“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。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。

对于问题一，使用 python 第三方库 pandas 对数据进行去重去空，得到不重复且完整的留言信息。利用 jieba 中文分词工具对留言详情进行分词去停用词，并通过 TF-IDF 算法生成词频向量然后生成 TF-IDF 向量，对数据进行特征值的抽取。完成文本 TF-IDF 的特征值抽取后采用 sklearn 中的先验为多项式分布的朴素贝叶斯分类器对群众留言内容进行模型训练。最后，使用 F1-score 对分类方法对模型进行评价，得出分类结果。

对于问题二，为提取热度前五的留言信息，我们对留言文本数据进行相似留言的归类，再把所有指标同向化，将极小型指标转化为极大型指标，将提取出的指标数据进行无量纲化处理。运用 TOPSIS 综合评价法建立评价模型对数据的热度进行综合评价，基于评价得分对数据进行排序，按格式输出热度前五的留言数据。

对于问题三，为了对答复意见的质量进行评价，我们首先确定并提取可用指标 ，构建答复意见质量评价模型，计算指标权重。运用 RSR 秩和比综合评价法对答复意见质量进行评价并分级，最后输出评价结果。

关键词：去重 中文分词 TF-IDF 算法 相似度分析 TOPSIS 评价法 RSR 秩和比综合评价

Abstract

In recent years, with the online questioning platforms such as WeChat, Weibo, Mayor ’s Mailbo x, and Sunshine Hotline, etc., it has gradually become an important channel for the government to unde rstand public opinion, gather people ’s wisdom, and gather public opinion. The work of the relevant de partments, which mainly relied on manual work to divide the message and organize hotspots, brought g reat challenges. At the same time, with the development of big data, cloud computing, artificial intellig ence and other technologies, the establishment of a smart government system based on natural languag e processing technology has become a new trend of social governance innovation and development, wh ich has a great impact on improving the government's management level and governance efficien cy Promote the role.

For question one, use the python third-party library pandas to deduplicate and empty the data, and ge t a unique and complete message. Use jieba Chinese word segmentation tool to segment the message de tails to stop words, and generate word frequency vector through TF-IDF algorithm and then generate T F-IDF vector to extract the feature value of the data. After the feature value extraction of the text TF-ID F is completed, a simple Bayesian classifier with a priori polynomial distribution in sklearn is used t o model the mass message content. Finally, use F1-score to evaluate the classification method to get th e classification results.

For question two, in order to extract the top five message information, we classify similar text messa ges on the message text data, and then all the indicators are isotropic, turning very small indicators int o very large indicators, and extracting the extracted indicator data. Dimensionless processing. Use the T OPSIS comprehensive evaluation method to establish an evaluation model to comprehensively evaluat e the heat of the data, sort the data based on the evaluation score, and output the top five message data a ccording to the format.

For question three, in order to evaluate the quality of the response opinion, we first determine and ex tract the available indicators, construct a response opinion quality evaluation model, and calculate the i ndex weight. Use the RSR rank sum ratio comprehensive evaluation method to evaluate and grade th e quality of the reply opinions, and finally output the evaluation results.

Keywords: deduplication Chinese word segmentation TF-IDF algorithm
Similarity analysis TOPSIS RSR

目录

1. 挖掘目标

本次建模目标是利用网络问政平台获得的群众建议留言、答复意见的文本数据，在对文本进行基本的数据预处理、中文分词、去停用词后，用高质量的数据通过无监督的朴素贝叶斯模型、classification_report 分类评价方法、TOPSIS 综合评价模型、RSR 综合评价模型等多种数据挖掘模型和评价模型，实现对文本的留言的分类、归并、评价和排序，以期望得到有隐藏的有价值的内容，并达到以下目标：

（1）利用文本分词和文本聚类方法对非结构性数据进行文本挖掘，建立关于留言内容的一级标签分类，以便后续处理相关问题时方便通过查找标签进行处理。

（2）对群众留言进行分析，在对留言文档进行相似度分析，对指标进行同向化出来后，建立热度评价体系，取出热度前五的留言，方便政府对热度高的留言进行处理。

（3）构建留言回复质量的评价体系，提取出文本中隐藏的指标，方便寻找出评价的方法，在找到评价方法后，对回复质量进行评级，如此可以更清晰明了的看出回复的质量。

2. 分析方法与过程

问题分析总体流程如下图

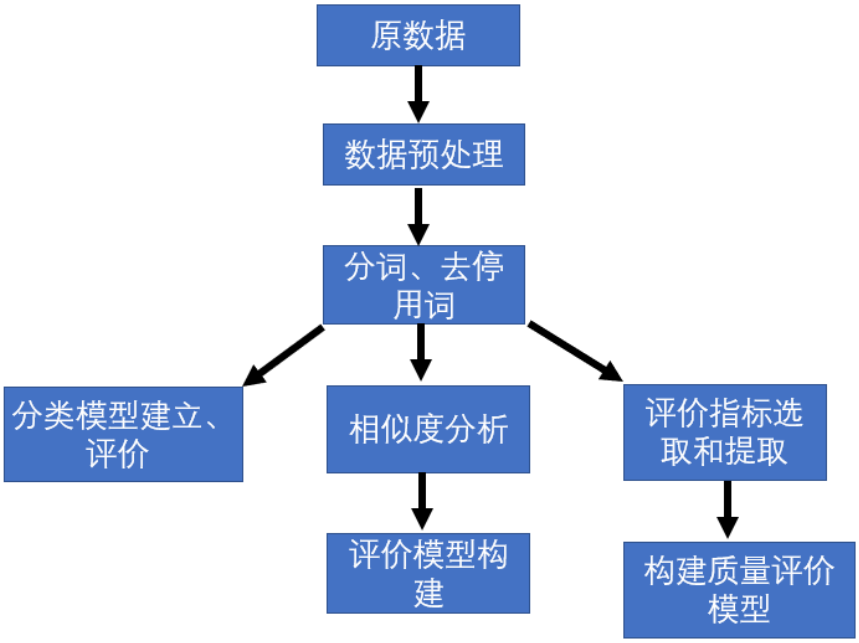


图 1 问题总体流程图

图例解析：

步骤一：对原数据进行分析，从中获取有用的数据。

步骤二：对获取的数据进行数据预处理，对文本数据进行中文分词、去停用词等处理将必要数据进行无量纲化处理。

步骤三：对进行过基本处理后的数据进行多角度分析，运用多种模型和方法对数据进行挖掘得到结果。

步骤四：分析结果中的隐藏价值。

步骤五：模型评估，对分类方法进行评价。

2.1 问题一的分析方法与过程

2.1.1 流程图

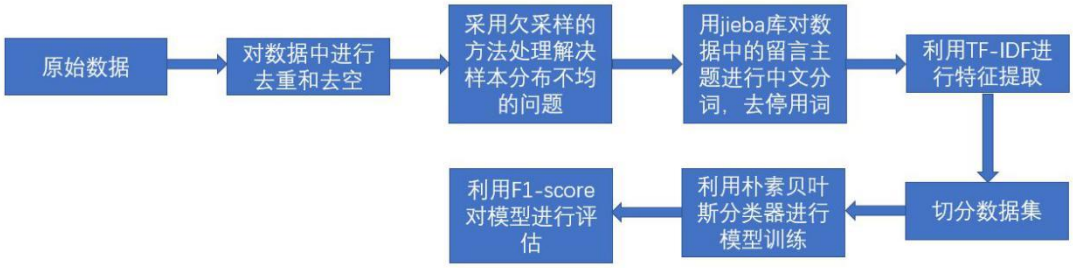


图 2 问题一流程图

2.1.2 数据预处理

2.1.2.1 对附件 2 群众留言信息进行欠抽样

在题目给出的数据中各个类别的数据数量不同，甚至有很大的差异，即数据不平衡。在数据不平衡时，默认的阈值会导致模型输出倾向与类别数据多的类别。为了解决该问题，采用欠抽样的方法，以数量最少的分类的数量为标准，随机抽样样本均衡保证模型记住各类样本特征。

2.1.2.2 对附件 2 群众留言信息的去重、去空

在题目给出的数据中，可能会出现重复的留言数据。例如同一 ID 或不同 ID 发布相同的留言内容。所以需要根据留言主题对数据进行去重处理。导入文件时使用的是 Python 的第三方库 pandas 库，返回 DataFrame 格式的数据，而针对 DataFrame 格式的数据可使用 drop_duplicates（）函数删除重复值。同时导入的文件中可能会有缺失值，即空的记录，干扰了问题的分析，采取 DataFrame.dropna()函数，将含有空值的行从文本中删除。对招聘数据去重、去空的 python 程序见附件。

2.1.2.3 对留言主题进行中文分词

在对留言信息进行挖掘分析之前，先要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 2 中，以中文文本的方式给出了数据。该题中，依据留言主题对留言内容进行分类。为了便于转换，先要对这些职位描述信息进行中文分词。这里采用 python 的中文分词包 jieba 进行分词。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(DAG)， 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，使得能更好的实现中文分词效果。

中文分词后，数据中还存有大量对挖掘分析无意义的词汇、符号等内容。因此导入了停用词文本文件，利用该文件去除数据中的大部分无意义内容。

本文采用 python 的 jieba 库对附件二的留言主题进行分词去停用词部分结果如下：

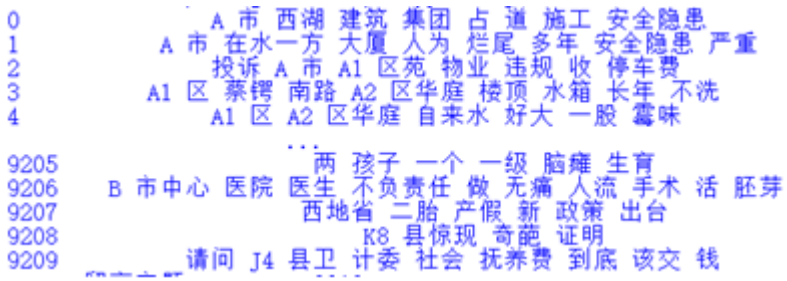


图 3 分词去停用词部分结果

2.1.2.4 TF-IDF 算法

对留言信息进行分词后，需要把这些词语转换为向量，以供后续挖掘使用。这里采用 TF-IDF 算法，把留言信息转换为权重向量。TF-IDF 算法的具体原来如下所示：

第一步：计算词频，即 TF 权重（Term Frequency）

词频（TF）=某个词在文本中出现的次数

考虑到留言有长短之分，为了便于不同留言的比较，对“词频”进行标准化，除以文章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：

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

或

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

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），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。

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

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）

$$TF-IDF = 词频(TF) \times 逆文档频率(IDF)$$

某个词文本的重要性越高，TF-IDF 值越大，计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为我们所需要的关键词。

2.1.3 多分类模型训练

完成文本 TF-IDF 的特征值抽取后，用 train test split 的方法对数据集进行切割其中 80%作为训练集，20%作为测试集。本题中采用 sklearn 中的先验为多项式分布的朴素贝叶斯分类算法以及支持向量机（SVM）算法对群众留言内容进行分类。最终根据预测准确率及评价得分选择最优模型。

1、朴素贝叶斯分类算法

贝叶斯算法的中心思想是：用贝叶斯定理计算需要预测的数据集的概率，根据训练数据

集进行分类，概率最大的就为预测数据集的类别。

假设 A、B 分别为两个事件,将事件 A 发生的概率记作 P(A),将事件 B 发生的概率记作 P(B),将事件 A、B 同时发生的概率记作 P(AB)

由此在 A 事件发生的条件下，B 事件发生的概率$\mathrm{P(B|A)} = \frac{P(AB)}{P(A)}$，也可以得出在 B 事件发生的条件下，A 事件发生的概率$\mathrm{P(A|B)} = \frac{P(AB)}{P(A)}$

由条件概率可以推出贝叶斯公式$\mathrm{P(A|B)} = \frac{P(B|A)P(A)}{P(B)}$

进一步假设有数据集 X、$\mathrm{Y(X} = (x_1, x_2, x_3, \dots, x_n), \mathrm{Y} = (y_1, y_2, y_3, \dots, y_n)$，且特征之间相互独立互不影响，则可以得到朴素贝叶斯的公式为：

$$\mathrm{P}(Y_m|\mathrm{X}) = \mathrm{P}(Y_m) \times \prod P(X_n|y_m)$$

在分子和分母上分别加上一个数，防止因为多项式一个概率为零从而使整个公式为零。

2、支持向量机（SVM）算法

支持向量机（SVM）算法是一种建立在统计学习理论基础上的机器学习方法。该算法建立于结构风险最小化原理之上，把数据集合压缩到支持向量集合，通过学习得到分类决策函数。该技术解决了以往需要非常大样本数量的问题，它只需要将一定数量的文本通过计算，抽象成向量化的训练文本数据，提高了分类的准确率。

支持向量机（SVM）算法是根据有限的样本信息，在模型的复杂性与学习能力之间寻求最佳平衡点，以获得最好的推广能力。

2.1.4 评价分类模型

使用metrics.scorer.accuracy_score（）函数计算模型整体的预测准确率，准确率的计算公式为：混淆矩阵中主对角线数字之和与所有数字之和的商。不过，该指标只能衡量模型的整体预测效果，却无法对比每个类别的预测精度、覆盖率等信息。为了计算各类别的预测效果，这里使用了classification_report(y_true, y_pred, target_names = None)函数，其中y_true是目标真实值 y_pred 是预测值，target_names是显示需要显示的样本名。该函数用于显示主要分类指标的文本报告，

下图为一个输出报告示例：

	precision	recall	f1-score	support
setosa	1.00	1.00	1.00	10
versicolor	0.83	1.00	0.91	10
virginica	1.00	0.80	0.89	10
micro avg	0.93	0.93	0.93	30
macro avg	0.94	0.93	0.93	30
weighted avg	0.94	0.93	0.93	30

图 4 F-score 示意图

其中列表左边的一列为分类的标签名，右边 support 列为每个标签的出现次数。avg / total 行为各列的均值（support 列为总和）。precision，recall，f1-score 三列分别为各个类别的精确度，召回率及 F1-score 值。

精确度：precision，正确预测为正的，占全部预测为正的比例，$\mathrm{precision} = \frac{TP}{TP+FP}$

召回率：recall，正确预测为正的，占全部实际为正的比例，$\text{recall} = \frac{TP}{TP+FN}$

F1-score：精确率和召回率的调和平均数，$f1 = \frac{2 \times \text{precision} * \text{recall}}{\text{precision}+\text{recall}}$，f1 值越靠近 1 分类模型越好。

同时还会给出总体的微平均值，宏平均值和加权平均值。

微平均值：micro average，所有数据结果的平均值。

宏平均值：macro average，所有标签结果的平均值。

加权平均值：weighted average，所有标签结果的加权平均值。

2.2 问题二的分析方法与过程

2.2.1 筛选出排名前五的热点问题

2.1.1.1 文档相似度分析

相似度分析如图所示：

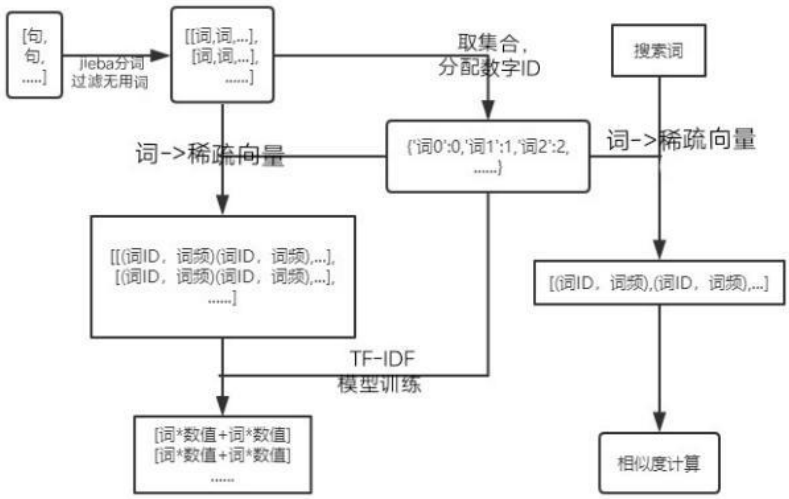


图 5 相似度分析流程图

首先读取附件 3 中留言主题的文字内容，对文档的内容进行分词，将文档进行整理成指定格式，方便后续进行计算；计算出词语的词频，对于词频低的词语进行过滤；建立语料库词典；加载要对比的文档；将要对比的文档 doc2bow 转化为词袋模型；对词袋模型进一步处理，得到新的语料库；将新语料库通过 tfidfmodel 进行处理，得到 tf-idf；最后通过 token2id 得到特征数、稀疏矩阵相似度，从而建立索引、得到最终相似度结果。

经研究发现，相似度大于 0.1 的留言主题可看作同一类留言，统计同一类问题的数量、点赞数、反对数，求出时间范围，将同一类问题的索引存入列表，为解决后续问题做准备。

2.2.1.2 指标属性同向化

TOPSIS 法使用距离尺标来度量原本差距，使用需要对指标属性进行同向化处理（如果一个维度的数据越大越好，而另一个维度的数据越小越好，就会造成尺度的混乱）。通常采用成本型指标向效益型指标转化（即数组越大评价越高）。在本问中由于反对对于评价来说是逆向指标，所以要进行正向化处理。在这里我们采用了极小型指标转化为极大型

指标公式：max − x。

2.2.1.3 计算指标的权重

其主要计算步骤如下：

1、归一化处理

对原始数据矩阵按列进行归一化处理，本文选用比值归一化

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}$$

2、计算各指标的熵值

$$e_j = -k \sum_{i=1}^{n} p_{ij} \ln p_{ij},\ (j = 1,2,\dots,m)$$

其中k 与样本数量有关，常取 $k = 1/\ln n$，此外，若 $p_{ij} = 0$，则令$p_{ij} \ln p_{ij} = 0$

3、计算各指标的权系数

$$h_j = \frac{1 - e_j}{\sum_{k=1}^{m}(1 - e_k)}, (j = 1,2,\dots,m)$$

其中熵权系数 h_j越大，则该指标所代表的信息量越大，表示其对综合评价的作用也就越大。

4、 构造归一化初始矩阵

为了消去不同指标量纲的影响，需要对已经正向化的矩阵进行标准化处理，具体流程如下：

（1）共有 n 个带评价对象，每个对象有 3 个指标，则原始数据矩阵构造为：

$$X = \begin{bmatrix} x_{11} & \cdots & x_{13} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nn} \end{bmatrix}$$

（2）构造加权规范矩阵，属性进行向量规范化，即每一列元素都除以当前列向量的范数（使用余弦距离度量）

$$z_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}}\ \left(\text{每个元素} \Big/ \sqrt{\text{其所在列的元素平方和}}\right)$$

（3）计算得分并归一化

我们有 n 个要评价的对象，3 个评价指标的标准化矩阵：

$$Z = \begin{bmatrix} z_{11} & \cdots & z_{13} \\ \vdots & \ddots & \vdots \\ z_{n1} & \cdots & z_{n3} \end{bmatrix}$$

2.2.1.4 确定最优方案和最劣方案

最优最劣方案Z^+由Z中每列元素最大值、最小值构成：

定义最大值：$Z^+ = (Z_1^+, Z_2^+, Z_3^+)$

$$= (max\{z_{11}, z_{21}, \dots, z_{n1}\}, max\{z_{12}, z_{22}, \dots, z_{n2}\}, max\{z_{13}, z_{23}, \dots, z_{n3}\})$$

定义最小值：$Z^- = (Z_1^-, Z_2^-, Z_3^-)$

$$= (min\{z_{11}, z_{21}, \dots, z_{n1}\}, min\{z_{12}, z_{22}, \dots, z_{n2}\}, min\{z_{13}, z_{23}, \dots, z_{n3}\})$$

2.2.1.5 计算各评价对象与最优方案的接近程度及最终得分，并进行排序

定义第$i(i=1,2,\ldots,n)$个评价对象与最大值的距离$D_i^+=\sqrt{\sum_{j=1}^{3}w_j(Z_j^+-Z_{ij})^2}$

定义第$i(i=1,2,\ldots,n)$个评价对象与最小值的距离$D_i^-=\sqrt{\sum_{j=1}^{3}w_j(Z_j^--Z_{ij})^2}$

其中 w_j为第 j 个属性的权重，本文中用熵值法求出

由此我们可以计算得出第i个评价对象未归一化的得分：$S_i=\frac{D_i^-}{D_i^++D_i^-}$

很明显的可以看出$0\le S_i\le 1$且S_i越大D_i^+越小，也就越接近最大值。根据S_i的大小进行排序，并给出评价结果。

2.2.2 相应热点问题的留言信息

将综合得分表中的排序、问题 ID、热度指数与留言归类表中的时间范围、地点/人群、问题描述合并为一个 DataFrame，根据排序结果对该 DataFrame 进行排序，排序后索引值不变，将其存入 result.xlsx 文件中。最后将前五个热点问题存入热点问题表.xlsx 文件中。

将问题 ID 作为索引值，查找 2.1.1.1 中记录的各类问题索引值列表中对应的元素，即各子问题的索引值，再根据该索引值从附件 3 中对应的数据，与问题 ID 相结合得出热点问题留言明细表。

2.3 问题三分析方法与过程

2.3.1 指标提取

指标提取是对答复意见进行分析的一项重要任务。本题中为了构建答复意见质量指标体系针对答复意见所需要提取的主要特征指标有三个：主题相关度、信息量、可解释性。对于以上三个特征指标可以通过以下不同方式获得：

（1）通过简单的计算可以获得信息量指标，其中信息量就是表中答复意见文本长度，通常是以文本内容的字数或是词数组成。

（2）通过参考、改进公式可以获得可读性指标，关于中文文本可读性的研究，目前还没有可参考的算法，而英文可读性研究相对比较成熟，可参考英文可读性的衡量公式，自动化可读指数 ARI 的值表示，即

$$\mathrm{ARI}=4.71\times\frac{\text{总字符数}}{\text{总字数}}+0.5\times\frac{\text{总字数}}{\text{总句数}}-21.43$$

其中字符数可按照中文一个汉字相当于两个字符来计算，从而获得答复意见的可读性指数.。

（3）获取相关性指标，可以运用 python 中的 difflib 库对两字符串进行比较，获得字符串的相关性。

1、预处理

根据答复意见内容，对数据经过一系列的计算与编程处理，提取出主题相关度指标因素。精确的高质量的模型必须依赖高质量的数据，因此为了保证数据挖掘的正常进行和模型的准确率，要在数据挖掘前进行数据预处理，提高数据的准确度。

（1）对数据中的缺失值进行整行删除

（2）对重复的数据进行去重处理

（3）使用 Python 中的 jieba 库对每条留言详情及答复意见进行分词。

（4）使用停用词文件，剔除答复意见文本内容中的无用字词，比如“的”，“在" 及一些无实际意义的副词、介词等。

（5）经过上述处理后，将分词结果整理为指定格式并保存下来。

2、相关度

获取相关度指标时，我们将答复意见的每一条答复和留言详情中的每一条留言做为对象用 python 中的 difflib 库对数据进行两两差异化比较，求出两数据之间的相似度数值结果。

研究表明,某个留言内容下的答复意见与留言内容相关度越高，该答复意见对留言内容的价值越大，其数据质量就越高。

2.3.2 构建答复意见质量评价指标体系

1、答复意见质量评价指标体系构建过程

（1）从信息质量评价角度出发，前辈们提出的信息质量评价指标把信息特质划分为内在特质、内容特质以及描述特质三个方面，内在特质强调信息的固有特性，内容特质反映信息的含量和完备性，而描述特质则体现信息的语言风格。本文通过分析各种信息质量评价指标并结合本文数据特征，把答复意见内容特征归属到相关度、信息量两个类别中，用于构建答复意见质量评价指标体系

其中，信息量可理解为答复意见信息的含量，如句子量和词汇量，也可表示为平均句长或单句长度。通常认为，答复内容的信息含量越大，答复意见的质量就越高，本研究中取答复意见长度代表信息含量。

相关度主要从内容相关和主题相关两方面分析，描述答复意见与留言内容的相关性，本研究主要计算的是内容相关度，文中也称其为相关度。

（2）从数据角度出发，《数据整理实践指南》的第十九章揭秘数据质量分析时，用“4 个 C"概括了数据质量分析，即完整性、一致性、准确性以及可解释性。同样的，答复意见质量狭义上指的就是答复意见的数据(包括文本和数值)质量，在此我们可以参考“4 个 C"概念，来分析答复意见的质量。由于答复意见中只存在文本数据，不存在数值数据，因此只参考可解释性作为评价指标，具体如下：

可解释性本来指的是可以追踪到数据的来源，但就答复意见而言，可解释性对于中文评论来说没有多大意义，可以将其理解为可读性，答复意见的可读性或读者的理解能力可以用自动化可读性指数 ARI 来表示，在过去的研究中，研究者用可读性来定性地检验一些 英文文本的特征，表示某文本内容吸引人的程度，而 ARI 依赖于文本的字符数，又比其他的可读性测试方法准确率高，因而经常被用于与此相类似的研究中。计算时可以直接采用公式：

$$\mathrm{ARI} = 4.71 \times \frac{\text{总字符数}}{\text{总字数}} + 0.5 \times \frac{\text{总字数}}{\text{总句数}} - 21.43$$

其结果近似等于人们可能理解一段文字的最低程度。国内关于中文可读性的研究并未成熟，没有可使用的公式，因此本文仅借鉴英文可读性计算公式（ARI）作为参考。

结合以上理论，我们可以得到本研究的答复意见质量评价指标体系，分别包含相关度、信息量和可读性，进而全面地衡量答复意见的质量。

那么，由此形成的答复意见质量评价指标如下表所示。

指标	说明
信息量	从内容上确保评论质量，以评论长度衡量(词/字数统计)
相关度	答复意见与留言内容主题的相关性
可读性	用于检测评论的被理解程度，由 ARI 表示

表 1 答复意见质量评价指标表

2、确定评价指标权重

关于答复意见质量评价模型的研究，采用不同的指标权重将会得到完全不同的实验结果，因此选择合适的权重闲的尤为重要。本文采用基于信息论的熵值法来确定指标权重。

基于信息论的熵值法是依据各指标所含的信息有序程度的差异性，来确定指标权重的一种客观赋权方法，仅依赖于数据本身的离散程度。熵用于衡量不确定性，指标的不确定性越大（离散程度越大），熵值越小，表明指标值提供的信息量越多，因此该指标的权重也应越大。主要计算步骤如下：

（1）归一化处理

对原始数据矩阵按列进行归一化处理，本文选用比值归一化

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}$$

（2）计算各指标的熵值

$$e_j = -k \sum_{i=1}^{n} p_{ij} \ln p_{ij},\ (j = 1,2,\dots,m)$$

其中k 与样本数量有关，常取 $k = 1/\ln n$，此外，若 $p_{ij} = 0$，则令$p_{ij} \ln p_{ij} = 0$

（3）计算各指标的权系数

$$h_j = \frac{1 - e_j}{\sum_{k=1}^{m}(1 - e_k)}, (j = 1,2,\dots,m)$$

其中熵权系数 h_j越大，则该指标所代表的信息量越大，表示其对综合评价的作用也就越大。

2. 3. 3 构建答复意见质量评价模型

根据前文提取的指标特征，用 RSR 综合评价法来构建本文的答复意见质量评价模型。RSR 法(Rank-sum ratio) 是古典参数统计和近代非参数统计优点于一体的统计分析方法。

其一般过程是将效益型指标按照从小到大的顺序进行排名、成本型指标按照从大到小的顺序进行排名，再计算秩和比，最后进行统计回归以及分档排序。通过秩转换，获得无量纲统计量 RSR；在这个基础上，运用参数统计分析的概念和方法，研究 RSR 的分布；以 RSR 值对评价对象的优劣进行直接排序或分档排序，由此对评价对象做出综合评价。其计算步骤如下：

1、 列出原始数据表并编秩

（1）整次秩和比法

将 n 个评价对象的 m 个评价指标排列成 n 行 m 列的原始数据表，编出每个指标各评价对象的秩，高优的指标需要从小到大进行编秩，低优指标则需要从大到小进行编秩。得到秩矩阵，记为 $R = \left(R_{ij}\right)_{m \times n}$

（2）非整次秩和比法

对于低优型指标：

$$R_{ij} = 1 + (n-1)\frac{X_{ij} - \min\left(X_{1j}, X_{2j}, \dots, X_{nj}\right)}{\max\left(X_{1j}, X_{2j}, \dots, X_{nj}\right) - \min\left(X_{1j}, X_{2j}, \dots, X_{nj}\right)}$$

对于高优型指标：

$$R_{ij} = 1 + (n-1)\frac{\max(X_{1j}, X_{2j}, \dots, X_{nj}) - X_{ij}}{\max(X_{1j}, X_{2j}, \dots, X_{nj}) - \min(X_{1j}, X_{2j}, \dots, X_{nj})}$$

2、 计算秩和比并排序

$$RSR_i = \frac{1}{n}\sum_{i=1}^{m} w_j R_{ij}$$

3、 确定 RSR 的分布(转化为概率单位)

RSR 的分布是指用概率单位[公式]表达的值特定的累计频率 。RSR 模型是一种广义的线性模型，服从正态分布。其转换方法为：

（1）编制 RSR 频数分布表，列出各组频数f，计算各组累计频数 $\sum f$

（2）确定各组 RSR 的秩次范围以及平均秩次。

（3）计算累计频率$\frac{\bar{R}}{n} \times 100\%$，最后一项记为 $1 - \frac{1}{4n}$进行修正。

（4）将累计频率换算为概率单位，$Probit, Probit$为累计频率对应的标准正态离差 μ 加 5。

4、计算直线回归方程

以累积频率所对应的概率单位 $Probit_i$为自变量，以 RSR 值为因变量，计算直线回归方程，即：

$$RSR = a + b \times Probit$$

4、 检验回归方程输出 RSR 矫正值，并对评价对象进行分档排序

对该回归方程，需要进行检验。

（1）回归系数b的有效性检验：t检验法和置信区间检验法

（2）拟合优度检验：决定系数、Pearson 相关系数、Spearman 秩相关系数、交叉验证法等。

6、输出“RSR 分析结果报告”

3、结果分析

3.1 问题一的结果分析

3.1.1 朴素贝叶斯分类模型评价

1、计算得出朴素贝叶斯分类模型的预测准确度为 0.9758158508158508

2、使用classification_report函数得出的评分报告如下：

	precision	recall	f1-score	support
交通运输	0.75	0.69	0.72	123
卫生计生	0.86	0.90	0.88	123
劳动和社会保障	0.77	0.84	0.81	122
城乡建设	0.83	0.86	0.84	123
商贸旅游	0.84	0.82	0.83	123
环境保护	0.78	0.73	0.75	123
教育文体	0.88	0.86	0.87	122
accuracy			0.82	859
macro avg	0.82	0.82	0.82	859
weighted avg	0.82	0.82	0.82	859

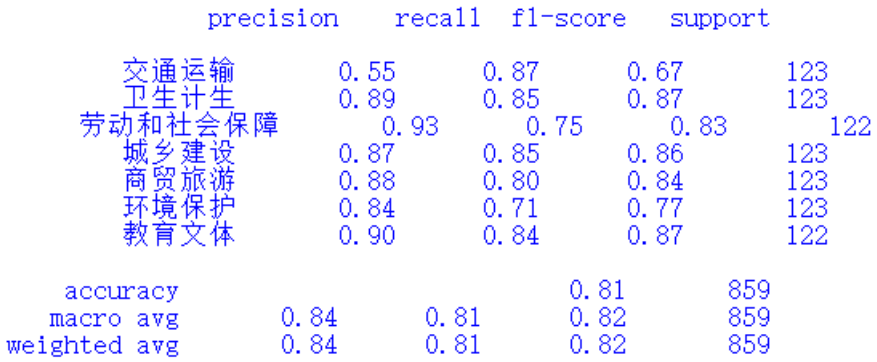
图 6 朴素贝叶斯模型评分报告

如报告所示，使用朴素贝叶斯分类模型对留言内容进行一级标签分类,得到结果的精确率平均值（accuracy）为 0.82，宏平均值（macro_avg）为 0.82，加权平均值（weighted_avg）也为 0.82。

3.1.2 支持向量机（SVM）分类模型评价

1、计算得出朴素贝叶斯分类模型的预测准确度为 0.810244470314319

2、使用classification_report函数得出的评分报告如下：



	precision	recall	f1-score	support
交通运输	0.55	0.87	0.67	123
卫生计生	0.89	0.85	0.87	123
劳动和社会保障	0.93	0.75	0.83	122
城乡建设	0.87	0.85	0.86	123
商贸旅游	0.88	0.80	0.84	123
环境保护	0.84	0.71	0.77	123
教育文体	0.90	0.84	0.87	122
accuracy			0.81	859
macro avg	0.84	0.81	0.82	859
weighted avg	0.84	0.81	0.82	859

图 7 SVM 模型评分报告

如报告所示，使用支持向量机（SVM）分类模型对留言内容进行一级标签分类,得到结果的精确率平均值（accuracy）为 0.81，宏平均值（macro_avg）为 0.82，加权平均值（weighted_avg）也为 0.82.

综合以上结果可以得出，朴素贝叶斯分类模型与支持向量机（SVM）分类模型相比较，朴素贝叶斯分类模型更适用于本题中的群众留言分类。使用 F-Score 对其进行评价，评价结果为 0.82.

3.1.3 朴素贝叶斯模型预测结果显示

将一级标签进行编号，使用朴素贝叶斯模型对测试集进行预测，得到留言主题所属的类别，输出留言主题所属编号，如图 8 所示，真实值标签与预测值标签相差不大，因此朴素贝叶斯模型用于分类比较可靠。

一级标签	一级标签 ID
城乡建设	0
环境保护	1
交通运输	2
教育文体	3
劳动和社会保障	4
商贸旅游	5
卫生计生	6

表 2 标签对应 ID

```
真实值标签 [4, 5, 5, 3, 6, 1, 4, 6, 1, 5, 4, 2, 0, 0, 3, 3, 6, 3, 5, 2]
预测值标签 [4, 5, 2, 3, 6, 1, 4, 6, 1, 2, 4, 2, 2, 0, 3, 3, 6, 4, 5, 2]
```

图 8 分类预测结果

3.2 问题二的结果分析

3.2.1 根据相似度对各类问题归类结果（节选，详见留言归类.xlsx）

根据相似度对留言主题进行归并，count 为相似度高的留言主题的总条数，结果如表 3：

	留言主题	时间范围	count	点赞数	反对数
0	A3区一米阳光婚纱艺术摄影是否合法纳税了?	2019-02-28至2019-10-22	7	5	1
1	咨询A6区道路命名规划初步成果公示和城乡门牌问题	2019-01-22至2019-11-21	12	23	10
2	反映A7县春华镇金鼎村水泥路、自来水到户的问题	2019-01-09至2020-01-08	24	21	0
3	A2区黄兴路步行街大古道巷住户卫生间粪便外排	2019-01-16至2019-12-12	7	2	0
4	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	2019-01-11至2020-01-25	164	85	5
5	A3区麓泉社区单方面改变麓谷明珠小区6栋架空层使用性质	2019-01-16至2019-12-18	32	15	2
6	对A市地铁违规用工问题的质疑	2019-01-01至2020-01-03	52	92	4
7	A市6路公交车随意变道通行	2019-01-07至2019-12-30	58	40	0
8	A3区保利麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2点施工扰民	2019-03-11至2020-01-06	50	40	10
9	A7县特立路与东四路口晚高峰太堵，建议调整信号灯配时	2019-01-01至2019-12-24	50	120	8
10	A3区青青家园小区乐果果零食炒货公共通道摆放空调扰民	2019-01-09至2019-12-27	26	23	10
11	关于拆除聚美龙楚在西地省商学院宿舍旁安装变压器的请求	2019-01-09至2020-01-08	33	19	3
12	A4区北辰小区非法住改商问题何时能解决?	2019-01-02至2019-12-26	107	133	7
13	请给K3县乡村医生发卫生室执业许可证	2019-01-08至2020-01-06	16	0	2
14	咨询异地办理出国签证的问题	2019-01-06至2020-01-05	40	32	12
15	投诉A市温斯顿英语培训学校拖延退费!	2019-02-10至2019-12-30	60	36	7
16	A6区乾源国际广场停车场违章乱建现象严重	2019-01-25至2020-01-06	38	23	2

表 3 留言主题归类结果表

3.2.2 指标属性同向化结果（节选）

	留言主题	时间范围	count	点赞数	反对数
0	A3区一米阳光婚纱艺术摄影是否合法纳税了?	2019-02-28至2019-10-22	7	5	62
1	咨询A6区道路命名规划初步成果公示和城乡门牌问题	2019-01-22至2019-11-21	12	23	53
2	反映A7县春华镇金鼎村水泥路、自来水到户的问题	2019-01-09至2020-01-08	24	21	63
3	A2区黄兴路步行街大古道巷住户卫生间粪便外排	2019-01-16至2019-12-12	7	2	63
4	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	2019-01-11至2020-01-25	164	85	58
5	A3区麓泉社区单方面改变麓谷明珠小区6栋架空层使用性质	2019-01-16至2019-12-18	32	15	61
6	对A市地铁违规用工问题的质疑	2019-01-01至2020-01-03	52	92	59
7	A市6路公交车随意变道通行	2019-01-07至2019-12-30	58	40	63
8	A3区保利麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2点施工扰民	2019-03-11至2020-01-06	50	40	53
9	A7县特立路与东四路口晚高峰太堵，建议调整信号灯配时	2019-01-01至2019-12-24	50	120	55
10	A3区青青家园小区乐果果零食炒货公共通道摆放空调扰民	2019-01-09至2019-12-27	26	23	53
11	关于拆除聚美龙楚在西地省商学院宿舍旁安装变压器的请求	2019-01-09至2020-01-08	33	19	60
12	A4区北辰小区非法住改商问题何时能解决?	2019-01-02至2019-12-26	107	133	56
13	请给K3县乡村医生发卫生室执业许可证	2019-01-08至2020-01-06	16	0	61
14	咨询异地办理出国签证的问题	2019-01-06至2020-01-05	40	32	51
15	投诉A市温斯顿英语培训学校拖延退费!	2019-02-10至2019-12-30	60	36	56
16	A6区乾源国际广场停车场违章乱建现象严重	2019-01-25至2020-01-06	38	23	61

表 4 指标属性同向化结果表

3.2.3 评价指标权重

熵值法求得的权重如下表所示：

指标	权重
count	0.16957936
点赞数	0.82705993
反对数	0.00336071

表 5 指标权重表

3.2.4 确定最优方案和最劣方案

在每一列中分别选取最大值、最小值构成最优最劣方案：

	count	点赞数	反对数
负理想解	0.001578	0.000000	0.000000
正理想解	0.258771	0.596251	0.043326

图 9 优劣方案图

3.2.5 各评价对象评价结果（节选）

	count	点赞数	反对数	正理想解	负理想解	综合得分指	排序
0	0.01086	0.001072	0.043225	0.550684	0.004707	0.008475	424.5
1	0.018617	0.004929	0.03695	0.546646	0.008625	0.015534	323
2	0.037235	0.0045	0.043922	0.545678	0.015502	0.027623	190
3	0.01086	0.000429	0.043922	0.551258	0.004644	0.008355	435.5
4	0.260645	0.01843	0.040436	0.525169	0.108267	0.17092	9
5	0.049647	0.003215	0.042528	0.545989	0.020219	0.03571	125
6	0.094639	0.020359	0.041134	0.527881	0.042713	0.074857	28
7	0.09619	0.009215	0.043922	0.537845	0.040032	0.069274	34
8	0.076021	0.008144	0.03695	0.539923	0.031689	0.055439	57
9	0.077573	0.025717	0.038345	0.524022	0.039192	0.069586	33
10	0.040338	0.004929	0.03695	0.545079	0.016765	0.029839	170
11	0.051198	0.004072	0.041831	0.545118	0.020967	0.037038	121
12	0.166006	0.028503	0.039042	0.51749	0.072689	0.123164	13
13	0.024823	0	0.042528	0.550593	0.009925	0.017706	297
14	0.062058	0.006858	0.035556	0.541922	0.025824	0.045485	82
15	0.093087	0.007715	0.039042	0.539361	0.038494	0.066615	40

表 6 评价结果表

3.2.6 排名前 5 的热点问题

根据评级结果提取出热度排名前五的问题并按照规定格式输出，结果如表 7：

	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
0	1	148	0.852045	2019-01-(A1区汇一城	A1区汇一城小区高层住宅与变电站距离不到15米	
1	2	101	0.720487	2019-01-1	A2区五矿紫湖香醍为何经常停水	
2	3	95	0.609769	2019-03-1	A7县	在A7县里沙水痘疫苗及腮腺炎疫苗是入学必须补种的吗？
3	4	344	0.578432	2019-01-1	A市	严惩A市58车贷特大集资诈骗案保护伞
4	5	74	0.301578	2019-01-(A市人民西		请清理A市人民西路137号人行道附近的僵尸车

表 7 问题结果表

根据表 6 所显示的内容可以得知，表中五项问题在民众所提出的问题中关注度最大，应针对

这五类问题尽早给出让民众满意的答复。

3.2.7 相应热点问题对应的留言信息（节选，详见热点问题留言明细表.xlsx）

按照热度排名结果，找出每排名中的每一条数据并输出，部分结果如表 8 所示：

	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
0	148	195252	A0001095	A1区汇一	2019-10-	最近在我	0	0
1	148	208636	A0007717	A市A5区汇	2019-08-	我是A市A	0	2097
2	148	212171	A0001557	A2区月塘	2019-08-	A市A2区怪	0	0
3	148	213145	A0009625	A市405公	2019-03-	A市405公	0	0
4	148	226753	A0002054	A7县新国	2019-07-	我是A7县	0	0
5	148	240074	A0003218	新买的纯	2019-03-	为响应国	0	1
6	148	240110	A0003502	A7县泉塘	2019-07-	尊敬的县	0	0
7	148	246708	A0001039	A市海德公	2019-01-	星沙海德	0	0
8	148	248942	A0009109	A7县九木	2019-08-	本人现居	0	0
9	148	256073	A0001888	反对在A1	2019-11-	我们汇一	0	0
10	148	263372	A0003264	A7县星通	2019-07-	星通15路	0	0
11	148	263672	A0004144	A4区绿地	2019-09-	您好，近	0	669
12	148	263679	A0004005	A市远鑫尚	2019-05-	远鑫集团	0	0
13	148	264311	A0004588	A市石马铺	2019-04-	1.石马铺	0	0
14	148	266131	A0001068	呼吁停止	2019-09-	沈书记：	4	0
15	148	274250	A909134	A市楚江新	2019-12-	A市楚江新	0	0

表 8 对应热点留言信息

3.3 问题三的结果分析

3.3.1 各评价指标原始数据表（节选）

	信息量	可读性	相似性
0	454	18.64919	0.283871
1	305	6.885591	0.15311
2	357	16.73467	0.201835
3	310	4.032388	0.152866
4	161	1.3615	0.285714
5	232	19.60558	0.122449
6	245	6.409083	0.195918
7	624	6.055538	0.189944
8	505	9.6863	0.172757
9	224	9.216537	0.23622
10	489	8.943136	0.137681
11	427	16.4949	0.130841
12	139	4.599286	0.133971
13	101	-1.20242	0.062257
14	210	7.596563	0.14966

表 9 提取评价指标结果

3.3.2 原始数据表编秩结果及秩序（节选）

	X1：信息量	R1：信息量	X2：可读性	R2：可读性	X3：相似性	R3：相似性	RSR	RSR_Rank
0	0.057234	437.0	0.257401	2099.0	0.429317	1683.0	0.359825	191.0
1	0.038325	292.0	0.171003	1119.0	0.231559	786.0	0.195555	1623.0
2	0.044924	344.0	0.243340	2017.0	0.305248	1184.0	0.292528	654.0
3	0.038959	297.0	0.150048	795.0	0.231190	783.0	0.177649	1770.0
4	0.020051	149.0	0.130431	512.0	0.432105	1687.0	0.204463	1535.0
5	0.029061	220.0	0.264425	2144.0	0.185188	553.0	0.219370	1391.0
6	0.030711	233.0	0.167503	1069.0	0.296301	1132.0	0.208673	1498.0
7	0.078807	584.0	0.164907	1029.0	0.287265	1087.0	0.280335	760.0
8	0.063706	482.0	0.191573	1411.0	0.261273	951.0	0.268426	868.0
9	0.028046	212.0	0.188123	1355.0	0.357252	1429.0	0.245547	1129.0
10	0.061675	468.0	0.186115	1329.0	0.208225	666.0	0.236619	1222.0
11	0.053807	413.0	0.241579	2003.0	0.197880	614.0	0.259091	972.0
12	0.017259	127.0	0.154211	857.0	0.202614	637.0	0.131305	2079.0
13	0.012437	91.0	0.111601	331.0	0.094155	216.0	0.057483	2432.0
14	0.026269	198.0	0.176225	1195.0	0.226341	749.0	0.176021	1788.0
15	0.082234	606.0	0.220845	1802.0	0.477236	1776.0	0.387885	104.0
16	0.018020	133.0	0.123859	435.0	0.072885	151.0	0.067346	2396.0
17	0.017513	129.0	0.118362	389.0	0.234931	803.0	0.118603	2145.0
18	0.023985	180.0	0.130350	509.0	0.207806	663.0	0.125082	2114.0
19	0.016497	121.0	0.111836	337.0	0.168041	477.0	0.086416	2311.0
20	0.024492	184.0	0.164332	1015.0	0.264940	966.0	0.180737	1746.0

图 10 秩序图

3.3.3 RSR 值分布表（节选）

	f	Σ f	\bar{R} f	\bar{R}/n*100%	Probit
0.000024	3	3	2.0	0.000727	1.816292
0.000041	1	4	4.0	0.001454	2.022704
0.000058	23	27	16.0	0.005816	2.476888
0.000065	1	28	28.0	0.010178	2.680284
0.000075	1	29	29.0	0.010542	2.693510
0.000091	1	30	30.0	0.010905	2.706344
0.000092	1	31	31.0	0.011269	2.718811
0.000110	1	32	32.0	0.011632	2.730933
0.000127	1	33	33.0	0.011996	2.742731
0.000136	1	34	34.0	0.012359	2.754223
0.000144	1	35	35.0	0.012723	2.765425
0.000156	1	36	36.0	0.013086	2.776354
0.000169	1	37	37.0	0.013450	2.787024
0.000178	1	38	38.0	0.013813	2.797448
0.000182	1	39	39.0	0.014177	2.807637

图 11 RSR 值分布

3.3.4 计算回归方程

t 检验统计量较大，说明模型的回归系数具有统计学意义；拟合优度达到了 98.4%，拟合效果较好，通过回归检验。

回归直线方程为：y = 0.10732898996824755 Probit - 0.32749530178062675

OLS Regression Results

Dep. Variable:	y	R-squared:	0.984
Model:	OLS	Adj. R-squared:	0.984
Method:	Least Squares	F-statistic:	1.647e+05
Date:	Mon, 04 May 2020	Prob (F-statistic):	0.00
Time:	21:08:26	Log-Likelihood:	7868.0
No. Observations:	2719	AIC:	-1.573e+04
Df Residuals:	2717	BIC:	-1.572e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.3275	0.001	-241.905	0.000	-0.330	-0.325
Probit	0.1073	0.000	405.815	0.000	0.107	0.108

Omnibus:	765.490	Durbin-Watson:	0.010
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16641.326
Skew:	-0.804	Prob(JB):	0.00
Kurtosis:	15.013	Cond. No.	28.0

图 12 回归计算结果

3.3.5 答复意见质量评价结果（节选）

对意见答复质量进行评级，最后可以的出没一条答复意见的等级，等级越高说明答复质量越高。

	X1：信息量	R1：信息量	X2：可读性	R2：可读性	X3：相似性	R3：相似性	RSR	RSR_Rank	Probit	RSR Regression	Level
0	0.057233503	437	0.257401024	2099	0.429317223	1683	0.359825103	191	6.482784827	0.368295446	1
1	0.038324873	292	0.171003058	1119	0.231558659	786	0.195554951	1623	4.773474128	0.184836855	2
2	0.044923858	344	0.243339772	2017	0.305248484	1184	0.292528428	654	5.714793812	0.285867746	2
3	0.038959391	297	0.150047654	795	0.231189935	783	0.177649491	1770	4.633406437	0.169803531	2
4	0.020050761	149	0.130431275	512	0.432104997	1687	0.204463288	1535	4.855073844	0.19359487	2
5	0.029060914	220	0.2644252	2144	0.185187856	553	0.219370251	1391	4.986787653	0.20773158	2
6	0.03071066	233	0.167503334	1069	0.29630057	1132	0.208673018	1498	4.889065748	0.197243187	2
7	0.078807107	584	0.16490672	1029	0.287265334	1087	0.280335256	760	5.595066013	0.273017482	2
8	0.063705584	482	0.191572911	1411	0.261272789	951	0.268425522	868	5.481281774	0.260805135	2
9	0.028045685	212	0.188122724	1355	0.357252163	1429	0.245546638	1129	5.227460821	0.233562788	2
10	0.061675127	468	0.186114732	1329	0.20822451	666	0.236619129	1222	5.141243974	0.224309221	2
11	0.053807107	413	0.241578809	2003	0.197879859	614	0.259090806	972	5.37733439	0.249648567	2
12	0.017258883	127	0.154211247	857	0.202613827	637	0.13130498	2079	4.308540186	0.134935965	2
13	0.012436548	91	0.111600539	331	0.094155175	216	0.057483218	2432	3.806420956	0.081044015	3
14	0.026269036	198	0.176224797	1195	0.226340713	749	0.17602062	1788	4.615807882	0.167914696	2
15	0.082233503	606	0.220844826	1802	0.477235964	1776	0.387885465	104	6.781187279	0.40032268	1
16	0.018020305	133	0.123859337	435	0.07288518	151	0.067346205	2396	3.870803563	0.087954135	3
17	0.01751269	129	0.118361587	389	0.234930872	803	0.118602556	2145	4.229990192	0.126505273	2

表 10 答复意见质量评价结果

4. 结论

对相关部门的群众留言信息及回复信息进行分析研究，对了解民智、汇聚民智、凝聚民气具有重要作用，我们可以通过群众留言了解居民生活需要解决的问题，让政府更贴近居民生活。在第一问中我们解决了对居民问题进行标签的分类，让政府在解决问题时更加方便，当它解决相应问题时只需要调出相应标签下的问题就可以查看，大大方便问题的解决进度；在第二问时我们找出来留言问题中热度最高的前五个，在思考问题前期我们考虑到留言问题的相似度会对问题产生的影响，所以使用了相似度分析剔除了留言会对结果产生的影响，然后使用 TOPSIS 法对热度进行排序，截取出前五热度的问题，可以看出哪些问题是民众亟待解决；在第三问我们对留言答复进行分析，从文本中提取出指标，通过相关

度、信息量和可读性对问题的回复质量进行分析。

近年来各种数据增多，所以使用大数据的思维对问题的分析尤其重要，我们把问题进行分类，然后对提出的问题进行解决，最后对解决问题答复进行质量分析就是解决一个问题的全部过程了。使用这种分析问题的方式，我们可以解决许许多多相似的问题，帮助人民可以更好的生活。

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[D].山东师范大学,2018.

[2]管述学,庄宇.熵权-TOPSIS 模型在商业银行信用风险评估中的应用[J].情报杂志,2008,27(12):3-5+10.

[3]孙海峰，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.

[4]Q.Ethan McCallum.数据整理实践指南[M].北京邮电出版社，2010，36(23):94-196.

[5] 茆诗松, 程依明, 濮晓龙, 等. 概率论与数理统计教程第二版[M].北京: 高等教育出版社, 2011

[6] 樊宏, 吉华萍, 杜宪明,等. 应用综合指数法及秩和比法综合评价某市 2008 年～2010 年的医疗服务质量[J]. 卫生软科学, 2012, 26(3):201-203.

[7] 韦懿芸, 颜艳, 胡宇峰,等. TOPSIS 法和秩和比法相结合综合评价城市老年人生存质量[J]. 中国老年学杂志, 2006, 26(4):440-442.