“智慧政务”中的文本挖掘应用

摘要

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、 汇聚民智、凝聚民气的重要渠道，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。因此，运用网络文本分析和数据挖掘技术对群众问政留言信息分类和热点整理的研究具有重大意义。

针对问题1，将留言详情文本数据分为训练集和测试集，利用Python的jieba中文分词工具对训练集数据进行分词并去除噪声数据，并通过TF-IDF算法提取每类一级标签的前50个特征词项，并得到每个特征词项的权重。同样对测试集的每个文档进行中文分词、并通过TF-IDF算法对分词进行加权处理，最后利用朴素贝叶斯文本分类法将待分类项归到所属的一级标签类别中。

针对问题2，对留言主题进行中文分词并去除噪声数据，采用TF-IDF算法生成TF-IDF向量，得到权重向量后采用K-means算法对留言主题进行聚类，最后采用余弦相似度计算聚类后每一类中主题的相似度，进一步提高聚类效果，并根据定义的热度评价指标对每一类主题进行评价，评价结果以热度指数形式呈现，利用excel对热度指数排序并选出热度指数前5名的类。

针对问题3，对留言详情和答复意见内容进行分词，利用Python实现分词后答复意见是否含有留言详情涉及的词语，以此来判断答复意见与留言详情的相关程度及完整性，最后构建解释性字眼列表来判断答复意见是否具有解释性。

关键词：Python　中文分词　TF-IDF算法　朴素贝叶斯分类　K-means聚类

Application of Text Mining in "Smart Government"

Abstract

In recent years, as WeChat, microblog, mayor's mailbox, sunshine hotline and other online political platforms have gradually become an important channel for the government to understand public opinions, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been continuously increasing, which has brought great challenges to the work of relevant departments that used to rely mainly on manpower to divide messages and sort out hot spots. Therefore, it is of great significance to research on message information division and hot spot sorting by using network text analysis and data mining technology.

In response to question 1, the text data of message details are divided into training set and test set. Python's jieba Chinese word segmentation tool is used to segment the training set data and remove noise data. TF-IDF algorithm is used to extract the first 50 feature word items of each class of first-level labels, and the weight of each feature word item is obtained. Similarly, each document in the test set is subjected to Chinese word segmentation, and the word segmentation is weighted by TF-IDF algorithm. Finally, the items to be classified are classified into the category of the first-level label by Naive Bayesian text classification.

In response to question 2, Chinese word segmentation is carried out on message topics and noise data is removed. TF-IDF algorithm is used to generate TF-IDF vectors. After weight vectors are obtained, K-means algorithm is used to cluster message topics. Finally cosine similarity is used to calculate the similarity of topics in each category after clustering to further improve clustering effect. Each category of topics is evaluated according to the defined heat evaluation index. The evaluation results are presented in the form of heat index. excel is used to sort the heat index and select the top five categories of heat index.

In response to question 3, word segmentation is carried out on the details of the message and the contents of the reply comments. Python is used to realize whether the reply comments after word segmentation contain words related to the details of the message, so as to judge the relevance and completeness of the reply comments and the details of the message. Finally, a list of explanatory words is constructed to judge whether the reply comments are explanatory.

Keywords: Python Chinese word segmentation TF-IDF algorithm
Naive Bayesian classification K-means clustering

目录

1 挖掘目标

本次建模目标是以互联网公开来源的群众问政留言记录及相关部门对部分群众留言的答复意见内容作为文本数据，利用 jieba 中文分词工具、朴素贝叶斯分类算法、TF-IDF 算法、K-means 聚类算法，达到以下三个文本挖掘目标：

（1）利用 jieba 中文分词工具、朴素贝叶斯分类算法、TF-IDF 算法对数据进行文本挖掘，把群众的留言详情归到已知的类别中。

（2）以群众的留言主题作为数据挖掘对象，通过聚类算法将相似度高的文本归为一类，并利用余弦相似度计算各类别中名词的相似度，即把地名相似度高的文本聚类。

（3）对留言信息和答复意见进行分词，计算答复意见中含有留言信息相关特征词项的个数，以此来识别答复意见的相关性程度以及完整性，最后识别答复意见内容是否含有权威性解释的相关字眼。

2 分析方法与过程

2.1 问题 1 的分析方法与过程

2.1.1 流程分析

首先，对给出附件 2 的全部数据进行预处理，在预处理之前，先把全部的数据分为训练集和测试集两部分[1]，这里选数据集的 80%作为训练集，其中的 20%作为测试集，然后在训练集上进行特征选取并对特征进行加权处理[1][3]，每个特征都有不同的权重，权重的大小表示该特征词项能使最终的分类达到更好的效果所占的比重的大小，目的是使不同的特征词项因为重要程度而被区分对待，提高分类性能。由此可初步构造分类器。然后将构造好的分类器应用到测试集上对测试集的文本进行分类，输出所属类别，并根据正例和负例个数计算查准率、查全率以及 F 常使用 F-Score 对分类方法进行评价[7]。

2.1.2 数据预处理

附件2的数据处理主要是对留言详情进行处理,也就是对中文文本进行处理,而提取特征词项的过程也就是过滤噪声数据，所谓的噪声数据是指图片、音频以及标点符号等等，根据留言详情的实际内容，主要是去除标点符号、空格和无用词汇等。

2.1.2.1 对留言详情信息进行中文分词

这里直接采用python的中文分词包jieba进行分词。分词代码见附件cutwords.py。jieba采用了基于前缀词典实现的高效词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图,同时采用了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，对于未登录词，采用了基于汉字成词能力的HMM模型，使得中文分词达到了更好的效果[2][5]。

2.1.2.2 TF-IDF算法加权处理

在对留言详情内容进行分词后，需要把词语向量化，这里采用TF-IDF算法对特征词项进行加权处理[3]，把特征词项转为权重向量。具体原理如下：

词频（TF）是指某个词在文本出现的频率，公式表示：

$$\mathrm{TF}=\frac{\text{某个词在该文中出现的个数}}{\text{文中的总词数的个数}} \tag{1}$$

逆文档词频（IDF）公式：

$$\mathrm{IDF}=\log\left(\frac{\text{语料库中的文档总数}}{\text{包含该特征词的文档数}+1}\right) \tag{2}$$

最终公式：

$$\mathrm{TF-IDF}=\mathrm{TF}*\mathrm{IDF} \tag{3}$$

从TF-IDF的定义式可知，特征词在某个文本的权重与它在当前文本出现的频率成正比，与整个文本集中包含该特征词的文档数量成反比。TF-IDF值越大，表明包含该特征词的文本数量比较少并且该特征词在某个文本内容中属于高频词。计算每个词的TF-IDF值并进行排序，次数最多的即为要提取的特征词项。结合实际运行结果，以及考虑到留言详情内容较多，特征词项的选取不宜过少，这里选取权重排名前50的50个特征词项。

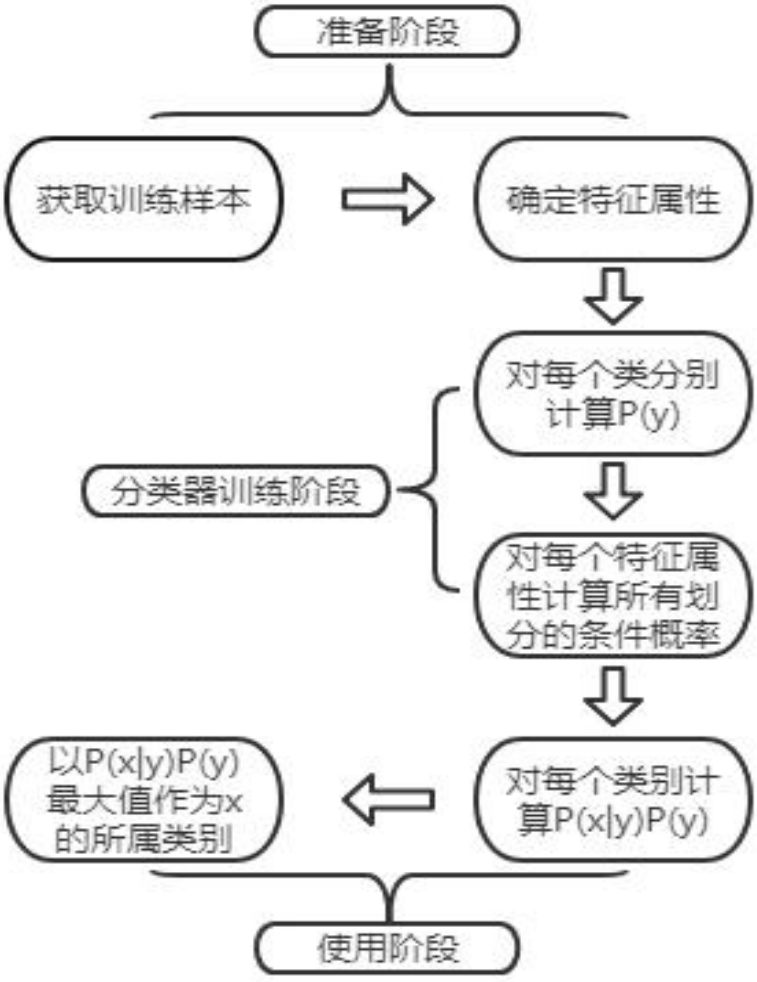
2.1.2.3 删除停用词

在进行分词加权之后，发现明显的对分类不起作用且权重较大的词，为了保证分类的准确度，需要将这些停用词删掉[1][4]，停用词一般包括语气助词（如：“啊”，“吧”，“呢”等），人称代词（如：“我”，“我们”，“你们”等）以及标点符号等，结合实际运行结果，将可能高频词但却对分类不起作用的词添加到停用表中，停用表见附加 stpowords.txt，然后利用停用表对文本进行一次删除操作，重新计算 TF-IDF 值。

2.1.3 建立分类模型

根据 TF-IDF 算法构造权重向量后，假设各个特征属性是独立的，采用朴素贝叶斯算法来构造分类器对留言详情内容进行分类。

流程图如下：



思想基础：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。

算法步骤：

（1）设 $x=\{a_1,a_2,...,a_m\}$ 为一个待分类项， a 为 x 的一个特征属性

（2）有类别集合 $C=\{\mathrm{y}_1,\mathrm{y}_2,...,\mathrm{y}_\mathrm{n}\}$

（3）计算 $P(y_1|x),P(y_2|x),...,P(y_n|x)$

（4）如果 $P(y_k|x)=\max\{P(y_1|x),P(y_2|x),...,P(y_n|x)\}$，则 $x\in y_k$

由于分类已知，即可得到各类别下的特征属性的条件概率估计，即

$$P(a_1|y_1),P(a_2|y_1),...,P(a_m|y_1)$$
$$P(a_1|y_2),P(a_2|y_2),...,P(a_m|y_2)$$
$$.$$
$$.$$
$$P(a_1|y_n),P(a_2|y_n),...,P(a_m|y_n)$$

所以

$$P(y_i|x)=\frac{P(x|y_i)P(y_i)}{P(x)} \tag{4}$$

又因为分母对于所以类别为常数，所以只需将分子最大化即可，即有

$$\begin{aligned}P(x|y_i)P(y_i)&=P(a_1|y_i)P(a_2|y_i)...P(a_m|y_i)P(y_i)\\&=P(y_i)\prod_{j=1}^{m}P(a_j|y_i)\end{aligned} \tag{5}$$

这样便可以得到所要分类项 x 属于每一个类别 y_i 的概率，从中选取最大的，则便得到分类项所属的类别。

2.2 问题 2 的分析方法与过程

2.2.1 流程分析

针对附件 3 给出的数据无法知道类别，所以应先对内容进行归类，这里主要对留言主题进行归类。首先，采用基于 TF-IDF 算法进行 K-means 聚类，目的是使相似的留言主题归为一类，为了进一步使分类效果更好，聚类后，二次计算每类文本中文本的余弦相似度，相似度高且占比较大的文本保留在该类中。最终聚类结束后，根据定义的热度评价指标对每一类主题进行评价，评价结果以热度指数形式呈现，最后选出热度指数前 5 名的类。

2.2.2 数据预处理

首先，与问题 1 的处理方式相同，对留言主题内容进行分词后采用 TF-IDF 算法生成 TF-IDF 向量，得到权重向量后采用 K-means 算法进行聚类[2][8]。通过观察附加 3 的留言主题，发现有表达同一个意思但表达方式不同的情况，且有 4326 条记录，综合考虑，这里取 K=20。

2.2.2.1 K-means 算法

K-means 聚类算法是使用误差平方和准则函数来评价聚类性能的。给定数据集 X，X 包含 K 个聚类子集，即 $X=(x_1,x_2,...,x_k)$，将此数据集划分为 K 类，m_i 是第 i 类的聚类中心，数据 x_i 属于第 i 类 C_i，则误差平方和准则函数，即代价函数公式为

$$V=\sum_{i=1}^{K}[\sum_{x_i\in C_i}||x_i-m_i||^2] \tag{6}$$

聚类过程就是寻找最佳聚类中心 $m_i(i=1,2,...,K)$ 使得代价函数 V 为最小的过程。即先计算每一类中各数据到该类中心地距离平方和，然后将 K 类地距离平方和相加，即可得到代价函数 V，当代价函数 V 最小时，K-means 算法收敛，即可实现对数据集的 K 个聚类划分。具体算法步骤如下：

（1）选定分类数 K。

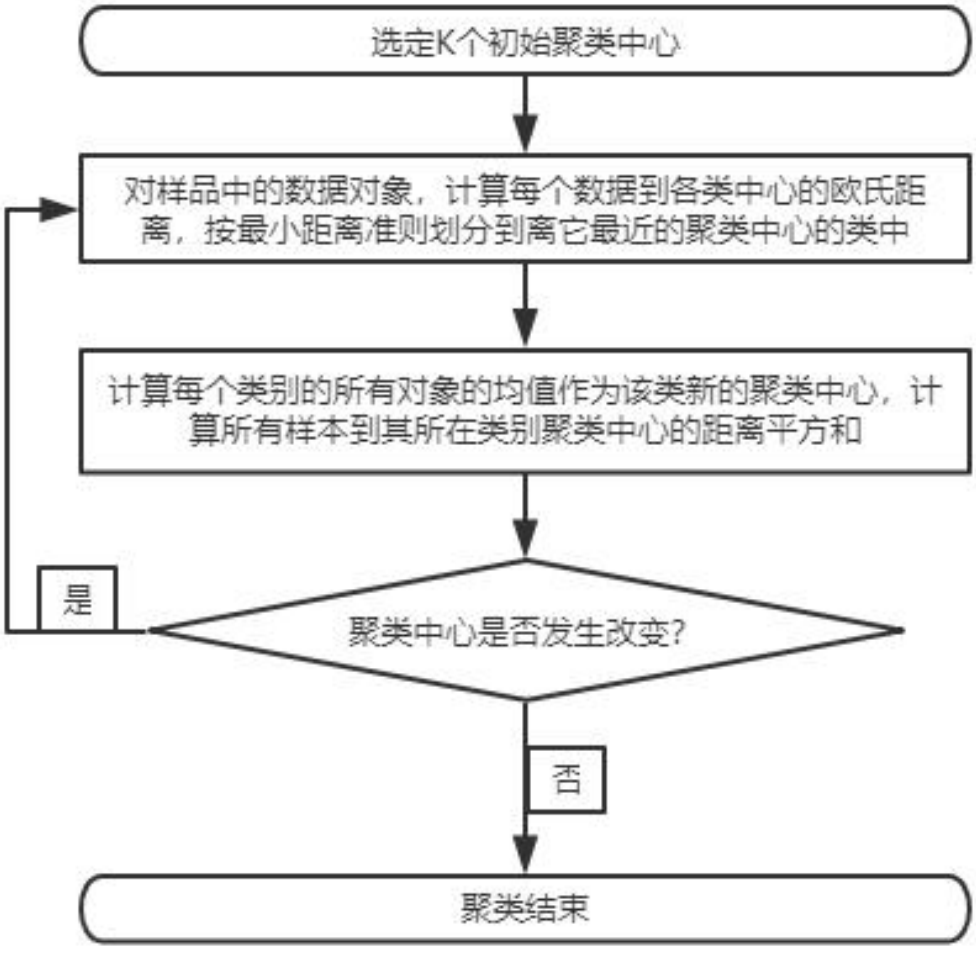
（2）从数据集 X 中选取 K 个元素，作为 K 类 $\{C_1,C_2,...,C_K\}$ 的初始聚类中心 $\{m_1(0),m_2(0),...,m_K(0)\}$，其中 $m_i(n)$ 表示第 i 类 C_i 在第 n 此迭代后新的聚类中心。

（3）根据欧式距离，依次计算每个数据到各类中心的距离，并比较这些距离的大小，按最小距离准则将个数据划分到离它最近的那个距离中心的类别中。

（4）重新计算 K 个聚类中心 $\{m_1(j), m_2(j), ..., m_K(j)\}$，其中 $m_K(j) = (\sum x_m)/n$ 表示第 j 次划分后聚类 C_i 的中心，n_i 为当前 C_i 中数据的个数。

（5）若第 $j-1$ 次划分后的聚类中心和第 j 次划分后的聚类中心不同，则 $m_i(j) \neq m_i(j-1)$，则继续执行步骤（3）（4）的聚类调整，直至连续两次迭代后的聚类中心没有变化，$m_i(j) = m_i(j-1)$，则算法收敛，聚类结束，此时，代价函数 V 最小，数据集 X 被划分为 K 个聚类。

流程图如下：



2.2.2.2 余弦相似度

$$Sim(d_i,d_j)=\frac{\sum_{K=1}^{M}W_{iK}*W_{jK}}{\sqrt{\sum_{K=1}^{M}W_{iK}{}^2}\sqrt{\sum_{K=1}^{M}W_{jK}{}^2}} \quad (7)$$

其中，d_i 为测试文本的特征向量，d_j 为 j 类中心向量，M 为特征向量维数：W_K 为向量的第 K 维[1]。

2.2.3 热度评价指标

由于最终聚类后留言主题有了较高的相似度，这里认为聚类后的主题已为某一特定人群或某一特定地点，所以聚类后的用户留言信息中对选取评价指标有用的信息可取留言用户的数量、反对数以及点赞数，考虑到反对数和点赞数可能是因为浏览信息用户的手误操作，所以赋予反对数和点赞数的权重不宜过大，但根据生活中的实际情况又考虑到两者具有一定的参考价值，根据附件 3 的数据特点，这里不考虑用户权威度[6]等影响，为了保证 index 在$(0,+\infty)$区间内，不会为负值，综合考虑，定义了以下的比较合理的热度指数评价公式：

$$index=m*10+\frac{i-j}{i+j}*9 \quad (8)$$

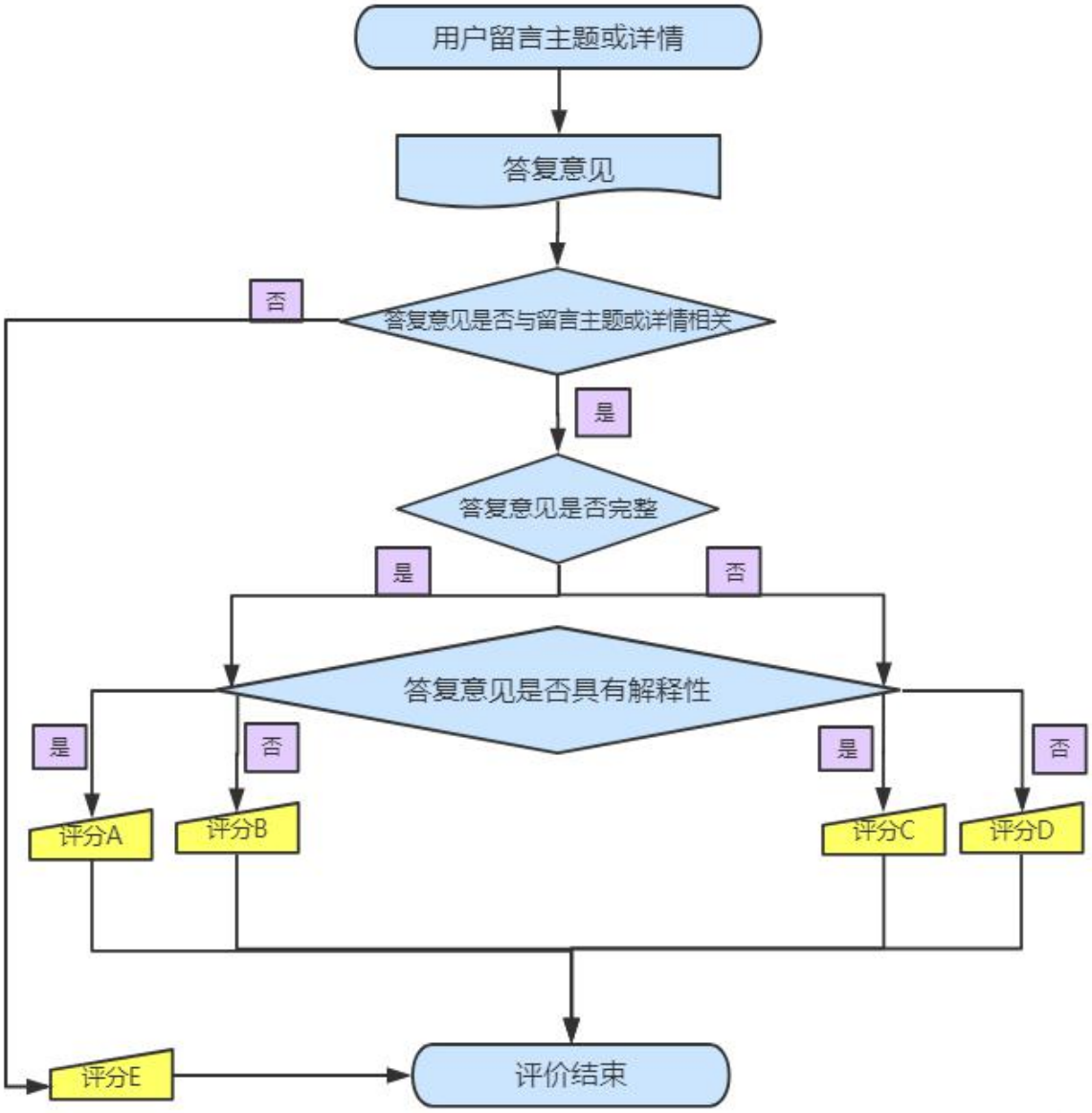
其中 m，i，j 分别代表用户数量、点赞数和反对数

2.3 问题 3 的分析方法与过程

2.3.1 流程分析

可从答复的相关性、完整性、可解释性依次判断答复意见的质量，这里定义了评价等级。首先，判断答复意见是否与用户留言的内容相关，如果不相关，即使答复内容完整且有解释性也是答非所问；若答复内容与用户留言相关，则先判断答复是否完整，最后再判断是否有可解释性。

评价流程图：



2.3.2 数据预处理

在判断答复意见内容是否与留言主题或详情相关时，同样采用分词步骤将留言和答复进行分词操作，这里取留言详情进行分词并去除噪声数据，取其中的特征名称作为识别特征，若对应的答复意见中含有留言详情的特征名词，则可认为答复意见与留言内容有所相关，这里同样采取 TF-IDF 算法（忽略权重）。

如若相关，判断答复意见含有留言详情的特征名词的个数与留言详情的特征名词的占比，若占比大于或等于原先设定的阈值 T，则可认为答复意见具有完整性。由于留言详情内容较多，提取词也较多，因此阈值可取 T = 0.4（经观察，阈值取 0.4 具有可行性）。

$$T=\frac{\text{答复意见中含有留言详情特征名词的个数}}{\text{留言详情中特征名词的总个数}} \quad (9)$$

可解释性即解释的权威性，判断答复意见是否具有权威性可从答复内容识别是否具有“根据”、“规定”、“法律”、“法规”、“政府”、“颁布”、“已经”“解决”以及书名号“《”、“》”来判定，若有出现以上标识，则可认为答复意见具有可解释性。由于具有解释性的答复有多种表达形式，而解释性的相关字眼列举有限，因此，评分等级达到 C 等级及以上的答复都可认为是质量良好的答复。

3 结果分析

3.1 问题 1 的结果分析

7 类一级标签的前 4 个特征词项及权重如下表，全部特征词项及权重见附件 Weight of keywords，代码见附件 TF-IDF.py

表 1 特征词项权重表

一级标签	特征词项及权重
城市建设	('业主', 0.0656), ('小区', 0.0566) ('建设',0.0178), ('房屋', 0.0261）
环境保护	('污染', 0.0615),('环保局', 0.0541) ('环保', 0.0258),('噪音', 0.0250)
交通运输	('出租车', 0.0901),('快递', 0.0654) ('的士', 0.0471),('司机', 0.0392)
教育文本	('学校', 0.0874),('教师', 0.0597) ('学生', 0.0583),('教育局', 0.0531)
劳动和社会保障	('职工', 0.0462),('退休', 0.0396) ('社保', 0.0361),('劳动合同', 0.0284)
商贸旅游	('景区', 0.0205),('旅游', 0.0191) ('游客', 0.0131),('消费者', 0.0109)
卫生共计	('医院', 0.1053),('医生', 0.0615) ('患者', 0.0293),('卫生院', 0.0182)

由训练集构造的模型（即分类器）应用在测试集中，代码见附件Classification.py，7类一级标签的运行结果整理如下：

表2 朴素贝叶斯文本分类

一级标签	precision	recall	f1-score
城乡建设	0.92	0.90	0.91
环境保护	0.89	0.92	0.91
交通运输	0.93	0.93	0.93
教育文体	0.96	0.94	0.95
劳动和社会保障	0.93	0.95	0.94
商贸旅游	0.94	0.96	0.95
卫生计生	0.92	0.95	0.94
Avg/total	0.93	0.94	0.93

其中f1的计算如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{9}$$

其中P_i为第i类的查准率，R_i为第i类的查全率。

根据TF-IDF计算结果选取的50个特征词项（每个类别50个）进行分类器的训练，并在测试集上预测，结果7个类别的的f1值的平均值为0.93，整体效果良好。

3.2 问题 2 的结果分析

由 K-means 聚类算法（代码见附件 K-means.py）得出的结果仍然含有噪声文本，如以下是部分整理的运行结果：

表 3 K-means 聚类结果

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
195917	A909119	A 市涉外经济学院组织学生外出打工合理吗？	2019/11/05 10:31:38	...	0	1
233759	A909118	A 市涉外经济学院强制学生实习	2019/04/28 17:32:51	...	0	0
242062	A00028889	西地省涉外经济学院变相强制学生“社会实践”	2019/11/27 23:14:33	...	0	0
264084	A00074365	西地省财政经济学院以报名名额已满拒绝让学生报名 cet-4	2019/3/19 23:11:44	...	0	0
266368	A00038920	A 市涉外经济学院寒假过年期间组织学生去工厂工作	2019/11/22 14:42:14	...	0	0
360110	A110021	A 市经济学院寒假过年期间组织学生去工厂工作	2019-11-22 14:42:14	...	0	0
360111	A1204455	A 市经济学院组织学生外出打工合理吗？	2019-11-05 10:31:38	...	1	0
360112	A220235	A 市经济学院强制学生实习	2019-04-28 17:32:51	...	0	0
360113	A3352352	A 市经济学院强制学生外出实习	2018-05-17 08:32:04	...	3	0
360114	A0182491	A 市经济学院体育学院变相强制实习	2017-06-08 17:31:20	...	9	0

可以看到A00028889和A00074365留言用户的留言内容显然与其他8位留言用户的留言内容不属于同一类，经过二次计算处理后（余弦相似度代码见附件sim.py），分类结果如下：

表4 重聚类结果

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
195917	A909119	A市涉外经济学院组织学生外出打工合理吗？	2019/11/05 10:31:38	...	0	1
233759	A909118	A市涉外经济学院强制学生实习	2019/04/28 17:32:51	...	0	0
266368	A00038920	A市涉外经济学院寒假过年期间组织学生去工厂工作	2019/11/22 14:42:14	...	0	0
360110	A110021	A市经济学院寒假过年期间组织学生去工厂工作	2019-11-22 14:42:14	...	0	0
360111	A1204455	A市经济学院组织学生外出打工合理吗？	2019-11-05 10:31:38	...	1	0
360112	A220235	A市经济学院强制学生实习	2019-04-28 17:32:51	...	0	0
360113	A3352352	A市经济学院强制学生外出实习	2018-05-17 08:32:04	...	3	0
360114	A0182491	A市经济学院体育学院变相强制实习	2017-06-08 17:31:20	...	9	0

由此可看出分类效果良好。由定义的合理热度评价指标计算出每一类的热度指数，按指数高低排序，排名前5的热点问题见附件热点问题表.xls，相应热点问题对应的留言信息见附件热点问题留言明细表.xls

3.3 问题 3 的结果分析

以下是对附件 4 其中两条答复意见的评价

表 5 评价结果

留言主题	答复意见	评价等级
希望相关部门治理一下中海国际社区一期旁边工地的噪音问题，早上很早就施工，严重影响居民的正常休息睡眠，晚上也有很大噪音，严重影响居民的睡眠质量。	网友“UU008201”您好！您的留言已收悉。现将有关情况回复如下：接到您的投诉后，A3 区城管执法中队于 2019 年 1 月 2 日晚至 A3 区国家大学科技城（东鹤）安置小区一期核实情况。据查，施工单位由于需要夜间连续作业，已办理《建筑工程夜间施工登记证明》。A3 区城管执法中队已告知施工方需降低噪音，文明施工，尽量不要影响到附近住户的休息。感谢您对我们工作的支持、理解与监督！2019 年 1 月 7 日	A
梅溪湖至今没有一个图书馆，这与梅溪湖品位极不相称。建议在艺术中心先期借一个小馆开办读书馆。方便住在梅溪湖的市民借阅。	网友“UU008706”您好！您的留言已收悉。现将有关情况回复如下：梅溪湖一期引进 A 市图书馆分馆，位于梅溪湖创新中心，已开馆营业。梅溪湖二期金菊路与雪松路东南角规划有西地省图书馆新馆，目前正在进行前期筹备工作，具体开馆时间待定。感谢您对我们工作的支持、理解与监督！2019 年 1 月 9 日	A

根据分词结果以及计算，两条留言的阈值分别为

$$T_1 = 0.4$$

$$T_2 = 0.6$$

因此符合相关性及完整性，经检验，答复意见具备解释性（方法实现代码见附件 authentic interpretation.py）。其余评价结果见附件“答复意见评价结果表.xls”。结果显示，答复意见评价等级中，A 等级占 47.9%，B 等级占 44.5%，C 等级占 1.8%，D 等级占 3.3%，E 等级占 2.5%，整体效果良好。

4 结论

现在，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，但同时也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。因此，对群众问政留言信息进行文本分析的研究对提升政府的管理水平和施政效率具有重大意义。本文主要采用朴素贝叶斯分类法和 K-means 聚类算法对群众问政留言信息进行分类和聚类。

由结果分析可知，在已知分类类别的基础上，且各个特征属性到各类别下的条件概率估计可直接得到，所以采用朴素贝叶斯分类法对群众问政留言信息进行分类，结果显示 F1=0.93，分类效果是比较好的。

利用 K-means 聚类算法对群众的留言信息进行聚类，聚类后利用定义的评价指标进行评价，评价结果显示热度排名前 5 的问题描述分别是小区临街餐饮店油烟噪音扰民、各种活动扰民、使用住房公积金出现问题、对小学建设及制度的不满以及学校强制学生去定点企业实习，而“扰民”话题就占了热度排名榜第一和第二，可见大部分群众对居住环境是比较重视的。

统计答复意见中具有留言详情特征词项的个数，以及判断是否具有解释性的字眼，从而给答复意见的质量进行评价。结果显示，答复意见评价 A、B、C 等级共占 94.2%，D、E 等级仅占 5.8%，整体的评价效果是较优的。

5 参考文献

[1] 张航.基于朴素贝叶斯的中文文本分类及 Python 实现.硕士学位论文.2018

[2] 孙海锋，郑中枢，杨武岳.网络招聘信息的数据挖掘与综合分析.北京林业大学.2016

[3] 石俊涛.中文文本分类中卡方特征提取和对 TF-IDF 权重改进.西华大学.2017

[4] 马治涛.文本分类停用词处理和特征选择技术研究.西安电子科技大学.2014

[5] 李原.中文文本分类中分词和特征选择方法研究.吉林大学.2011

[6] 张昭，艾中良.一种基于用户关联分析的热点话题识别算法.华北计算技术研究所总体部.2014

[7] 王欣杰，李海峰，马琳等.基于 F-score 的大数据公共空间模式选择方法.哈尔滨工业大学.2014

[8] 黄韬，刘胜辉，谭艳娜.基于 K-means 聚类算法的研究.哈尔滨理工大学.2011