“智慧政务”中的文本挖掘应用

摘 要

“智慧政务”是近些年来传统政府部门与大数据、互联网等相互融合产生的现代信息技术，即整合信息服务资源，通过应用各种平台，提高政府服务和管理的质量。在过去，政府是按管理职责分别设定的，存在严重的部门壁垒，极大地阻碍了部门之间和部门内部的运营效率。随着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。同时，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。

在本文中，基于大数据与机器学习的角度出发，利用传统模型与深度学习模型对于“智慧政务”中的一些关键文本数据挖掘问题进行处理分析。

针对问题一，构造以朴素贝叶斯分类器、SVM 支出向量机分类器等传统模型为基础的多分类模型。由于留言内容在所给数据中体现为留言主题和留言详情两部分，在此，我们对于留言主题及留言详情均进行分类处理，并结合相应指标综合对比分析。首先将数据的分类标签由文本型转为数值型，再进行数据预处理，即对于非结构化数据进行去空去重、中文文本分词处理、停用词设置以及 TF-IDF 权重法向量化处理后构造文本矩阵，由于这种方法具有高维度，高稀疏度以及同义词影响的缺点，因此，本文进一步利用基于潜在语义（LSA）分析的奇异值分解算法（SVD）对词文本矩阵进行空间语义降维。最后，通过分层抽样划分训练集和测试集后在多种分类模型下对留言主题及留言详情根据准确率、召回率以及 F-Score 对分类方法进行横向以及纵向综合评价衡量。

针对问题二，构造以文本相似度为基础的热度指标评价体系模型。首先对于数据进行一个基本的预处理，然后在此之上通过文本相似度来进行文本聚类，再通过所定义的热度指标评价体系模型以时间跨度和文本数量为一级指标，点赞数和反对数为二级指标来对每个文本进行一个基本的界定，然后通过定义的热度指数对每个文本的热度计算，最后综合到所处的类别中并进行排名。

针对问题三，本文从相关性、及时性、完整性以及可解释性四个方面来对答复意见的质量给出了一套评价方案。首先构造以基于答复意见的相关性与及时性出发的特征值，并利用多种聚类分析模型对特征值进行聚类分析，聚类为合适的簇后，分析不同簇的共性，最后来对答复意见的完整性和可解释性得到一个评价结果，再在聚类结果之上利用分类模型来对未来答复意见的质量作为最终评价方案。

关键词：TF-IDF；奇异值分解；多分类模型；文本相似度；聚类分析

Abstract

"Smart government" is a modern information technology produced by the integration of traditional government departments, big data, Internet, etc. in recent years, that is to integrate information service resources and improve the quality of government services and management through the application of various platforms. In the past, the government was set up according to the management responsibility, there were serious Department barriers, which greatly hindered the operation efficiency between departments and within departments. With the network politics platform gradually becoming an important channel for the government to understand the public opinion, gather people's wisdom and gather people's spirit, the amount of text data related to all kinds of social situation and public opinion is increasing, which has brought great challenges to the work of relevant departments that mainly rely on manual to divide messages and sort out hot spots in the past. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of smart government system based on natural language processing technology has become a new trend of social governance innovation and development, which has a great role in promoting the management level and efficiency of the government.

In this paper, from the perspective of big data and machine learning, traditional model and deep learning model are used to process and analyze some key text data mining problems in "smart government".

To solve the first problem, a multi classification model based on the traditional models such as naive Bayes classifier and SVM expenditure vector machine classifier is constructed. As the message content is embodied in two parts of message subject and message details in the given data, here, we classify the message subject and message details, and make a comprehensive comparative analysis combined with the corresponding indicators. Firstly, the classification label of data is transformed from text type to numerical type, and then the data preprocessing is carried out, that is, the unstructured data is processed by de empting and de duplicating, Chinese text segmentation, stop words setting and TF-IDF weight normal quantization to construct the text matrix. Because this method has the disadvantages of high dimension, high sparsity and synonym

influence, this paper further uses Singular value decomposition (SVD) algorithm based on latent semantics (LSA) analysis is used to reduce the dimension of word text matrix. Finally, after dividing the training set and test set by stratified sampling, the topic and details of the message are evaluated and measured according to the accuracy, recall rate and F-score.

In order to solve the second problem, a thermal index evaluation system model based on text similarity is constructed. Firstly, the data is preprocessed, and then the text is clustered by text similarity. Then, each text is defined by the defined heat index evaluation system model, which takes time span and text quantity as the first level index, and likes and dislikes as the second level indexes. Then, each text is defined by the defined heat index Heat calculation, finally integrated into the category and ranking.

In view of the third question, this paper gives a set of evaluation scheme for the quality of the reply from four aspects: relevance, timeliness, integrity and interpretability. Firstly, construct the eigenvalues based on the relevance and timeliness of the reply, and use a variety of clustering analysis models to cluster the eigenvalues. After clustering into appropriate clusters, analyze the commonness of different clusters, and finally get an evaluation result of the integrity and interpretability of the reply. Based on the clustering results, the classification model is used to evaluate the quality of future replies as the final evaluation scheme.

Keywords: TF-IDF; Singular value decomposition; Multi classification model; Text similarity；Cluster analysis

目　录

1. 问题重述

一、问题的背景：

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。同时，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。

二、题目的所给信息及参数：

附件给出了收集自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。

三、所要解决的问题：

1、群众留言分类：在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。建立关于留言内容的一级标签分类模型。

通常使用 F-Score 对分类方法进行评价：

$$F = \frac{1}{n}\sum_{i=1}^{n}\frac{2R_iP_i}{P_{i+R_i}}$$

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

2、热点问题挖掘：某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。请根据所给数据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，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

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，并尝试实现。

2. 问题分析

2.1 问题一

问题一实际上是一个文本分类(多分类)问题。文本分类是自然语言处理中普遍具有挑战性的任务，其主要原因是文字的符号不具有像图片和语音能够进行量化的比较的性质。就本问解题思路：以附件 2 中的“留言主题”和“留言详情”为分类依据，以“一级分类”作为分类标签进行分类，其中“留言主题”和“留言详情”均是文本，可以从中选取一个或者全部作为分类依据。通过文本表示和分类模型两方面进行如下分析：

第一，文本预处理。文本预处理是文本分类前的准备工作，主要包括分词、停用词等，本文使用 python 下的 Jieba 分词工具进行分词，去停用词是为了避免一些无意义词语的加入会降低分类的准确度，同时也要对占位符进行处理，例如：附件 2 的留言详情中带有的“n/、t/”等。

第二，特征提取。特征提取是将原有特征空间进行某种形式变换，以得到新的特征。理论上, 特征越多越能提供比较好的识别能力, 但 对于有限的训练数据, 过多的特征会导致分类器对训练数据的“过适应”问题， 尤其是那些与类别不相关的特征和冗余特征, 会导致参数估计的准确率下降, 进而影响分类器的性能。在此利用文本表示模型 TF-IDF 将文本转换为计算机可理解和计算的形式——向量。

第三，分类器设计。分类器设计是文本分类的核心部件，通过相应的分类算法将未知文本划分到已知类别的文本是分类器的主要功能。本文采用朴素贝叶斯分类器以及 SVM 支持向量机分类模型。同时采取分类器后要对分类器分类结果的好坏做一个大体的评判，这时需要以精准率以及召回率等来反映模型的分类效果，同时采用综合了查准率以及查全率的评判标准 F –Score。

2.2 问题二

问题二类似于一个文本回归问题。其一，对于问题识别，如何从众多留言中识别出相似的留言；其二，对于问题归类，把特定地点或人群的数据归并，即把相似的留言归为同一问题；其三，热力评价，热度评价指标的定义和计算方法。

根据留言的文本信息，使用文本相似度等，设定一个阈值，将多个文本判断为同一个文本，然后依据一定的指标设定来判断哪些留言为热点问题，如时间跨度，相似问题频数等，并分配出热度指数值。最后依据热度指数值排名按照题意要求来输出结果。

2.3 问题三

问题三可以转换为聚类问题进行处理，我们首先我们先对题中提出的三方面进行理解，相关性:答复意见的内容是否与问题相关；完整性:是否满足某种规范；可解释性:答复意见中内容的相关解释能否解决这个问题，同时还可依据回复和提问时间差值来判断答复意见内容的及时性。

根据提供的数据，构建评价完整的评价指标。例如:回复时间和提问时间的差值，回复和提问的文本相似度，等等。以这些评价指标为特征，使用 K-Means 等聚类方法进行聚类。聚类为合适的簇后，分析不同簇的共性，最终得到评价结果，再在聚类基础之上根据分类模型来作为未来答复意见的评价方案。

3. 符号说明

符号	定义说明
TF	词频
IDF	逆文档频率
TF-IDF	TF 与 IDF 的乘积
X	待分类项
a_i	X 的某一特征属性
y_k	标签项
A_{m*n}	m×n 的矩阵
U_{m*m}	m×m 的矩阵

符号	定义说明
V	n×n 的矩阵
I	单位矩阵
Σ	除对角线外其它全为 0 的矩阵
w_i	文本向量
S	相似矩阵
W	邻接矩阵
D	度量矩阵
L	拉普拉斯矩阵
F	特征矩阵

4. 模型建立与求解

4.1 问题一

4.1.1 模型建立

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系对留言进行分类，以便后续将群众留言分派至对应的职能部门处理。根据已有的数据来构建分类模型是在传统的文本挖掘基础上的延伸。

文本挖掘指的是从海量的文本数据当中提取人们无法预知但是能够理解的且最终能够为自己可用的信息的一个过程，利用这些所提取到的信息，可以为自己将来的行动做一定的参考。文档本身属于一种非结构化的数据，这种文本的形式随机且机器很难理解它的准确定义，所以对于以文本形式所存储的数据信息必须提取其特征，从所提取的特征当中分析得出我们所需要的信息，根据这些信息搭建相应的模型。

我们首先需要对于原文本数据进行基本的预处理，即相应的去空去重，同时在 python 中导入数据发现留言详情文本中存在大量转义符，需要处理去除。（部分数据如图 1 所示）

	留言编号	留言用户	留言主题	留言时间	留言详情	一级分类
0	24	A00074011	A市西湖建筑集团占道施工有安全隐患	2020/1/6 12:09:38	\n\t\t\t\t\t\n\t\t\t\t\tA3区大道西行便道，未管所路口至加油站路段，...	1
1	37	A000107866	A市在水一方大厦人为烂尾多年，安全隐患严重	2020/1/4 11:17:46	\n\t\t\t\t\t\n\t\t\t\t\t位于书院路主干道的在水一方大厦一楼至四楼人为...	1
2	3742	A00013884	A3区杜鹃文苑小区外的非法汽车检测站要开业了！	2018/12/26 10:13:37	\n\t\t\t\t\t\n\t\t\t\t\tA市政府、市交管支队、市安监局、市环保局、A...	1
3	5315	A0009647	民工在A8区明发国际工地受伤，工地方拒绝支付医疗费	2018/7/17 7:13:58	\n\t\t\t\t\t\n\t\t\t\t\t胡书记，您好，感谢您百忙之中查看这份留言。我...	1
4	100026	A00047862	K8县丁字街的商户乱摆摊	2017/10/14 14:25:14	\n\t\t\t\t\t\n\t\t\t\t\t K8县丁字街的商户乱摆摊，前段时间丁字街...	1

图 1 数据导入显示

然后在此基础上，我们利用了 jieba 中文分词将留言主题与留言详情进行了分词处理。（部分分词内容如图 2 所示）

图 2 词云

同时还利用停用词表对于文本信息中包含的很多功能词进行处理，这类词语对于文本分类并没有什么实际性的含义，如“的”、“啊”、“等”，功能词主要用来描述名词和表达概念，文本中充斥大量的停用词不仅会降低后期处理的正确率，且在文本进行向量化时，会产生高维稀疏矩阵，严重影响处理的效率。因此，需要对这些词进行过滤，保留主要词语作为特征词。最后经过 TF-IDF 向量化和奇异值分解（SVD）处理，再利用分层抽 样按一定比例对于各层样本划分为训练集和测试集，最后通过朴素贝叶斯分类器和 SVM 分类器来对数据进行训练测试，并以查准率和 F-Score 来对不同分类器，以及选取的不同的文本数据进行纵向和横向评价。

（1） TF-IDF

在对留言主题和留言详情信息分词和停用词处理后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法，把描述信息转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重。

词频（TF）=某个词在文本中出现的次数

考虑到文本有长短之分，为了便于不同文本的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即：

$$\text{词频（TF）} = \frac{\textit{某个词在文本中出现的次数}}{\textit{文本的总词数}}$$

或者

$$\text{词频（TF）} = \frac{\textit{某个词在文本中出现的次数}}{\textit{该文本出现次数最多的词的出现次数}}$$

第二步，计算 IDF 权重，即逆文档频率，需要建立一个语料库，用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$\text{逆文档频率（IDF）} = \log\left(\frac{\textit{语料库的文本总数}}{\textit{包含该词的文本数}+1}\right)$$

第三步，计算 TF-IDF 值。

$$\text{TF-IDF} = \text{词频（TF）} \times \text{逆文档频率（IDF）}$$

（2） 潜在语义分析之奇异值分解

浅层语义分析（LSA）是一种自然语言处理中用到的方法，其通过“矢量语义空间”来提取文档与词中的“概念”，进而分析文档与词之间的关系。LSA 的基本假设是如果两个词多次出现在同一文档中，则这两个词在语义上具有相似性。LSA 使用大量的文本上构建一个矩阵，这个矩阵的一行代表一个词，一列代表一个文档，矩阵元素代表该词在该文档中出现的次数，然后再此矩阵上使用奇异值分解（SVD）来保留列信息的情况下减少矩阵行数。

为处理 TF-IDF 向量化处理文本所带来的文本矩阵过大、文本矩阵过于稀疏等确定，所以需要用到 LSA 对文本矩阵降维处理，通常采用的一种方法为奇异值分解（SVD）。

现假设存在矩阵 A 是一个 m×n 的矩阵，存在矩阵 X 的一个分解，即矩阵 A 可分解成正交矩阵 U 和 V，和对角矩阵Σ的乘积。其中 U 是一个 m×m 的矩阵，Σ是一个 m×n 的矩阵，除了主对角线上的元素以外都是 0，主对角线上的每个元素都称为奇异值，V 是一个 n×n 的矩阵。U 和 V 都是酉矩阵，即满足:

$$U^T U = I \qquad V^T V = I$$

这种分解就叫做奇异值分解（SVD）即：

$$A = U\Sigma V^T$$

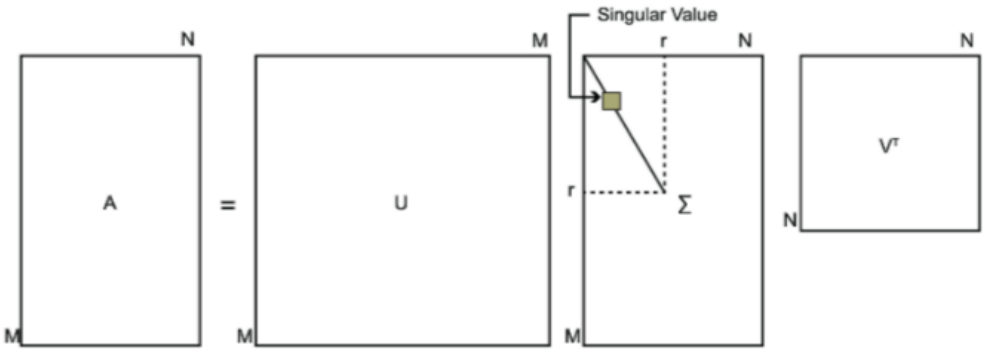

图 3 SVD 定义图

（3） 朴素贝叶斯分类模型

实际上朴素贝叶斯原理就是求解后验概率，而且朴素贝叶斯方法的一个前提假设是对于给定的训练集，各文本属性之间是相互独立。

基本步骤：

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、 同时还存在标签集合 C={y1,y2,...,yn}。

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

4、 若 P（$y_k|x$）=max{ P（$y_1|x$），P（$y_2|x$），……，P（$y_n|x$）}，则表示 x∈y_k，即 x 属于第 k 类。

分类流程：

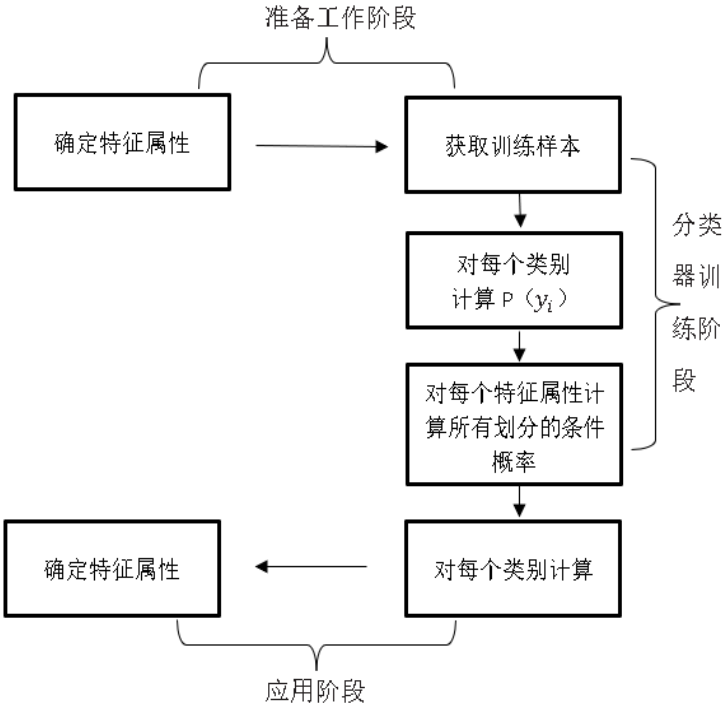

图 4 朴素贝叶斯分类流程图

可以看到，整个朴素贝叶斯分类分为三个阶段：

第一阶段——准备工作阶段，这个阶段的任务是为朴素贝叶斯分类做必要的准备，主要工作是根据具体情况确定特征属性，并对每个特征属性进行适当划分，然后由人工对一部分待分类项进行分类，形成训练样本集合。这一阶段的输入是所有待分类数据，输出是特征属性和训练样本。这一阶段是整个朴素贝叶斯分类中唯一需要人工完成的阶段，其质量对整个过程将有重要影响，分类器的质量很大程度上由特征属性、特征属性划分及训练样本质量决定。

第二阶段——分类器训练阶段，这个阶段的任务就是生成分类器，主要工作是计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，并将结果记录。其输入是特征属性和训练样本，输出是分类器。这一阶段是机械性阶段，根据前面讨论的公式可以由程序自动计算完成。

第三阶段——应用阶段。这个阶段的任务是使用分类器对待分类项进行分类，其输入是分类器和待分类项，输出是待分类项与类别的映射关系。这一阶段也是机械性阶段，由程序完成。

在没有其他信息的情况下，我们会选择条件概率最大的类别，这就是朴素贝叶斯的思想基础：对于给出的待分类项，在此项出现在的条件下各个类别的出现概率，哪个最大，就认为此待分类项属于哪个类别。从朴素贝叶斯分类的原理来看，对于一个未知类别的样本 X,可以先分别计算出 X 属于每一个类别的概率，选择其中概率最大的类别作为其分类。

（4） SVM 支持向量机分类模型

SVM 是一种基于结构风险最小化的二分类器，给定两类训练样本，$\{(x_i, y_i)\}_{i=1}^{l}$，$x_i \in R^m$，y_i属于$\{\pm 1\}$，SVM 首先通过非线性映射 φ：$R^m \rightarrow R^n$将数据投影至高维可分空间，然后构造最大间隔分类超平面 w * φ（x）+ b =0 对高维空间进行划分，判断被测样本属于正类或负类。其中 w 为法向量，决定了超平面的方向，b 为位移量，决定了超平面与原点的距离。由于该文本问题为多分类问题，因此需先对 SVM 进行多分类扩展。

目前扩展方法大多采用各类树状结构将多组 SVM 分类器有序排列并形成决策路径，结合 SVM 的实际输出，给出最终判断结果。且基于 python 中指定的包可实现。

4.1.2 总体流程

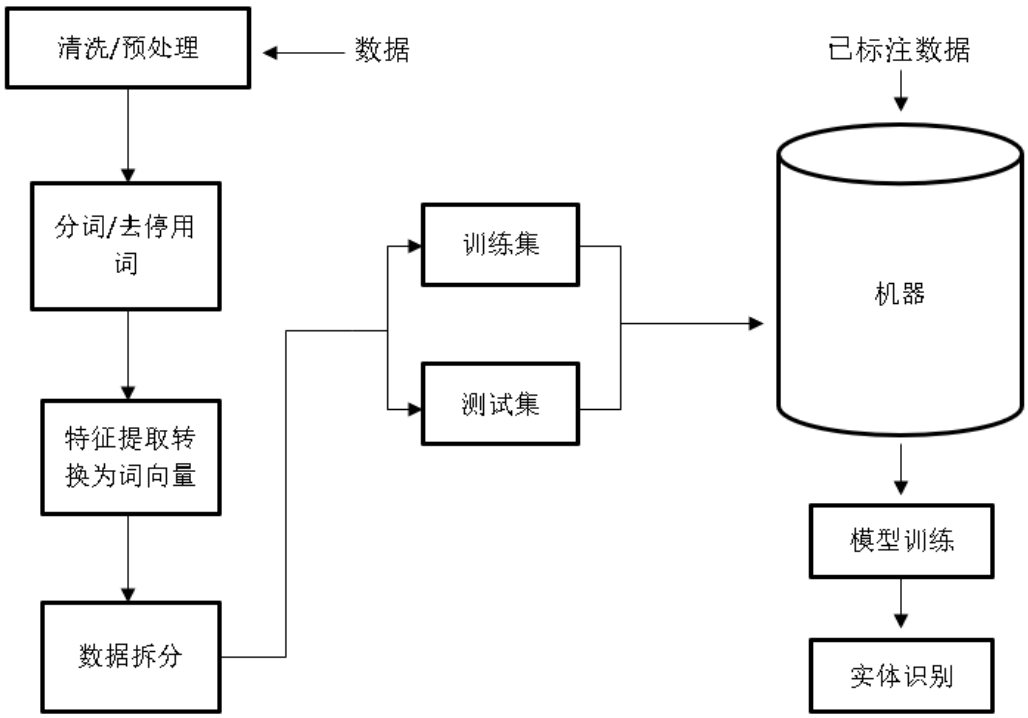

图 5 分类总流程图

4.1.3 模型求解

借用 python 工具完成对于留言主题和留言详情的分类，并通过不同文本下以及不同模型下对分类模型评价分别进行横向以及纵向比较，经运行代码得出结果如下：

横向比较：

表 1 朴素贝叶斯模型下留言详情和留言主题相应指标

	留言详情	留言主题
准确率	0.6233	0.6204
召回率	0.6233	0.6204
F-Score	0.5815	0.5796

表 2 SVM 支持向量机模型下留言详情和留言主题相应指标

	留言详情	留言主题
准确率	0.6233	0.5694
召回率	0.6233	0.5204
F-Score	0.5815	0.4705

由表 1 和表 2 可看出，朴素贝叶斯分类模型，在整体看来，不论是针对于留言详情进行分类处理，亦或是对留言主题进行分类处理，在准确率、召回率以及 F-Score 上均好于 SVM 支持向量机分类模型。

纵向比较：

表 3 留言主题在不同模型下的相应指标

	朴素贝叶斯	SVM 支持向量机
准确率	0.6204	0.5694
召回率	0.6204	0.5204
F-Score	0.5796	0.4705

表 4 留言详情在不同模型下的相应指标

	朴素贝叶斯	SVM 支持向量机
准确率	0.6233	0.6233
召回率	0.6233	0.6233
F-Score	0.5815	0.5815

由表 3 和表 4 可看出，在相同分类模型处理下，通过对于留言主题进行分类处理，对于分类模型的评价在准确率、召回率以及 F-Score 上均好于留言详情。

综上，我们针对问题一主要采取基于朴素贝叶斯分类器对于留言主题进行分类处理，因其能取得一个更好的结果。

4.2 问题二

4.2.1 模型建立

某一时段内群众集中反映的某一问题可称为热点问题。在众多文本数据中发现热点问题，就要对于文本数据的其他相关信息指标来进行评价，如时间跨度，相似类别文本频数等。在这基础之上，我们首先需要对于相似文本进行划分，即

利用文本相似度来将文本在指定的阈值上归为一类。

（1） 文本相似度

通过文本模型来表示提取文本的特征项，以向量的方式表示文档的内容，将文档的相似度计算简化为向量的相似度计算，它主要是用某种度量方法来计算两个文本之间的相似性，该方法也可以用于计算一个文本与某个类别的相似性和用于热点话题的发现。针对文本间相似度计算的方法主要有 Dice 系数法、Jaccard 系数法和余弦相似度等。

文本采取余弦相似度的方法来计算文本相似度，公式如下：

设任意两个文本的向量分别为

$$w_1=（d_1,d_2,d_3,\dots,d_n）$$

$$w_2=（q_1,q_2,q_3,\dots,q_n）$$

其中 n 表示维数。余弦相似度计算公式为：

$$\text{SIM}（w_1，w_2）=\frac{w_1\cdot w_2}{|w_1|\times|w_2|}=\frac{\sum_{i=1}^{n}(d_i\times q_i)}{\sqrt{\sum_{i=1}^{n}{d_i}^2\times\sum_{i=1}^{n}{q_i}^2}}$$

SIM（w_1，w_2）表示文本w_1与w_2之间的相似度，d_i和q_i分别表示文本w_1和w_2的第 i 个特征项的权限值，n 表示文本的特征项数，SIM 值越大表示两个文本越相似，反之两个文本越不相似。

（2） 热度指标评价体系

将不同文本按照文本相似度来归类后，从数据中首先看来，第一明显特征便是哪一类文本数量频数最高，则哪一类相对于更偏向于热点问题，这是直接反映给我们的。同时，在所给数据中，不同的文本信息均有留言时间信息标注，那么在归类后的文本集群存在一个时间跨度，进而我们可得知若两个文本集群在文本数量相同时，那么再来比较哪个文本集群的时间跨度更短，时间跨度更短则说明在较短的时间内达到了一个相对数量更高的水准，则更偏向于热点问题。简而言之，若是单一将文本数量来当做一级指标对各个文本集群评判，则不免带有片面因素。因此，将文本数量与时间跨度均作为一级指标来衡量更符合逻辑。

再基于数据本身看来，可见对于每个文本数据有点赞数和反对数两个直观的评价指标。点赞数通俗来讲是各个用户对于此文本信息从喜爱或认可的角度基于

一个数字的客观反馈，同样而言，反对数则是各个用户对于此文本信息从厌恶或对立的角度基于一个数字的客观反馈。二者在文字上而言是对立关系，可在对于热点问题发现中我们可看到，不论是点赞数还是反对数，均能使得文本的本身信息更加丰富，更可能被大众所观测，所以点赞数与反对数均能使得文本更偏向于热点问题，鉴于点赞数与反对数在所给数据中反馈情况，与实际生活中我们所在各信息公告平台上看到的点赞数与反对数相比，数据中反映出的点赞数以及反对数不太具备反馈热度的信息，因此将其作为二级指标搭建在热度指标评价体系当中，仅作为比较参考。

此处热度指数以文本集群中的文本数量与其时间跨度的比值来衡量。即：

$$\text{热度指数} = \frac{\textit{文本数量（单位：个）}}{\textit{时间跨度（单位：年）}}$$

表 5 文本热度综合评价指标体系

	一级指标	二级指标	指标内涵
文本热度综合评价指标体系	文本数量		某一类文本集群的计数（单位：个）
	时间跨度		某一类文本集群中，距今最早年份与最晚年份的差值（单位：年）
	传受众热度特征影响力	点赞数	在某个时段对某一文本持有认可态度的数目
		反对数	在某个时段对某一文本持有反对态度的数目

考虑到可能存在时间极短，文本数量也不多，但热度指数较高的情况，在此，我们对于相同或近似热度指数的，优先考虑文本数量较多的一方，而对于这种因为时间跨度极短而导致异常大的值的主题不予考虑，其也并不符合现实生活中热点话题的反馈标准。

4.2.2 总体流程

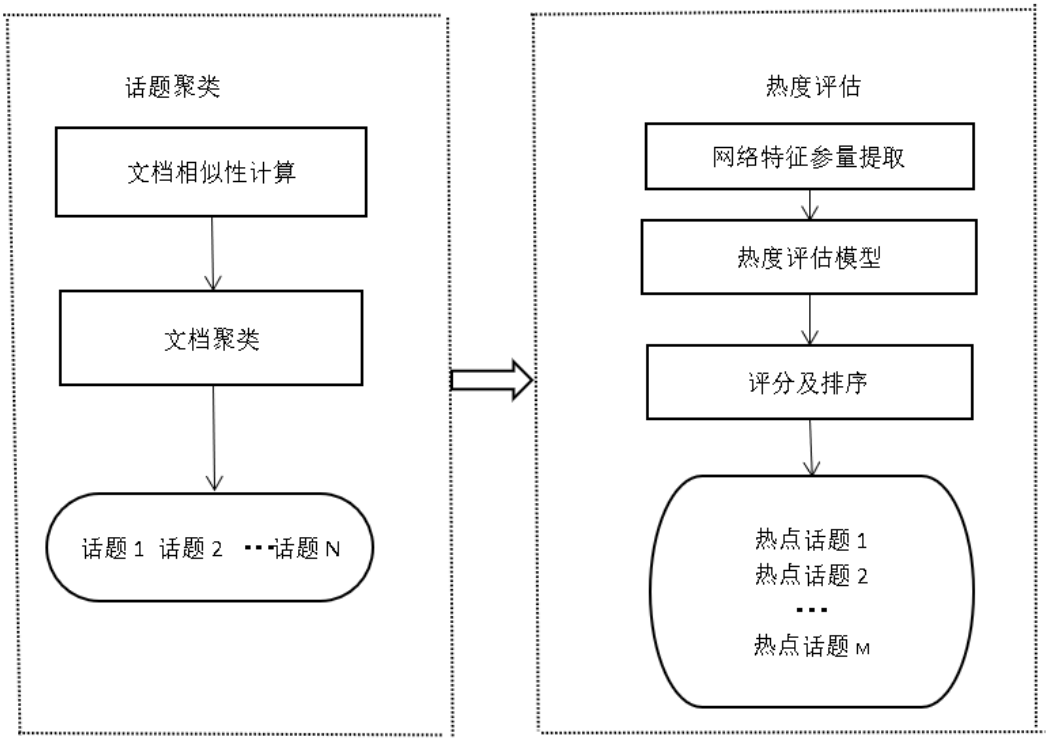

图 6 热点话题挖掘流程图

依据附件 3 所给出的数据，在此我们假定认为留言主题能够代表留言者所反映的一个客观事实主题，所以在此选定用留言主题进行文档相似度的计算。再通过循环查找出每一行数据两两之间相互的相似度大小，进而在此基础上，以文档相似度为 0.6 来对所反映出来的数据指标进行筛选，最后进行综合比对，来将属于同一主题的话题聚类并通过热度评估模型计算出相应的热度指标，最后进行排序求解。

4.2.3 模型求解

表 6 热点问题表（部分）

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	131.0053	2019/07/21 至 2019/09/25	A 市 A5 区魅力之城小区	街餐饮店油烟扰民

2	2	121.7378	2019/11/2 至 2020/1/25	A 市丽发新城小区	小区搅拌车噪音及环境污染

表 7 热点问题留言明细表（部分）

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	236798	A00039089	A5 区劳动东路魅力之城小区油烟扰民	2019/07/28 12:49:18	尊敬的政府：A5 区劳动……	0	4
1	350108	A0283523	A5 区劳动东路魅力之城……	2019/08/01 16:20:02	局长：你好，A5 区劳动东路……	6	0
…	…	…	…	…	…	…	…
2	267050	A909227	噪音、灰尘污染的 A2 区丽……	2019/11/02 10:18:00	A2 区丽发新城附近修……	0	0
2	264944	A0004260	A2 区丽发新城附近修建……	2019/11/02 14:23:11	A 市 A2 区丽发新城小区附近，作为……	0	0
…	…	…	…	…	…	…	…

通过求解我们得出以上热点问题表和热点问题留言明细表，其结果我们发现在排名前五的热点话题中，时间跨度一般基于一年左右，同时文本数量占比在整个数据各个主题中也具有一定的优势，结合现实生活看来，我们熟知的微博等主流社交客户端的热点话题均是在较短的时间内同时也拥有较多的文本数量，并不会存在较短的时间的文本数量较少，因此我们对于此类异常值的处理也具有一定的意义，依据此指标我们以“越短越多”作为最后的指标依据进行衡量热点话题是具有一定的指导性，其结果也符合我们的预期。

4.3 问题三

4.3.1 模型建立

根据附件所给留言相关部门进行了答复。所给数据中包含用户留言主题、详情和时间，以及相关部门进行的答复意见及时间。有效地识别答复内容的相关性、可解释性和及时性等对于用户以及相关部门双方均起到一定的积极意义。

相关性简单而言就是部门答复意见和用户留言内容二者的相互关联程度。在此，我们在基于留言主题与答复意见之上考虑使用文本相似度来对于答复意见质量之一的相关性进行评价。

同时基于所给数据中答复意见和提问主题的各自时间，可认为二者的时间差直观地呈现出答复意见质量之一的及时性进行评价。

在依据已经构建好的相关性指标和及时性指标来重新构建出一个基于数据本身特征所反映出来的特征指标。在此，特征指标定义如下：

相关性指数 = 文本相似度指数

及时性指数 = 答复意见时间 — 留言时间

基于上述流程，每条数据均有其特征指标来侧面反映出其特征程度。基于特征评价指数，我们采取 K-Means 聚类、高斯混合聚类和谱聚类三种聚类模型来对其进行聚类分析，聚类为合适的簇后，分析不同簇的共性，再来对相关部门答复意见的完整性和解释性进行综合性评判。

综上，我们构建了一个以充分利用数据本身特征，并通过其特征指标来进行聚类分析的有关答复意见的相关性、及时性、完整性和可解释性四个方面的质量评价方案。

再通过此现有的质量评价基础上，对于未来的答复意见构建一个以分类模型进行预测的一个评价模型。

（1）K-Means 聚类

K-Means 算法是硬聚类算法，是典型的基于原型的目标函数聚类方法的代表，它是数据点到原型的某种距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。K-Means 算法以欧式距离作为相似度测度，它是求对应某一初始聚类中心向量最优分类，使得评价指标最小，算法采用误差平方和准则函数作为聚类准则函数。

算法流程：

1）计算欧式距离

2）随机选取 k 个初始聚类中心点

3）更新簇的中心点

4）迭代，直到收敛

通常停止迭代的条件其一是达到指定的迭代次数，再者就是质心不再发生明显变化，即可认为收敛。

（2）高斯混合聚类

高斯混合聚类通过选择成分最大化后验概率来完成聚类。与 K-Means 聚类相似，高斯混合模型也使用迭代算法计算，最终收敛到局部最优。使用高斯混合模型的聚类属于软聚类方法，各点的后验概率提示了各数据点属于各个类的可能性。

算法流程：

1）初始化高斯混合分布的模型参数

2）计算x_j由各混合成分生成的后验概率

3）计算新的模型参数

4）按照新的模型参数重复 2-3 步，直到满足条件

5）样本来自哪个分模型的概率大就划入哪个分模型的簇中，最终就得到了 K 个聚类

（3）谱聚类

谱聚类的主要思想是把所有的数据看做空间中的点，这些点之间可以用边连接起来。距离较远的两个点之间的边权重值较低，而距离较近的两个点之间的边权重值较高，通过对所有数据点组成的图进行切图，让切图后不同的子图间边权重和尽可能的低，而子图内的边权重和尽可能的高，从而达到聚类的目的。

谱聚类算法的流程：

输入：样本集 D =（$x_1, x_2, x_3, \dots, x_n$），相似矩阵的生成方式，降维后的维 k1，聚类方法，聚类后的维度 k2

输出：簇划分 C =（$c_1, c_2, c_3, \dots, c_{k2}$）

1）根据输入的相似矩阵的生成方式构建样本的相似矩阵 S

2）根据相似矩阵 S 构建邻接矩阵 W，构建度量矩阵 D

3）计算出拉普拉斯矩阵 L

4）构建标准化后的拉普拉斯矩阵$D^{-1/2}LD^{1/2}$

5）计算$D^{-1/2}LD^{1/2}$最小的 k1 个特征值所各自对应的特征向量 f

6) 将各自对应的特征向量 f 组成的矩阵按行标准化，最终组成 n×k1 维的特征矩阵 F

7）对 F 中的每一行作为一个 k1 维的样本，共 n 个样本，用输入的聚类方法进行聚类，聚类维数为 k2

8）得到簇划分 C =（$c_1, c_2, c_3, \ldots, c_{k2}$）

4.3.2 总体流程

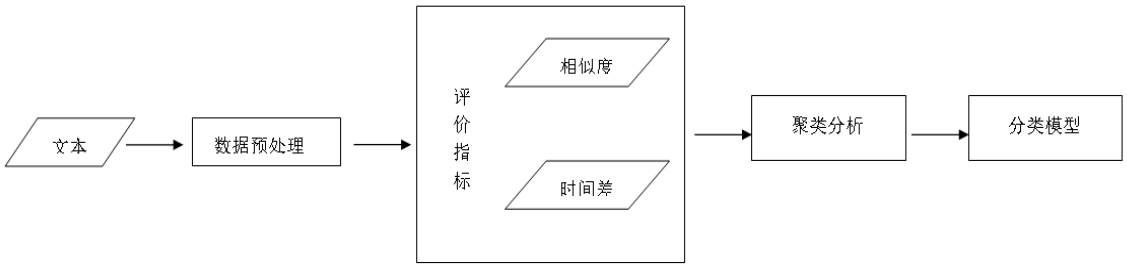

图 7 答复意见质量评价流程图

4.3.3 模型求解

因留言主题能反映出留言详情的一个大体特征，我们在此选用以每一条数据的留言主题与答复意见的相似度以及答复意见与留言的时间差作为两个评价指标来进行聚类分析。（部分评价指标计算结果如下表）

表 8 评价指标结果（部分）

序号	相似度	时间差
1	0.8452	15.22551
2	0.9129	14.73993
3	0.7276	14.75637
4	0.6301	14.77931
5	0.5909	15.70013
6	0.9623	31.05889

（1） K-Means 聚类

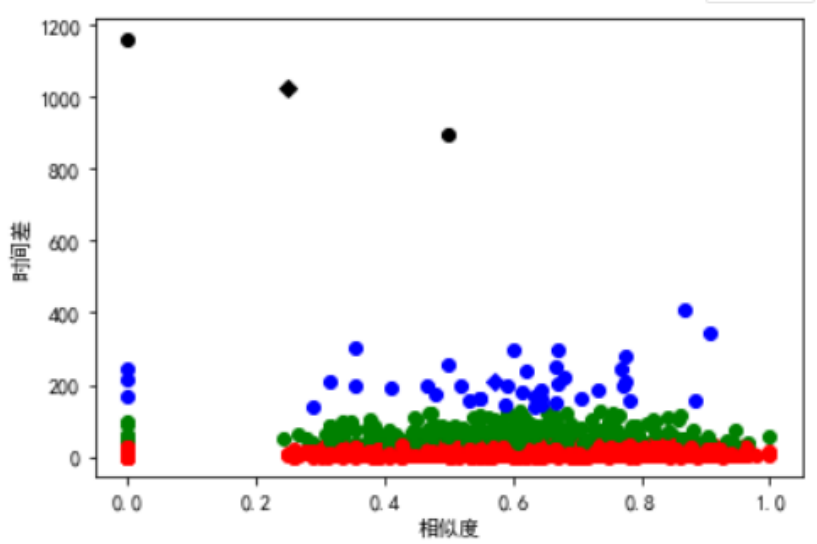

图 8 K-Means 聚类

由 K-Means 聚类结果我们可以看出，数据被划分为 4 大类，而且红色、绿色以及蓝色这三类的中心点均位于偏右下方，同时整体数据普遍集中在右下方，也就是在较短的时间内回复且答复意见与留言主题的相似度较高。并且图像也直观呈现出红色及绿色部分占比较高，答复意见整体质量中等偏上，且数据分割较为清楚，我们在此将红色这一类别判断为这组数据中质量偏高的一类，但我们在此处红色仍有部分答复时间间隔虽短，且答复意见相似度并不高的情况，在这里我们就不能够清楚地界定不同类别之间的划分关系，因此，我们认为 K-Means 聚类没有很好地对于此部分数据进行划分聚类。

（2） 高斯混合聚类

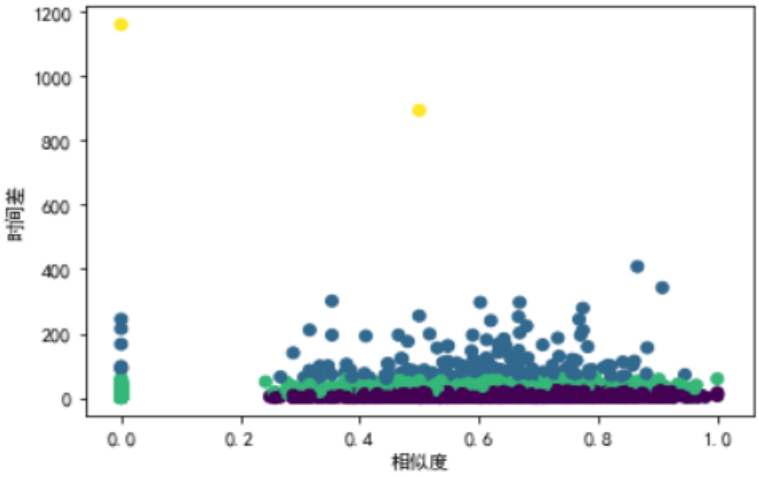

图 9 高斯混合聚类

在高斯混合聚类中我们也看到数据在整体上被划分为 4 大类。相比于 K-Means 聚类，我们认为在高斯混合聚类中，对于黑色这一类，有着更好的划分。黑色这一类在这组数据中心，明显属于高质量的那一部分，而且对于时间差较短，但相似度很低甚至为 0 的数据，它并没有将其划分黑色这一类，而是以一种明显的层级递进给了绿色和蓝色这一类。在高斯混合聚类中，我们可发现，这 4 个类别部分，单以相关性和及时性来看，质量的高低可看作：黑色>绿色>蓝色>黄色，相比于 K-Means 我们认为高斯混合聚类有着更明显的划分特征呈现特性。

（3） 谱聚类

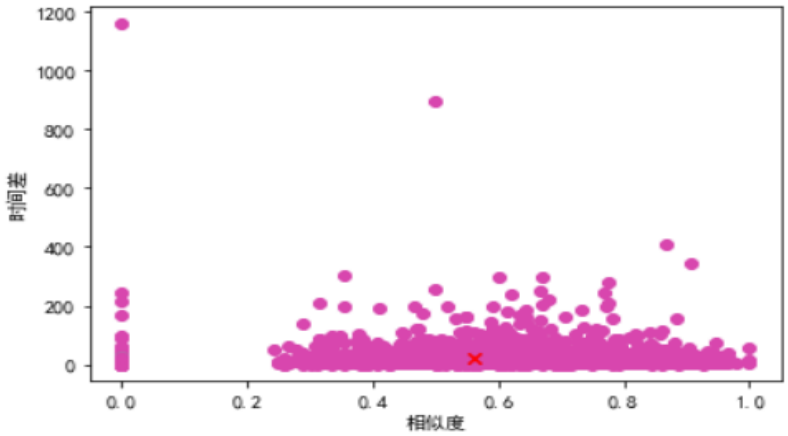

图 10 谱聚类

基于谱聚类我们可以看出，它并没有很好地满足我们的需求，它将所有数据归于一类，且经了解谱聚类在大量数据进行聚类划分的情况下，目前仍然存在较大的改进空间。

由此，我们最终选用高斯混合聚类来对整体的答复意见从相关性、及时性、完整性以及可解释性四个方面进行评判：

答复意见在时间跨度上的范围极广，最快的答复有不足一个小时就答复给询问者，而最慢的答复也有以年为单位来计算的，可由高斯混合聚类得出的散点图我们发现，黑色部分时间分布在 15 天左右，绿色部分时间分布在 40 天左右，而蓝色与黄色部分较之分布在 100 天以上，同时大量的答复意见普遍位于黑色与绿色部分，时间普遍少于 40 天，且这一时间基于当代人而言，因体谅到网络工作效果任务的繁重，此时间段内是处于能够理解接受的范围之内。因而答复意见在答复的及时性上较为人意。

在对于答复意见的相关性上采取文本相似度来进行评判，经高斯混合聚类后

得出的散点图直观显示出，大量的答复意见与其对应的留言主题的相似度分布在0.6及其以上，少部分位于以下，但也仍存在极少数的相似度为0的情况，一般这种情况一种是留言者的留言主题不明确，甚至并无留言主题而导致，另一种情况则是，意见答复人员题不对意，或是以过于模糊或客观的说法去进行答复。就本文数据而言，很多情况则是在进行答复时并未提及到有关留言的相应信息，仅以“我们会第一时间将XXX反映到XXX部门”等不确定的答复造成了相似度为0的情况。但从整体数据趋势上看来，答复意见的相关性是较好的。

针对于答复意见的完整性，答复意见的完整性不仅仅是对留言的答复，也体现在基本的回复礼仪这一方面，固定的用户称呼，来信悉知，感谢用语，回复时间等都是必不可少的。基于对本数据的了解，我们发现普遍的答复意见都有一些基本的固定格式，如“网友 XXX 您好！您的留言已收悉。XXXXXX。感谢您对我们工作的关心，监督与支持。XXX年XXX月XXX日”等。但也仍存在敷衍，以及没有基本的回复礼仪仅有答复内容等答复意见的存在。

且根据高斯混合聚类的聚类结果我们可以看出，以黑色和绿色部分的答复意见的完整性结合到数据看来，是明显好于蓝色及黄色部分的答复意见。所以基于本文数据而言，在大体上看来可认为时间差越短且相似度越高的答复意见的完整性就越好，又由于黑色和绿色部分占主要程度的数据占比，因此对于答复意见的完整性在整体上虽没有很完美，但也处于较好的程度。

对于答复意见的可解释性，我们发现此处的可解释性可分为三类。第一类是明确地对于留言内容进行了相对应的回复，能够使得留言者得到一个较明确的解决方向，可解释性强。第二类是不太清晰留言内容所提出问题的相应解决方案，以“已反映给某部门此类事件”作为答复标准，可解释性中，较模糊。第三类是答不对意，甚至连基本的回应都无，可解释性差。将此三类可解释性标准投射到高斯混合聚类得出的四类数据看来。并结合实际数据，每个类别中均存在三类可解释性，但不同的是，越是时间差越短以及文本相似度越高的，出现第一类可解释性的程度越高，反之出现第三类可解释性的程度越高。同样而言，黑色部分及绿色部分普遍分布较密，且更多，所以可认为对于答复意见的可解释性是处于中等偏上的。

综上，我们分别对应于答复意见的及时性、相关性、完整性以及可解释性作

了一个整体的评价，同时也发现对于聚类所分出的四个类别，不仅仅在数据的相关性以及及时性上的质量体现出：黑色>绿色>蓝色>黄色，同时在大多数数据中耳炎，它们也代表着数据的完整性和可解释性的部分特征，可对此进行说明。所以从四个方面看来，答复意见的评价总体而言是呈现中等偏上的趋势。

基于经过高斯混合聚类后的数据，也就是在很大程度上而言，黑色、绿色、蓝色和黄色在整体答复意见质量上是由高到低的，所以分类后并对其数据标签化，在此之上再通过 SVM 支持向量机分类模型来作为对未来答复意见的一个评价模型。

accuracy: 91.48%

	true Z	true T	true O	true D	class precision
pred. Z	561	57	0	0	90.78%
pred. T	12	167	2	0	92.27%
pred. O	0	0	45	1	97.83%
pred. D	0	0	0	0	0.00%
class recall	97.91%	74.55%	95.74%	0.00%	

图 11 SVM 支持向量机分类结果

在此处，Z 代表的高斯混合聚类中的黑色部分，T 代表绿色的部分，O 代表蓝色的部分，D 代表黄色部分。由上图可看出，基于聚类分析得出的对各类数据进行标签后再通过 SVM 支持向量机分类模型得出的准确率为 91.48%，而对于 D 类结果的召回率与预测的概率均为 0，可能是由于 D 类数据在整体中占比过少导致的（在整体数据中，D 类仅有两组数据），整体结果符合预期。

整体而言，首先通过聚类分析得出聚类标准及其结果再通过分类模型而构建的评价模型在一定程度上能够满足我们对于答复意见的评价。

5. 模型评价与推广

5.1 模型优缺点

1、朴素贝叶斯模型发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率。对大数量训练和查询时也具有较高的速度。即使使用超大规模的训练集，针对每个项目通常也只会有相对较少的特征数，并且对项目的训练和分类也仅仅是特征概率的数学运算而已。同时对缺失数据不太敏感，算法也比较简单，常用于文本分类而且结果解释容易理解。

2、文本热点指标评价体系模型充分利用了文本信息来构造文本向量以进行文本相似度评价，同时依据一级指标和二级指标将附件所给数据中的其它相关信息也结合利用来综合反映热点话题，具备一定的充分性和理论性。

3、答复意见质量评价模型利用基于文本自身属性，来构建新的评价指标，有理有据。同时，依据聚类分析模型，其结论简洁明了，直观。高斯混合聚类应用非常广泛，且算法运行较快。它还可以近似拟合任意形状的概率分布，并且能做到连续可微。高斯函数具有一定的对数线性特征，适合把非线性问题进行一定程度的线性化。

5.2 模型推广

1、准确率是最为常用的分类器判断标准，能体现出分类器的实际分类效果。本文以分类准确率作为性能评价指标，衡量分类器的性能。问题一中的分类模型不仅可以运用于本文的多分类问题，也可以运用于处理传统的二分类或多分类的文本问题当中来，甚至可以运用于自动应答，突发公共事件识别预警等，与文本处理相关的均可适用。

2、在问题二中所构造的热度指标评价体系模型中，是针对于现有的数据来进行一个热点问题的划分。可在此基础上基于现有数据对于未来一定时段内新增话题的热点属性判断进行一个概率预测，即在现有的已划分好热点话题数据的基础上，将其作为训练集，然后往数据挖掘分类模型中靠拢，对于新增的文本话题进行分类，但输出的结果为是否是热点话题的概率指标。但基于数据更新频率的加快，还需要增加一个固定时间点，不断地重复上述模型，以此构造成一个动态的热点话题挖掘模型。所以在此基础上，该模型可使用于多领域的文本热点话题或频率的相关研究。

3、聚类模型对大规模数据集有较高的可伸缩性和高效性，在许多领域发挥着越来越重要的作用，如模式识别、机器学习和图像处理等，同时了使聚类模型能更好达到对问题分析，拿 K-Means 模型来说，K-Means 算法与层次聚类结合、K-Means 算法与遗传结合、K-Means 算法与 SOM 结合等，它们的结合达到了对 K-Means 算法缺点得弥补和改进，可以使问题达到更优化，所以聚类模型在不断的研究进展下，在未来该模型的应用将更加广泛，满足不同的问题需求。

参考文献

[1]黄春梅,王松磊.基于词袋模型和 TF-IDF 的短文本分类研究[J].软件工程,2020,23(03):1-3.

[2]何伟. 基于朴素贝叶斯的文本分类算法研究[D].南京邮电大学,2018.

[3]谢志炜,冯鸿怀,许锐埼,李慧夫.电力基建施工问题文本分类研究[J].现代信息科技,2019,3(17):17-19.

[4]艾楚涵,姜迪,吴建德.基于主题模型和文本相似度计算的专利推荐研究[J].信息技术,2020,44(04):65-70.

[5]吴柳,程恺,胡琪.基于文本挖掘的论坛热点问题时变分析[J].软件,2017,38(04):47-51.

[6]宋逸群,王玉海,聂梅,冯翰钊.大数据透视下的京津冀协同发展民生热点问题探究[J].领导之友,2017(05):61-68.

[7]姜玉坤. 舆情热点信息挖掘技术的研究与应用[D].天津大学,2017.

[8]陈莉萍,杜军平.突发事件热点话题识别系统及关键问题研究[J].计算机工程与应用,2011,47(32):19-22.

[9]张旭华,任蔚,李欣.基于 k-measns 的大学生健康数据分类方法研究[J].信息与电脑(理论版),2019,31(24):10-12.