“智慧政务”中的文本挖掘应用

摘要

随着互联网时代的急速发展，互联网的便利性为各行各业都带来了“互联网+”的改革模式，政府部门也不例外，“互联网+”背景下的民意收集模式——智慧政务平台在这个时代应运而生，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。在这样的条件下，各类社情民意相关的文本数据量急速膨胀，为人工分类整理留言增加了非常庞大的工作量。由于目前正处于大数据人工智能时代，建立基于自然语言处理技术的智慧政务系统对提升政府的管理水平和施政效率具有极大的推动作用，因此研究“智慧政务”中的文本挖掘应用有很重要的意义。

对于问题一，为解决依靠人工经验处理时工作量大、效率低且差错率搞等问题，本文首先对数据进行预处理，并对处理好的文本数据进行 TF-IDF 算法训练，将向量化后的文本数据进行朴素贝叶斯分类器训练，由于朴素贝叶斯、IDB 模型和线性支持向量机模型 SVM 的模型结果的精确度中最高的是 SVM 模型，故选择该模型作为最终分类模型，并利用 F1-Score 分数评估模型。

对于问题二，需要从附件 3 中挖掘出热点问题，并将一段时间内的留言按照特定地点或人群进行归类后建立合理的热度评价体系，给出排名前 5 的热点问题和相应的留言信息。首先，利用 BERT 预训练模型提取留言主题的句向量，利用 DBSCAN 聚类算法进行基于留言主题含义的一次归类；接着通过人工标注的 BIO 训练集和测试集微调 BERT 语言模型进行命名实体识别，主要识别出留言主题中的地点实体；然后通过训练好的模型识别出一次归类后的留言数据的地点实体，根据地点实体进行二次归类，实验中发现二次归类结果中有少部分偏差，利用补充规则消除偏差得到最后归类结果；为了建立评价体系，利用主题关键词抽取技术抽取出每个类别的前 10 关键词，进而通过建立评价公式：某类问题的热度指数=该类问题总留言数+与该类问题主题有关的总关键词数。热度指数越高，该类问题越热点。最终得到热点问题排名第一和第二的是发生在 A 市丽发新城小区的搅拌站污染环境，噪音扰民问题以及发生在 A 市伊景园滨河苑出现的捆绑销售车位，欺诈消费者问题。

对于问题三，本文从答复数据的相关性，完整性，可解释性三个方面来建立评价指标体系。相关性：对比留言详情（留言主题）与答复意见的相似度；完整性：留言详情（留言主题）中的关键词，答复意见中是否能够提到，计算出相应的概率；可解释性：留言详情（留言主题）与答复意见中是否存在因果关系。将上述的三种关系赋予相应的权重，得到一个评价分数，使用正态分布将其划分为合理的“优”，“良”，“差”三个区间，比例分别为 2:6:2。将划分好的区间带入到原本的数据集中查看，可以发现很好的划分效果；被划分为“优”的答复意见可以很好的解决留言详情中提出的问题；“良”则只能回答出一部分的问题，或者不能完全解决群众的问题；“差”则是拖延问题，推卸责任。因此本文所建立的评价指标体系可以很好评价答复意见。

关键词：TF-IDF　SVM　BERT　TextRank　评价指标体系

Application of Text Mining in Intelligent Government

Abstract

With the rapid development of the Internet era, the convenience of the Internet has brought the "Internet +" reform mode to all trades and professions. The government departments are no exception. The "Internet plus" mode of public opinion collection -- the intelligent government platform has emerged in this era. The Internet plus political platform of WeChat, micro-blog, mayor's mailbox, sunshine hotline has gradually become the government's understanding of public opinion. An important channel to gather people's wisdom and spirit. Under such conditions, the amount of text data related to social situation and public opinion is expanding rapidly, which increases a huge workload for manual classification and sorting of messages. At present, it is in the era of big data artificial intelligence. The establishment of intelligent government system based on natural language processing technology has a great role in promoting the management level and governance efficiency of the government, so it is of great significance to study the application of text mining in "intelligent government".

For the first problem, in order to solve the problems of heavy workload, low efficiency and error rate, this paper first preprocesses the data, and trains the processed text data with TF-IDF algorithm, and trains the quantized text data with naive Bayes classifier, because of the model knot of naive Bayes, IDB model and linear support vector machine model SVM SVM model is the most accurate one, so this model is chosen as the final classification model, and F1 score score is used to evaluate the model.

For question 2, we need to dig out the hot issues from Annex 3, and set up a reasonable heat evaluation system after classifying the messages in a period of time according to the specific location or population, and give the top 5 hot issues and corresponding message information. First, the sentence vector of the message subject is extracted by the Bert pre training model, and the DBSCAN clustering algorithm is used to classify the meaning of the message subject at a time; then the named entity is identified by the manually annotated bio training set and the test set fine tuning the Bert language model, which mainly identifies the location entity in the message subject; then the retention of the message subject after a classification is identified by the trained model The location entities of speech data are classified twice according to the location entities. In the experiment, it is found that there are a few deviations in the secondary classification results, which are eliminated by supplementary rules to get the final classification results. In order to establish the evaluation system, the top 10 keywords of each category are extracted by the subject keyword extraction technology, and then the evaluation formula is established: the heat index of a certain type of problem = this type The total number of questions + the total number of key words related to the topic of the question. The higher the heat index, the hotter the problem. Finally, the first and second hot issues are the pollution of the mixing station in Lifa new town community of city a, the noise nuisance and the bundling sales parking space in Binhe garden of Yijingyuan, city a, which cheat consumers.

For the third question, this paper establishes the evaluation index system from three aspects of the relevance, integrity and interpretability of the response data. RELEVANCE: compare the similarity between message details (message subject) and response opinions; integrity: key words in message details (message subject), whether they can be mentioned in response opinions, and calculate the corresponding probability; interpretability: whether there is causal relationship between message details (message subject) and response opinions. The above three relationships are given corresponding weights, and an evaluation score is obtained. The normal distribution is used to divide them into three reasonable intervals of "excellent", "good" and "bad", with the proportion of 2:6:2 respectively. The divided interval can be brought into the original data set to view, and good division effect can be found; the reply opinions classified as "excellent" can solve the questions raised in the message details very well; "good" can only answer part of the questions, or can not completely solve the questions of the masses; "poor" is to delay the questions and shirk the responsibility. Therefore, the evaluation index system established in this paper can evaluate the reply opinions very well.
Keywords: TF-IDF SVM BERT TextRank Evaluation

目录

一、挖掘目标

随着信息时代的高速发展，互联网的便利性与不可替代性的优势逐渐展现，越来越多的寻常工具以“互联网+”的模式诞生，在这样的背景下，智慧问政平台应运而生。在智慧问政平台上，老百姓可以自主自由地向政府反映自己社会生活中的意见与看法，因此在该平台上汇集了大量的社情民意线管的文本数据。为更快捷地利用大数据技术处理数据，如何建立类模型文本数据按等级标签进行分类是非常重要的一个课题。同样，通过文本挖掘找到在某一段时间内集中反映的热点问题，可以有效地提高政府服务人民的效率。此外，群众最期待的一定是政府根据群众的留言给出相关的完整的有解释性的答复。因此本文进行文本挖掘的主要目标有以下三点:

1. 问题一属于建立分类模型进行预测，并对模型效果进行评价的问题。对于该问题，本文期望达到的目标是根据附件 2 给出的数据建立效果良好的留言内容一级标签分类模型，并利用 F-Score 对分类方法进行评价。
2. 问题二属于建立文本挖掘模型分类与分析的问题。对于该问题，本文期望达到的目标是建立模型并根据附件三将某一时间段内反映特定地点或特定人群问题的留言进行归类，同时定义合理的热度评价指标，评价出热点问题得到最终的结果。
3. 问题三属于综合评价类的问题。对于该问题，本文期望达到的目标是建立一套完整的评价方案对附件四中相关部门对留言的答复意见做出评价，评价的角度从答复的相关性、完整性、可解释性等角度出发。

二、分析方法与过程

2.1 问题一分析方法与过程

2.1.1 流程图

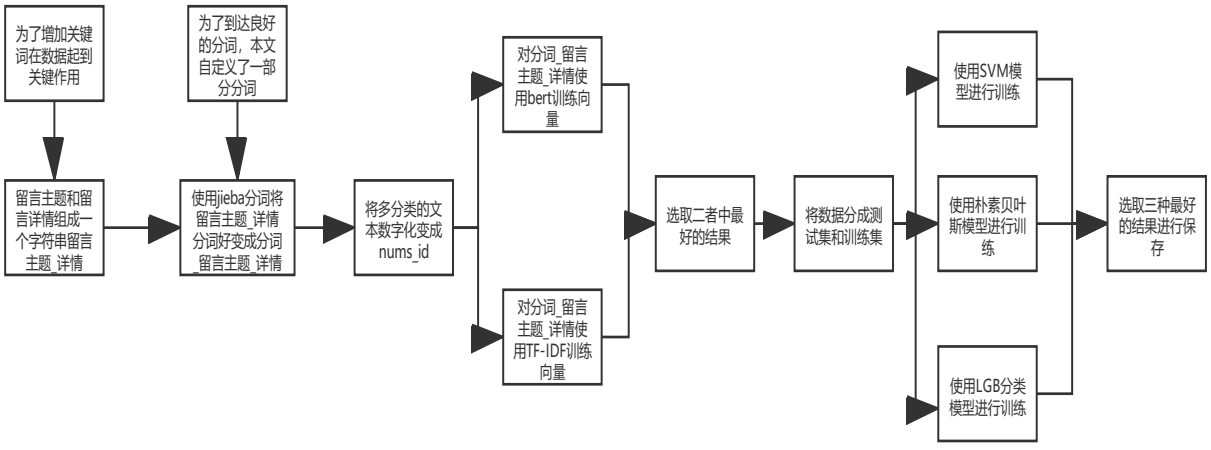

图 1 问题 1 流程图

2.1.2 数据预处理

题目提供了 4 个附件，附件中给出了收集自互联网公开来源的群众问政留言

记录，以及相关部门对部分群众留言的答复一件。第一中主要运用到附件 1 和附件 2 的数据，为了便于后续建立标签分类模型并对模型采用 F-Score 进行评价，本文首先对数据进行预处理，步骤如下：

1. 首先对 9210 条数据进行排查与筛选，未发现缺失值，故不作删除处理；
2. 附件 1 中所给的一级标签分别是：城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游和卫生计生。为便于分类模型的训练，将上述一级标签转换为 0、1、2、3、4、5、6；
3. 对于附件 2 中的文本数据，由于文本中存在一些无意义的词组和符号，这些词组和符号不但对文本的含义不构成直接的意义，还会增加算法运行的复杂度，因此本文在做数据预处理时选择删去文中的标点符号、中文停用词以及并未直接影响文本含义的词组；
4. 对清洗后的文本数据绘制词云图，从整体上对每一级标签所属的文本高频词汇有一个直观的了解。

2.1.3 模型的准备

1. TF-IDF 算法

TF-IDF（term frequency–inverse document frequency）是一种常用于信息检索于数据挖掘的加权技术，其中 TF 的含义是词频（Term Frequency），IDF 是逆文本呢频率指数（Inverse Document Frequency）。TF-IDF 算法是一种非常重要的统计方法，可以用来计算与评估一个词在一个文本资料中的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。具体计算方式如下。

a. TF词频

词频（TF）代表的是关键词或字在文本中出现的频率，计算公式如下：

$$tf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

代表的含义如下：

$$TF_\omega = \frac{\text{在某一类中词条}\omega\text{出现的次数}}{\text{该类中所有的词条数目}} \tag{2}$$

其中，$n_{i,j}$是该词在文件d_j中出现的次数，分母则是文件d_j中所有词汇出现的次数总和。

b. IDF 逆向文件频率

逆向文件频率（IDF）是指总问卷数目除以包含某一特定词语的文件数目后，再将商取对数得到的最终结果。计算公式如下：

$$idf_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|} \tag{3}$$

其中，$|D|$是语料库中的文件总数。$|\{j:t_i \in d_i\}|$表示包含词语t_i的文件数目（即$n_{i,j} \neq 0$的文件数目）。如果该词语不在语料库中，就会导致分母为 0，因此一般情况下使用$1 + |\{j:t_i \in d_j\}|$。

IDF 逆向文件频率的具体公式如下：

$$IDF = \log\left(\frac{\text{语料库的文档总数}}{\text{包含词条}\omega\text{的文档书}+1}\right) \quad (4)$$

其中，分母加 1 是为了避免分母为 0 的情况出现。

如果包含词条 t 的文档较少，IDF 越大，则说明词条有非常良好的类别区分能力。

c. TF-IDF

在计算了 TF 和 IDF 后，对 TF-IDF 做出定义，计算公式如下：

$$TF - IDF = TF \times IDF \quad (5)$$

根据公式，由于在特定文件内的高词语频率，以及该词语在整个文件集和中的低文件频率，可以得到高权重的 TF-IDF，因此 TF-IDF 的用法在于过滤掉常见的词语，保留重要的词语。

2. jieba 分词算法

Jieba 分词是 Python 中文分词组件，该组件的功能是可以对中文文本进行分词、词性标注和关键词抽取。Jieba 有三种分词模式，具体情况如下所示：

a. 精确模式，试图将句子最精确地切开，适合文本分析；

b. 全模式，可以很快速地将所有成词的词语都扫描出来，但不能解决歧义；

c. 搜索引擎模式，该模式是再精确模式的基础上，对长词进行再分割，提高召回率，适用于搜索引擎分词。

3.卡方检验

卡方检验是一种用途十分广泛的加黑色检验方法，它常在分类资料统计推断中应用，其中包括：两个率或两个构成比比较的卡方检验；多个率或多个构成比比较的卡方检验以及分类资料的相关分析等。卡方检验的具体步骤如下：

（1）提出原假设；

（2）将总体X的取值范围划分为k个互不相交的小区间$A_1, A_2, A_3, \cdots, A_k$看，可取$A_1 = (a_0, a_1], A_2 = (a_1, a_2], \cdots, A_k = (a_{k-1}, a_k)$；

（3）把落入第i个小区间的A_i样本值的个数记为f_i，表示组频数（真实值），所有组频数之和$f_1 + f_2 + \cdots + f_k$等于样本容量n；

（4）当H_0为真时，根据假设的总体分布，可以计算出总体X的值落入区间A_i的概率p_i，于是np_i就是落入第小区间A_i的样本值得理论频数；

（5）当H_0为真时，n次试验中样本值落入第i个小区间A_i的概率为p_i，当H_0不为真时，则f_i/n与p_i相差很大，基于这种思想，皮尔逊引进如下检验量：

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \quad (6)$$

在零假设成立的情况下服从自由度$k-1$为得卡方分布。

4. 朴素贝叶斯分类器

朴素贝叶斯分类器是一系列以假设特征之间强（朴素）独立下运用贝叶斯定理为基础的简单概率分类器。该分类器模型会给问题实例分配用特征值表示的类标签，类标签取自有限集合。它不是训练这种分类器的单一算法，而是一系列基于相

同原理的算法：所有朴素贝叶斯分类器都假定样本每个特征与其他特征都不相关。

朴素贝叶斯分类算法的详细情况如下:

（1）$X=(x_1,x_2,\cdots x_D)$表示含有 D 维属性的数据对象，训练集 S 含有 K 个类别，表示为$y=(y_1,y_2,\cdots,y_k)$；

（2）已知待分类数据对象，预测该对象的所属类别，计算方法如下：

$$y_k=\underset{y_k\in y}{arg\max}(P(y_k|X)) \tag{7}$$

即所得y_k即为x所属类别。已知待分类数据对象的情况下，分别计算x属于$y_1,y_2,\cdots,y_k$的概率，选取最大值对应的y_k为所属类别；

（3）根据贝叶斯定理，$P(y_k|x)$的计算方法如下：

$$P(y_k|x)=\frac{P(x|y_k)P(y_k)}{P(x)} \tag{8}$$

（4）假设数据对象各个属性之间相互独立，$P(x|y_k)$计算方式如下：

$$P(x|y_k)=\prod_{d=1}^{D}P(x_d|y_k) \tag{9}$$

（5）若属性A为分散属性或分类属性。训练集中属于类别y_k的数据对象在属性A下的相异属性有n个；训练集中属于类别y_k，且在属性A下的属性值x_d的对象有m个，此时计算公式如下：

$$P(x_d|y_k)=\frac{m}{n} \tag{10}$$

如果属性A是连续属性，通常假设属性均服从均值为μ，标准差为σ的高斯分布，即

$$G(x,\mu,\sigma)=\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{11}$$

因此$P(x_d|y_k)$计算方式如下：

$$P(x_d|y_k)=G\left(x_d,\mu_{y_k},\sigma_{y_k}\right) \tag{12}$$

其中，μ_{y_k}和σ_{y_k}表示在属性A下的均值和标准差。

5. 支持向量机模型

SVM 可以用来进行分类学习，分类学习就是针对给出的训练集数据，在样本空间中找到一个划分超平面，用该超平面将样本分开。而如图所示：

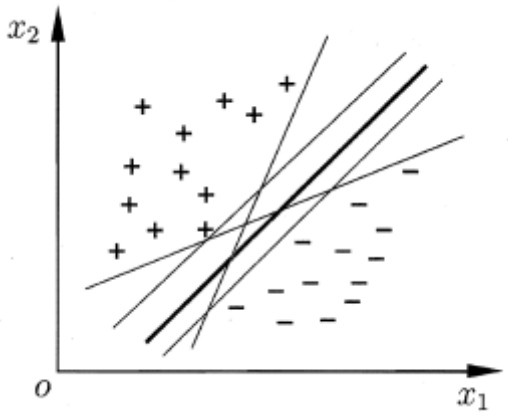

图 2 存在多个超平面划分两类训练样本

这样的超平面有很多，可根据如下的线性方程来确定：

$$w^T x + b = 0 \tag{13}$$

其中$\omega = (\omega_1, \omega_2, ..., \omega_d)$是法向量，确定超平面的方向，$b$为位移项，决定超平面与原点之间的距离。我们将超平面记为$f(\omega, b)$,则样本空间中任意点到$f(\omega, b)$的距离为

$$r = \frac{|\omega^T x + b|}{\|\omega\|} \tag{14}$$

假设超平面$f(\omega, b)$能将训练样本正确分类，即对于$(x_i, y_i) \in D$，若$yi = +1$，则有$\omega^T x_i + b > 0$；若$y_i = -1$，则有$\omega^T x_i + b < 0$。令

$$\begin{cases} \omega^T x_i + b \geqslant +1, y_i = +1 \\ \omega^T x_i + b \leqslant -1, y_i = -1 \end{cases} \tag{15}$$

如下图所示，距离超平面最近的几个训练样本点使得上式的等号成立，他们就是“支持向量”，两个不同类别的支持向量到超平面的距离的和为：

$$\gamma = \frac{2}{\|\omega\|} \tag{17}$$

它被称为“间隔”。

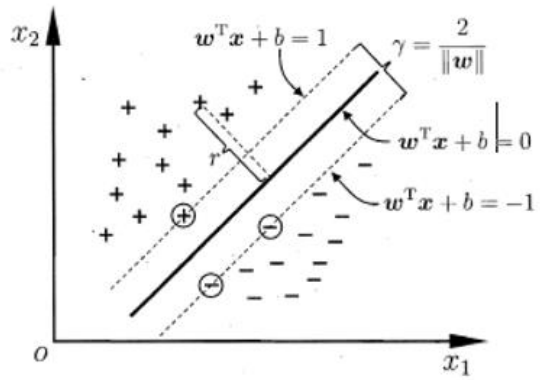

图 3 支持向量与间隔

欲找到具有“最大间隔”（maximum）的划分超界面，也就是要找满足（17）式中约束的参数ω和b，使得γ最大，即

$$\max_{\omega,b}\frac{2}{\|\omega\|} \tag{18}$$

$$s.t.\ y_i(\omega^T x_i+b)\geqslant 1,\quad i=1,2,...,m$$

上式可等价于

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2 \tag{19}$$

$$s.t.\ y_i(\omega^T x_i+b)\geqslant 1,\quad i=1,2,...,m$$

即得到 SVM 的初始模型。

针对实际的问题，由于现实中很难确定训练样本在特征空间中是否线性可分，就算有也不确定是不是由于过拟合所造成的。这时需要允许 SVM 模型在一些样本上出错。即允许某些样本不满足约束

$$y_i(\omega^T x_i+b)\geqslant 1 \tag{20}$$

但是，需要不满足约束的样本应尽可能地少，于是有优化目标：

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2+c\sum_{i=1}^{m}\iota(y_i(\omega^T x_i+b)-1) \tag{21}$$

ι 是损失函数，常用的是凸的连续函数：

$$\text{hinge 损失：}\ \iota_{hinge}(z)=\max(0,1-z) \tag{22}$$

使用 hinge 损失函数后，式 21 变成：

$$\min_{\omega,b}\frac{1}{2}\|\omega\|^2+c\sum_{i=1}^{m}\iota(y_i(\omega^T x_i+b)-1) \tag{22}$$

引入松弛变量$\xi_i\geqslant 0$，可将式 22 重写为：

$$\min_{\omega,b,\xi_i}\frac{1}{2}\|\omega\|^2+c\sum_{i=1}^{m}\xi_i \tag{23}$$

可使用拉格朗日乘子法得到下式：

$$L(\omega,b,\alpha,\xi,\mu)=\frac{1}{2}\|\omega\|^2+c\sum_{i=1}^{m}\xi_i+\sum_{i=1}^{m}\alpha_i(1-\xi_i-y_i(\omega^T x_i+b))-\sum_{i=1}^{m}\mu_i\xi_i \tag{24}$$

其中$\alpha_i\geqslant 0,\mu_i\geqslant 0$是拉格朗日乘子。

令$L(\omega,b,\alpha,\xi,\mu)$对$\omega,b,\xi_i$的偏导为零可得

$$\omega=\sum_{i=1}^{m}\alpha_i y_i x_i \tag{25}$$

$$0=\sum_{i=1}^{m}\alpha_i y_i \tag{26}$$

$$C=\alpha_i+\mu_i \tag{27}$$

将式（25）-（27）代入式（24）即可得到式（23）的对偶问题

$$\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j {x_i}^T x_j \tag{28}$$

$$s.t.\ \sum_{i=1}^{m}\alpha_i y_i=0,\quad 0\leqslant\alpha_i\leqslant C,\quad i=1,2,...,m$$

这时使用 SMO 算法求解上述优化问题的参数ω后，再求解参数b。

$$b=\frac{1}{|S|}\sum_{s\in S}\left(y_s-\sum_{i\in S}\alpha_i y_i {x_i}^T x_s\right) \tag{29}$$

其中$S=\{i|\alpha_i>0,i=1,2,...,m\}$为所支持向量的下标集。

2.1.4 模型的建立过程

在对附件中的数据进行预处理后，开始建立分类模型并进行评分。

第一步：首先使用 TF-IDF 的算法对处理后的附件 2 中的文本使用 TF-IDF 训练向量。

第二步：将得到的向量数据进行分类，分为测试集和训练集两类。

第三步：分别使用 SVM 模型、朴素贝叶斯模型、LGB 模型进行训练，将三种模型中训练结果最好的一种模型结果进行保存。

第四步：对这三种方法中的最好的一种方法的准确性进行模型评估。

以上四个建模步骤即为整个问题一解决的分析方法。

2.2 问题二分析方法与过程

问题二要求从附件 3 中挖掘出热点问题，需要将留言进行归类，并建立合理的热度评价指标，给出排名前 5 的热点问题和相应的留言信息。

由于需要对留言进行归类，并且目的是为了发现和挖掘热点问题，因此本文首先使用 DBSCAN 聚类算法将向量化后的附件 3 留言主题进行聚类得到一次归类结果。此时，同一类别下仍会出现多个地点。为解决该问题，本文基于 BERT 预训练模型来训练针对此题数据中地点和人群的命名实体识别模型，利用训练好的模型识别出留言中的地点或人群，再按识别结果进行二次归类得到最终的归类结果。在最终的归类结果中，将利用自然语言处理技术中的主题关键词抽取技术提取出各类的主题关键词，最终根据每一类别的留言次数和留言主题关键词建立热度评价指标体系，评价出热度指数得到排名前 5 的热点问题。

2.2.1 流程图

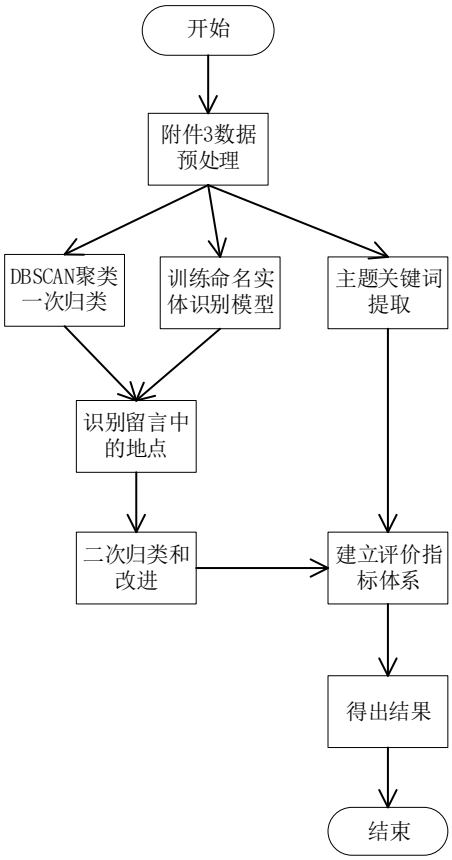

图 4 问题 2 流程图

2.2.2 数据预处理

数据预处理分为特殊字符删除、人工构建训练数据集和基于 BERT 中文预训练模型的句向量提取。BERT 中文预训练模型是谷歌提出的基于双向 Transformer 构建的语言模型。本文将该预训练模型和下游任务模型结合在一起，提升最终的模型效果。

通过附件 3 数据观察，发现在留言详情中出现大量的转义字符'\t', '\n'和空格等，干扰了后续操作，故需要在数据预处理过程中删除上述特殊字符。同时，提取 BERT 中文预训练模型的 Transformer 倒数第二层的输出值作为输入句子的句向量，作为例如实体识别、聚类等下游任务的输入值。

根据题目要求，需要提取出热点问题发生的地点或人群，故实验中利用 BIO 标注方法人工标注部分数据，以训练出基于 BERT 的命名实体识别模型。

2.2.3 DBSCN 聚类算法一次归类

首先，目的为找出某段时间的特定地点或人群的热点问题，但是附件 3 中有大量的留言和问题。为了初步筛选出可能的热点问题，实验利用基于密度的 DBSCAN 算法得到第一次的归类结果。

DBSCAN 算法第一次归类：

输入：样本集$D=(x_1,x_2,\ldots,x_m)$，邻域参数$(eps,MinPts)$，样本距离度量方式

输出：簇划分C.

1) 初始化核心对象集合Ω，初始化聚类簇数$k=0$，初始化样本集合$\Gamma=D$，簇划分$C=0$

2) 对于$j=1,2,\ldots m$，按下面的步骤找出所有的核心对象：

a) 通过欧式距离度量方式，找到样本x_j的eps邻域子样本集$N_{eps}(x_j)$

b) 如果子样本集样本个数满足$|N_{eps}(x_j)| \geq MinPts$，将样本$x_j$加入核心对象样本集合：$\Omega = \Omega \cup x_j$

3) 如果核心对象集合$\Omega = 0$，则算法结束，否转转入步骤 4.

4) 在核心对象集合Ω中，随机选择一个核心对象o，初始化当前簇核心对象队列$\mathrm{Omege}_{cur} = o$，初始化类别序号$k = k + 1$，初始化当前簇样本集合$C_k = o$，更新潍坊问样本集合$\Gamma = \Gamma - o$

5) 如果当前簇核心对象队列$\Omega_{cur} = 0$，则当前聚类簇C_k生成完毕，更新簇划分$c = \{C_1, C_2, C_3 \ldots, C_k$，更新核心对象集合$\Omega = \Omega - C_k$，转入步骤 3，否则更新核心对象集合$\Omega = \Omega - C_k$。

6) 在当前簇核心对象队列$\Omega_{cur} = 0$中取出一个核心对象o'，通过邻域距离阈值eps找出所有的eps邻域子样本集$N_{eps}(o')$,令$\Delta = N_{eps}(o') \cap \Gamma$，更新当前簇样本集合$C_k = C_k \cup \Delta$，更新未访问样本集合$\Gamma = \Gamma - \Delta$，更新$\Omega_{cur} = \Omega_{cur} \cup (\Delta \cap \Omega) - o'$，转入步骤 5.

由于算法中的邻域参数eps难以确定，故实验中发现 eps<5 或 eps>6 时，类簇数为 1，初步判断为 eps 在[5,6]之间。再进行步长为 0.01 的网格化搜索，得到图 5：

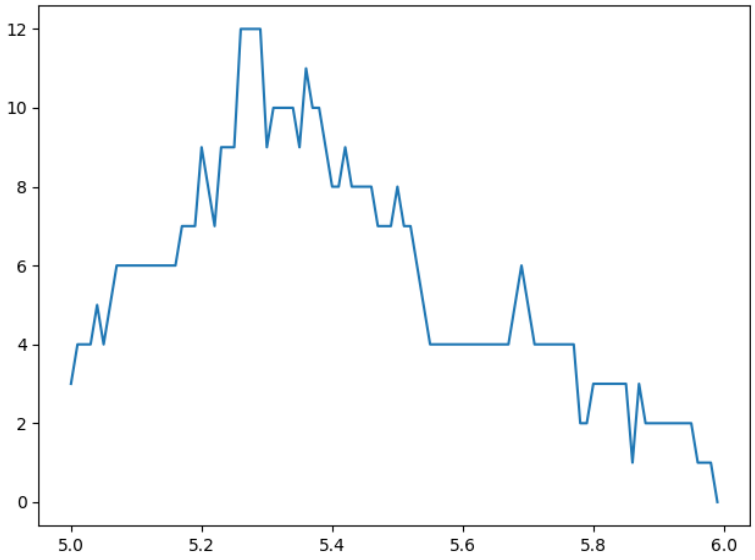

图 5 eps 趋势图

由于最终需要筛选出前 5 的热点问题，故此时选择 eps=5.05，类簇数为 5，此时聚类类簇和数量结果如下表：

表 1 一次归类结果

label	计数
-1	4101
0	140
1	46
2	17
3	12
4	10

其中 label 为-1 表示其为噪音数据，即这些数据无法聚类，表明其相互之间联系稀疏或只有极少数人反应的问题。label 值>-1 时，每一个值代表一个类别。可以看到

类别 label=0 的有 140 条留言记录，部分如下显示：

表 2 一次归类部分数据

留言编号	留言主题	label
188809	A 市万家丽南路丽发新城居民区附近搅拌站扰民	0
189950	投诉 A2 区丽发新城附近建搅拌站噪音扰民	0
190108	丽发新城小区旁边建搅拌站	0
190523	A 市丽发新城违建搅拌站，彻夜施工扰民污染环境	0
190802	A 市丽发小区建搅拌站，噪音污染严重	0
195095	魅力之城小区临街门面油烟直排扰民	0
199379	A2 区丽发新城附近修建搅拌厂，严重污染环境	0
200576	A2 区刘家冲路附近工地昼夜施工，严重扰民	0
203393	A 市丽发新城小区侧面建设混泥土搅拌站，粉尘和噪音污染严重	0
205329	A3 区天顶街道燕联小区 3 栋一 KTV 噪音扰民	0
205483	A2 区金石蓉园菜市场噪音扰民	0
208285	投诉小区附近搅拌站噪音扰民	0
208714	A2 区丽发新城附近修建搅拌站，污染环境，影响生活	0

表中可以看出该类别的大致主题都为噪音扰民，但是注意标红的三行中的地点与其他的留言中出现的地点不同，因此此处只作为第一次归类的结果。接下来需要利用命名实体识别技术识别出留言主题中出现的主要地点，并按照地点进行二次归类。

2.2.4 基于 BERT 预训练模型的命名实体识别技术

BERT 是一个预训练语言模型，要利用其进行命名实体识别，需要进行 Fine-tune 微调。即利用提前训练好的神经网络参数直接拿来用作微调模型的初始参数，进而针对自己的数据集进行训练，训练时模型只在原有的网络上增加少量的神经元，并只更新分类参数，缩短了微调模型的时间并大大提高了预测结果的准确性。

本题中通过人工标注附件数据，构建实体识别模型的训练和测试数据，其中部分数据如下表 3 所示：

表 3 BIO 标注示例

char	label	char	label	char	label
A	B-LOC	合	O	丰	I-LOC
3	I-LOC	法	O	路	I-LOC
区	I-LOC	纳	O	米	I-LOC
一	B-LOC	税	O	兰	I-LOC
米	I-LOC	了	O	春	I-LOC
阳	I-LOC	？	O	天	I-LOC
光	I-LOC	座	O	G	I-LOC
婚	I-LOC	落	O	2	I-LOC
纱	I-LOC	在	O	栋	I-LOC

艺	I-LOC	A	B-LOC	3	I-LOC
术	I-LOC	市	I-LOC	2	I-LOC
摄	I-LOC	A	B-LOC	0	I-LOC
影	I-LOC	3	I-LOC	，	O
是	O	区	I-LOC	一	O
否	O	联	B-LOC	家	O

最终构建了训练集 train.txt: 1287333 行，测试集 test.txt: 321833 行。利用 BERT-BASE 开源库进行微调训练。对训练的模型在测试集中利用 conlleval 评测标准进行准确度（accuracy）、精度（precision）、召回率（recall）、FB1 值（FB1）评测如下表：

表 4 命名实体识别测试结果

accuracy	precision	recall	FB1
97.53%	71.35%	79.85%	75.36

表中可以看出准确度比精度、召回率、FB1 值高很多。通过观察数据发现，测试集的标注结果并没有标注完全，有较多的真实标签未标注出来，但是模型预测时却成功预测出来了，因此才导致该现象的出现。

2.2.5 基于地点的二次归类

在 2.2.3 小节中利用 DBSCAN 聚类得出了一次归类结果，此时利用 2.2.4 小节训练好的实体识别模型识别出留言主题中出现的地点或人群，部分识别结果如下：

表 5 命名实体识别部分数据

留言编号	留言主题	label	实体识别结果
188809	A 市万家丽南路丽发新城居民区附近搅拌站扰民	0	a 市万家丽南路丽发新城居民区
189950	投诉 A2 区丽发新城附近建搅拌站噪音扰民	0	a2 区丽发新城
190108	丽发新城小区旁边建搅拌站	0	丽发新城小区
190523	A 市丽发新城违建搅拌站，彻夜施工扰民污染环境	0	a 市丽发新城
190802	A 市丽发小区建搅拌站，噪音污染严重	0	a 市丽发小区
195095	魅力之城小区临街门面油烟直排扰民	0	魅力之城小区
199379	A2 区丽发新城附近修建搅拌厂，严重污染环境	0	a2 区丽发新城
200576	A2 区刘家冲路附近工地昼夜施工，严重扰民	0	a2 区刘家冲路
203393	A 市丽发新城小区侧面建设混泥土搅拌站，粉尘和噪音污染严重	0	a 市丽发新城小区
205329	A3 区天顶街道燕联小区 3 栋一 KTV 噪音扰民	0	a3 区天顶街道燕联小区
205483	A2 区金石蓉园菜市场噪音扰民	0	a2 区金石蓉园菜市场
206431	A7 县橄榄城小区新安路货车严重扰乱了周边居民的生活	0	a7 县橄榄城小区新安路
208285	投诉小区附近搅拌站噪音扰民	0	
208714	A2 区丽发新城附近修建搅拌站，污染环境，影响生活	0	a2 区丽发新城

表中结果可以看出，实体识别结果非常准确，但是只利用留言主题来识别地点的话，由于主题中可能未出现地点信息，故地点实体识别不出来。此时在实验中将利用留言详情中出现的地点作为补充。

利用实体识别模型识别出所有类别中留言的地点实体后，根据地点利用 DBSCAN 算法二次归类，得到最终的归类结果。此时二次归类时要选择合适的聚类参数 eps，初步观察出区间[5,6]，然后进行步长 0.01 的网格化搜索得到最终的 eps=5.7，利用地点的词向量进行 DBSCN 聚类的部分结果如下：

表 6 二次归类部分结果

loc	label
a 市万家丽南路丽发新城居民区	0
a2 区丽发新城	1
丽发新城小区	1
a 市丽发新城	1
a 市丽发小区	1
魅力之城小区	2
a2 区丽发新城	1
a2 区刘家冲路	3
a 市丽发新城小区	1
a3 区天顶街道燕联小区	4
a2 区金石蓉园菜市场	5
a7 县橄榄城小区新安路	6
丽发新城小区	1
a2 区丽发新城	1

表中可以看出，二次归类的结果中标红的行应该在类别 1 中，但是由于其字数太长的原因导致算法无法很好的进行归类。因此，二次归类还有很大的改进空间。

2.2.6 主题关键词抽取算法

在二次归类完成后便得到了特定地点或人群的候选的热点问题，为了得到热点问题的主题，将利用$TextRank$算法提取关键词：

输入：文本集$D = \{T_1, T_2, \cdots, T_n\}$

输出：关键词集$R = \{K_1, K_2, \cdots, K_n\}$ 其中$K_i = \{K_{i1}, K_{i2}, \cdots, K_{in}\}$，$K_{ij}$为一个关键词

（1）遍历文本集D得到T_i

（2）把给定的文本 T_i按照完整句子进行分割，即：$T = [s_1, s_2, s_3, \cdots, s_n]$

（3）对于每个句子，进行分词和词性标注处理，并过滤掉停用词,只保留指定词性的单词，如名词，动词，形容词，得到$S_i = t_{i0}, t_{i1}, \cdots, t_{in}$ 其中t_{ij}是保留后的候选关键词。

（4）构建候选关键词图 $G = (V, E)$，其中 V 为节点集，由（2）生成的候选关键词组成，然后采用共现关系构造任两点之间的边，两个节点之间存在边，仅当它们对应的词汇在长度为 K 的窗口中共现，K 表示窗口大小，即最多共现 K 个单词。

（5）根据 $TextRank$ 的公式，迭代传播各节点的权重，直至收敛。

$$WS(v_i) = (1-d) + d* \sum_{v_j \in (V_j)} \frac{W_{ji}}{\sum_{v_k \in Out(V_j)} W_{jk}} WS(V_j)$$

（6）对节点权重进行倒序排序，从而得到最重要的 T 个单词，作为候选关键词。

（7）由（5）得到最重要的 T 个单词，在原始文本中进行标记，若形成相邻词组，则组合成多词关键词。将得到的K_{ij}存到K_i中

利用上述算法对二次分类的每个类别分别提取排名前 10 的关键词。

2.2.7 评价指标体系的建立

根据二次分类结果中各类别出现的留言次数以及留言关键词出现的次数确定指标体系评价热度指数hot。具体公式如下：$hot_i = msg_i + key_i$。

其中hot_i为第i类问题的热度指数msg_i为第i类的一条留言，key_i为与第i类主题有关的一个关键词。

利用公式计算出每个类别的热度指数，热度指数越大，代表该类问题越突出。最终选出热度指数排名前 5 的结果如表 7：

表 7 热度指数前 5 结果

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	313	2019-07-03 至 2020-01-25	A 市丽发新城小区	搅拌站污染环境，噪音扰民
2	2	247	2019-07-07 至 2019-09-01	A 市伊景园滨河苑	捆绑销售车位，欺诈消费者
3	3	98	2019-01-09 至 2019-09-12	a3 区西湖街道茶场村	何时启动拆迁
4	4	82	2019-01-03 至 2019-11-13	工人	被一些公司拖欠工资
5	5	58	2019-07-28 至 2019-09-25	a5 区劳动东路魅力之城小区	油烟扰民，烧烤店扰民

2.3 问题三分析方法与过程

2.3.1 流程图

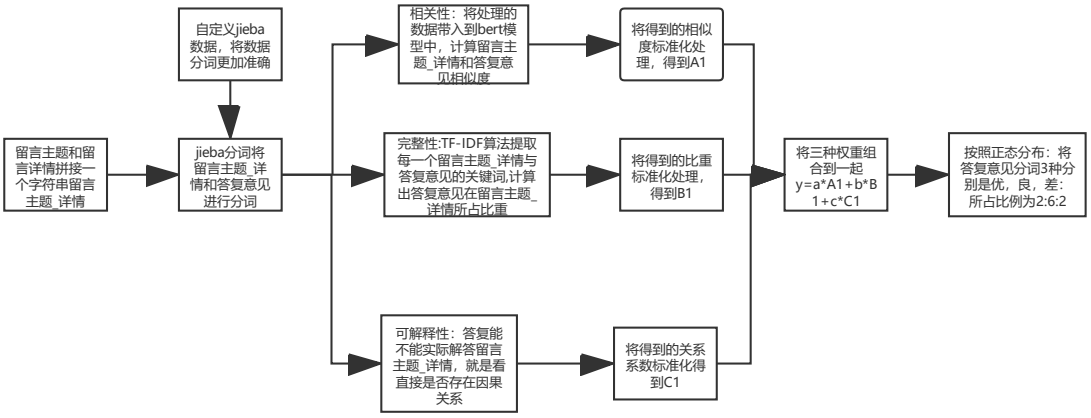

图 6 问题 3 流程图

2.3.2 评价指标体系结构

随着互联网时代到来，网络问政体系在政府部门发挥着重要作用，能够有效的知道群众的热点问题，但是政府的回复仍然存在一些推卸责任，回复不当的现象。因此建立适合的评价指标体系结构有利于提高政府的回复效率，以此来更好的监督政府工作。

本文将从留言答复数据的相关性、完整性、可解释性三个方面来建立评价指标体系：

1. 相关性：对比留言详情（留言主题）与答复意见的相似度。
2. 完整性：留言详情（留言主题）中的关键词，答复意见是否都有提及到。
3. 可解释性：答复意见是否能够真正的解决用户的留言

2.3.3 数据预处理

在对留言详情进行挖掘分析之前，先把非结构化的文本信息转换为计算机能够识别的结构化信息。在附件四中，以中文文本的方式给出了数据。为了便于转换，先对这些数据进行中文分词，这里采用 python 的中文分词包 jieba 进行分词。

在分词的同时，采用 TF-IDF 算法提取每一个留言详情（留言主题）与答复意见的关键词，这里采用的 jieba 自带的语义库。

2.3.4 数据对照筛选

以附录四中编号为 2549 数据为例，将留言主题和留言详情合并成留言主题_详情与答复意见带入 bert 模型中计算出相似度为 700.6，标准化为 0.99；提取其中的关键词可以得到['物业公司', '小区', '投票', 'A2 区', '业主', '景蓉华苑', '业委会', '高昂', '水电', '收费', '业主大会', '公安干警', '投票箱', '2019', 'A 市', '美顺', '4.23', '0.64', '以交', '20']，['业主大会', '业委会', '业主', '停车', '胡华衡', 'A2 区', '景蓉', '2019', '花苑', '反映', '感谢您', '建设局', '物业管理', '物业公司', '留言', '现将', '桂花', '意见', '来信', '栏目']，计算出完整性值为 0.3，标准化为 0.3；计算出可解释性为 698.2，标准化为 0.99。带入权重方程，可以得到评价分数为 0.72;分析总体数据可以看出 0.58 在总体数据中是属于较高的数据，因此该回答的评价为优。同样的，我们计算出附录四所有的答复评价指标（见附录四）。

三、结果分析

3.1 问题一的结果分析

3.1.1 分类模型选择结果分析

本文尝试使用了三种不同的机器学习模型进行分类，其中包括 SVM 模型、朴素贝叶斯模型和 LGB 分类模型，经过训练后发现这三类模型均具有良好的分类效果，其中 SVM 模型的分类效果最有，整体准确率达到了 0.93，因此本例最终选择 SVM 模型作为最终的分类模型。

3.1.2 分类模型评估结果分析

在选取 SVM 支持向量机模型作为本问的最终分类模型后，利用混淆矩阵，生成显示预测标签和实际标签之间的差异如图 7 所示：

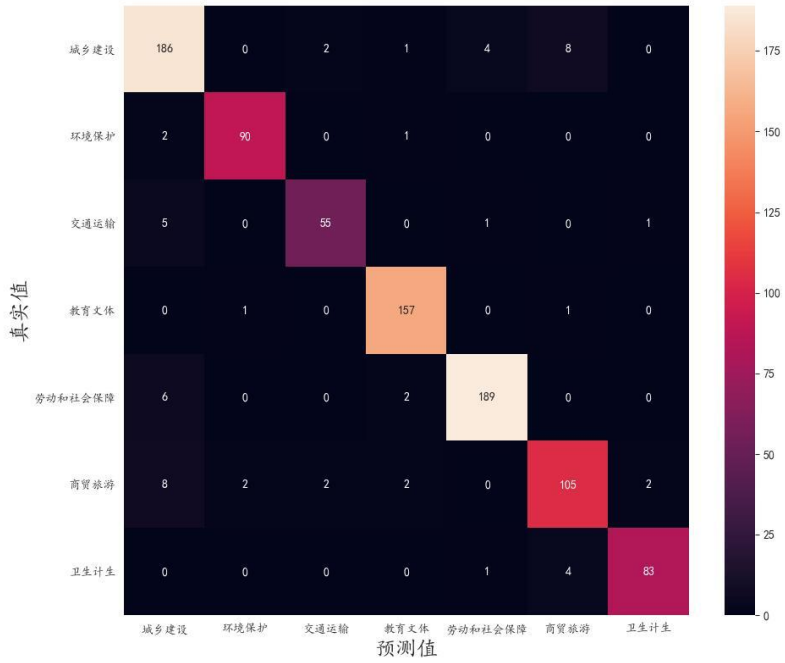

图 7 差异图

由图可知，模型对交通运输类的文本分类预测较为准确，对城乡建设类预测的错误较多。由于多酚类数据模型的准确率不能正确反映出每一个分类的准确性，当训练数据不均衡是，准确率并不能反应出模型的实际预测精度，故增加F_1分数指标进行评价，计算公式如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \tag{30}$$

其中P_i为第i类的查准率，R_i为第i类的查全率。得到最终结果如表 9 所示：

表 9 模型验证结果

	precision	recall	f1-score	support
城乡建设	0.87	0.9	0.88	201
环境保护	0.98	0.95	0.96	94
交通运输	0.92	0.95	0.94	61
教育文体	0.95	0.98	0.97	159
劳动和社会保障	0.93	0.94	0.94	197
商贸旅游	0.88	0.83	0.85	121
卫生计生	0.95	0.9	0.92	88
平均值	0.93	0.92	0.92	131.57

根据F_1指标可知，环境保护和教育文体的得分非常好，分别为 0.96 和 0.97，交通运输、劳动和社与卫生计生的得分次之，分别是 0.94、0。94 和 0.9，得分最差的是城乡建设和商贸旅游。因为这七类的数据数量分布不均衡，所以混淆矩阵并不准确的原因较为合理。

3.2 问题二的结果分析

3.2.1 对命名实体识别的分析

利用自己人工构建的数据进行模型训练，最后利用测试集进行判断模型好坏，发现最终的结果既有好的一面又有差强人意的一面。由于时间和繁琐问题，自行构建的数据集并没有将全部的地点、人群信息使用 BIO 方法标注出来，导致最后测试集的真实答案可能并不完美，在对模型的预测结果进行评测时也就无法很好的从评测结果中判断模型的好坏，但是总体评测结果还是比较满意的，并且从最后的识别效果来看，也是很好的。

3.2.2 对第二次聚类的结果分析及改进

第二次聚类仍然采用 DBSCAN 聚类算法进行地点的聚类，发现其中有少部分的结果聚类错误，例如：

表 9 二次归类误差示例

留言编号	loc_result	loc_label	留言编号	loc_result	loc_label
188809	a 市万家丽南路丽发新城居民区	0	235362	暮云街道丽发新城小区	0
189950	a2 区丽发新城	0	239336	a 市 a2 区丽发新城小区	0
190108	丽发新城小区	0	239648	a 市 a2 区丽发新城小区	0
190523	a 市丽发新城	0	244335	a 市暮云街道丽发新城社区	0
192898	a7 县星沙街道吾悦广场	0	253040	a2 区丽发新城	0
199379	a2 区丽发新城	0	258242	a 市暮云街道丽发新城社区	0
203393	a 市丽发新城小区	0	258378	丽发新城社区	0
208714	a2 区丽发新城	0	259788	a 市暮云街道丽发新城社区搅拌厂	0
213464	丽发新城小区	0	260979	a 市暮云街道丽发新城小区	0
214282	a 市丽发新城小区	0	262090	a7 县星沙商业乐园	0
215563	a2 区丽发新城小区	0	264944	a2 区丽发新城	0
216824	丽发新城小区	0	267050	a2 区丽发新城	0
217700	丽发新城小区	0	268109	a 市 a2 区丽发新城小区	0
219174	a2 区丽发新城小区	0	268300	a2 区丽发新城	0
220482	a6 区月亮岛街道乾源国际广场	0	272224	丽发新城小区	0
222279	a7 县星沙尚都花园城	0	272361	丽发新城小区	0
222831	a2 区丽发新城	0	273282	a2 区丽发新城	0
225217	a2 区丽发新城	0	274004	a 市暮云街道丽发新城社区	0
230375	a3 区东方红街道金南家园三期	0	281943	a2 区丽发新城小区	0
231136	a2 区丽发新城	0	284485	a 市丽发新城小区	0
233158	丽发新城小区	0	190802	a 市丽发小区	1
281546	丽发新城小区	0	247160	a 市丽发小区	1
243692	丽发新城小区	0			

从表中可以看出，二次聚类后标红的 5 处地点应该与其他地点属于不同的类别，而标蓝色的 2 处地点的 local_label 应该为 0，说明 DBSCAN 并不能很完美的解决根据地点二次归类的问题，此时应该结合规则来初步筛选之后再进行 DBSCAN 聚类。制定规则：1. 删除地点中出现的‘小区’、‘街道’、‘市’等通用词语；2. 统计地点中出现的最频繁的两个词 a 和 b，若 a 和 b 都未在识别的地点 loc_result 中出现，则将

该 loc_result 从该类别中剔除；3. 若一次聚类结果的 loc_result 中出现了 a 或 b，则将其加入到该类别中。例如上表：可以知道 a=丽发，b=新城。而标红的 5 处地点均为出现 a 和 b，故从类别中剔除该 5 个地点的留言，标蓝的两个地点由于均出现了丽发，加入到该类别。按照规则处理后得到如下表格 10：

表 10 二次归类误差消除结果

留言编号	loc_result	loc_label	留言编号	loc_result	loc_label
188809	a 市万家丽南路丽发新城居民区	0	235362	暮云街道丽发新城小区	0
189950	a2 区丽发新城	0	239336	a 市 a2 区丽发新城小区	0
190108	丽发新城小区	0	239648	a 市 a2 区丽发新城小区	0
190523	a 市丽发新城	0	244335	a 市暮云街道丽发新城社区	0
199379	a2 区丽发新城	0	253040	a2 区丽发新城	0
203393	a 市丽发新城小区	0	258242	a 市暮云街道丽发新城社区	0
208714	a2 区丽发新城	0	258378	丽发新城社区	0
213464	丽发新城小区	0	259788	a 市暮云街道丽发新城社区搅拌厂	0
214282	a 市丽发新城小区	0	260979	a 市暮云街道丽发新城小区	0
215563	a2 区丽发新城小区	0	267050	a2 区丽发新城	0
216824	丽发新城小区	0	268109	a 市 a2 区丽发新城小区	0
217700	丽发新城小区	0	268300	a2 区丽发新城	0
219174	a2 区丽发新城小区	0	272224	丽发新城小区	0
222831	a2 区丽发新城	0	272361	丽发新城小区	0
225217	a2 区丽发新城	0	273282	a2 区丽发新城	0
231136	a2 区丽发新城	0	274004	a 市暮云街道丽发新城社区	0
233158	丽发新城小区	0	281943	a2 区丽发新城小区	0
281546	丽发新城小区	0	284485	a 市丽发新城小区	0
243692	丽发新城小区	0	190802	a 市丽发小区	0

可以看到，现在的二次归类非常准确了。

3.2.3 对热点问题结果的分析

通过对排名前 5 的热点问题的数据进行挖掘，发现人们普遍抱怨的问题是扰乱群众正常生活和家乡建设发展的问题，例如噪音问题，环境污染问题，拖欠工资，家乡建设等。这些问题都反应出人们越来越注重生活质量，越来越注重幸福生活，紧跟着习近平特色社会主义核心价值观，表达出自己对幸福生活的期望。同时人们也希望相关部门能够关心人们生活，积极解决出现的扰民等各种问题，给人们一个更加和谐的居住和生活环境。

3.3 问题三的结果分析

分析所得到的结果，可以得到下表 11:

表 11 答复评价结果

答复评价指标	数目	评价分数平均值
优	563	0.74

良	1716	0.67
差	564	0.60

可以看出”优”与”差”的评价分数平均值优很大的差别，所以答复意见中还是有许多并不能真正的去解决群众的实际问题的。比如附录四中编号为 2549 的数据，可以看出答复意见很明显是在回答留言中所提出的问题，因此评价分数给出了一个很高 0.72 分。而附录四中编号为 179880 的数据，答复意见并不能很好回答留言中所提出的问题的，而说“您好，您所反映的问题，已转交相关部门调查处置。”，因此我们将这种回答放在了“差”的答复意见中。建立好的答复评价指标可以很好的监督政府的工作能力。回复是推进工作落实的基本要求回复作为工作中的基本程序,在推进工作高效落实中起着重要作用。好的答复意见能够有效的帮助人们群众去解决生活中的问题，但是“差”的答复意见只能让群众感觉政府部门“工作没落实,不敢回复”等这种情况。因此一个好的答复评价指标能够提升政府工作效率。

3.4 结合研究结果，提出建议

1. 针对群众提出的各种留言问题，将每一条留言都看到并给出满意答复是不现实的，此时可以利用本文提出的解决方法发现留言中的热点问题，相关部门一定要对热点问题进行解决，遵从为人民服务的宗旨，合理解决相关问题。

2. 可以采取相应的宣传措施让群众了解和积极进行留言，确保广大人民群众都有发言的机会，这样才能真正发现留言中的热点问题，切实保护好人民群众的利益，提高人民群众的幸福度。

3. 政府人员能够使用的本文中所建立的评价指标体系能够对自己答复意见进行评价，将自己的评价分成三类，以此来提高政府的工作效率。

4. 希望以后政府部门在提出答复意见之后，能够让群众对答复意见有所反馈。

四、结论

1. 通过对留言中热点问题的挖掘，本文发现热点问题中大都与群众生活息息相关的，比如噪音、工资、车位等问题。这些问题往往发生在基层，若无群众反映，政府很难主动发现，因此利用群众的留言挖掘出热点问题是非常有必要的，能够帮助政府更好的建设城市，更好的为人民服务。

2. 及时发现热点问题并进行解决很重要，因为有的问题持续时间长，严重并长时间干扰了群众的正常生活。

3. 针对第三问所得出的结果，我们可以看出答复意见中还是存在一部分“优”，说明答复意见还是很大程度上能够帮助人们解决生活的问题。

4.在问题三的结果中，我们可以看出答复意见中还是存在一部分“差”，而评价指标不能够单纯只看答复意见与留言的关系，我们还应该关注群众对留言的反馈是什么样的，才能够提高评价指标的精度。

参考文献

[1] 石凤贵.基于 TF-IDF 中文文本分类实现[J].现代计算机,2020(06):51-54+75.

[2] 朱衍丞,蔡满春,芦天亮,石兴华,丁祎姗.基于 SVM 的融合多特征 TextRank 关键词提取算法[J].软件导刊,2020,19(02):88-91.

[3] 卢佳伟,陈玮,尹钟.融合 TextRank 算法的中文短文本相似度计算[J/OL].电子科技 ,2020(10):1-8[2020-05-08].http://kns.cnki.net/kcms/detail/61.1291.TN.20191122.1644.032.html.

[4] 张松. 基于 SVM 和 Adaboost 的多分类算法研究[D].山东师范大学,2019.

[5] 黄勇,罗文辉,张瑞舒.改进朴素贝叶斯算法在文本分类中的应用[J].科技创新与应用,2019(05):24+27.

[6] Devlin, Jacob and Chang,et a1.BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[7] Erik F. Tjong Kim Sang, Fien De Meulder.Introduction to the CoNLL-2003 Shared Task: Language Independent Named Entity Recognition[C].Proceedings of the Seventh Conference on Natural Language Learning at {HLT}{NAACL} 2003, 2003, 142-147.

[8] 俞鸿魁, 张华平, 刘群. 基于层叠隐马尔可夫模型的中文命名实体识别[J]. 通信学报, 2006, 27(2): 87—93.

[9] Vlachos A, Gasperin C. Bootstrapping and evaluating named entity recognition in the biomedical domain[C]/Proceedings of the HLTNAACL BioNLP Workshop on Linking Natural Language and Biology. New York: Association for Computational Linguistics Morristown, 2006: 138—145.