第八届“泰迪杯”数据挖掘大赛

作品名称：基于网络平台群众留言的数据挖掘分析

摘要

随着网络的快速发展，群众在执行政府决策的同时，可以对政府本身，执行中的问题和身边的不良现象，通过网络发表留言来分享自己的感受和认识。因此，通过大数据、人工智能等技术对群众的留言进行分析挖掘就变得尤为重要。

为了更好地对文本数据进行分析挖掘，我们首先进行了文本预处理。其中包括对相同内容的留言进行适当的数据清洗查重、对文本的分词及去除停用词以及基于词云的可视化分析。此外，我们将文本信息向量化并使用 TF-IDF 方法将空间向量的不同特征项进行赋权，方便后续的研究挖掘。

对于题目一，为了将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，我们引入了朴素贝叶斯分类算法下的高斯朴素贝叶斯分类模型,将多分类问题转化为二分类问题，以概率的形式判断留言属于哪个一级标签类别。同时，我们使用 F-Score 对分类方法进行检测评价。对于题目二，我们基于 K-means 聚类算法将相似的留言信息聚类为同一类，归类出不同的留言主题，引入了留言点赞数、留言持续时间和同话题相同用户数等 7 个热度评价指标并且构建热度指标评价体系。基于上述工作，我们挖掘出了留言中有关噪音扰民、违法乱纪等五个热点问题并针对这五个问题构建了热点问题留言明细表。对于题目三，考虑到对有关部门回复质量的评估，我们引入了三个评估指标即回复的相关性、完整性以及留言-回复的时间系数，通过 RE 概念语义方法将指标量化，以此构建了回复质量评估模型，进而确定回复质量的度量和对其质量评估的评价方案。

通过上述方法，可以有效帮助相关部门进行留言的处理。从而提高其工作效率。

关键词：朴素贝叶斯分类模型、F-Score 方法、K-means 聚类算法、评估模型

Abstract

With the rapid development of the Internet, while implementing the government's decision, the masses can share their feelings and understandings on the government itself, the problems in the implementation and the adverse phenomena around them through the Internet. Therefore, it is particularly important to analyze and mine the messages of the masses through technologies such as big data and artificial intelligence.

In order to better analyze and mine the text data, we first do text preprocessing. It includes appropriate data cleaning and duplication checking for messages with the same content, word segmentation of the text and removal of stop words, and visual analysis based on word clouds for follow-up work. In addition, we vectorize the text information and use the TF-IDF method to weight the different feature items of the space vector to facilitate subsequent research and mining.

For topic one, in order to categorize the messages so that the mass messages can be dispatched to the corresponding functional departments for processing, we introduce the Gaussian Naive Bayes classification model under the Naive Bayes classification algorithm. Besides we converts the multi-classification problem into a two-class classification problem, and judges which first-class label category the message belongs to in the form of probability. At the same time, we use F-Score to detect and evaluate the classification method. For topic two, we cluster similar message information into the same class based on the K-means clustering algorithm and classify different message topics. Moreover, we introduce 7 popular evaluation indicators such as the number of message likes, the duration of the message, and the number of users with the same topic on the same topic and build a heat evaluation system. Based on the above work, we have dug out five hotspot questions about noise disturbing the people and breaking the law and discipline, etc.. Moreover, we construct a list of hotspot questions based on these five questions. For topic three, considering the evaluation of the response quality of relevant depart- ments, we have introduced three evaluation indicators, namely the relevance, integrity of the response, and the time coefficient of the message-reply. By quantifying the indicators through the concept of RE concept semantics, a response quality assess- ment model was constructed.Then we determine the quality of the response and the evaluation plan for its quality evaluation.

Through the above methods, it can effectively help relevant departments to process messages, and improve their work efficiency.

Key words: Naive Bayes classification algorithm, F-Score, K-means clustering algorithm, Evaluation model

目录

1 问题介绍

1.1 问题背景

随着网络的快速发展，互联网在生活中占据着越来越重要的地位，各个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。对于群众来说，他们不再只是被动的接受政府的政策，而是可以通过网络发表留言来分享自己的感受和认识，而留言中所包含的丰富信息，对政府管理也具有重要的价值。

但是，与此同时，由于在问政平台留言的方便与快速，越来越多的人倾向于在网络平台给政府提供留言意见。这就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。

网络留言的数量与日俱增，单单使用人工进行对这些留言的处理已经无法满足政府的需要。此时，大数据、 云计算、人工智能等技术的用途就凸显了出来。大数据的处理可以更加快速的对大量数据进行精准高效的处理，从而大大提高了政府的工作效率。

1.2 主要工作

根据题目给到的各类留言，从以下几个方面对数据进行分析挖掘：

（1）分析各个留言所属的标签分类体系。我们采用高斯朴素贝叶斯分类算法模型对已经文本预处理的留言进行一级分类。

（2）将相似的留言信息进行归类，并分析其热度挖掘出热点问题。为了将相似的留言信息进行归类，我们引入了 K-means 聚类算法将相似度大的留言信息进行聚类。同时，我们提出了 7 个热度指标评价并构建评价指标体系以反映留言主题热度的程度高低。

（3）分析政府对各个留言的答复的质量，给出评价标准。我们将相关性、完整性以及留言-回复时间间隔三个方面作为评价指标，建立出回复质量评估模型，借此对答复质量进行合适的评估。

2 文本预处理

2.1 数据清洗查重

数据清洗就是指对缺失、错误、重复的数据进行适当的增删改减，使整个数据具有规范性和一致性，便于后续的研究应用。数据清洗是数据处理中很重要的部分，若清洗不当，则会对后面的研究造成影响。

对于给出的文本数据，我们发现，其中出现了完全相同的留言信息，而且相同的留言可发现留言的时间间隔非常近，只间隔几分钟甚至几秒。因此，我们可以认为，这种时间间隔非常近的相同的留言可能是用户错误留言导致的，这种数据对于我们的研究挖掘来说是无用的，对于我们需要进行的文本分类也是累赘的。因此我们需要将相同的留言删去部分，只保留其中的一条。

2.2 分词以及去除停用词

所谓分词，就是将连续的字序列按照一定的规范重新组合成词序列的过程。通过将句段进行分词，可以提取很多有用的信息，同时对词频统计也具有很大的帮助。

本文中，我们使用 python 中的 jieba 分词工具进行分词。jieba 采用的算法为: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成词情况所构成的有向无环图 (DAG)；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，使用了 Viterbi 算法。

本文，我们针对不同的问题，为了更好的解决问题，我们对每个附件中不同的部分进行了 jieba 分词。如语句“K8 县丁字街的商户乱摆摊”通过 python 进行分词后的效果为“K8”“县”“丁字街”“的”“商户”“乱”“摆摊”。

通过分词后的结果可以发现，分出来的词组中有许多词组或者汉字是我们不需要的，也就是对我们的数据挖掘没有帮助意义的词组，如果保留这些词不对这些词或汉字进行处理，对于我们后续的词频统计就会造成较大的影响。因此，在这里我们需要引入停用词。停用词指类别特征很弱但出现次数很多的词。

以下是部分停用词表：

图 1 停用词表

为了去除停用词，我们首先扩展了传统的停用词表，针对本题即有关政府平台留言的问题，我们基于分词后的结果作出选择选出扩展的停用词，并将其放入传统的停用词表以构建新的停用词表。然后经过分词工作后得到的词与构建的停用词表进行匹配，若匹配成功则将该词删去，若不成功则保留。将文本进行分词和去除停用词后部分结果如下图所示：

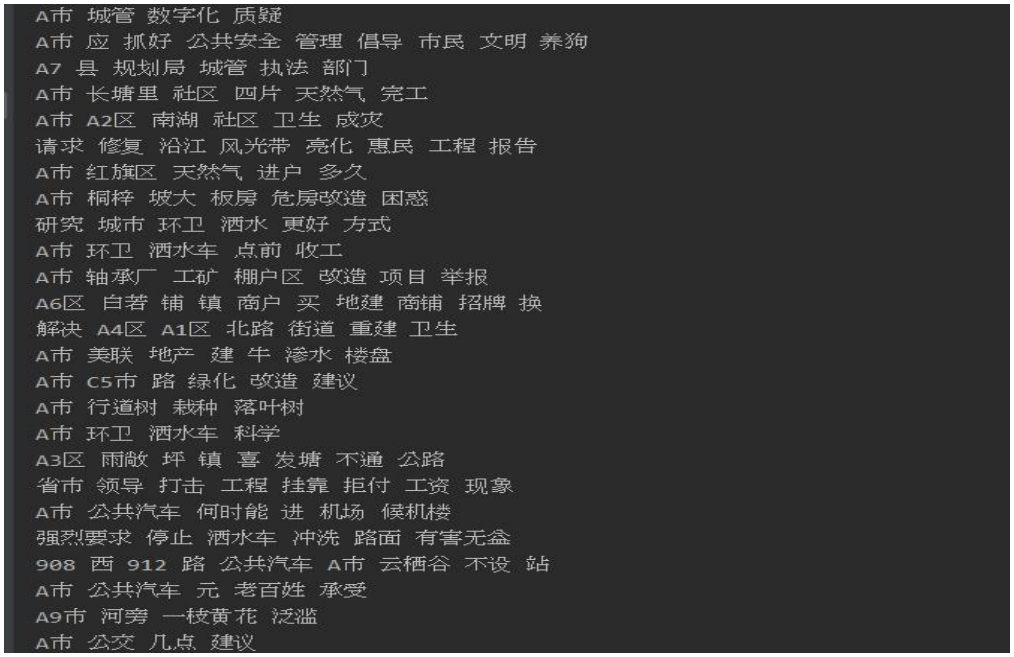

图 2 分词及去停用词后部分结果

2.3 基于词云的可视化分析

词云是一种对海量文本集进行词频（term frequency，简称 TF）统计分析，

然后生成可视化画像的文本分析技术。词云图可以直观清晰的展示语料库的词频。通过 python 中的 wordcloud 库能够有效对词袋进行词云分析绘图。

而词频（term frequency，简称 TF）统计是一种用于文本挖掘与情报检索的常用加权技术，用以评价一个词对于一个文件或者一个语料库中的一个领域文件集的重复程度。在实际操作中我们基于前面 python 的 jieba 切词处理后的词表，运用 Count 函数将分词后的词表出现次数进行统计，再返回词袋字典包含其词频数，并对统计后的词袋字典进行排序。我们作出的词频统计的部分结果如下面的词云图：

图 3 词云图

2.4 文本的空间向量模型及 TF-IDF 算法

由于计算机不能够直接处理文本信息，我们需要对文本进行处理，将文本表示成为计算机能够直接处理的形式，即文本数字化。文本表示也称为文本特征表达，目前常用的文本表示模型有向量空间模型、布尔模型和概率模型等。

将每一个文本表示为向量空间的一个向量，并以每一个不同的特征项（词条）对应为向量空间中的一个维度，而每一个维的值就是对应的特征项在文本中的权重，这里的权重可以由 TF-IDF 等算法得到。向量空间模型就是将文本表示成为一个特征向量方便后续计算机进行算法实现。

为了合理的计算出权重的值，本文我们均采用 TF-IDF 算法将空间向量的不同特征项进行赋权。其中，词频 TF，词频是一个词语在文章或句子中出现的次数。如果一个词很重要，很明显是应该在一个文章中出现很多次的，那么它的权

重的就越高，即 TF 的值越大。但是这也不是绝对的。因此又可以假设，如果某个词比较少见（在我们准备的词库中出现较少），但是它在这篇文章中多次出现，那么它很可能反映了这篇文章的特性，正是我们所需要的关键词。在此，在词频 TF 的基础上又引出了反文档频率 IDF 的概念。反文档频率（IDF）=log（语料库的文章数/包含该词的文档数+1）。最后可以得出$w_{ij} = TF \times IDF$。这就是词的权重。如果w_{ij}的值越大，说明这个词对文章的重要性越高，这个词的权重就越大。由此，我们可以避免简单的词频统计赋权所造成的纰漏。

通过上述分析，我们可以将文本信息向量化并进行权重赋值。

3 模型算法的构建

3.1 基于高斯朴素贝叶斯算法的一级标签分类模型

在处理网络问政平台的群众留言时，需要对其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。但目前大部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。为此，基于附件所给的部分数据，我们引入高斯朴素贝叶斯分类算法建立关于留言内容的一级标签分类模型。

（1）朴素贝叶斯思想

贝叶斯分类是一类分类算法的总称，这类算法均以贝叶斯定理为基础，故统称为贝叶斯分类。而朴素贝叶斯分类是贝叶斯分类中最为简单高效的一种。贝叶斯分类是基于贝叶斯定理实现的。朴素贝叶斯的思想基础是：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，哪个类别的概率最大，就认为此待分类项属于哪个类别。

根据上述分析，朴素贝叶斯分类的流程可以由下图表示：

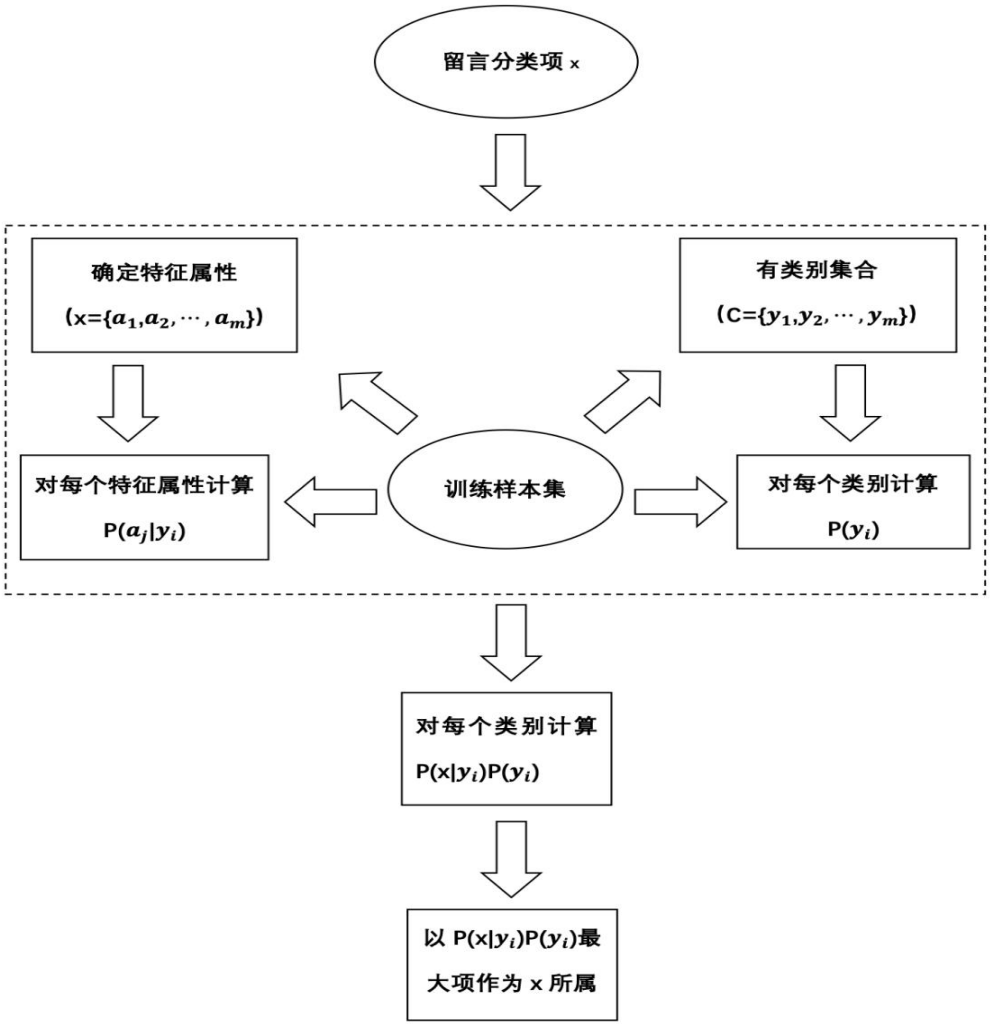

图 4 朴素贝叶斯分类流程图

对于特殊情况，当某个类别下某个特征项划分没有出现时，即 $P(a_j \mid y_i) = 0$ 时，这会令分类器质量大大降低。为了解决这个问题，我们引入 Laplace 校准，它的思想非常简单，就是对没类别下所有划分的计数加 1，这样如果训练样本集数量充分大时，并不会对结果产生影响，并且解决了上述频率为 0 的尴尬局面。针对不同的情况，朴素贝叶斯分类下也有三种不同的模型，分别有：高斯模型、多项式模型和伯努利模型。本题我们选用高斯朴素贝叶斯分类模型进行文本分类。

此外，由朴素贝叶斯分类的定义可以看出，每一个待分类项的特征属性也是关键因素。特征属性就是每一个文本具有的最明显的特征并将其表示出来的属性，因此特征属性、特征属性划分及训练样本质量在很大程度上决定了后面分类

器的质量。通过数据处理后，我们已经合理的将附件 2 中的留言主题文本转化为空间向量，并利用 TF-IDF 方法将向量转化为权值向量。因此，后续的朴素贝叶斯算法均是基于该权值向量进行分析。我们查阅资料发现，朴素贝叶斯分类算法模型主要针对的是二分类问题，即将数据分类分为两大类，而本问题是将留言划分给不同的部门进行处理，即本问题是多分类问题。因此，我们需要将本题的多分类问题转化为多个二分类问题。

（2）一级标签分类模型

基于上述高斯朴素贝叶斯分类思想，我们将待分类与类别的映射关系记作 F，则该分类模型可表示为：

$$F：X \xrightarrow{f(x)} C;$$

$$\text{其中，}\ f(x)=\max\left\{P(y_i \mid x)=\frac{P(x \mid y_i)P(y_i)}{P(x)}\right\}$$

（3）模型的实现与结果

基于上述思想和模型，我们将附件二所提供的 9210 条数据，通过 python 中的 sklearn 库切分数据集，以 8：2 的比例随机挑出训练集与测试集，其中训练集占 80%，测试集占 20%。后续我们对训练集进行训练，对其进行数据预处理以及将文本进行向量化表示并生成 TF-IDF 权值矩阵。测试集在对其进行数据预处理后，以训练集的 TF-IDF 权值矩阵为基准，生成对应的 TF-IDF 权值矩阵。本题使用 sklearn 库中的高斯朴素贝叶斯算法(GaussianNB)，通过训练集的 TF-IDF 权值矩阵及其对应标签，生成一特定模型，通过此模型对测试集的标签进行预测以及对预测结果的好坏进行判断。

我们先将随机选取的 20%数量的留言进行测试，例如判断留言是否属于城乡建设这一类。即将城乡建设的标签定义为 1，而不是城乡建设的其他所有的标签定义为 0，这样我们就将本题的多分类问题转化为多个“是”或“不是”的二分类问题。

综上所述，我们将留言进行分类所得到的部分结果如下：

id	message_num	label	message
4699	271348	0	其他
8663	106007	0	其他
1100	108894	0	其他
6348	208263	0	其他
144	16260	0	其他
...	...	...	...
2393	114659	1	环境保护
4425	199205	0	其他
2718	161515	1	环境保护
5996	128888	0	其他
5719	87134	0	其他

[1831 rows x 3 columns]

id	message_num	label	message
2307	87301	0	其他
8409	29570	0	其他
1682	160988	1	城乡建设
2005	259832	1	城乡建设
5749	92820	0	其他
...	...	...	...
8624	97985	0	其他
8243	160162	0	其他
1652	160912	1	城乡建设
1152	109673	1	城乡建设
2393	114659	0	其他

[1827 rows x 3 columns]

图 5 部分分类结果

（4）使用 F-Score 对分类方法进行评价

F-Score 是对测试准确性的度量，是度量特征在不同类别间的区分度的一种指标，它考虑了测试的精确率 Precision 和召回率 Recall 来计算分数，F-Score 值越大代表该特征在不同类别之间的区分度越强，F-Score 是衡量特征类别间分辨能力的有效方法。所以我们选择 F-Score 对上述所建立的分类模型进行评价，以最终所得分数 Score 作为对分类合理性的一个评判标准。以下便是综合考虑 Precision 和 Recall 的调和值的 F-Score 指标：

$$F\text{-}\mathrm{Score} = (1+\beta^2) \bullet \frac{\mathrm{Pr}\,ecision \bullet \mathrm{Re}\,call}{\beta^2 \bullet \mathrm{Pr}\,ecision + \mathrm{Re}\,call}$$

基于本文所需解决的问题，我们将精准率称为查准率 P，召回率称为查全率 R，

而公式中的β为召回率（查全率）的权重是精准率（查准率）的β倍，即召回率与精准率的重要程度的衡量。本题我们选取β为0.5，1和2，分别计算出每个分类项β不同时所对应的F-score。通过对各个类别下的β不同时对应的F-score求均值，可得到最后三种评测结果：

$$Score = \frac{1}{n}\sum F_{\beta,\ i}$$

其中，P_i为第i类的查准率，R_i为第i类的查全率，β取值为0.5，1或2。

基于此，我们得到F-score的结果如下图所示：

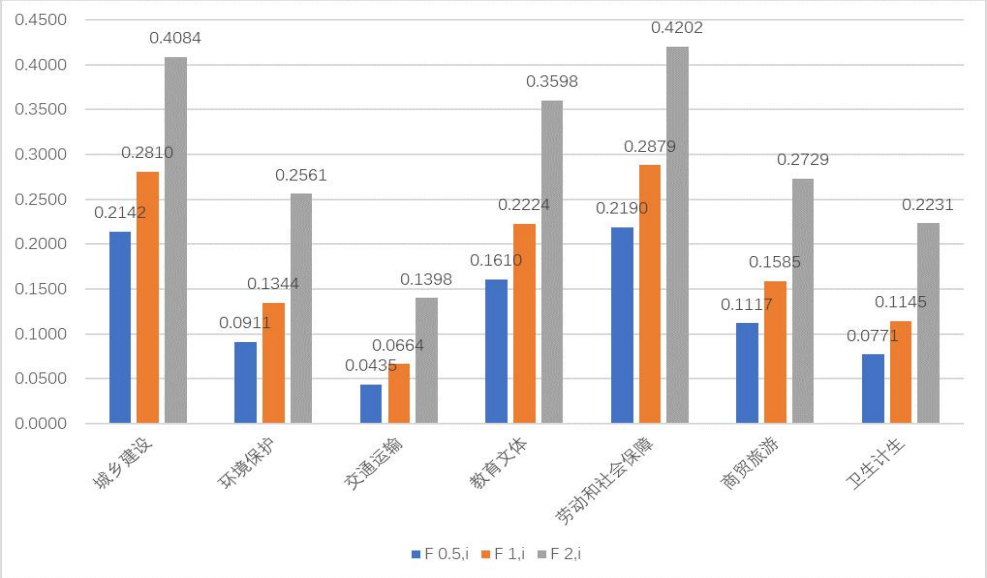

图 6 β不同时各分类项对应的F-score

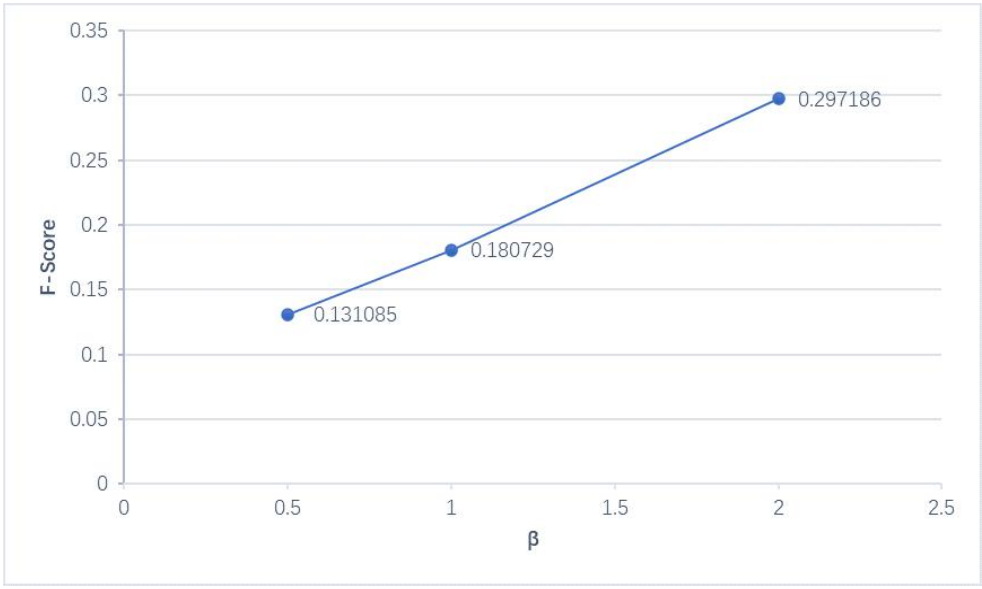

图 7 β不同时模型所对应的F-score

从图中看，分类模型的效果并不是特别理想，可能是对数据处理不到位导致的。但是，在考虑精准率和召回率的重要程度之后发现，如果认为召回率比精准率更重要，即以β=2 来评价该分类方法，其 F 分数的值越大，即相对来说分类效果更好。

3.2 基于 K-means 算法的留言聚类及热度指标评价体系

3.2.1 基于 K-means 算法的留言聚类聚类

（1）文本聚类

所谓文本聚类就是一种无监督的机器学习算法，即将无类别标记的文档信息根据不同的特征，利用文本相似度计算将有相似特征属性的文本聚类在一起。因此，通过文本聚类的方法，可以根据每一条留言的内容，对相似的留言内容进行聚类，从而保证同类间的留言内容相似度很大，不同类的留言内容相似度尽量小。基于留言内容文本聚类，可以为后续热点问题挖掘提供基础。

（2）文本相似度计算

相似度是用来衡量文本间相似程度的一个标准。在文本聚类中，需要研究文本个体间的差异大小，而差异的大小则可以根据文本间相似度的大小来衡量。最后将根据相似特性的信息进行归类。本文我们采用基于距离度量的欧几里得距离计算留言间的差异的大小。

欧几里得距离也即欧式距离（Eucliden Distance），衡量的是多维空间中两个点之间的绝对距离。其计算公式为：

$$dist(X,Y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

其中 $X=(x_1,x_{2,}...x_n)$ 和 $Y=(y_1,y_2...y_n)$ 为两个被 n 个数值属性标记的对象。

（3）K-means 聚类原理及方法

K-means 聚类算法是一种应用很广泛的基于划分原理的聚类算法。它是一种迭代求解的聚类分析算法。K-means 聚类原理为随机选取 K 个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。

其具体的算法描述如下：

a. 从数据中选择 k 个对象作为初始聚类中心；

b. 将每个样本数据集划分离它距离最近的类；

c. 根据每个样本所属的类，更新簇类的均值向量；

d. 重复 b、c 步，当达到设置的迭代次数或簇类的均值向量不再改变时，输出聚类算法结果。

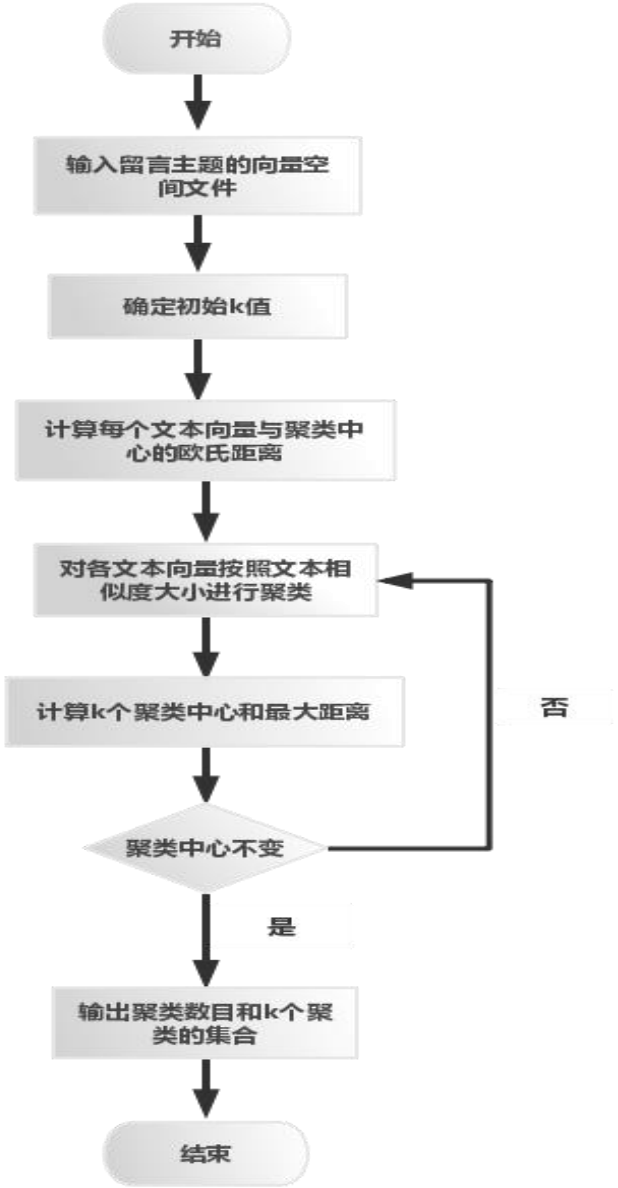

图 8 K-means 聚类算法流程图

K-means 聚类算法需要选择合适的数据集。因此，我们根据 TF-IDF 方法计算出的各个分词的权值，筛选出 TF-IDF 值高于我们选定的阈值的词组成新的词库。根据新的词库可以得到我们需要的词汇-文本矩阵作为输入的样本数据集。基于 Python 中的 collections 库，我们可以得到每个词语的词频，其中，词频排名前十的词语如下图：

图 9 词频排名前十的词语

并且，从上述描述可以看出，K-means 聚类算法需要人为给定 k 的值。而本问题由于数据量较大，无法人工判断出最佳的 k 的大小是多少，因此，我们使用“手肘法”利用 Python 作图确定最佳聚类数 k。

手肘法的思想为：随着聚类数 k 的增大，每个簇类的聚合程度会逐渐提高，那么所有样本的聚类误差平方和 SSE 自然会逐渐变小。当 k 小于真实聚类数时，由于 k 的增大会大幅增加每个簇的聚合程度，故 SSE 的下降幅度会很大，而当 k 到达真实聚类数时，再增加 k 所得到的聚合程度回报会迅速变小，所以 SSE 的下降幅度会骤减，然后随着 k 值的继续增大而趋于平缓。由此可以得到，所作的 SSE 的图形的拐点处（也即“手肘处”）就是 k 的最佳值。

利用“手肘法”作图得到部分截图如下：

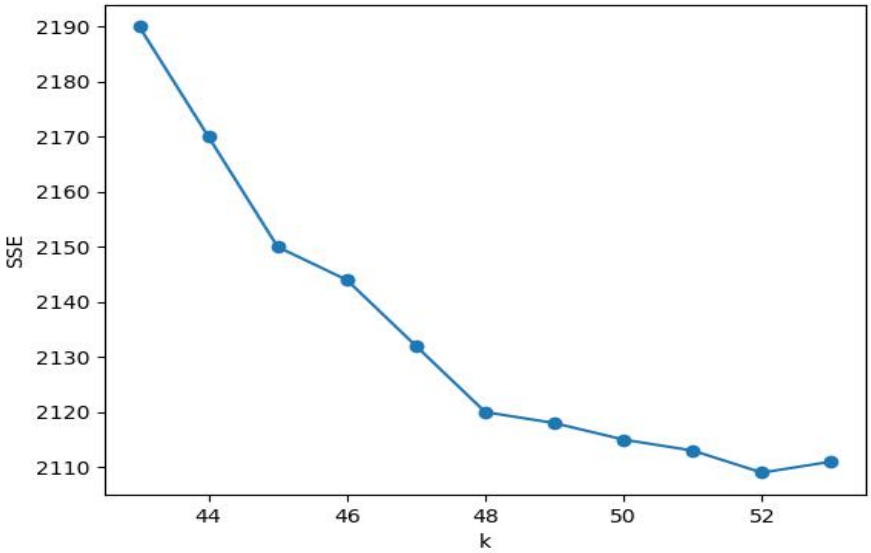

图 10 “手肘法”示意图

从图中可以看出，k=48 时，k 值之前 SSE 的下降幅度很大下降速度快，k 值之后 SSE 的下降幅度减少较多，下降速度变慢。因此，本题我们选取 k 值为 48。

（4）K-means 聚类结果

划分成48类的结果如下：
第1类共71个分别是84 152 220 302 344 356 486 496 518 645 755 870 906 938
第2类共54个分别是34 69 78 99 106 154 206 463 472 632 848 1010 1048 1070
第3类共1个分别是4126
第4类共40个分别是21 43 45 151 478 753 856 977 1057 1144 1229 1256 1641 1791
第5类共70个分别是39 47 134 193 275 286 300 350 404 562 591 715 731 810
第6类共9个分别是447 672 834 999 1189 1679 2380 3481 3891
第7类共55个分别是12 97 175 191 322 451 514 517 625 677 746 747 752 787
第8类共98个分别是23 50 86 132 153 164 365 423 507 528 588 653 734 745
第9类共34个分别是108 110 307 430 432 458 551 555 628 807 951 1779 1787 2014
第10类共2个分别是1426 3065
第11类共184个分别是13 16 28 31 71 85 90 105 119 169 183 215 223 225
第12类共54个分别是147 198 269 585 604 687 700 709 764 774 858 901 974 1115
第13类共1个分别是3661
第14类共4个分别是446 3070 3205 3269
第15类共41个分别是2 44 235 255 371 442 577 744 811 900 935 1079 1122 1128

图 11 K-means 聚类部分结果

通过上述的流程计算，我们可以通过 K-means 聚类算法将相似的留言信息进行聚类，归类出不同的留言主题。同时也方便我们后续挖掘出热点问题。

3.2.2 热度指标评价体系

针对政府平台的群众留言的热度评价，我们制定了评价体系及度量方案如下表：

表 1 评价体系及度量方案

H	评价要素	评价指标	度量方案
针对留言热度定向评价结果	用户指标 A_1	已发留言率 C_1	用户已发留言率 M_1
		同话题留言数 C_2	同一类留言问题的留言数量 M_2
		同话题相同用户数 C_3	同话题相同用户数 M_3
	传播指标 A_2	留言反对数 C_4	同类留言的反对总数 M_4
		留言点赞数 C_5	同类留言的点赞总数 M_5
		留言持续时间 C_6	同类留言时间跨度 M_6
		受众地域分布 C_7	同类留言覆盖的地理区域数量 M_7

最终评价结果 H 的大小由评价指标 C_1～C_7 的度量值加权计算得到。本题中，权重的计算有单一权重与组合权重两种。热度指标计算方法具体如下。

（1）单一权重计算

首先，我们构造判断矩阵。对于本题，我们构造了三个判断矩阵 RH、RA_1、RA_2。其中 RH 代表用户指标 A_1、传播指标 A_2 相对于留言热度评价结果 H 重要性的两两判断矩阵。RA_1 代表已发留言数 C_1、同话题留言数 C_2、同话题相同用户数 C_3 相对于 A_1 重要性的两两判断矩阵。而 RA_2 代表留言反对数 C_4、留言点赞数 C_5、留言持续时间 C_6、受众地域分布 C_7 相对于 A_2 重要性的两两判断矩阵。由于指标 C_1～C_{14} 对留言热度评价结果的贡献不同，为了结果更加具有可信性及准确性，我们参考《政府负面网络舆情热度定量评价方法》这篇论文中提到的打分方法，构造的 3 个两两比较判断矩阵如下所示。

$$RH \to \begin{pmatrix} 1 & \frac{1}{5} \\ 5 & 1 \end{pmatrix} \qquad RA_1 \to \begin{pmatrix} 1 & \frac{1}{7} & \frac{1}{5} \\ 7 & 1 & 3 \\ 5 & \frac{1}{3} & 1 \end{pmatrix} \qquad RA_2 \to \begin{pmatrix} 1 & \frac{1}{5} & 3 & \frac{1}{3} \\ 5 & 1 & 7 & 3 \\ \frac{1}{3} & \frac{1}{7} & 1 & \frac{1}{5} \\ 3 & \frac{1}{3} & 5 & 1 \end{pmatrix}$$

根据我们进行的一致性检验判定可发现，这 3 个判断矩阵均通过一致性检验，因此矩阵是合格的。

（2）最大特征值对应权重向量计算

为了后续组合权重的计算，我们需要计算上述 3 个单一矩阵的权重向量。因此首先，我们需要计算 3 个判断矩阵所对应的最大特征值，根据最大特征值计算出矩阵的权重向量 γ， α_1， α_2。

（3）组合权重计算

上述表 1 中的评价指标 A_1～A_2 及 C_1～C_7 之间没有交叉，因此上述的 3 个判断矩阵是完全的。基于此及根据上述权重向量的计算，我们计算出 C_1～C_7 各评价指标相对于 H 的组合权重 $c\omega_i$ 即：

$$cw_i = \begin{cases} \gamma_1 \alpha_{1j} & j = 1,2,3 \\ \gamma_2 \alpha_{2j} & j = 1,2,3,4 \end{cases}$$

其中，γ_1和 γ_2为向量 γ的第 1，2 个分量。

（4）留言热度评价结果的计算

通过上面的研究分析可得，本题中的留言热度的最终评价结果 H 由 C_1～C_7 的度量值 M_1～M_7 加权计算得到，公式即为：

$$H = \sum_{i=1}^{7} c\omega_i \times M_i$$

通过上述分析，我们便可以建立一套基于 7 个热度评价指标以及系生出的热度指数的热点问题评价体系，通过该体系便可得出不同聚类后的留言问题下的热度指数，我们选出排名前五的热点问题汇总成表即热点问题表和热点问题留言明细表（详细见附件）。

表 2 热点问题表

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	287.592	2019/1/3 至 2020/1/7	A7 县	噪音扰民
2	2	173.269	2019/1/2 至 2019/12/31	A 市	违规违章
3	3	123.775	2018/10/27 至 2020/1/6	A7 县星沙	社区生活问题
4	4	114.615	2019/1/1 至 2020/1/6	A3 区	违法乱纪
5	5	68.573	2019/1/14 至 2020/1/6	A 市楚江	管理者管理不当

表 3 部分热点问题留言明细表

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	239262	A00053043	A7县松雅湖东北角项目建设噪音极大!	2019/3/28 5:51:13	松雅湖位于黄兴大道东北角莲花塘公交站对面最近新建了一处施工项目部，项目	0	1
…	…	…	…	…	…	…	…
1	225191	A00017755	A7县泉塘第三小学大放广播扰民	2019/9/4 15:03:54	我们是一大群居住在泉塘三期安置区的打工一族，工作在远大集团、菲亚特、北	0	0
2	221380	A00063751	举报A市维也纳智好酒店涉黄	2019/8/9 7:23:00	A市高新区谷园路39号维也纳智好酒店涉黄从2018年开始至今一直营业的红红火	0	0
…	…	…	…	…	…	…	…
2	257402	A00074795	举报A市时代年华精装房欺骗购房者	2019/5/15 8:30:29	尊敬的领导您好，我是A市时代年华的业主，出于对时代年华的信任我买了人生中	0	2
3	269143	A00042014	A7县星沙文化公园东南门附近广场舞影响附近	2019/8/29 17:17:31	A7县星沙文化公园东南门附近，有人跳广场舞，严重影响附近居民生活！！！	0	0
…	…	…	…	…	…	…	…
3	198673	A00077225	A7县星沙联络线桥下有人敲鼓严重影响居民休	2019/10/6 7:46:49	在A7县碧桂园翘楚棠生活区外面的星沙联络线桥下有一对练习敲鼓的早上六点	0	1
4	241886	A00083878	投诉A市中交中央公园绿城物业违规收费	2019/5/11 13:04:14	1.本人于2018年8月份购买了A1区中交中央公园第四期的一套房子，购房合同注	0	0
…	…	…	…	…	…	…	…
4	263525	A00026153	利社区拆迁，居民被强制	019/4/18 10:32:1	建那么多120㎡户型给谁？3、为什么5年前	0	0
5	191474	A000110333	12123上申请驾驶证期满换证，一个星期了都无	2019/4/26 15:28:42	说是推出了便民服务，可以直接网上换证，听起来是挺好的。可是我在12123上	0	0
…	…	…	…	…	…	…	…
5	253478	A00059726	A3区山景区管理不作为，“南辕北辙”路标继续	2020/1/1 19:00:42	去年12月26日在A3区山风景区看见有外地游客被A3区山脚下的“南辕北辙”路标所	0	0

3.3 回复意见质量评估的 RISQE 模型

群众的留言可以体现出群众所关注的问题，而相关部门对留言的回复则可以反应政府对该问题的回应程度。针对问题三答复意见的质量如何进行评价，我们预想从答复的相关性、完整性以及留言-回复的时间间隔的角度对答复意见进行评价。

（1）回复质量指标体系

首先，针对答复内容与留言内容之间的相关性，我们通过答复与留言间的概念语义相似度判断他们之间是否存在相关性。目前，概念性语义相关性计算大多

依赖于本体或语料库，我们从周春等人《基于概念语义相关性和 LDA 的文本标记算法》一文中得到，本题我们可通过 RE 方法对概念语义进行相关性计算。

RE 方法是一种基于特征的方法，基于特征的方法是使用函数计算两个概念之间公共特征及非公共特征的比例以决定概念的相关性。在这里，我们对留言内容与有关部门的回复这两个本体的共同特征和差异特征作相似度计算。从以往的研究论文中可以得到，RE 方法计算本体的相似度需要更多地考虑两个文本间的共性而非差异性。两个文本的特征语义相似度的计算公式为：

$$S（a,b）=\frac{|A\cap B|}{|A\cap B|+\rho(a,b)|A/B|+(1-\rho(a,b)|B/A|}$$

其中，a 为目标，即需要我们进行评价的回复内容。b 为比较基准，即群众的留言内容。而 A、B 分别为对应于 a、b 的特征词集合（在这里，特征词的选择与题目一的方法相同）。B/A 表示选取 的特征词在 B 中存在但在 A 中不存在，A/B 表示选取的特征词在 A 中存在但在 B 中不存在，而其中的 ρ 为一个定义非共同部分的相对重要性函数。ρ的计算方法如下：

$$\rho（a,b）=\begin{cases}\dfrac{depth(a)}{depth(a)+depth(b)}(depth(a)\le depth(b)),\\ 1-\dfrac{depth(a)}{depth(a)+depth(b)}(depth(a)>depth(b))\end{cases}$$

其中，depth 函数表示本体库中实体所在树的深度。而 ρ 的取值范围为（0，1）。

RE 方法可以根据上述计算出文本各部分分类的语义相关度。而各部分分类可以组合成一个总体的文本。我们可以将文本基于概念分类成三大部分，分别文本的组成部分、文本的内容功能和文本的定义属性。对此，将这三大部分赋予合适的权重，并通过公式将这三大部分相结合计算出两文本间最终的语义相似度。最终相似度计算方法为：

$$S_F(a,b)=\omega_1S_1(a,b)+\omega_2S_2(a,b)+\omega_3S_3(a,b)$$

从公式可以看出，相关性的取值不会为 0.其中，SF 为最终两个文本间的语义相似度，而 S1、S2、S3 分别为文本中组成部分、功能和属性的相似度的值，而 ω_1、ω_2、ω_3 分别为相对应的部分的权重，且 $\omega_1+\omega_2+\omega_3=1$

通过上述的分析计算，我们就避免了仅从字面上考虑文本相似度的问题，而可以更深层次的分析文本的语义相似度。从而可以研究回复的相关性问题。

其次，对于答复的完整性判断，通过分析我们发现一个完整回复内容的开头、结尾通常会包含着一些特征词语类似您好、已获悉以及谢谢这样的词汇。通过整理，我们分别构建了开头特征词汇包以及结尾特征词汇包这样两个词包。通过python，我们对答复内容进行迭代选取，使用 jieba 分词工具对答复内容进行切词，我们将切词后的答复内容依次对词包进行匹配。这里，我们构建一个二维向量（m,n），m 代表开头词汇，n 代表结尾词汇。若答复内容匹配到开头词汇，则 m 返回 2，反之返回 1；同理，若匹配到结尾词汇，则 n 返回 2，反之返回 1。于是我们可以得到 3 个等级的完整性，以向量表示为（1，1）；（1，2）和（2，1）；（2，2），取平均值转化为数值为 1、1.5、2（这样可以避免完整性取值为 0 对后续计算造成干扰）。于是关于每条答复的完整性我们给予了量化处理，方便了我们后续对答复质量的评价判断。

最后，考虑到用户对答复的满意度，时间跨度也是十分重要的。答复的及时一方面可以给留言用户带来充足的解决后续问题的时间，另一方面也可以体现答复的效率高低。所以显然的，时间间隔也是我们考虑的指标之一，我们记时间系数为 T，作为时间间隔的度量：$T=\frac{1}{t'-t}$。其中，t’为答复时间，t 为留言时间，如此我们得到了时间这一指标。

（2）回复质量评估的 RISQE 模型

RISQE 模型的概述及其基本思想：为了从答复的相关性、完整性以及留言-回复的时间间隔的角度对答复意见的质量给出一套评价方案，我们提出了基于回复意见质量评估的 RISQE 模型。本模型认为答复的相关性 R、完整性 I 以及留言-回复的时间间隔 T 等因素对于评估答复质量均有不同程度的影响。所以对于处理该方面的工作人员而言，他们会根据自己的主观想法给影响决定评估质量的因素赋予不同的权重，以确定自己的评价方案。RISQE 模型首先根据工作人员提供的各个因素的权重，反馈给工作人员一套基于所给权重的评价方案，然后系统便会根据评价方案收集评估指标 RIT 的数据，然后计算出所处理留言回复的质量综合评估得分，最后反馈给工作人员。

模型的定义和表示：

因素变量：相关性（R）、完整性（I）和时间系数（T)

参数变量：α，β

模型表示：

$$Q_{\alpha\beta}\text{-}Score=(1+\alpha^2+\beta^2)\bullet\frac{R\bullet I\bullet T}{\alpha^2\bullet IT+\beta^2\bullet RT+RI}$$

表达式构造是源于 F-Score 的思想，但又与其有涵义上和本质上的区别，我们考虑了三个评估指标，引入了两个权重参数度量，以计算评估分数。选择参数α，β，以改变三个指标对总体评估质量影响的评价方案。

（2）方案的确定

基于 RISQE 模型的参数α，β的设置，便可决定各个因素变量的权重，从而确定一套评价方案。我们以几套方案为例：

Q_{21}：$\alpha=2$，$\beta=1$；即该方案认为相关性 R 比另外两个因素更重要些

Q_{11}：$\alpha=1$，$\beta=1$；即该方案认为相关性 R 与另外两个因素同等重要

$Q_{1,0.5}$：$\alpha=1$，$\beta=0.5$；即该方案认为时间系数 T 比完整性 I 更重要些

4 结论

本文通过自然语言处理技术将留言内容和政府的答复内容进行合理分析，运用高斯朴素贝叶斯算法对留言进行分类并综合 F-Score 方法对模型进行评估。其次，根据 K-means 聚类算法将留言进行聚类，再提出了热度指标评价体系计算留言热度。最后，我们确定了回复质量评估体系构建了 RISQE 模型。通过对这些算法和模型的运用，可以帮助政府部门大大提高了对留言的分类、处理能力。

5 参考文献

[1]周春,蒋运承.基于概念语义相关性和 LDA 的文本标记算法[J].华南师范大学学报(自然科学版),2018,50(04):121-128.

[2]李少温. 基于网络问政平台大数据挖掘的公众参与和政府回应问题研究[D].华中科技大学,2019.

[3]王丽坤,王宏,陆玉昌.文本挖掘及其关键技术与方法[J].计算机科学,2002(12):12-19.

[4]孙飞显,程世辉,靳晓婷,倪天林.政府负面网络舆情热度定量评价方法——以新浪微博为例[J].情报杂志,2015,34(08):137-141.

[5]陈湘辉.基于朴素贝叶斯算法的社交网络数据挖掘技术研究[J].计算机测量与控制,2017,25(06):199-202.

[6]算法杂货铺——分类算法之朴素贝叶斯分类(Naive Bayesian classification)
https://www.cnblogs.com/leoo2sk/archive/2010/09/17/naive-bayesian-classifier.html.2010-09-17

[7]K-means 聚类最优 k 值的选取
https://blog.csdn.net/qq_15738501/article/details/79036255.2018-01-11