“智慧政务”中的文本挖掘应用

摘要

近年来，随着信息时代的发展，各种网络技术的出现，为大家提供了越来越多的平台去交流沟通，群众可以向相关部门进行留言，反应遇到的问题，政府也可以通过网络平台了解民意，了解民情。但同时也随之而来的带来一些不可避免的问题，留言信息数量多，内容涉及范围广泛，以往靠人工来进行留言分类和热点问题整理的难度加大，因此对于处理网络问政平台上的群众留言时，基于自然语言处理，大数据等技术的出现和发展，为留言的分类，热点问题的挖掘以及留言回复的分析产生推动作用提升政府等的施政效率。

对于问题一，通过 Positionld 对于留言信息统计表进行去重，得到不重复的留言信息，通过数据增强有效的解决留言问题同一级别中每个类别分配不均的问题。利用 jieba 中文分词工具对具体留言信息进行中文文本分词，并通过 TF-IDF 算法提取关键词，建立多分类模型，将多个二分类结合，根据一级标签，对留言信息进行多分类，有效的将留言信息进行了系统的分类。利用 F-Score 评价方法从召回率和准确率方面对建立的多分类模型进行评价。

对于问题二，通过对留言信息的数据预处理，利用 excel 对去重后的信息进行处理，利用 jieba 对留言进行中文分词，通过 TF-IDF 算法得到每个留言内容类型的 TF-IDF 权重变量，采用 K-means 算法对 IF-IDF 权重向量进行聚类，得到 7 个质心，随后根据 KNN 最邻近分类算法筛选热点问题。对于热度评价指标的构建，通过指标的选取原则，我们选取了 3 个二级指标以及 6 个三级指标，根据指标的性质，分为效应型指标和成本型指标，对指标进行定性和定量分析，利用熵权法对三级指标进行赋权，并利用简单加权平均算法对数据进行融合，建立影响热度的定量评价模型。最后，综合定性和定量分析判断哪些留言属于热点问题。

对于问题三，通过 pandas，jieba 等工具对文本进行分词，且向量化转化为结构化语言，进行数据预处理；采取 LDA 方法进行主题抽取，利用 match 查询，处理留言全文字段和精确字段，对留言回复的相关性进行相应解释，进而对收集到的数据进行编码，对结果进行频数统计与交叉表分析，解释留言回复的完整性，最后，从留言回复的可解释性方面进行说明。综合考虑留言回复应具备的内容，礼仪规范特征等方面，给出了一套相对较好的评价方案。

关键词：jieba 中文分词　F-Score 评价方法　K-means 聚类　KNN 算法　TF-IDF 算法　多分类模型

目录

1.挖掘目标

本次建模目标是利用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数据，利用 jieba 中文分词工具对于留言内容进行分词，权重策略 TF-IDF 进行分类，F-score 评价方法，K-means 聚类的方法以及 KNN 算法，达到以下三个目标：

（1）对数据进行数据增强，按照一定的划分体系对留言进行分类，利用 TF-IDF 算法计算权重变量进行分类，建立多分类模型，得出关于留言内容的一级标签分类模型，使用 F-score 评价方法对所建立的分类模型进行评价，对模型进行验证和调优。

（2）利用文本分词和文本聚类的办法对于非结构化的数据进行文本挖掘，根据所得出来的聚类结果，结合问题一所得出的多分类模型，定义热度评价指标，对热度评价指标进行定性和定量分析，构建基于熵权法的热度评价影响的指数体系进行总结出热点问题。

（3）从相关性，完整性，可解释性等角度对于所给出的答复意见的质量进行评价，构建相关指标，通过对指标进行定量和定性的分析，得出一套完整的评价方案。

2.分析方法与过程

总体流程图

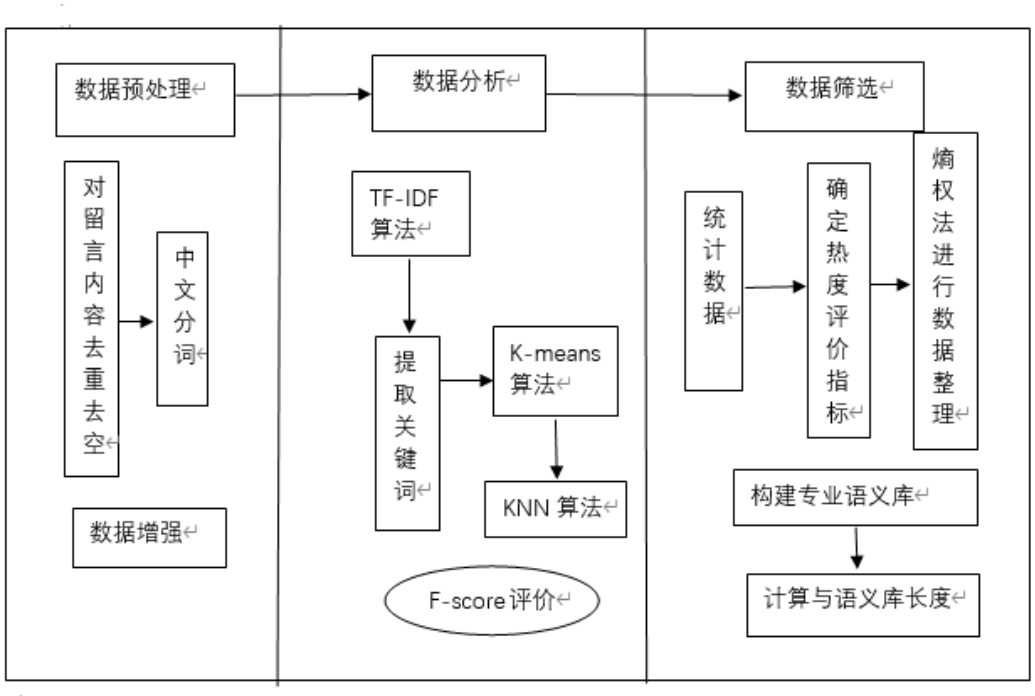

图 1：总体流程图

主要包括以下步骤：

步骤一：数据预处理，在题目所给的数据中，出现了很多重复且无用的留言内容数据，在原始的数据上进行去重去空处理，在此基础上进行了中文分词，由于数据不平衡会带来很大的影响，因此进行了数据增强。

步骤二：数据分析，在数据预处理的基础上进行数据分析，采用 TF-IDF 算法将分词后的留言内容转化为权重向量，以供后续的文本挖掘分析使用，采用 TF-IDF 算法找出留言内容的关键词，构造多分类模型，利用 word2vev 工具进行文本分类，利用 F-score 算法进行评价。采用 K-means 算法对职业进行分类，利用 KNN 算法找出与各中心相似的元素，判定所属类别。

步骤三：数据筛选，统计相关留言数据，分类筛选汇总，定义热度评价指标，对热度指标进行定性和定量分析，构建基于熵权法的热度评价影响的指数模型，得出当下的热点问题。

步骤四：对答复意见进行评价，从几方面得出一套评价方案。

2.1 问题一分析方法和过程

2.1.1 解决问题流程图

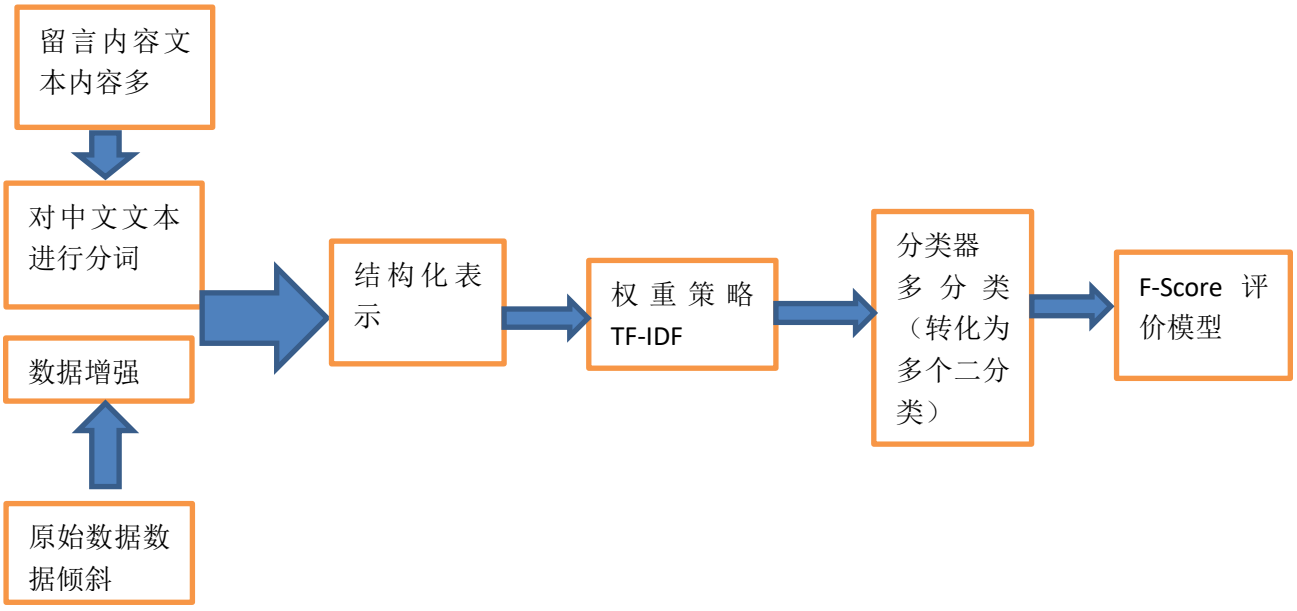

图 2：问题一流程图

2.1.2 数据预处理

2.1.2.1 留言信息进行数据增强

对于题中给出的留言信息，留言的随机性会带来社会问题的暴露，但也正是留言的广泛性和随机性，致使集中性的问题在数据中会集中出现，集中性问题的出现会导致数据发生严

重倾斜，致使出现过拟合现象，所以，对已有数据进行数据增强。

2.1.2.2 文本数据增强

相较于图像数据增强，文本是离散的，图像是连续的，文本数据增强更像同义句的生成过程。从大方向上，文本的数据增强方法有 EDA、BackTranslation、生成对抗网络、语境增强等等。本体采用的是 EDA 中的四种形式，同义词替换（SR）在同义词替换中不考虑 stopwords，而是在留言句子中随机抽取 n 个词，然后从同义词词典中随机抽取同义词，进行替换；随机插入（RI）不考虑 stopwords，随机抽取一个词，然后在该词同义词集合中随机抽取一个插入原来留言中的随机位置，重复 n 次；随机交换（RS）是指在留言中随机选词的位置交换、随机删除（RD）是指留言中的每个词以概率 p 随机删除。

2.1.2.3 对留言文本进行中文分词

在对留言文本进行挖掘分类之前，需要对留言的非结构性特质进行结构化，转化为计算机可以识别的语言。为了转化为计算机可识别的结构性语言，需要对留言文本进行中文分词，在本题中，使用 python 中的 jieba 库，对文本的语句进行中文分词。Python 中 Jieba 分词的方法一共支持三种分词模式（精确模式、全模式、搜索引擎模式），并且支持自定义字典，对于留言文本从词性等方面最大限度上进行分词。对中文分词后的语句，为快速提取有效信息，我们结合 TF-IDF 算法，从每条留言中抽取的 5 个关键词，进行后续分析。

2.1.2.4 TF-IDF 算法

在对给定文本进行中文分词后，为了便于后续的挖掘，将这些词转换成向量。我们将使用 TF-IDF 算法将消息文本转换为权重向量。

TF-IDF 实际上是 TF 和 IDF 的产物。TF（term frequency）是单词频率，IDF（inverse document frequency）是反向文件频率。具体算法如下所示：

第一步：计算词频，即 TF 权重。TF 表示文档中单词的频率

为了便于不同文章的比较，“词频”是标准化的，除以正文中的总字数或最多的正文中的字数，即如下公式：

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

或

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

第二步：要计算 IDF，我们需要建立一个语料库来模拟语言环境。

计算 IDF 的主要思想是：如果包含条目 T 的文档越少，即 n 越小，然后 IDF 越大，说明条目 T 具有较好的分类能力。

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

实际上，一个词条在一个类的文档中频繁出现，则就恰恰说明该词条可以代表这个类的文本的特征，那这种表达性强的词条，在计算时就要赋予较高的权重，并选来作为该类文本的特征词以区别与其它类文档。

第三步：TF-IDF 值的计算

TF-IDF=词频（TF）×逆文档频率（IDF）

实际分析得出 TF-IDF 值与一个词在职位描述表中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本分词中的 TF-IDF 值，进行排序，次数越高，则说明为要提取的每条留言的关键词。

2.1.2.5 生成 TF-IDF 向量

生成过程如下所示：

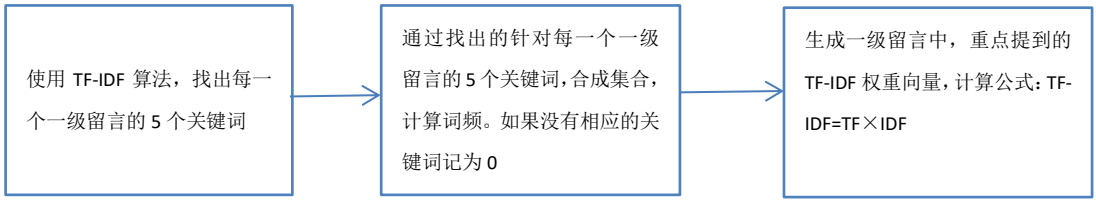

图 3：TF-IDF 生成向量过程图

根据留言者留言中的关键的情感词语进行情感分析，情感分析就是用户的态度分析，情感分析系统对文本进行“正负二项分类的”，即只需要判断文本向量是正向还是负向的。我们也就可以通过情感词典等工具判断文本出现的次数成正比。

2.1.3 数据分析

该问题中的多个留言对应于七个一级留言标签中的一个，这是一个多对多的问题，因此有必要建立一个多分类模型。本文将一级标签的七个分类器分别训练成一个分类器，然后综合多个分类器的结果。通常在机器学习中，多重累积分类或多项式分类是将实例分配给一个类别，而不是两个以上的类别。实例分为两类，即二进制分类。二元问题通常扩展到以下方法：神经网络、k 近邻、决策树、支持向量机、朴素贝叶斯。将多分类问题的输出空间划分为一棵树，将每个父节点划分为多个子节点，然后重复该过程，直到每个子节点只代表一个类。

在本题中，我们将使用朴素贝叶斯算法来分析二进制分类问题。朴素贝叶斯分类器基于一个简单的假设：当给定目标值时，属性相互独立。

朴素贝叶斯模型：

$$vmap=\arg\max P(V_j|\ a_1, a_2 \ldots a_n)\quad V_j \in A$$

其中 vmap 是给定一个 example,得到的最可能的目标值.，$a_1, a_2 \ldots a_n$是这个给定的 example 里面的属性之一.，Vmap 目标值,是后面计算得出的概率最大的一个.所以用 max 来表示。贝叶斯公式应用到 $P(V_j|\ a_1, a_2 \ldots a_n)$中.可得 Vmap=

$\frac{arg\ max\ P(a1,a2 \ldots an|Vj)P(Vj)}{P(a1,a2 \ldots an)}$ 。

又由于朴素贝叶斯分类器默认$a_1, a_2 \ldots a_n$之间是互相独立的。所以 $P(a_1, a_2 \ldots a_n)$对于结果并不会发挥作用. 由于所有的概率都要除同一个东西之后再进行大小比较,最后结果也似乎影响不大，于是就可以得到 Vmap= arg max P （ $a_1, a_2 \ldots a_n V_j$），然后观察到联合$a_1, a_2 \ldots a_n$的概率正好对应每个单独属性的概率乘积：

$$P(a_1, a_2 \ldots a_n|\ V_j) = \Pi i\ P(a_i|\ V_j\ )。$$

2.1.4 F-Score 模型

在机器学习中，大的数据集在进行评估时，分类模型的精准率（Precision）和召回率（Recall）的评估指标两个指标相互制约，一般情况下，P 高，R 就低，R 高，P 就低。这时就需要我们对两个指标进行综合权衡，引入 F-Score 评价模型，综合考虑二者的调和值。

$$F-Score=\frac{\cdot Precision \cdot Recall}{\beta^2 \cdot Precision+Recall}(1+\beta^2)$$

当β=1 时，我们称其为F_1-Score，此时，精确率和召回率都很重要，权重近乎相同。然而在有些情况下，我们认为精确率更重要些，那就调整β的值小于 1，如果我们认为召回率更重要些，那就调整β的值大于 1。

2.2 问题二分析方法与流程

2.2.1 流程图

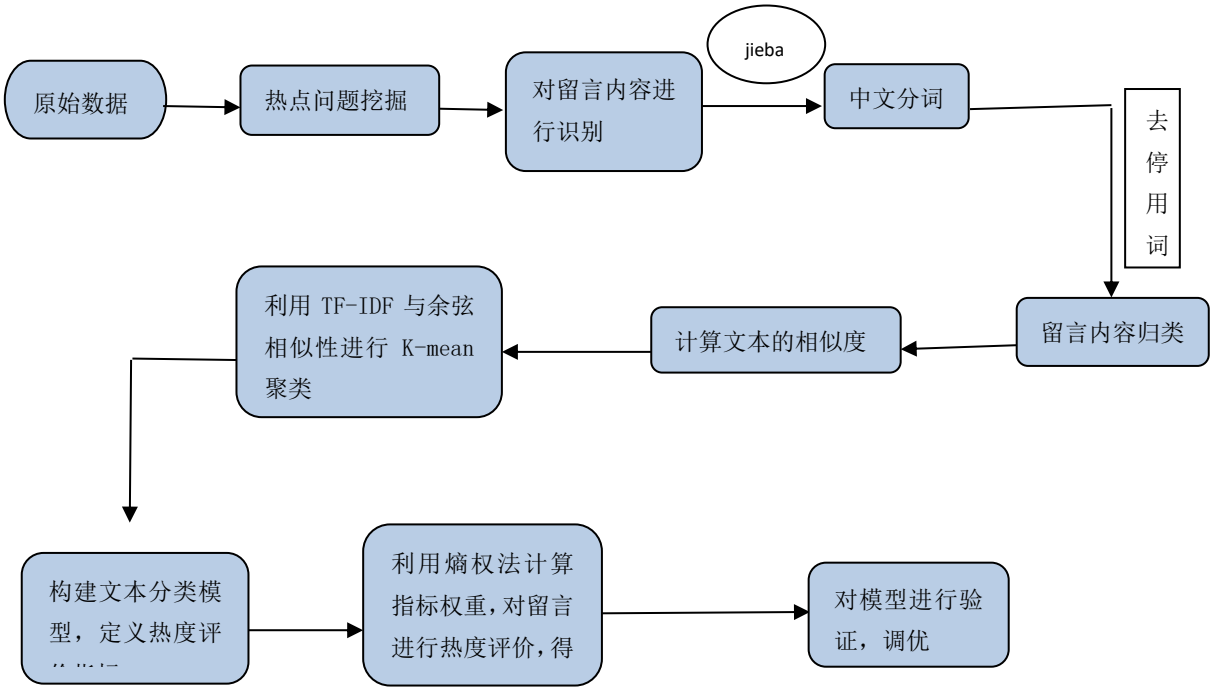

图 4：问题二分析流程图

分析流程说明：进行热点问题挖掘的同时，从以下三个方面进行分析，第一个方面是问题的三要素即留言的特定时间，事件发生的地点，以及发生的问题来进行留言的分类，通过使用中文分词，去停用词等方法进行数据的预处理；第二个方面是问题的归并，通过对文本的特征选择，文本相似度的计算对留言内容进行 k-mean 聚类，构建文本分类模型；第三个方面是热度评价指标的选取，通过对热度评价指标的定义以及量化方法的阐述，对群众留言进行热度评价。

拟解决的关键问题是：

（1） 表达多样化的问题，即相同地点相同事件的不同表达方式；

（2） 留言内容的特征较多，；两两之间计算相似计算量大。

2.2.2 数据的预处理

2.2.2.1 留言内容的去空、去停用词

在留言统计表中会出现留言内容为空的情况，干扰了对问题的分析，采用直接滤过的方法，从文本中直接删除。对现有的词库进行去停用词操作，利用停用词词库进行去重，使用jieba 分词的搜索引擎模式 cut_for_search()。

2.2.2.2 对留言内容表进行中文分词

在对群众的留言内容进行挖掘和分析之前，必须将文本数据这些非结构化的语言转换为计算机能够识别的结构化信息。在附件 3 中，数据以中文文本的形式给出。为了便于转换，首先要对这些信息内容进行中文分词。在这里，我们使用 python 的中文分词包 Jieba 进行中文文本分词。Jieba 分词的原理是利用基于前缀字典的高效构图扫描，生成由句子中所有汉字构词条件组成的 DAG。同时，利用动态规划方法，找出基于词频的最大概率路径和最大分割组合。针对于对于没有注册的词语，便会采用基于汉字构词能力的 HMM 模型，以更好地实现中文文本的切分效果。

在中文分词的同时，采用 TF-IDF 算法提取每条消息内容的前五个关键词。这里使用的是jieba 的语义数据库。

2.2.2.3TF-IDF 算法

在对留言内容进行分词后，需要将这些词语转化为向量，以供后续的挖掘分析使用，这里将采用 TF-IDF 算法，将留言内容转化为权重向量。TF-IDF 算法的具体原理如下：

（1） 计算词频，考虑到留言有长有短，为了便于不同文章的比较，进行“词频”标准化，除以留言内容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。

（2） 计算 IDF 权重，即逆文档频率，需要建立一个语料库（corpus）,用来模拟语言的使用环境。IDF 越大，这个特征性在文本中的分布就越集中，说明该分词在该文本内容属性能力增强。

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

（3）计算 TF-IDF 值。

TF -IDF=词频（TF）×逆文档频率（IDF）

实际分析可得 TF-IDF 值与一个词在文本描述中出现的次数成正相关，若一个词在文本中的重要性越高，TF-IDF 值越大。通过计算文本中每一个词的 TF-IDF 值，次数最多的就是我们要提取的关键词，然后进行分类。

2.2.3 留言内容的分类

生成留言内容的 TF-IDF 权重向量之后，根据每一个类型的 TF-IDF 权重向量对留言内容进行分类。这里将采用 K-mean 算法将留言内容分类。

K-mean 聚类的原理如下：

假设有一个包含 n 个 d 维数据点的数据集 $X=\{x_1,x_2,\cdots\cdots x_i,\cdots\cdots x_n\}$,其中 $x_i \in R^d$,K-means 聚类方法是将数据集 X 组织为 K 个划分 $C=\{c_k,i=1,2,3\cdots\cdots K\}$。使得每个划分的区域代表一个类 c_k，每一个类 c_k 有一个类别中心 μ_i。选取了欧氏距离作为相似性和距离判断准则，计算该类内个点到聚类中心 μ_i 的距离平方和

$$J(c_k)=\sum_{x_i \epsilon c_k} ||x_i-\mu_k||^2$$

聚类目标是使各类总的距离平方和$C_j = 最小$

$$J(C)=\sum_{k=1}^{K} J(c_k)=\sum_{k=1}^{K}\sum_{x_i \in C_i}\left\|x_i-\mu_i\right\|^2=\sum_{k=1}^{K}\sum_{i=1}^{n} d_{ki}\left\|x_i-\mu_i\right\|^2$$

其中，$d_{ki}=\begin{cases}1, 若x_i \in c_i\\ 0, 若x_i \notin c_i\end{cases}$，所以根据最小二乘法和拉格朗日原理，聚类中心 μ_k 应该取为类别 c_k 类各数据点的平均值。

K-mean 均值聚类的算法步骤如下；

（1） K 元素是从 X 元素中随机选取的，X 元素被视为 K 簇的中心。

（2） 接下来，我们计算剩余元素到 K 个簇中心的相异程度，并将这些元素分类为相异程度最低的簇。

（3） 根据聚类结果，重新计算每个新的中心点。计算方法是取每个簇中所有元素的每个维度的算术平均值。

（4） 所有元素根据重新选择的中心再次聚集。

（5） 重复步骤 4，直到聚类结果不再改变。

（6） 输出结果。

K-mean 聚类的算法流程图如下：

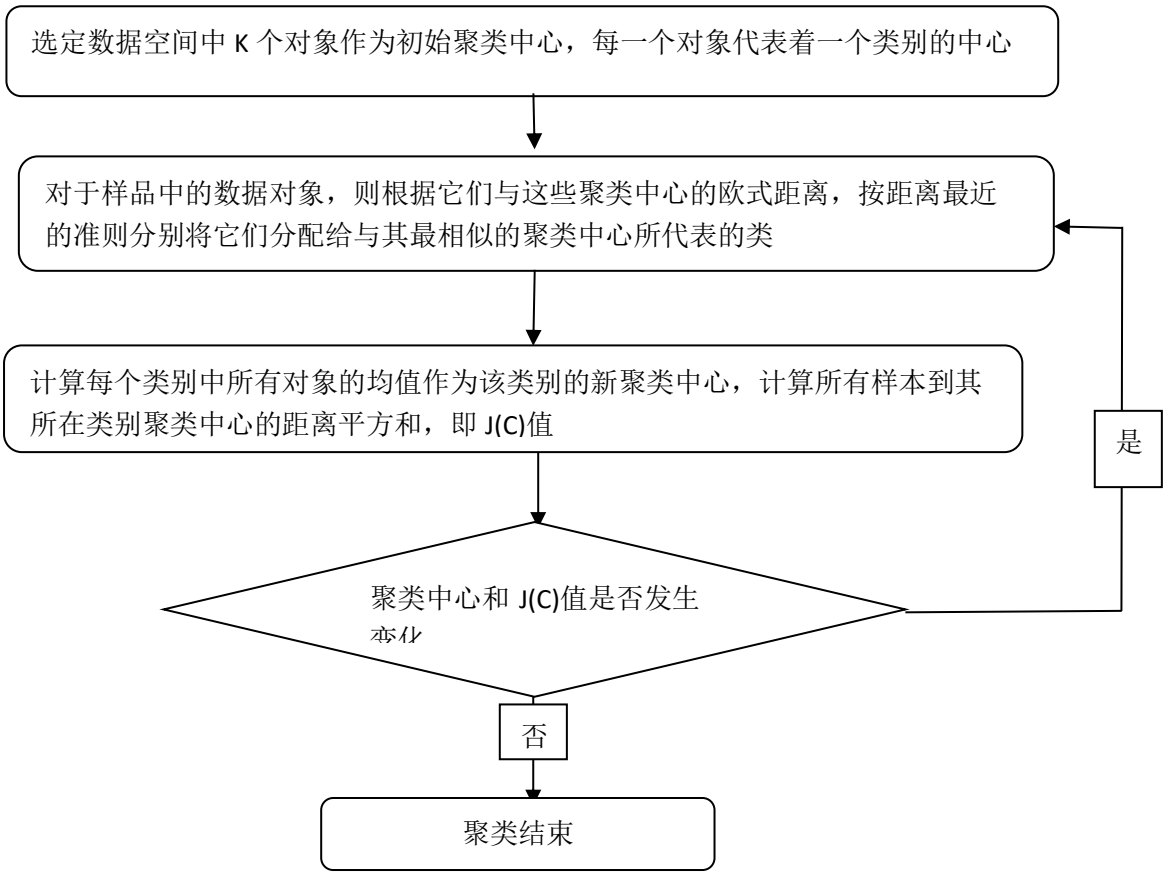

图 5：K-mean 聚类的算法流程图

由于留言统计表中给出了 4326 条数据，去重操作后还有 4224 条数据记录，假如将所有的留言内容记录都用来进行挖掘分析，会占用很大的机器性能与操作时间，为了高效的完成相关工作，获得结果和有关结论，我们将从记录数据中随机抽取一部分数据记录，保存抽样结果，对于所选样本进行分词，TF-IDF 向量的求解，利用 K-mean 聚类,将数据分为几大类，计算出每步迭代出来的 J(C)值。

2.2.4Knn 最邻近分类算法

采用 K-means 分类方法得到聚类中心，然后用 Knn 算法找出与每个中心相似的元素，并根据元素数目确定分类。根据向量空间模型，将每一类别的中心向量记录为C_j（W_1，W_2，W_n），将要分类的文本 7 以 n 维向量的形式表示，然后将文本内容形式化为特征空间中的加权特征向量，即 d=d（T_1, W_1；T_2，W_2 T_n，W_n），对于测试文本数据，计算其与训练样本集中每个文本的相似度，找出 k 个最相似的文本，并根据加权距离判断测试文本数据的类别。具体算法步骤如下：

（1） 对于一个测试文本数据，根据前边确定的特征词形成测试文本向量。

（2）计算测试文本数据与训练集中每个文本数据的文本相似度，计算公式为:

$$Sim(d_i, d_j) = \frac{\sum_{k=1}^{M} W_{ik} \times W_{jk}}{\sqrt{\sum_{k=1}^{M} W_{ik}^2} \sqrt{\sum_{k=1}^{M} W_{jk}^2}}$$

式中，d_i 为测试文本数据的特征向量，d_j 为 j 类的中心向量；M 为特征向量维数；W_k 为向量的第 k 维。k 值的确定一般先采用一个基本初始值，然后根据实验测试 K 的结果进而来调整&值。

（3）按照文本的相似度，在训练文本集中送出与测试文本最相似的 k 个文本。

（4）在测试文本的 k 个近邻值中，以此计算每类的权重，计算公式如下：

$$p(X, C_j) = \begin{cases} 1, 若 \sum_{d \epsilon Knn} Sim(x, d_i) y(d_i C_j) - b > 0 \\ 0, 其他 \end{cases}$$

式子中，x 为测试文本数据的特征向量：Sim（x，d_i）为相似度计算公式；b 为阈值，有待于进一步优化选择：而 y（d_i，C_j）的值为 1 或者 0，如果 d_i 属于 C_j，则函数值为 1，否则为 0。

（5）比较类的权重，将文本数据划分到权重最大的那个类别中。

2.2.5 热度评价指标的构建

2.2.5.1 指标的构建原则

指标体系的构建原则通常根据要求和对象的不同分为三个层面：客观性原则，系统性原则和敏感性原则。

客观性原则：指指标体系的选择必须要从客观实际出发，全面反映留言内容的热度情况，解决因人而异的主观因素的影响。

系统性原则：指指标体系的设计应从系统整体出发，能够包括形成热门留言内容的各个因子，各个指标间既相互独立又相互联系，共同形成一个有机整体。

计算与操作层面一般采用数据的可得性和可操作性原则，是指在设计热度评价指标的时候用较少的指标反应相对来说更多的实质性内容，以便于后续指标的收集和量化。除了以上所提及的三个基本原则，还应该新增趋势性原则和导向性原则。由于留言内容的丰富性，决定其是一个不断变化的指标，趋势性原则体现热度较高的留言内容的变化趋势；导向型原则是指该套热度评价指标的构建不仅仅要对留言内容进行检测，更是要对判断热门的留言提供方向指导。

2.2.5.2 构建理论依据

群众问政留言其实在本质上是一种信息传播活动。根据新闻传播学的理论，信息传播活动包括四个要素，即来源、渠道、目的地和信息。影响信息传播效果的因素主要有四个，即传播者、接受者、渠道和内容。从这个角度看，它类似于马尔科姆·格拉德威尔提出的流行的三要素理论。马尔科姆·格拉德威尔认为，一个物体要想流行，就必须具备流行的基本要素，即关键人物法则、环境力量法则和内容附着法则”。

群众留言和流行的事务是类似的。关键字符规则具体是指在消息的主题和内容中起关键作用的字符，即群众所留留言的的环境和主题。环境功率定律是指在不同的环境、不同的时间出版不同的内容所产生的热量效应肯定是不同的。消息生成器发布的消息受环境影响。内容附着规律意味着热点问题的热度还取决于信息内容是否简洁、不可预测、具体、可信、情感、叙事性等方面，在表现形式上是否具有鲜明的主题性，长度是否合适，能否赢得眼球等。

根据三大流行元素和新闻传播理论，从信息的内容、信息的喜欢程度、信息的发生率三个维度来判断其是否已经成为一个热点问题。我们将从信息内容特征热度的影响、传播特征热度的影响、受众反映特征热度的影响三个方面对这一热点问题进行量化评价。

（1） 信息内容特征的影响。根据流行的三因素理论，信息的内容特征对其流行具有相对来说较强的影响。我们将选择两个二级指标来反映信息的内容和形式，即信息的内容和内容、问题信息的及时性和两个二级指标。消息的长度用于反映消息的内容和信息的丰富性。信息的字数越多，内容就越丰富，越容易引起讨论。信息问题的及时性主要体现在信息的新鲜性上。一手资料比二手资料更受关注。

（2） 通信的特征的影响。留言的传播特性可以最直观地反映群众在网络平台上引起的注意程度。一般来说，大众信息传播的方式是重复信息。留言信息被复制的次数和节点数将影响群众留言问题的流行性。因此，信息问题的重复性和相似热词的出现频率也可以成为评价留言其是否是热点问题的重要方法。

（3） 受众特征的影响。受众特征是指受众在收到信息后对信息的反应和态度。其他受众的积极性也会对热点问题产生很大影响。在这里，我们将用能表达观众态度的反对和喜欢的数量来衡量群众的观点。

据此构建出来的留言热度评价指标体系以及各指标之间内涵如图所示：

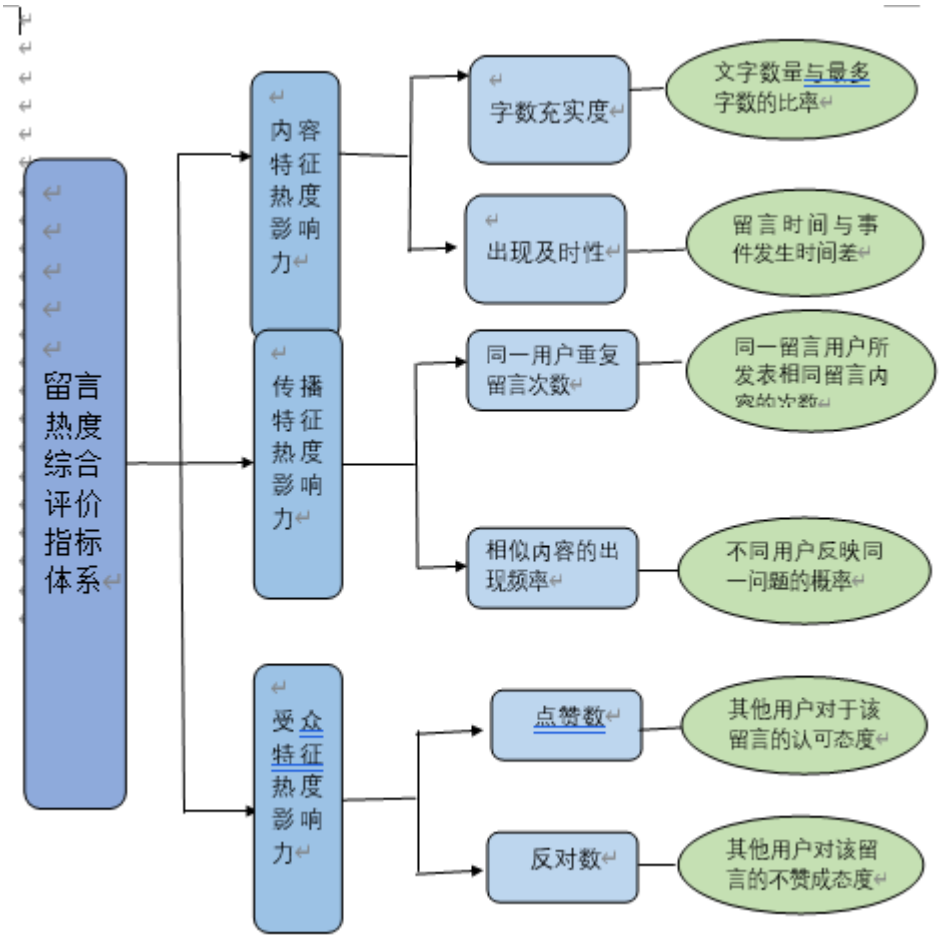

图 6：留言热度评价指标体系

2.2.5.3 数据的预处理

Step1：在上文构建的热度评价指标体系中，存在多个指标，而且这些指标的量纲不同的，并且有些指标为效应型指标，即指标越大越好，有些指标为成本型指标，即指标越小越好，所以在处理前必须对其进行一定预处理。

设 r_j(i=1,2,3……,n)是效益型指标，则其归一化公式为：

$$r_i' = \frac{r_i - \min_i r_i}{\max_i r_i - \min_i r_i}, (i = 1, 2, \ldots, n).$$

设 e_j(j=1,2,……,m)是成本型指标，则其归一化公式为：

$$e_j' = \frac{\max\limits_j e_j - e_j}{\max\limits_j e_j - \min\limits_j e_j}, (j = 1, 2, ..., m).$$

根据热度评价指标体系中各个指标的定义，我们可以清楚的判断出每一个指标的类型，其中属于成本型的指标有：反对数；效应型指标有：字数充实度，出现及时率，同一用户频繁留言数，相似内容出现的次数以及点赞数。对于上述效应型指标我们可以利用效用型指标的公式对数据进行预处理归一化处理。

Step2：利用熵权法确定各三级指标的权重系数

在信息论的基本原理中，信息是系统有序度的度量，是系统无序度的度量。如果一个指数的信息熵越小，它所反映的信息就越大。否则，如果一个索引的信息熵越大，它所反映的信息就越小。信息熵越大，指标在评价体系中的作用越小，相应的权重也就越小。本文采用熵权法确定了热度评价指标体系中三级指标的权重系数。

（1）设有 n 个指标， m 个时间段的观测值，指标值为 $r_{ij}(1\leqslant i\leqslant m，1\leqslant j\leqslant n)$，得到原始数据矩阵为：

$$R' = (r_{ij}')_{m\times n} = \begin{pmatrix} r_{11}' & \cdots & r_{1n}' \\ \vdots & \ddots & \vdots \\ r_{m1}' & \cdots & r_{mn}' \end{pmatrix}.$$

（2）利用 step1 中的将原始数据 R 进行归一化处理，得到标准化的指标矩阵为：

$$R' = (r_{ij}')_{m\times n} = \begin{pmatrix} r_{11}' & \cdots & r_{1n}' \\ \vdots & \ddots & \vdots \\ r_{m1}' & \cdots & r_{mn}' \end{pmatrix}.$$

（3）计算道路通行能力评价指标体系中指标的熵值，得到熵值 $X=（x_i,x_2,x_3……x_n）$.

$$x_j = -\frac{1}{\ln(m)}\sum_{i=1}^{m} r_{ij}' \ln(r_{ij}'), 1\leq i\leq m, 1\leq j\leq n.$$

（4）计算热度评价指标体系中各指标的权重系数 $W=（w_{11},w_{12},w_{13}……w_{1n}）$。

$$w_{1j} = \frac{1-x_j}{n-\sum\limits_{j=1}^{n} x_j}, 1\leq j\leq n.$$

Step3：基于熵权法的留言内容热度评价指标体系的指数模型

在热度评价指标体系中，我们构建了 3 个二级指标，我们通过对内容特征热度影响力，受众特征热度影响力以及传播特征热度影响力三个指标进行综合，为了方便计算，将 3 个二级指标的权重取一个相等的值。

假设指标所对应的权重系数 $W_2=（w_{21},w_{22},w_{23}……w_{2k}）$.

$W_{2l}=1/k，(1\leqslant l\leqslant k)$

最后得出综合评价公式：

$$G=\sum_{j=1}^{n}\frac{\left[\frac{1}{\ln(m)}\sum_{i=1}^{m}r_{ij}^{'}(\ln r_{ij}^{'})\right]^{2}+\left[\frac{1}{\ln(m)}\sum_{i=1}^{m}r_{ij}^{'}(\ln r_{ij}^{'})\right]}{k(n-\sum_{j=1}^{n}x_{j})}$$

2.3 问题三的分析方法和过程

2.3.1 问题分析

网络问政作为政府与大众沟通的桥梁，其功能不仅能伸张民意，监督舆论，更是解决群众问题的有效途径。因此网络问政中政府或有关部门的答复尤为重要。

回应主要包括以下四个方面：第一，信息社会中的政府回应是指社会主导互动。二是对社会期望的理解。社会期望涵盖了需求的多个方面，每个期望都是不同的。回应社会期待的实质是社会利益的选择和整合，但并非所有的期待都需要政府的回应。当社会期望无法整合或整合成本较大时，政府需要对这种社会期望做出回应。第三，回应的形式是公共政策。公共政策是政府对社会领域的权威输出，是政治制度对整个社会价值的权威分配。回应的最终形式是政府制定公共政策以满足社会期望。第四，反应效应的理解。回应是对社会需求的回应，但并不简单地等同于满足社会利益的需求。回应还需要考虑社会需求合理表达的可能性，考虑社会的成熟性和自主性。此外，答复还应包括关于要求的指导。因此，正确认识应对效果应该能够满足预期，引导预期。

因此制定了从答复的相关性、完整性、可解释性等角度对答复意见的质量的评价方案，使网络问政朝向更规范化的方向发展。

2.3.2 相关性

相关性通常是指评价查询与评价结果间之间的相关联程度，并根据这种相关程度对结果排名的一种能力。

2.3.2.1 流程图

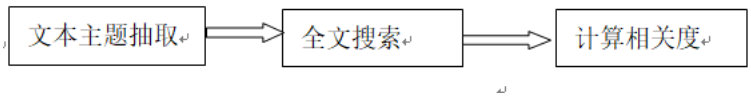

图 7：计算相关度流程图

2.3.2.2 步骤执行

Step1:利用 python 对文本主题抽取

（1）数据准备

为了对表格数据进行处理，使用数据框工具 Pandas。先调用它。读入数据文件，对数据框的头几行进行观察，以确认读取是否正确。其次再看看数据框的长度，以确认数据是否读取完整。当行列数都与爬取到的数量一致，确认通过。

（2）分词

下面进行一件重要工作——分词。因为需要对每篇文章的关键词进行有效提取，但是中文本身并不使用空格在单词间划分，因此使用 python 中的 jieba 库来进行分词。首先调用 jieba 分词包，把这项工作并行化。编写一个函数，处理单一文本的分词。执行完毕之后，需要查看一下，文本是否已经被正确分词。

（3）文本的向量化

将留言中的关键词转换为一个个特征（列），然后对每一条留言关键词出现个数统计。处理的文本里面可能会有大量的词汇，所以会造成处理时间太长，而且那些很不常用的词汇对主题抽取意义不大。因此在这里做了一个限定，只从文本中提取 1000 个最重要的特征关键词，然后停止。

（4）用 LDA 方法进行主题抽取

应用 LDA 方法，指定主题个数。如果我们对划分的结果不够满意，文本数据的可读性不够高的情况下，可以通过继续迭代，调整主题数量来进行进一步的优化。

Step2:全文搜索

利用 match 查询，match 既能够处理全文字段，又能处理精确字段，了解字段映射的信息。要查询一个已分析的全文字段，它们会先将查询字符串传递到一个合适的分析器，然后生成一个供查询的词项列表。一旦组成了词项列表，这个查询会对每个词项逐一执行底层的查询，再将结果合并，然后为每个文档生成一个最终的相关度评分。

match 查询的步骤是：

1.检查字段类型

2.分析查询字符串

将查询的字符串传入标准分析器中。

3.查找匹配文档

用 term 查询在倒排索引中查找某字符串，然后获取一组包含该项的文档。

4.为每个文档评分 。

用 term 查询计算每个文档相关度评分 _score，这是种将词频和反向文档频率，以及字段的长度相结合的计算方式。

Step3:计算相关度

Lucene 的实用评分函数

使用 TF-IDF 以及 向量空间模型，随后将它们组合到单个高效的包里进而收集匹配文档并进行评分计算。

向量空间模型：

设文档和表示 VSM 中的两个向量：

$$D_1 = D_1(w_{11}, w_{12}, ..., w_{1n})$$

$$D_2 = D_2(w_{21}, w_{22}, ..., w_{2n})$$

借助于 n 维空间中两个向量间的某种距离来表示文档之间的相似度，使用向量之间的内积来计算：

$$Sim(D_1, D_2) = \sum_{k=1}^{n} w_{1k} \times w_{2k}$$

考虑到向量的归一化，可以使用两个向量的余弦值来表示相似系数：

$$\mathrm{Sim}(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{n} w_{1k} \times w_{2k}}{\sqrt{\sum_{k=1}^{n} w_{1k}^2 \sum_{k=1}^{n} w_{2k}^2}}$$

采用向量空间模型进行文本表示时，需要经过以下两个主要步骤：

（1）根据训练样本集生成文本表示所需要的特征项序列；

（2）依据文本特征项序列，对训练文本集和测试样本集中的各个文档进行权重赋值、规范化等处理，将其转化为机器学习算法所需的特征向量。

2.3.2 完整性

完整性是指答复意见是否合乎规范

2.3.2.1 规范

1、留言时间与答复时间间隔长短

留言答复是否及时反映了相关部门的工作态度

2、全文使用尊称

不管对方是什么身份，反映的事情是否合理，都应以尊称相待，以表示回复方的诚意和服务姿态。

3、首段说明收件和处理安排

简要说明收件情况，如“您于 XX 年 XX 月 XX 日反映.....的情况已收悉”。

4、另起一段说明调查情况

简要阐述工作人员前往现场调查处理核实情况，应向对方说明调查结果是否属实，根据需要可以简述被调查对象的准确信息，如“经查，被投诉人在 xx 时间 xx 行为，与您所述的情况属实”。

5、现场处理情况

工作人员对该留言是如何处理的，提出了怎样的整改意见，对方是否接受整改。

6、整改到位情况

无论被调查人是否整改到位，在信访回复中，都应该说明进展情况，保障留言网友的知情权，让对方有心理准备。

7、下一步工作计划

简述下一步工作思路，表明态度，传达正能量，如“我单位将加大日常监管和巡查力度，避免此现象的发生”

8、表示感谢

文末向投诉人表示感谢，并欢迎对方继续监督和提供宝贵意见。

2.3.2.2 通过对答复内容的分析，生成如下分析

	变量	案件数	最小值	最大值	平均值	标准差
年份						
答复时间间隔						
问题类型						
处理结果						
解决问题形式						
未解决问题形式						

年份：年份最小值为 1，随年份增长，数字按依次增长

答复时间间隔：五天为一个区间，1-5 天回复记为 1,6-10 天记为 2，以此类推。

问题类型：建议记为 1，咨询记为 2，投诉记为 3。

处理结果：解决记为 1，未解决记为 2。

解决问题形式：按群众要求解决记为 1，按政府的合理方式解决记为 2。

未解决问题形式：不愿或不能解决记为 1，不必解决记为 2。

2.3.2.3 分析数据

对收集到的数据进行编码，对编码结果进行频数统计与交叉表分析

2.3.3 可解释性

解释学最初来源于解释学。它关注的是人们如何创造和维持他们生活的社会世界来理解和解释。它侧重于对社会现象的解释学理解。从解释的起源来看，可解释性的内容主要来源于与情境相关的影响因素，如个人过去的经验、现状和未来的预期目标。对可解释性的需要主要是由于对问题和任务的理解不足。

2.3.3.1 流程图

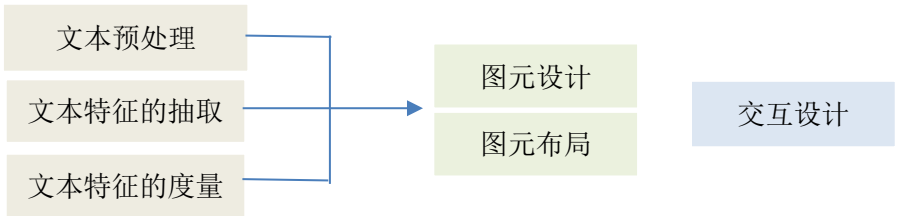

图 8：对数据可视化分析流程图

利用文本数据可视化来建立我们对数据的直观理解，特别是当文本数据量非常大或者数据维度非常高的时候，如果可以建立一些交互式的可视化方法将会极大地帮助我们从各个层次角度理解数据的分布和状态。

通过交互设计，不仅对当下情形做出分析，还可以对未来进行粗略预测。

3.结果分析

3.1 问题一结果分析

3.1.1 多分类结果分析

通过对留言文本进行数据增强，随后使用 TF-IDF 进行关键词的寻找，根据朴素贝叶斯，进行二分类，再对二分类进行综合分析，将给出的数据根据一级标签进行分为 7 类，其中个别组别的留言数量相对较少。然仍得以筛选出来。

3.1.2F-score 评估模型结果分析

通过 F-Score 评估模型对有多分类根据一级标签分类的留言从精准度和召回度进行了评估，评估结果显示精准度较高，相对于召回率来看。

3.2 问题二结果分析

3.2.1 聚类中心得分类结果

通过去重之后对文本进行分词，提取五个关键词后由 K-Means 分类得到聚类中心，利用 Knn 算法找到离各个聚类中心点最近得五个元素，根据“少数服从多数”的原则判定聚类中心所属类别。Knn 算法的大致步骤如下所示：

（1） 算距离：根据给定的聚类中心，计算聚类中心与样本中每一个 TF-IDF 权重向量的距离；

（2）找邻居：以聚类中心为圆心圈定距离最近的 15 个样本，作为聚类中心的近邻点；

（3）做分类：根据这 5 个近邻归属的主要类别，对聚类中心进行有效分类。

3.2.2 热度评价指标体系的结论

根据上述热评价指标体系模型，通过定性和定量分析对群众留言的热度进行评价。定性分析主要是哪些指标对信息热度有正面影响，哪些指标对信息热度有负面影响。我们在本题中定量分析中，我们建立了基于熵权法的指标模型来评价热指数的影响。通过量化指标，综合各指标体系，得出综合评价指标，进而得出一套热点评价指标体系的结论。。

3.3 问题三结果分析

本题将综合答复的相关性、完整性、可解释性来指定评估方案。相关程度是回复和留言的关联的反应程度，完整性方面我们则是对回复的留言从不同角度上进行了解释，通过交互设计对可解释性，进行可视化，三者有机结合给出评价方案。对于相关性，在本题中，我们觉得是至关重要的，民生反应民意，留言本意是希望所反映的问题得到解决，所以将对相关性，做出评价方法——归一化折损累计增益 NDCG。

这个指标通常用来衡量和评价搜索结果算法。

DCG 的两个思想：

1、高关联度的结果比一般关联度的结果更影响最终的指标得分；

2、有高关联度的结果出现在更靠前的位置的时候，指标会越高；

由于搜索结果随着关键词的不同，返回的数量是相对来说并不一致的，而 DCG 是一个累加的值，进而我们没法针对两个不同的搜索结果进行比较，需要对此进行归一化处理。

$$NDCG_p = \frac{DCG_p}{IDCG_p}$$

IDCG 为理想情况下最大的 DCG 值。

$$IDCG_P = \sum_{i=1}^{|rel|} \frac{2^{rel_i}-1}{log_2(i+1)}$$

其中|rel|表示，结果按照相关性从大到小的顺序排序，取前 P 个结果组成的集合。也就是按照最优的方式对结果进行排序。

据此我们，根据留言回复的相关性、完整性、可解释性做出了评价。得出一套评价方案

3.3.1 评价方案

随着互联网的不断发展，网络问政逐渐在人民群众中活跃起来，但群众获得的回复信息质量从现有数据看优劣区分明显。对互联网上的问答内容进行评价的问题可以追溯到 Google 的 PageRank 算法。在判断回复质量的问题上，An-swerBus 使用问题类型和答案的匹配、命名实体识别、指 代 消 解、冗 余 度、词匹配等指标来为候选答案排序。Cong 等利用频繁项集挖掘算法从论坛上提取问题与答案组，包括对问 题的识别和在同一主题中对答案的识别。Su 等指出问答社区中的回复平均信息质量较高，但其信息质量差异很大。 他们通过在 Yahoo！Answers 中 进 行 分 析 发 现，一个问题的所有回复中正确回复的比例只有 17%～45%，而一个问题至少有一个优质回答的比例为 65%～90%。所以，对现有回复信息作出评价尤为重要。

一、合理的答复

1.事实验证

对答复意见的质量进行评价的第一步为事实验证，从数据中筛选出哪些留言是有事实依据的，而非因为传言或者群众效应才留言的信息，来确认反映的对象是否客观存在。

2. 内容规范

（1）全文使用尊称

不管对方是什么身份，反映的事情是否合理，都应以尊称相待，以表示回复方的诚意和服务姿态。

（2）首段说明收件和处理安排

简要说明收件情况，如“您于 XX 年 XX 月 XX 日反映.....的情况已收悉”。

（3）向留言群众说明调查情况

简要阐述工作人员前往现场调查处理核实情况，应向对方说明调查结果是否属实。

（4）处理情况

工作人员对该留言是如何处理的，提出了怎样的整改意见，对方是否接受整改。

（5）整改到位情况

在回复中，向留言群众说明进展情况，保障留言网友的知情权。

（6）下一步工作计划

简述下一步工作思路，表明态度，传达正能量，如“我单位将加大日常监管和巡查力度，避免此现象的发生”，让群众放心。

（7）表示感谢

文末或者文初向留言者表示感谢，并欢迎对方继续监督和提供宝贵意见。

3.建立回复留言评估指标体系

如今许多的学者都在涉及问答网站的回答内容质量评估领域，其中 Zhu 等采用直观法、经验法、以及比较法，提炼了回答质量的十三个维度，并建立了多维质量评估模型。如下

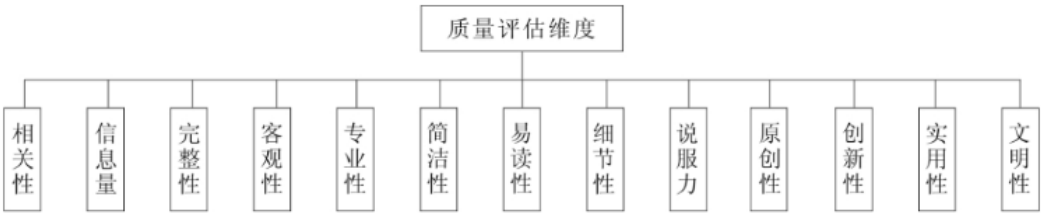

图 9：质量评估模型

从表中我们可以看出，每个维度之间不是一一独立的，而是存在一定的相关性。鉴于网络问政留言内容不似网络问答社区留言问题涉及面广，问题较杂，故在此十三个维度我们不需要一一分析。

二、回复现状

1.回复特征

（1）回复内容具有专业性，即回复内容中是否引用了一些专业性词语或者例子来解读留言内容。。

(2)内容充实，有事实依据。

(3)内容过于简短，回复信息没有解答留言群众的疑问或只是解决一小部分。

(4)回复内容完全套用模板，没有表达任何有用信息。

(5)对不同的留言问题都采取相同的回复，例如“已经移交相关部门处理”、“您的建议我们已经收到”等。

(6)容只表示感谢，例如：感谢您的留言，谢谢等；答复内容中包含对留言群众的继续追问，没有解决问题；答复内容与留言完全不相关等。

2.回复现状

（1）留言时间与回复时间间隔不等，回复是否及时反映了相关部门对网络问政的重视程度，网络问政的目的就在于希望通过互联网这一方便途径来聆听不同群众的声音，而长时间多打几个月才回复的会让留言群众认为相关部门不重视这件事。

（2)大部分答复首段都会说明收件和处理安排,例如“您于XX年XX月XX日反映.....的情况已收悉”。在礼貌用语方面表现较为出色。

（3）部分答复中只有“已经移交相关部门处理”类似字眼，群众可能缺乏某方面的专业知识，所以需要相关部门对群众做一个简单的解答，而不是用“已经移交相关部门处理”这种话来敷衍。

（4）采用社会化问答平台质量评估模型，在细节性、原创性、创新性这三个维度上表现较差，从数据中看出大部分回复内容较为相似，缺乏原创性。对于群众的留言不能全部说明原因或不能大部分说明原因。

（5）答复意见中无效答复较少，答复意见语句中礼貌用语。

三、改善与展望

答复质量优劣区分明显归根结底即为不同工作对待工作态度不同造成的,在如今网络信息化时代，网络问政无疑是一个大趋势，所以相关部门应该加强重视网络问政。

（1）加大宣传网络问政这一政策，截止 2020 年 3 月，中国网民数量已经达到 9.04 亿，加大宣传能够让更多基层百姓了解网络问政，并参与其中。

（2）对网络问政的留言群众设立门槛以及限制每人每天的留言数量，随着网络问政的人数逐渐增多，其中不乏有扰乱工作的人，通过设立门槛筛选出真正想通过网络问政这一形式得到答复的人。

（3）加强答复工作部门的建设，像培养面对面解决群众疑问的工作人员一样培养网络人工客服，使回复更具有专业性，创新性。能让群众对网络问政这一政策有更好的体验感。

（4）定期发放线上调查问卷，调查此阶段群众对答复的满意度，中和群众意见，若有不足之处进行整改。

4.结论

针对于不同问题，每个人根据自身情况，从不同媒体渠道进行留言，面对错综复杂的留言，将其进行分类整理，通过对文本中文的分词，对数以万计的留言进行了大的分类，同时，进行关键词的删选，Knn 聚类分析，选出热点问题，这有利于热点问题的解决，对人工的工作量来说，是一个大的筛选，同时结合相关性、可解释性、完整性对留言的回复进行评估，进而逐步改进留言回复，对于政府和人民来说，都是利用现代科技技术进一步解决问题。

5.参考文献

（1） 邢娟韬，白金牛。《基于改进 ML-KNN 算法的文本分类研究》。2020.3.25
（2） 张波，黄晓波。《基于 TF-IDF 的卷积神经网络新闻文本分类优化》.2020
（3） 卜凡军。KNN 算法的改进及其在文本分类中的应用。2009
（4） 杨虎。面对海量短文文本去重技术的研究和实现.国防科技大学.2007
（5） 翟东海，鱼江等。最大距离法选取初始簇中心的 K-means 文本聚类算法的研究。西南交通大学.2014.
（6） 张旭，孙玉伟等.不同特征对文本聚类效果的比较研究—以新闻文本为例子.2019
（7） 王千，王成，冯振元，叶金凤.K-means 聚类算法研究综述.2012
（8） 曹卫峰.中文分词关键技术研究.南京理工大学.2009
（9） 赵琳瑛.基于隐马尔科夫模型的中文命名实体识别研究.西安电子科技大学.2007