$$P(x|y_i)P(y_i)=P(a_1|y_i)P(a_2|y_i)...P(a_m|y_i)P(y_i)=P(y_i)\prod_{j=1}^{m}P(a_j|y_i) \quad (3)$$

对每个类别计算一个概率 $p(y_i)$，然后再计算所有特征的条件概率 $p(x_j|y_i)$，那么分类的时候我们就是依据朴素贝叶斯找一个最可能的类别：

$$p(a_i|y_0,y_1,...,y_n)=\frac{p(a)}{p(y_0,y_1,...,y_n)}\prod_j^n p(y_j|a_i) \quad (4)$$

5.3 模型的检验

查看部分留言信息如图 6：

留言主题	留言详情	一级标签
希望L市能给小学生教室装空调保暖	天气越来越冷了，可是小学生们上课的教室没有暖气...	城乡建设
A7县北山镇新桥集镇棚户提质改造工程存在质量问题	屋面材料、外墙防水工程提质改造验收交付续后使用质量无...	城乡建设
J8县爱莲名邸闲置五年仍未动工	J8县爱莲名邸（2号小区）作为J8县行政中心搬迁的配套项目...	城乡建设
A市坪塘大道施工及大车噪音严重扰民	坪塘大道施工及大车噪音严重扰民近几个月，之前环评造假...	环境保护
A8县市花明楼镇郑家冲的贵庭住工沥青味太重	位于花明楼镇郑家冲的贵庭住工，每到凌晨一两点偷偷摸摸...	环境保护

图 6 话题留言信息

利用 R 语言的 e1071 包中 naiveBayes()函数进行朴素贝叶斯的实现，预测情况如图 7：

劳动和社会保障	城乡建设	环境保护	城乡建设	教育文体	城乡建设	劳动和社会保障	商贸旅游	交通运输	城乡建设
劳动和社会保障	城乡建设	环境保护	城乡建设	教育文体	城乡建设	卫生计生	商贸旅游	交通运输	城乡建设

图 7 部分预测情况图

其中第一行是预测情况，第二行是真实情况，可以看出分类的准确率挺高。下面对模型进行更详细的检验，下面给出所有类整体的预测情况，如图 8。

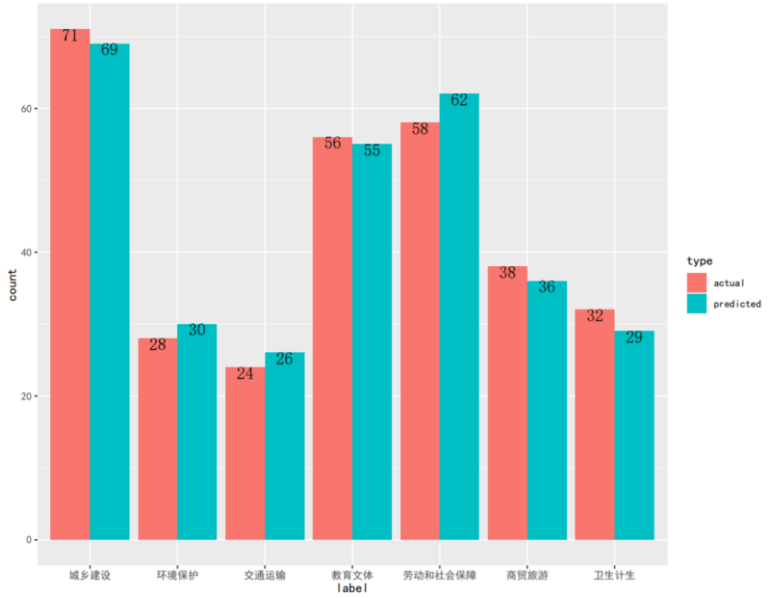

图 8 整体预测图

由图可以发现预测值和真实值很接近，下面使用 F-Score 对分类方法进行更直观准确的评价。

在二元分类模型中，个案预测有四种结局[2]，以“城乡建设”为例：

1. 真阳性（true positive， TP）：诊断为城乡建设，实际上也是；
2. 伪阳性（false positive，FP）：诊断为城乡建设，实际却不是；
3. 真阴性（true negative，TN）：诊断为不是城乡建设，实际上也不是；
4. 伪阴性（false negative，FN）：诊断为不是城乡建设，实际上却是。

这四种结局可以画成 2 × 2 的混淆矩阵：

表 1 混淆矩阵

		真实性		总数
		p	*n*	
预测输出	*p*`	TP	FP	P`
	n`	FN	TN	N`
总数		P	N	

其中，阳性（正例）：Positive（用 P 表示），阴性（反例）：Negative（用 N 表示）。

样本的数量 Z =正例 + 反例 = P + N = TP + FP + FN + TN

利用软件求的混淆矩阵详情如图 9：

predicted \ actual	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	
城乡建设	66	2	1	0	0	0	0	0
环境保护	3	26	0	0	0	1	0	
交通运输	0	0	23	1	0	2	0	
教育文体	0	0	0	51	1	3	0	
劳动和社会保障	1	0	0	4	55	1	1	
商贸旅游	1	0	0	0	0	31	4	
卫生计生	0	0	0	0	2	0	27	

图 9 混淆矩阵详情图

首先分别计算查准率(Precision)、查全率(Recall)[3]。

查准率 P_i 为第 i 类的查准率，即正真正确占预测正确的比重：

$$\Pr ecision = \frac{TP}{TP + FP} \qquad (4)$$

查全率 R_i 即正真正确占实际正确的比重：

$$\mathrm{Re}\,call = \frac{TP}{TP + FN} \tag{5}$$

综合可得到 F-Score 的具体计算方法，如下：

$$F_i = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{6}$$

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。利用 R 软件实现，得到每一类留言的“召回率”、“准确率”和 F 值如表 2：

表 2 模型检验信息表

	P	R	F
城乡建设	0.9565	0.9296	0.9429
交通运输	0.8846	0.9583	0.9200
教育文体	0.9273	0.9107	0.9189
劳动和社会保障	0.8871	0.9483	0.9167
环境保护	0.8667	0.9286	0.8966
卫生计生	0.931	0.8438	0.8852
商贸旅游	0.8611	0.8158	0.8378

由表 2 可知，7 类的 F 值前四类均大于 0.9，环境保护和卫生计生接近 0.9，只有商贸旅游在 0.85 以下，在误差允许的范围内可认为模型检验效果很好，其中城乡建设的 F 值是最高的，达到了 0.9429，在经济发展的时代，城乡建设是促进经济增长的重要原因之一，而商贸旅游最低为 0.8378，这一类检验得到的模型效果偏低，与 FN 的预测偏多有关。

对所有类综合处理，得到 P、R 和 F 的整体效果，画出箱线图如图 9：

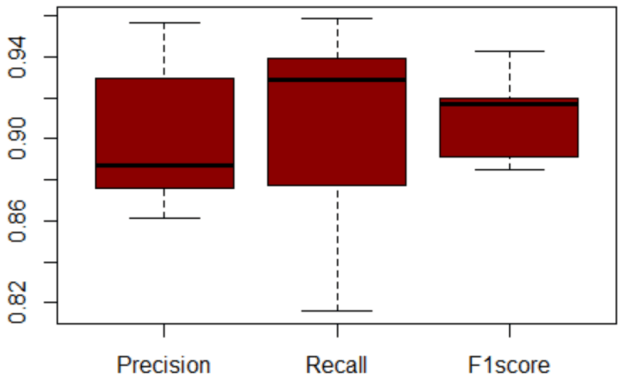

图 10 箱线图

对所有类检验，得到整体的 P、R 和 F 的分别是 0.9020、0.9050 和 0.9026，均大于 0.9，由图可以直观的看出整体值是很高且集中的，说明分类模型的质量很好。

6、热点问题的文本挖掘

本文前面进行了留言文本直接处理建模，这里在之前的基础上对留言文本进行中文分词处理，就是将汉字序列切分成有意义的词，以字为单位，句和段则通过标点等分隔符来划界[4]。

基于 TDT 技术设计出中文留言热点话题识别流程，主要环节如图 10 所示。

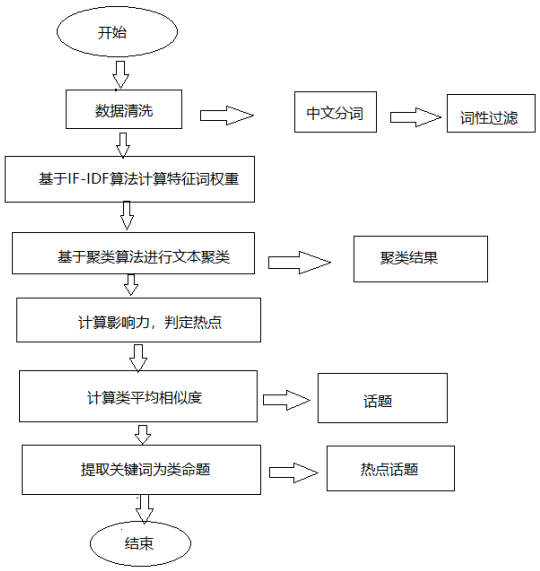

图 11 整体环节流程图

利用中文分词处理过滤数据；对预处理后的留言中的每个特征词，利用特征词权值计算方法 TF-ID（TermFrequency-InverseDocumentFrequency）计算特征权重并建立向量空间模型，分别利用 Word2Vec、Doc2Vec 词语将和短文的

稀疏矩阵转化成为容易计算的稠密矩阵，再利用 K-means 文本聚类来归纳出多个话题；最后对多个话题的影响力进行计算并分析，通过效果验证识别出热点话题。

6.1 中文分词

中文分词系统对留言进行分词处理，同时标注词性，并过滤留言内容，保留名词及名词性词语，然后将所有的单字过滤，再去除所有的英文字符、数字和一系列数学符号等非中文词，只留下有意义的中文词语。图 11 为提取出的留言文本示例。

实训公司休息影响不到物流记录店铺回馈小区油烟 原因经营扰民补交土地生活声音没人缴纳管道返回

图 12 话题文本示例

做分词前后的词量的频数分析，如下图：

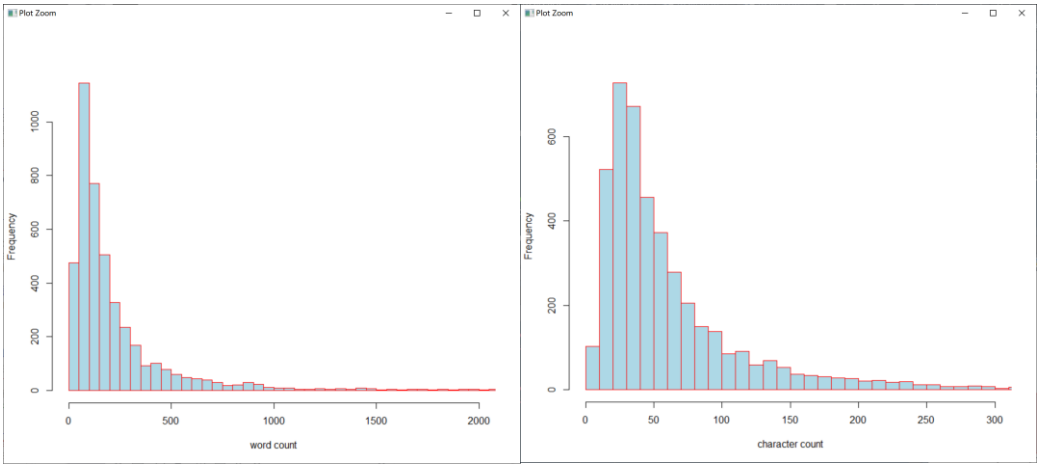

图 13 分词处理前后对比图

由图 12 可知分词前后留下的有效词量相比分词前少了很多，且更加集中；分词前的词汇量集中在区间[0,500]，词频在区间[0,1100]，明现更加稀疏；分词后词汇量集中区间[0,100]，词频在区间[0,700]，说明处理后的效果较好。

6.2TF-IDF 算法计算特征权重

清洗之后的数据，一共是有 9210 条，随机选取 8903 条作为训练集，剩下的为测试集，满足 29：1 的条件，在朴素贝叶斯文本分类器里，计算 TF-IDF 算法。

根据句子上下文环境给句中的每个词标记一个正确的词性[5]，进行分词，选用 TF-IDF 算法的算法计算特征权重。

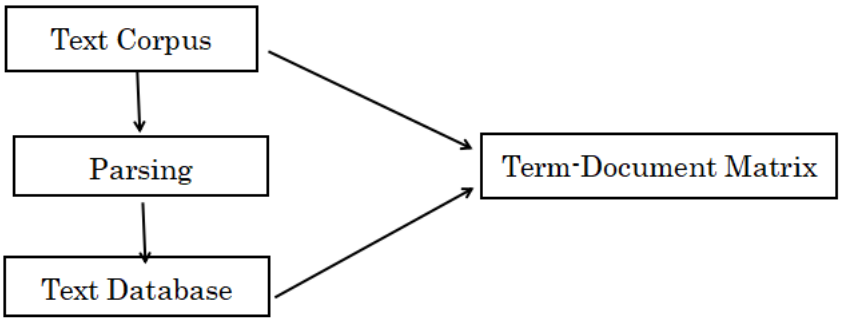

图 14 文本挖掘的处理流程图

经过分词处理后得到留言内容的离散符号向量如下：

1. A 市[0,0,0,0,0,0,0,1,0,.....,0,0,0,0,0,0,0]
2. J4 县[0,0,0,0,1,0,0,0,0,.....,0,0,0,0,0,0,0]
3. J7 县[0,0,0,1,0,0,0,0,0,.....,0,0,0,0,0,0,0]
4. A2 区[0,0,0,0,0,0,0,0,0,.....,1,0,0,0,0,0,0]

上述我们可以知道在语料库中，A 市、J4 县、J7 县、A2 区各对应一个向量，向量中只有一个值为 1，其余都为 0。但是 One-HotEncoder 有以下问题。一方面，城市编码是随机的，向量之间相互独立，看不出城市之间可能存在的关联关系。其次，向量维度的大小取决于语料库中字词的多少。如果将世界所有城市名称对应的向量合为一个矩阵的话，那这个矩阵过于稀疏，下面解决解决这个问题。

6.2Word2Vec

训练每个词映射成 k 维实数向量（k 一般为模型中的超参数），通过词之间的距离来判断语义相似度。。

首先将高维分词后的关键词表示的词语映射成低维向量，将 4201 维转换成 100 维，再在保留单词上下文的同时，从一定程度上保留其意义。通过 CBOW 和 Skip-gram 两种方法实现 Word2Vec。

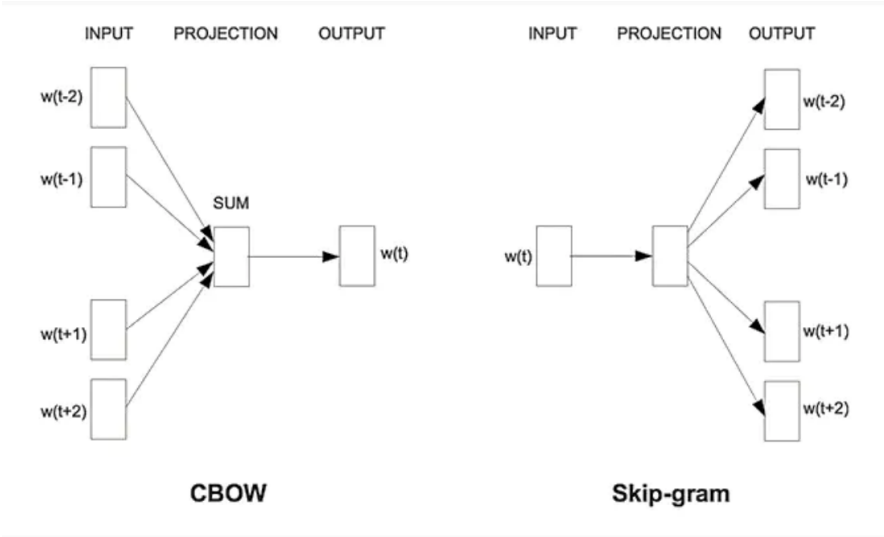

图 15Word2Vec 高维低维转换图

Word2Vec 将分词后的关键词转化为低维度的连续值，也就是稠密向量，并且其中意思相近的词将被映射到向量空间中相近的位置。

将 embed 后的留言向量通过 PCA 降维后可视化展示出来，如图 15：

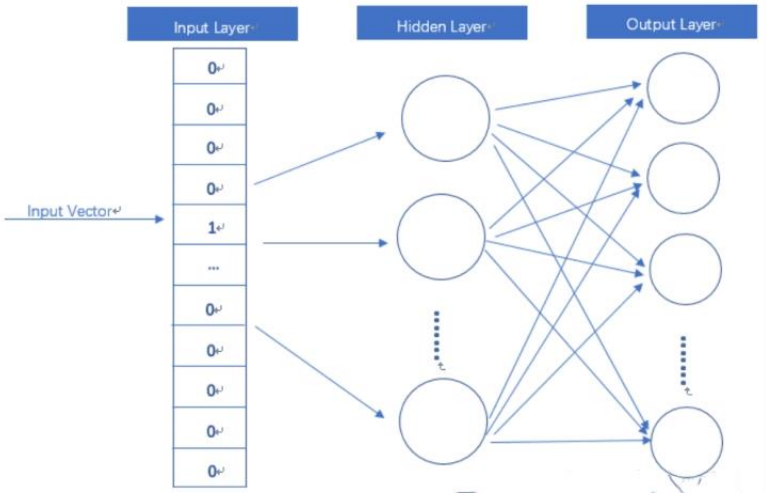

图 16 PCA 降维可视图

InputLayer 是分词后的高维向量，HiddenLayer（隐藏层）没有激活函数。OutputLayer 维度跟 InputLayer 的维度一样，用的是 Softmax 回归。我们要获取的低维向量其实就是 HiddenLayer 的输出单元。

$$[0 \quad 0 \quad 0 \quad 1 \quad \dots \quad 0 \quad 0 \quad 0]_{1\times 4201} = \begin{bmatrix} 2 & 11 & \dots & 7 & 16 \\ 6 & 21 & \dots & 12 & 18 \\ 17 & 5 & \dots & 7 & 1 \\ 10 & 12 & \dots & 19 & 7 \\ \dots & \dots & \dots & \dots & \dots \\ 3 & 9 & 10 & 23 & 14 \\ 23 & 15 & 8 & 9 & 14 \\ 8 & 17 & 13 & 11 & 19 \end{bmatrix}_{4201\times 100} = [10 \quad 12 \quad \dots \quad 19 \quad 7]_{1\times 100}$$

权重矩阵就是每个词的向量编码一行一行的堆叠起来，所以说之前的只有一个 1 其他全是 0 的超长稀疏表达矩阵就转变为了稠密的矩阵，将这个稠密矩阵定义的隐藏层的宽度为 100。

6.3 Doc2Vec

6.3.1 Doc2Vec 模型训练

前面完成了 Word2Vec 训练词向量的步骤，通过 Word2Vec 模型训练出唯一的向量来表示词。接下来，利用 Doc2Vec 对留言句子和短文训练唯一的向量。

Doc2Vec 模型类似 Word2Vec，预测词向量时，预测出来的词是含有词义的，比如上文提到的词向量“项目”会相对于“噪音”离“规划”距离更近，在 Doc2Vec 中也构建相同的结构。现在存在训练样本，每个句子是训练样本。使用 PV-DM 训练方式，如图 16：

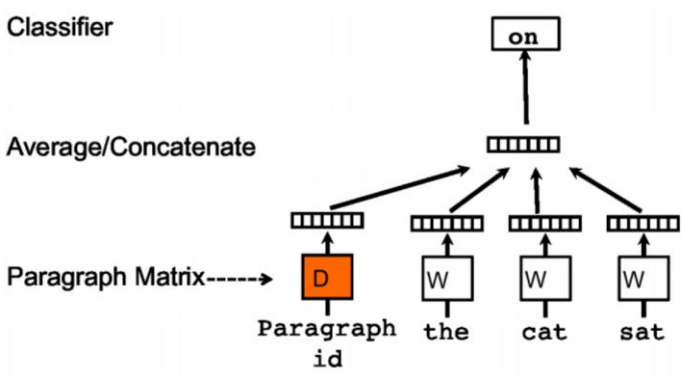

图 17 PV-DM 训练图

6.3.2 Doc2Vec 模型训练结果

在 python 中使用 gensim 包调用 Doc2Vec，使用给定数据，里边每一条都是群众的留言信息。具体的 Doc2Vec 训练 Paragraph vector 步骤如下：

- 导包：导入包，使用 jieba 给文本进行分词。

```
import gensim
import jieba
import pandas as pd
import os
from gensim.models.doc2vec import Doc2Vec
```

图 18 jieba 文本分词

- 导入数据集，提取留言列，根据第 4 节提到的数据预处理方法进行处理。
- 将提取好的留言列中的内容进行分词，并去除停用词。
- 改变成 Doc2Vec 所需要的输入样本格式，由于 gensim 里 Doc2Vec 模型需要的输入为固定格式，输入样本为：[句子，句子序号],这里需要用 gensim 中 Doc2Vec 里的 TaggedDocument 来包装输入的句子。
- 加载 Doc2Vec 模型，并开始训练。
- 模型训练完毕以后，就可以预测新的句子的向量了，这里用 gensim 里 infer_vector()预测新的句子，alpha（学习步长）设置小一些，迭代次数设置大一些。可以看到训练出来的结果与测试的新句子是有关联的。

6.4 K-means 文本聚类

K-means 算法以欧式距离作为相似性的评价指标[6]，即认为两个对象的距离越近，其相似度就越大，得到紧凑且独立的簇是聚类的最终目标。K-means 算法中距离的计算公式如下：

$$v=\sum_{i=1}^{k}\sum_{x_i\in S_I}(x_j-u_i)^2 \qquad (8)$$

K-means 算法流程如下：

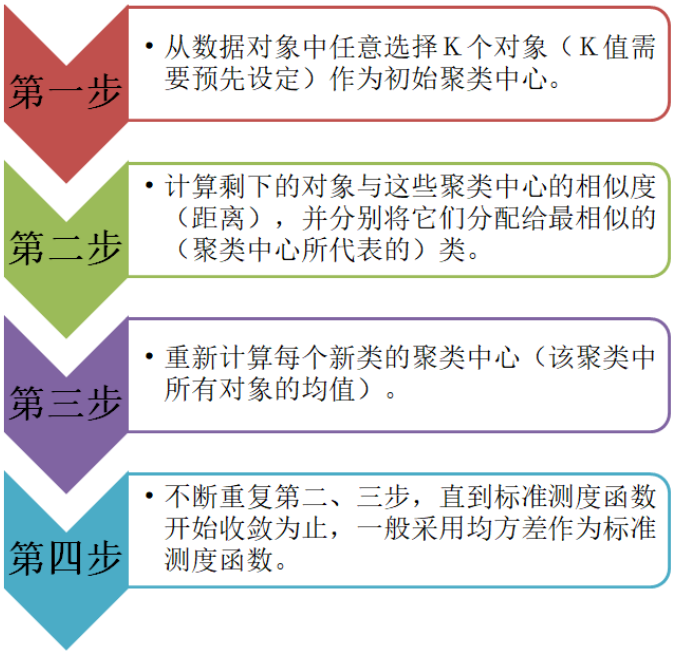

图 19 K-means 算法流程图

该算法在处理大数据集时是相对高效和可伸缩的，计算的复杂度为$O(N_{kt})$，其中N是数据对象的数目，t 是迭代的次数（一般K≤N，t≤N，同时算法对顺序不太敏感，因此较适合对 VSM 表示的文本集进行聚类。本文聚类效果的验证采用类平均相似度，公式为：

$$AVG_T(SIM)=\frac{\sum C_{T_{i=1}} f(avg(sim))}{C_T}(t\in T) \qquad (9)$$

其中$AVG_T(SIM)$表示类T的平均相似度；C_T表示类T所包含的微博条数；$f(avg(sim)$表示类T中单条留言t的个体平均相似度，即t与类T中其余留言的相似程度之和取平均值。将类中所有留言的个体平均相似度之和取一次平均值，从而得到类的平均相似度数据如图 19，具体见附件 mean_cor.csv。

	1	2	3	4	5	6	7	8
mean_cor	0.1400613	0.1333322	0.143611	0.144516	0.1399589	0.1474039	0.1738926	0.

图 20 分类信息详情图

	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数	kc$cluster
1	188006	A00102948	A3区一米阳光摇纱艺术摄影是否合法纳税了？	2019/2/28 11:25:05	座落在A市A3区联丰路米兰香天G2栋320，一家名叫一米阳...	0	0	47
2	188007	A00074795	咨询A6区道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	A市A6区道路命名规划已经初步成果公示文件，什么时候能...	0	1	123
3	188031	A00040066	反映A7县鲁华镇金鼎村水泥路、自来水到户的问题	2019/7/19 18:19:54	本人系鲁华镇金鼎村七里组村民，不知是否有相关水泥路到...	0	1	13
4	188039	A00081379	A2区黄兴路步行街大吉道巷住户卫生间臭便外排	2019/8/19 11:48:23	靠近黄兴路步行街，城南路街道、大吉道巷、一步两搭桥小...	0	1	124
5	188059	A00028571	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	2019/11/22 16:54:42	A市A3区中海国际社区三期四期中间，即蓝天唤和利幼儿园...	0	0	104
6	188073	A909164	A3区翡翠社区单方面改变翡翠明珠小区6栋底层住房使用性质	2019/3/11 11:40:42	作为翡翠社区翡翠明珠小区6栋居民，我们近期感觉到很震...	0	0	147
7	188074	A909092	A2区富绿新村房产的性质是什么？	2019/1/31 20:17:32	“二离一邻”发出关于针对非法集资的打击的通知中是针对的...	0	0	199
8	188119	A00035029	对A市地铁违规用工问题的质疑	2019/5/27 16:04:44	我是一名在A市某地铁站上班的安检员，我是由中介公司介...	0	0	169
9	188170	A88011323	A市6路公交车随意变道通行	2019/12/23 8:50:24	12月21日下午17时52分许，6路公交车（自机座位旁边的汽...	0	0	95
10	188249	A00084085	A3区银利翡翠林语枫阵坡路与翡松路交汇处地铁施工噪音2点施...	2019/9/17 4:25:00	银利翡翠林语枫阵坡路与翡松路交汇处地铁施工噪音2点施工，...	0	0	152
11	188251	A00013092	A7县特立路与东四路口晚高峰太堵，建议调整信号灯配时	2019/10/19 11:02:40	近来，下午晚高峰五点半左右，经过特立路与东四路口时，...	0	0	113
12	188260	A00053484	A3区青青家园小区乐果果零食炒货公共通道程放空调扰民	2019/5/31 17:06:13	还我宁静我爱夏习迎考，太半年底商空调/冰拒外机轰响扰民...	0	0	152
13	188396	A00047580	关于拆除聚英龙是在西地省南学院宿舍旁安装变压器的请求	2019/4/15 16:23:09	枫阵坡589号白鹅咀停车场，由聚英龙是新能源公司建的“商...	2	1	186
14	188399	A00097934	A市利保壹号公馆项目夜间噪声扰民	2019/7/3 6:23:25	也好，我想举报 A市利保壹号公馆项目 夜间噪声扰民A市利...	0	0	49
15	188409	A0003274	A市地铁3号线星沙大道站地铁出入口设置极不合理！	2019/6/19 10:14:39	尊敬的领导也好，我是A7县星沙街道的一个有两个孩子，上...	0	4	90
16	188414	A00096844	A4区北辰小区非法住改商问题何时能解决？	2019/8/1 7:20:31	也好！我是北辰D1区一名业主，近年来，我小区住改商数量...	0	0	15
17	188416	A00029753	请独K3县乡村医生发卫生室执业许可证	2019/06/20 20:38:47	K3县的乡村卫生室现在大多处于无证行医的状态。造成这种...	0	0	189
18	188451	A00013004	A7县鲁华镇石磨铺村有党员家开麻将馆	2019/4/11 17:54:25	我是鲁华镇一名村民，最近接到政府的通知，Q29县道旁不...	0	2	166
19	188455	A00035902	咨询异地办理出国签证的问题	2019/5/16 15:20:43	书记也好！我是外地人，2018年毕业后落户A市，请问申请...	0	0	30
20	188467	A00050188	投诉A市盛斯顿英语培训学校恶延退费！	2019/3/28 19:57:19	退费之日起3个月之内要退还一切学习费用！于今日已经超...	0	1	13
21	188475	A00055810	A6区敏源国际广场停车场隐患乱建现象严重	2019/12/3 15:04:58	A市A6区月亮岛街道的敏源国际广场停车场隐患乱建现象严...	0	0	3
22	188535	A00061775	A7县时代星城4栋有非法经营的家庭旅馆	2019/6/13 15:28:44	尊敬的省位领导A7县时代星城小区4栋1321室一直非法存在...	0	0	73
23	188546	A0006817	A2区佳兆业水质都小区垃圾无人处理	2019/1/23 13:09:19	敬爱的领导：你好，我是佳兆业水质都一期小区业主。我小...	0	0	22

图 21 分类示例图

由图 20 可知，在聚出的类中，类平均相似度有很多小于 0.01，所以剔除这些类，留下的类作为最终结果。

6.5 热点问题留言识别

针对每条留言内容，进行上面的处理后，实验中的K值在最大值范围内通过多次实验结果验证来选取。经过多次试验，最终将留言内容聚 200 类，并对各类进行类关键词提取，结果如表 21 所示。

	length.keyword
1	287,['实训', '公司', '老师', '学院', '管理', '工?
2	219,['提供', '购房', '材料', '领导', '人才', '医?
3	111,['药品', '癌症', '建议', '关爱', '生活', '光?
4	139,['油烟', '经营', '扰民', '关闭', '小区', '烧?
5	227,['证件', '原因', '状态', '村医', '作为', '造?
6	457,['居民', '柴火', '店面', '油烟', '烧烤', '小?
7	457,['居民', '柴火', '店面', '油烟', '烧烤', '小?
8	211,['学院', '南塘', '公交车', '暮云', '节省', '?
9	275,['溪湖', '游玩', '桃花', '形成', '商业空间', >
10	111,['网络', '覆盖', '基础', '完成', '基本', '城?
11	66,['经营', '处理', '烧烤', '上能', '地方', '没有>
12	108,['油烟', '直排', '无法', '市政府', '请求', '?

图 22 关键词图

为保证数据的准确性，在以上 200 类中，进行剔除弱信息。本文对剩余的数据求类平均相似度，此时设置类平均相似度阈值为 0.3，剔除类平均相似度小于 0. 3 的弱实时性的留言话题，剩下 10 类实时性较强的话题。

得到最终的分类情况如上，可以发现存在以一条为一个类，所以剔除这些类，避免出现误差，留下其他 10 类计算他们的类平均相似度。

表 3 类平均相似度

	1	2	3	4	5	6	7	8	9	10
类平均相似度	0.487	0.440	0.995	0.997	0.996	0.996	0.996	0.996	0.996	0.997

整理得出留言信息如图 22：

留言主题	留言时间	留言详情
问问A市经开区朱六线以西泉塘昌和商业中心以南的有关规划	2019/1/2 20:27:26	A市经开区朱六线以西，泉塘昌和商业中心以南，新建品阁...
问问A市经开区朱六线以西泉塘昌和商业中心以南的有关规划	2019/1/11 15:46:04	A市经开区朱六线以西，泉塘昌和商业中心以南，新建品阁...
A市一师三附小龙湾小学门口迫切需要搭建人行天桥	2019/10/17 14:12:29	您好！希望您在百忙之中，能收到来自龙湾小学2000余名师...
A市木莲中路百米人行道路狭窄	2019/6/21 12:48:56	木莲中路益光幸福锦庭至涧井路交汇段有百米人行道非常狭...
问问A市经开区朱六线以西泉塘昌和商业中心以南的有关规划	2019/1/2 20:27:07	A市经开区朱六线以西，泉塘昌和商业中心以南，新建品阁...
咨询A5区涧井镇鄱阳村拆迁安置房问题	2019/9/16 22:58:33	书记您好：本人系A5区涧井镇鄱阳村村民，有三个疑问待书...
咨询移动通信业务问题	2019/10/18 23:52:58	尊敬的领导：你好，本人前期向西地省通信管理局反馈，本...
A3区洋湖湿地附近空域飞机轰鸣声打扰学生学习	2019/4/15 10:42:28	A市A3区洋湖湿地附近空域经常有飞机声，噪音不言而喻，...
J4县供销合作社在岗失业职工追缴社保	2020/01/06 10:20:31	关于J4县供销合作社在岗失业职工追缴社保问题的诉求尊敬...

图 23 留言信息示例图

提取出的 10 类情况如上，所有情况见附件 theme_classed_all.csv，从图中可以看出这 10 类话题的具体情况，他们是实时性最强的类，根据留言主题、信息点赞数以及反对数等可以得出话题的影响力得分。

6.6 热度评价指标体系建立及实现

大多关于热点发现的算法认为，在聚类后出现留言影响力排序的热点词频率较高，则该话题即为热点话题。这种原理是基于热点词与话题的附属关系，但却忽略了当话题较分散的情况下聚类也能进行，同时在聚类结果中，可能有些话题只是局部较热的小话题，整体来讲算不上热度很高[7]，因此本文设置一个阈值来区分话题冷热，话题热度（本文中以话题影响力来衡量）高于阈值则表示聚类出来的话题为“热点话题”，低于阈值则视为“非热点话题”。

定义热度评价指标，如图 23：

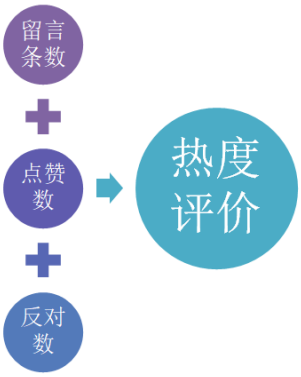

图 24 热度评价指标体系

热点与非热点的概念是相对的，因此建立影响力得分公式如下：

$$F = \alpha x_1 + \beta（x_2 + x_3）\qquad（10）$$

其中F是影响力得分，α、β是权重，x_1, x_2, x_3分别是留言条数、点赞数和反对数，然后按照热度分数排序，分数越高表示话题影响力越大，热度越高。

根据话题影响力公式计算出前 5 类留言的热度，得到热度指数以及相应热点问题截取部分信息如图 25，详情见附件热点问题表.xls，

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	58.4	19/1/2至2019/10/	A市经开区	对闲置土地的申请规划
2	2	4.4	19/4/15至2020/1	J4县供销合作社失业职工	在岗失业职工追缴社保
3	3	3.8	19/7/18至2019/7/	A7县新国道107	马路距离住户太近，相关政府部门为何不同意拆迁
4	4	1.6	19/7/7至2019/8/	A市伊景园滨河苑	捆绑车位销售
5	5	1.8	9/10/31至2019/10	A市南塘城轨	能否设立公交站

图 25 热点问题示例图

得到留言详细信息截取部分信息如图 24 ，对应的留言详细信息见附件热点问题留言明细表.xls。

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
1	233542	A00080329	线以西泉塘昌和商业中	2019/1/2 20:27	A市经开区东六线以西，泉塘昌和商	0	24
1	239670	A00080329	线以西泉塘昌和商业中	2019/1/11 15:46	A市经开区东六线以西，泉塘昌和商	0	41
1	246891	A000112858	龙湾小学门口迫切需要	2019/10/17 14:12	您好！希望您在百忙之中，能收到来	0	0
1	248550	A00019345	木莲中路百米人行道路	2019/6/21 12:48	木莲中路蓝光幸福	0	0
1	256358	A00080329	线以西泉塘昌和商业中	2019/1/2 20:27	A市经开区东六线以西，泉塘昌和商	0	29
1	287451	A00045746	区洞井镇鄱阳村拆迁安	2019/9/16 22:58	书记您好：本人系A5区洞井镇鄱阳村	0	0
2	209864	A00094706	咨询移动通信业务问题	2019/10/18 23:52	尊敬的领导：你好，本人前期向西	0	0
2	225682	A00028989	附近空域飞机轰鸣声	2019/4/15 10:42	A市A3区洋湖湿地附近空域经常有飞	0	0

图 26 留言详细信息示例图

考虑到留言信息的扩散性，相对于评论其影响力更大，因此公式中 α 取值为 0.4，β 取值为 0.6。最终得到问题的留言归类见表 4。

根据话题影响力计算公式：

$$影响力=留言条数\times 0.4+(点赞数+反对数)\times 0.6 \qquad (10)$$

得到所提取的 10 个话题在当前时段的影响力评分及排名，如表 4 所示。

表 4 热点问题影响力排行表

话题	留言数	点赞数	反对数	总评分	影响力排行
土地规划	6	94	0	58.8	1
搅拌站噪音扰民	4	33	2	22.6	2
公园项目规划优化	4	29	0	19	3
小区门口建设医院	4	16	0	11.2	4
新国道 107 附近居民住房拆迁	4	5	0	4.6	5
在岗失业职工追缴社保	5	2	2	4.4	6
建立公交站	3	1	0	1.8	7
捆绑车位销售	4	0	0	1.6	8
建设施工噪音扰民	3	0	0	1.2	9
候车厅增设改签窗口	2	0	0	0.8	10

根据表中信息，我们得出热度评价指标为土地规划、搅拌站噪音扰民公园项目规划等十项。可以发现影响力最强的是有关土地规划的话题，总评分达到了58.8，说明这类现象在群众意见中比较普遍，业内人士指出，在资金流不充裕的情况下，房企选择参拍地块更为谨慎，非优质地块较此前更易流拍，但对热点城市、优质土地的争抢依然激烈[8]。

继而是搅拌站噪音扰民的情况，评分达到 22.6，说明该类工厂严重影响到了居民的日常生活，这与现在在居民区建立搅拌站的违法工厂有关；排名第三的是公园项目规划的问题，评分为 19，紧接第二之后，说明群众对公园的建设问题也很在意，这和全民健身意识加强有关；排名最低的是候车厅增设改签窗口，候车厅，候车厅的增加容易出现在发展比较快的地区，虽然这些地区较少，但是近年来经济发展较快，这类问题也逐渐增多。综上，相关部门应该重视此类问题，加强在这些方面的改善与创新。

7、答复意见的质量评价方案

7.1 评价指标体系的建立

答复意见的评价针对相关部门的答复意见，我们知道有效的答复意见可以很好的帮助公民解决需求，而答复意见的好差本文建立[9]答复的相关性、完整性、可解释性、时效性、客观性、友好性六个指标进行分析评价，下面我们将给出一套评价方案来对答复进行评价[10]，评价体系如图 26：

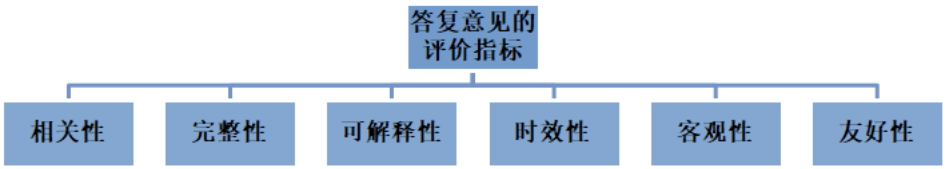

图 27 答复意见评价指标体系

确定好评价指标后，下面就这六个指标进行逐一分析处理。

7.2 相关性

7.2.1 词袋模型

对于留言和答复意见（下文统称文本）中出现频率越高的词项，越能描述文本（不考虑停用词）。统计每个词项在每个文本中出现的次数即词项频率，记为

$$tf_{t,d}$$

其中 t 为词项，d 为文本。获得文本中每个词的 tf 权重，一个文本转换成词-权重的集合，称词袋模型（bag of words model）即 VSM 模型[11]。词袋就是说，像是把一个文本拆分成一个一个的词条，然后将它们扔进一个袋子里，在袋子里的词与词之间是没有关系的，因此词袋模型中，词项在文本中出现的次序被忽略，出现的次数被统计。例如，“噪音” 和“噪声” 具有同样的意义。将词项在每个文本中出现的次数保存在向量中，得到文本向量。

6.2.2 余弦相似度

前面提出了文本向量的概念，通过 VSM 模型，将文本进行检索、文本聚类、文本分类等。考虑到两条相似的文本，由于文本长度不一样，他们的向量差值会很大。下面采取余弦相似度的方法去除文本长度的影响。

$$sim(d_1,d_2)=\frac{\vec{V}(d_1)\cdot\vec{V}(d_2)}{\left|\vec{V}(d_1)\right\|\vec{V}(d_2)\right|} \tag{11}$$

公式的分母是两个向量的欧几里得长度之积，分子是两个向量的内积。这样计算得到的 sim 是两个欧式归一化的向量之间的夹角的余弦。如图 27：

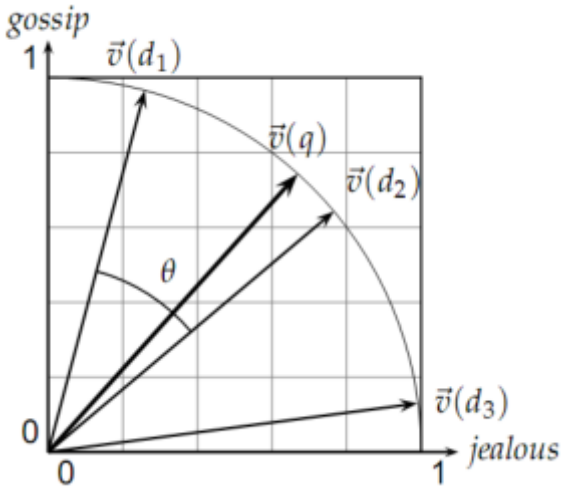

图 28 余弦相似度展示图

得到留言和答复意见的余弦相似度如下表，具体数据见附件 indicator_all.csv 的 E 列，表 5 展示部分数据：

表 5 相似度

序号	1	2	3	4	5	6	7	8
相似度	0.981	0.954	0.965	0.907	0.876	0.799	0.934	0.957

由表可以看出该指标值非常大，均大于 0.9，说明该部分每一条答复与留言的相关性较强。

7.3 完整性

对于答复的完整性，把留言和答复的关键词做在一个数据库中，将留言与答复进行一对一求相关度，根据 5.3 节提到的 Word2Vec 中提到的词向量的方法计算，过程如图 28：

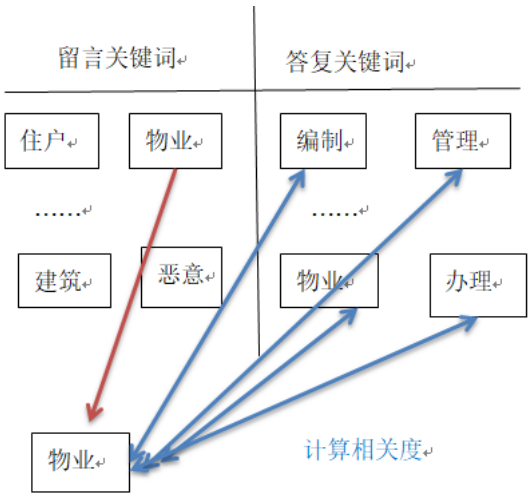

图 29 Word2Vec 计算过程图

计算得到每条留言和所有答复关键词的相关度，定义阈值为 0.95，统计相关度是否存在大于 0.95 的值，若存在至少一个值大于则记作 1，反之，若均小于 0.95 则记作 0，得到所有留言关键词的个数与上述的结果，见附件 result_comp.csv，示例图如下：

length_revi	count_all
81	39
69	15
73	18
62	12
38	10

图 30 关键词个数示例图

对得到的结果进行如下计算：

$$完整度=\frac{count\_i}{length\_i} \quad (12)$$

得到最终的完整性得分完整度，具体数据见附件 indicator_all.csv 的 F 列，表 展示部分数据如表 6：

表 6 完整度

序号	1	2	3	4	5	6	7	8
完整度	0.481481	0.217391	0.246575	0.193548	0.263158	0.0625	0.357143	0.210084

由表可以看出该指标值较小，均小于 0.5，说明该部分每一条答复的完整性较差。

7.3 可解释性

利用已经得到的关键词，利用 R 软件进行处理，得到相关已发布了的措施文件及数量数量，如图 30：

```
[[978]]
[1] "《鹊山村土地合作经营实施细则》" "《鹊山村土地合作经营实施细则》"
[3] "《鹊山村土地合作经营实施细则》" "《村民委员会组织法》"

[[979]]
[1] "《建筑用轻质隔墙条板》"      "《内隔墙–轻质条板（一）》"  "《住宅使用说明书》"
[4] "《住宅使用说明书》"          "《住宅使用说明书》"          "《西地省房产面积测算规则》"

[[980]]
[1] "《A市城市市容和环境卫生管理办法》"

[[981]]
NULL

[[982]]
[1] "《A8县市农村公共资源交易管理办法》"
```

图 31 措施文件图

答复意见有理可依，数据越大说明答复意见的可解释性却强，其质量越好。对得到的数据进行归一化处理，得到所有答复意见可解释度，具体数据见附件 indicator_all.csv 的 D 列，表 7 展示部分数据：

表 7 可解释度

序号	1	2	3	4	5	6	7	8
可解释度	0	0	0.05	0.05	0	0	0.05	0.05

由表可以看出该指标值非常小，均小于 0.1，说明该部分每一条答复的可解释性非常差。

7.4 时效性

答复意见的质量考虑时间效应，因为答复的机构人员处理留言是有时间变化的。留言者一年前留言的问题现在不一定感兴趣，相比于推荐过去问题的解决方案，推荐留言者问题的进行方案更有参考价值。每个系统时间效应的大小不同，比如时间对电影的作用就没有新闻那么明显。要考虑时效性，必须加入时间参数，比如三元组(留言者,问题,时间)代替简单的二元组(留言者,问题)。给定时间 T ，一般化的时间衰减公式为：

$$n_i(T)=\frac{1}{1+\alpha(T-t)} \tag{13}$$

其中 $n_i(T)$ 是留言 i 的时效性，α 是时间衰减参数，由最大时间差的倒数计算求得，T 是发布问题的时间，t 答复时间。

得到最终的时效度数据，具体数据见附件 indicator_all.csv 的 A 列，表 8 展示部分数据：

表 8 时效性

序号	1	2	3	4	5	6	7	8
时效性	0.987	0.987	0.987	0.987	0.986	0.973	0.965	0.975

由表可以看出该指标值非常高均大于 0.95，说明时效性很强。

7.5 客观性

客观性又称真实性，与主观性相对，客观性，即客观实在性，它指事物客观存在，唯有将主观性通过实践与客观性统一才能获得客观真理。

在分词后的数据中，找出主观词，例如：我，我们等，记作 x，对数据进行

归一化处理，利用以下公式计算客观度：

$$客观度=\frac{\max x-\min x-x}{\max x-\min x} \quad (14)$$

得到客观度数据，具体数据见附件 indicator_all.csv 的 B 列，表 9 展示部分数据：

表 9 客观性

序号	1	2	3	4	5	6	7	8
客观性	0.958333	0.986111	0.972222	1	0.986111	0.986111	0.972222	0.986111

由表可以看出该指标值非常大，均大于 0.9，甚至还有达到 1 的，说明该部分每一条答复的客观性很强。

7.6 友好性

7.6.1 语料库分析

为了对带有情感的答复意见进行分类，我们利用已有标签数据的文本进行训练，作为语料库，在这里，我们对答复意见文本做一个简单的分析，包括数据集中的情感分布，数据集中的答复意见句子长度分布。其中 0 为负面评价，1 为正面评价。

利用训练好的库[12]，对答复意见进行评分。

7.6.2 中文情感分析

情感倾向可认为是主体对某一客体主观存在的内心喜恶，内在评价的一种倾向，它由两个方面来衡量：一个是情感倾向方向，一个是情感倾向度[13]。

中文情感分析，句子是由词语根据一定的语言规则构成的，应该把句子中单词的依赖关系纳入到句子感情的计量中，不同的依存关系，进行情感倾向计算是不一样的。句子的情感，根据单词对句子的重要程度赋予不同权重，调整词语对句子情感的贡献程度。

7.6.3 构建语料库对评级量化

我们选取的是知网发布的情感分析用词语集，利用 python 的 snownlp 包和基于词典的情感分析方法，并通过语料库的训练，得到答复意见的友好度具体数据见附件 indicator_all.csv 的 C 列，表 10 展示部分数据：

表 10 友好性

序号	1	2	3	4	5	6	7	8
友好性	0.824023	0.037856	0.959154	0.589035	0.105884	0.116427	0.470899	0.001481

由表可以看出该指标值有大有小，说明该部分每一条答复的友好性不均一，存在较大的不同。

7.7 答复意见的质量评价方案确立

得到六个指标的信息后，根据熵权法计算指标权重[14]，具体流程如下：

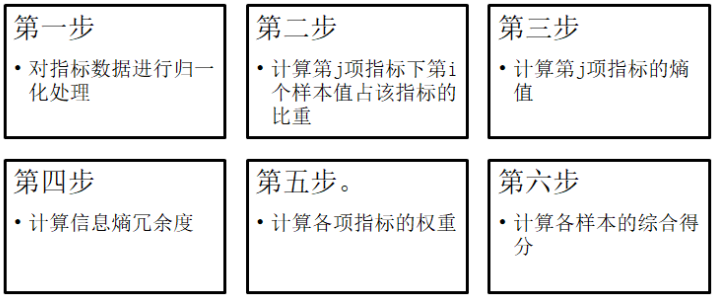

图 32 熵权法计算权重过程图

所有的答复意见的综合得分见 score_all.csv 文件，得到答复意见的评价得分模型：

$$评价得分=0.004x_1+0.006x_2+0.121x_3+0.785x_4+0.003x_5+0.089x_6 \quad (18)$$

其中，$x_i,(i=1,2,...,6)$分别是时效性、客观性、友好性、可解释性、相关性(cor)、完整性。可知答复意见的质量与可解释性相关性最强，权重达到了 0.785，因为回复里提到的相关出版文献是比较权威的答复，其次是友好性和完整性，它们的权重为 0.121、0，089，次于可解释性，因为答复人员的情感、完整性与答复质量比较相关，情感度、完整度越高说明答复质量越好，客观性、时效性以及相关性权重较低，在整体的答复质量上的贡献度较小，可能与其他指标之间存在相关性。

通过对得分加权后聚类得到评价答复意见好坏的定性结果，聚类后的可视化如下：

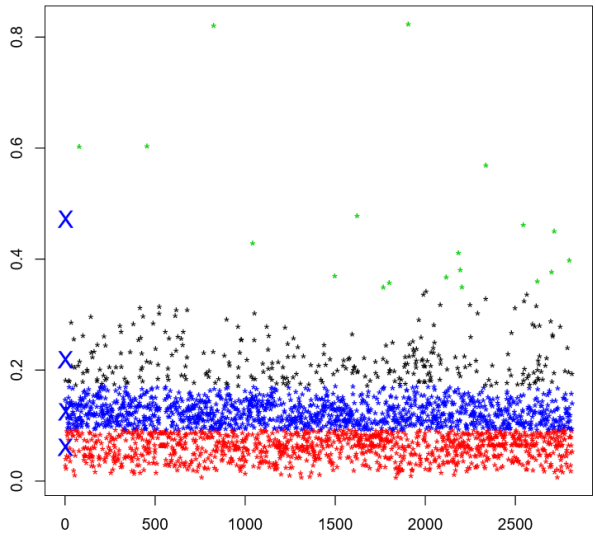

图 33 聚类可视化图

给出答复质量评价的区间如下：

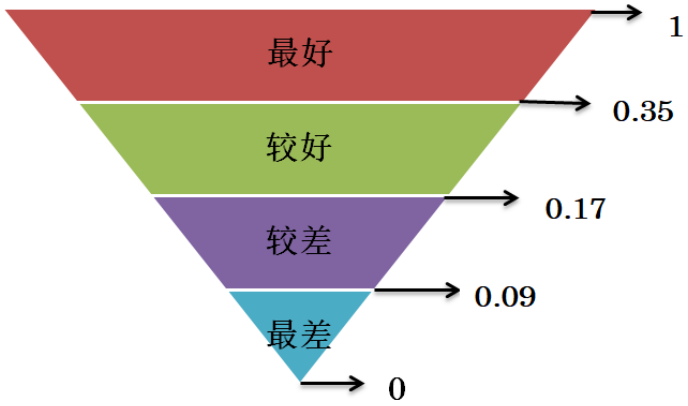

图 34 答复意见的定性结果图

定义答复意见质量的定性评价区间为：[0,0.09)，[0.09,0.17)，[0.17,0.35)，[0.35,1]分别是质量由最低到最高的区间。

通过建立的答复意见评价指标体系可以对任何一条满足本文要求的留言进行定性与定量描述，我们对所给每一条留言进行评价，评价情况如下，详情见答复质量评价表.xlsx：

留言时间	留言详情	答复意见	答复时间	score	eval
2019/4/25 9:32:09	业公司却以交20万保证金，不	收取停车管理费，在业主大会结束后业委会	2019/5/10 14:56:53	0.146885	较差
2019/4/24 16:03:40	面的生意带来很大影响，里	需整体换填，且换填后还有三趟雨污水管	2019/5/9 9:49:10	0.027873	最差
2019/4/24 15:40:04	同时更是加大了教师的工作压	办幼儿园聘任教职工要依法签订劳动合同，	2019/5/9 9:49:14	0.181511	较好
2019/4/24 15:07:30	落户A市，想买套公寓，请问	年龄35周岁以下（含），首次购房后，可分	2019/5/9 9:49:42	0.131729	较差
2019/4/23 17:03:19	“马坡岭小学”，原“马坡岭	保留“马坡岭”的问题。公交站点的设置需	2019/5/9 9:51:30	0.039983	最差
2019/4/8 8:37	再把泥巴冲到右边，越是上下	于您问题中没有说明卫生较差的具体路段，	2019/5/9 10:02:08	0.023094	最差
2019/3/29 11:53:23	为老社区惠民装电梯的规范	A市A3区人民政府办公室下发了《关于A市A3	2019/5/9 10:18:58	0.132042	较差
2018/12/31 22:21:59	好远，天寒地冻的跑好远，	修前期准备及设施设备采购等工作。下一步	2019/1/29 10:53:00	0.062079	最差
2018/12/31 9:55:00	也没得到相关准确开工信息。	单位落实分户检查后，西地省楚江新区建设	2019/1/16 15:29:43	0.045201	最差
2018/12/31 9:45:59	立交桥等地方做立体绿化，取	部分也按规划要求完成了建设，其中西边绿	2019/1/16 15:31:05	0.090085	较差
2018/12/30 22:30:30	规划局审批通过《温室养殖	局支付一笔耕地征收补偿款给原大托村，但	2019/3/11 16:06:33	0.140992	较差
2018/12/29 23:27:51	区安置房地下室近两万平方米	续，按长人防发[2014]7号文件要求，鄱阳	2019/1/29 10:52:01	0.082216	最差

图 35 答复质量评价信息图

本文给出了一套完整的关于答复意见的质量评价方案，为其他类似质量评价方面的问题提供了一个很好的参照。

8、参考文献

[1]https://blog.csdn.net/gzt940726/article/details/80256011

[2] https://blog.csdn.net/Joseph__Lagrange/java/article/details/90813885

[3]https://blog.csdn.net/XiaoYi_Eric/article/details/86726284

[4]张启宇,朱玲,张雅萍.中文分词算法研究综述[J].情报探索,2008(11).

[5]杨冠超. 微博热点话题发现策略研究[D].杭州：浙江大学硕士学位论文，2011.

[6] 李梅.改进的 K 均值算法在中文文本聚类中的研究[J].安徽大学硕士学术论文.2010，4.

[7]程军军，刘云. 基于新闻评论的热点话题发现系统研究[J]. 网际网路技术学，2008，9（5）.

[8] http://house.jschina.com.cn/lsyw/202004/t20200426_6619052.shtml

[9] 候玉林.基于文本意见挖掘的快递服务质量研究.[J].北京交通大学，2019（6）.

[10] 成全、王火秀、骈文景.基于证据推理的医疗健康网站信息质量综合评价研究[J].福州：福州大学经济与管理学院.2020，4（191）.

[11] 基于改进的 TF-IDF 权重的短文本分类算法[J]. 杨彬,韩庆文,雷敏,张亚鹏,刘向国,杨亚强,马雪峰. 重庆理工大学学报(自然科学). 2016(12)

[12] 徐志浩. 基于维基百科的中文命名实体语料库构建研究[D].苏州大学,2016.

[13] 冯鑫,王晨,刘苑,杨娅,安海岗.基于评论情感倾向和神经网络的客户流失预测研究[J].中国电子科学研究院学报,2018,13(03):340-345.

[14] https://zhuanlan.zhihu.com/p/28067337