“智慧政务”中的文本挖掘应用

摘要

在互联网的时代，随着微信、微博、市长信箱等网络问政平台逐步成为政府了解民意、汇聚民意等重要途径。相关的民意文本数量不断攀升，给工作人员整理资料带来许多挑战。本文基于自然语言处理技术和文本挖掘对于互联网的群众留言进行研究。

针对问题一，通过中文分词的基本原理以及 jieba 分词的方法对留言主题进行分词。过滤停用词，构建词袋模型，计算 $TF-IDF$ 并提取关键词作为文本特征，建立特征矩阵。最后构建朴素贝叶斯分类器，求解对于给出的留言主题在此项条件下各个类别出现的概率，得到一级标签分类模型。最后用 $F-Score$ 对分类进行检验，最终平均 $F-Score$ 为 0.727，验证了分类的有效性。

针对问题二，利用 excel 去重同个用户描写一模一样的留言主题，数据预处理后，本文基于肘部法则得到最佳聚类数 K，利用 K-means 对其进行聚类，即将某一时段内反映特定地点或特定人群问题的留言进行归类。在聚类和去重的基础上，以反对数、点赞数、留言条数、留言与中心问题的相关程度作为影响因素，赋予一定权重，建立热度评价模型，得到排名前五的热点问题并加以描述。

对于问题三，为了更好地评价相关部门对留言的答复情况，本文采用了模糊综合评价对留言详情进行分析。先用 excel 统计留言详情的字数情况，以答复的字数作为检验完整性的因素之一，其次还有答复有效程度，礼貌程度等作为评价的参考因素。同时结合 python 计算文本的相似度，对于每个留言都按照多个因素进行模糊综合评价，最后根据总评分可得一套评价方案。

关键词：留言主题　贝叶斯分类器　K-means 聚类　热度评价　模糊综合评价

Text mining application in "smart government"

Abstract

In the era of the Internet, with the development of wechat, microblog, mayor's mailbox and other online political platforms, the government has gradually become an important way to understand and gather public opinion. The number of relevant public opinion texts is increasing, which brings many challenges to the staff in sorting out information. Based on natural language processing technology and text mining, this paper studies the public message on the Internet.

In order to solve the problem one, we use the basic principle of Chinese word segmentation and the method of Jieba word segmentation to segment the message topic. Filter the stop words, construct the word bag model, calculate the $TF-IDF$, extract the keywords as the text features, and establish the feature matrix. Finally, the naive Bayes classifier is constructed to solve the probability of each category for the given message subject under this condition, and the first level label classification model is obtained. Finally, $F-Score$ is used to test the classification, and the final average $F-Score$ is 0.727, which verifies the validity of the classification.

Aiming at the problem as like as two peas, we use Excel to duplicate the same message topic. After data preprocessing, we get the best clustering number K based on the elbow rule. We use K-means to cluster the two groups, and classify the words that reflect the specific location or specific group problem in a certain period. On the basis of clustering and de duplication, taking the inverse logarithm, the number of likes, the number of messages and the degree of correlation between messages and the central problem as the influencing factors, giving a certain weight, establishing a heat evaluation model, getting the top five hot issues and describing them.

For the third question, in order to better evaluate the response of the relevant departments to the message, this paper uses the fuzzy comprehensive evaluation to analyze the details of the message. Firstly, excel is used to count the number of words in the details of the message, and the number of words in the reply is one of the factors to test the integrity. Secondly, the validity of the reply and the politeness of the

reply are used as the reference factors for evaluation. At the same time, combining Python to calculate the similarity of text, each message is evaluated according to multiple factors. Finally, a set of evaluation scheme can be obtained according to the total score.

Key words: message subject Bayesian classifier K-means clustering
Evaluation of heat fuzzy comprehensive evaluation

目录

一、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，以及相关部门对部分群众留言的答复意见。结合中文分词的基本原理以及 jieba 分词的用法对附件的群众留言的标签进行分类，达到以下三个目标：

（1）附件一过滤停用词之后，构建词袋模型，通过建立特征矩阵，使用朴素贝叶斯分类器进行分类，最终实现对文本留言内容的一级标签分类，以便后续将群众留言分派至相应的职能部门处理。

（2）使用 k-means 聚类的方法附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，可得排名前 5 的热点问题。

（3）根据相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，大大减轻政府人员的工作量。

二、总的分析方法与过程

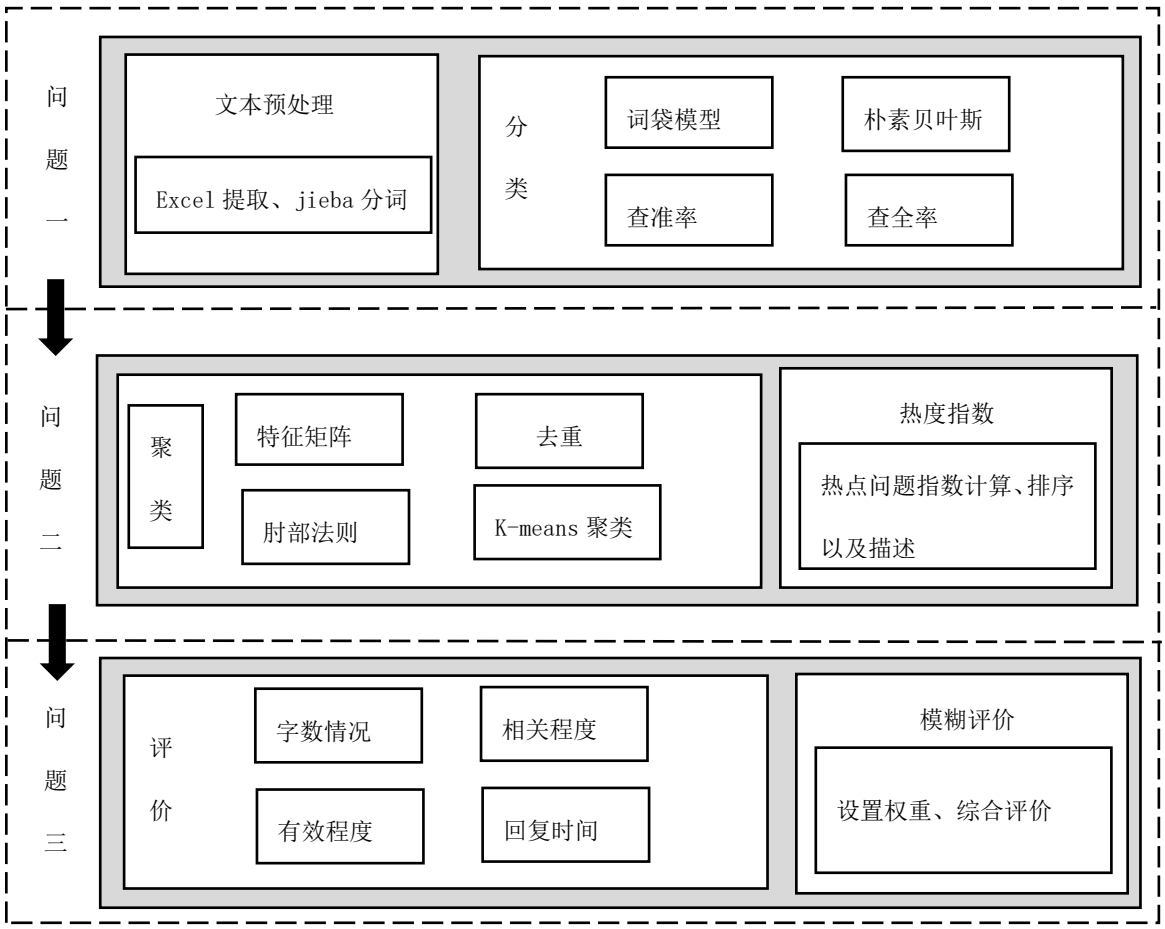


图 1 总体的流程图

题目要求的是对于附件给出了收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见。利用自然语言处理和文本挖掘的方法解决以下 3 个问题[1]。下面本文详细讲解问题的分析与过程。

2.1 问题 1 分析方法与过程

2.2.1 流程图

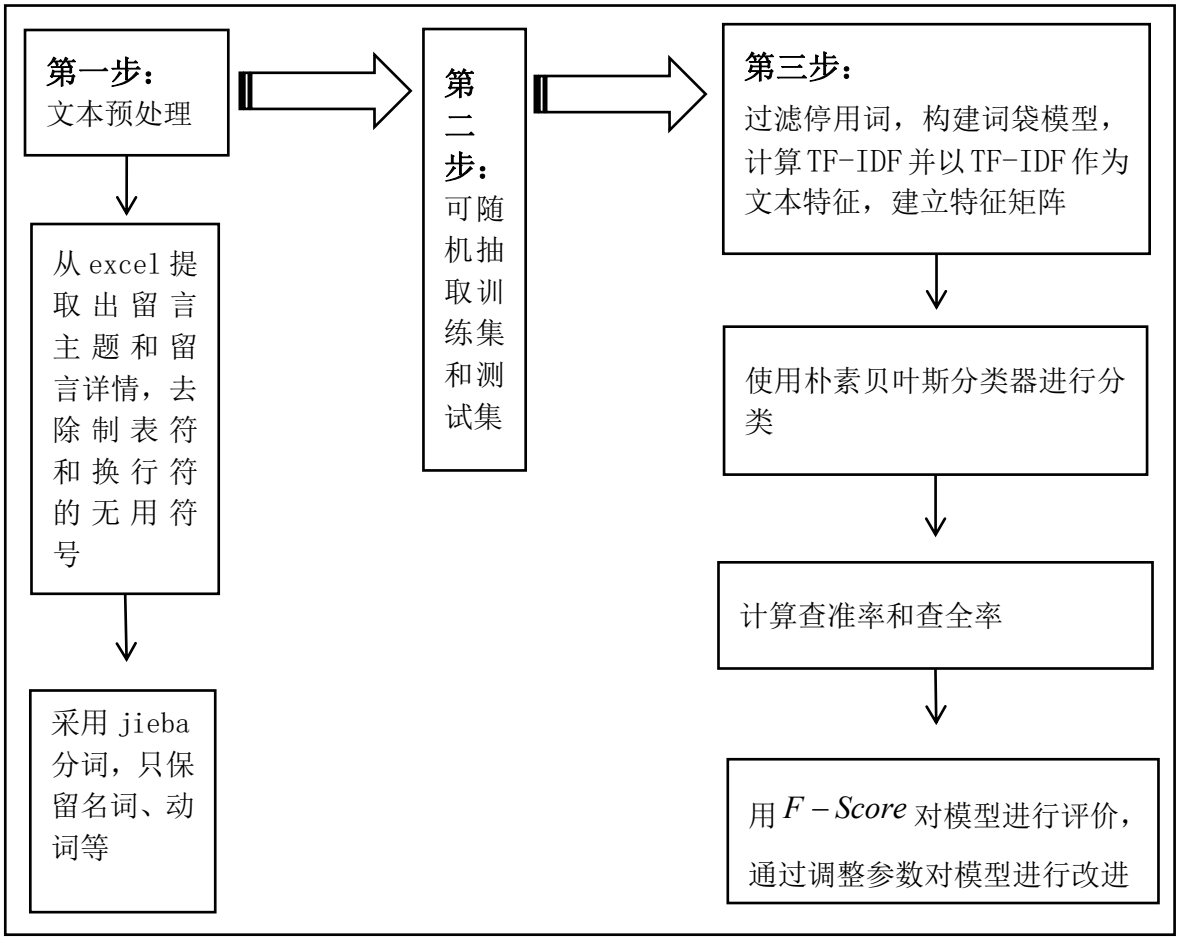


图 2：问题 1 流程图

2.1.2 对留言主题进行中文分词

将留言主题序列切分成一个一个单独的词，即将连续的留言按照一定的规范重新组合成词序列。应用 python 中的文本分词包 jieba 进行分词，试图将句子最精确地切开。基于 Trie 树结构实现高效的词图扫描，每个留言主题使用 DAG(查字典)和动态规划，可得到最大概率路径。当然不排除出现未登录词，可以采用基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

例如：

A 市西湖建筑集团占道施工有安全隐患

A 市在水一方大厦人为烂尾多年，安全隐患严重

投诉 A 市 A1 区苑物业违规收停车费

A1 区 A2 区华庭自来水好大一股霉味

... ...

这些留言主题经过分词成:

```
A/市/西湖/建筑/集团/占/道/施工/有/安全隐患
A/市/在水一方/大厦/人为/烂尾/多年/，/安全隐患/严重
投诉/A/市/A1/区苑/物业/违规/收/停车费
A1/区/A2/区华庭/自来水/好大/一股/霉味
```

2.1.3 构建词袋模型

具有相似意义的词在向量空间中的位置也是相似的。这样，我们根据词向量进行分类训练，从本质上对意义相似的句子进行分类。根据出现的频率，文本被视为向量。将词袋模型的结果从词或短语的频率向量转换为词或短语的重要性向量。

2.1.4 权重策略：计算 TF-IDF

在对留言主题信息分词后，本文需要将这些词语转换为向量，这里应用 if-idf 算法，把留言主体的内容转为权重向量。具体的步骤原理如下[2]：

Step1：统计词频，即 TF 的权重(Term Frequency)

词频（TF）即为某个词在该类文本中出现的频率。这个数字经常会被归一化，以防止它偏向长的文件。所以对“词频”进行标准化。

$$\text{词频的重要性（TF）}=\frac{\text{某词在该类文本中出现的次数}}{\text{文本的总数词}} \qquad (1)$$

Step2:计算 IDF，即反文档频率(Inverse Document Frequency)

反文档频率（IDF）包含条目的文档越少，IDF 就越大，这意味着具有很好的区分类别的能力。如果条目经常出现在类的文档中，则意味着条目可以很好地表示此类文本的特征。这些条目应该赋予它们更高的权重，并被选为这类文本的特征词，以区别于其他文档。|D|:语料库中的文件总数。

$$\text{反文档频率（IDF）}=\log\left(\frac{|D|}{\text{包含该词的文本数}+1}\right) \quad (2)$$

Step3：计算 tf-idf 值

特定文件中的高频字和整个文件集中的低频字可以产生高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常用词，保留重要词。

$$\text{TF-IDF} = TF * IDF \quad (3)$$

权重策略——TF-IDF：使用 TF-IDF 发现特征词，并抽取为反映文档主题的特征

2.1.5 构建朴素贝叶斯分类器

朴素贝叶斯是指，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，哪个最大，就认为此待分类项属于哪个类别。这是一种经典的高精度机器学习算法。它可以作为一个分类器，以获得良好的效果。用一个 fit 方法用来模型训练，有一个 predict 方法用来模型预测，在此我们就传入了训练特征和训练目标进行了模型的训练。

故做了以下定义分类。

Step1：设 $x=\{a_1,a_2,...,a_m\}$ 作为留言主题的待分类项，而每个 a 为 x 的一个特征属性。

Step2：有类别集合 $C=\{y_1,y_2,...,y_n\}$

Step3：计算 $P(y_1|x),P(y_2|x),...,P(y_n|x)$

统计得到在各类别下各个特征属性的条件概率估计。即

$$p(a_1|y_1),p(a_2|y_1),...,p(a_m|y_1),...,p(a_m|y_n) \quad (4)$$

假如各个特征属性是条件独立的，则根据贝叶斯定理有如下推导：

$$p(y_i|x)=\frac{p(x|y_i)p(y_i)}{p(x)} \quad (5)$$

最后是应用阶段。 此阶段的是使用分类器对要分类的项目进行分类。 输入为分类器和要分类的项目，输出是要分类的项目和类别之间的映射关系。 该阶段也是机械阶段，也是由 python 程序完成的。

2.1.6 分类结果评估

为了评估算法的准确性，通常使用$F-Score$对分类方法进行评价，我们定义有三个基本指标：

Step1：召回率（recall rate,查全率）设为R_i：

$$召回率=\frac{系统检索到的相关文件}{系统所有相关文件综述} \tag{6}$$

Step2：准确率（Precision，查准率）设为P_i：

$$查准率=\frac{系统检索到的相关文件}{系统所有的检索到的文件数} \tag{7}$$

准确率和召回率是存在一定的联系的，理想情况下是二者都高，但是一般情况下会成反例。即准确率高，召回率就低；召回率高，准确率就低。

Step3：对分类进行评价：

$$F_1=\frac{1}{\mathrm{n}}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{8}$$

2.2 问题 2 分析方法与过程

2.2.1 文本预处理

我们从附件 3 中提取出留言主题和详情，去除制表符和换行符的无用符号。以每条留言为一行的形式存为 txt 纯文本，便于程序的运行。通过 Python 使用 jieba 分词，只保留名词、动词等关键词汇。最后计算 tf-idf 并以 tf-idf 作为文本特征，建立特征矩阵。这个跟问题一的处理方式相似。

本文后续的热度评价模型中，不同问题的留言次数将成为热度评价的指标之一。为使得热度评价模型更加准确，我们找出相同 ID 给出的相同留言进行去重，减少了无关因素的干扰。同样的为使得整个模型更加简便，不占用过多机器性能，加快运算速度，我们利用 excel 去重同个用户描写一模一样的留言主题，因为里面有些留言发重复了，会影响热度的评价，进行数据预处理。

2.2.2 流程图

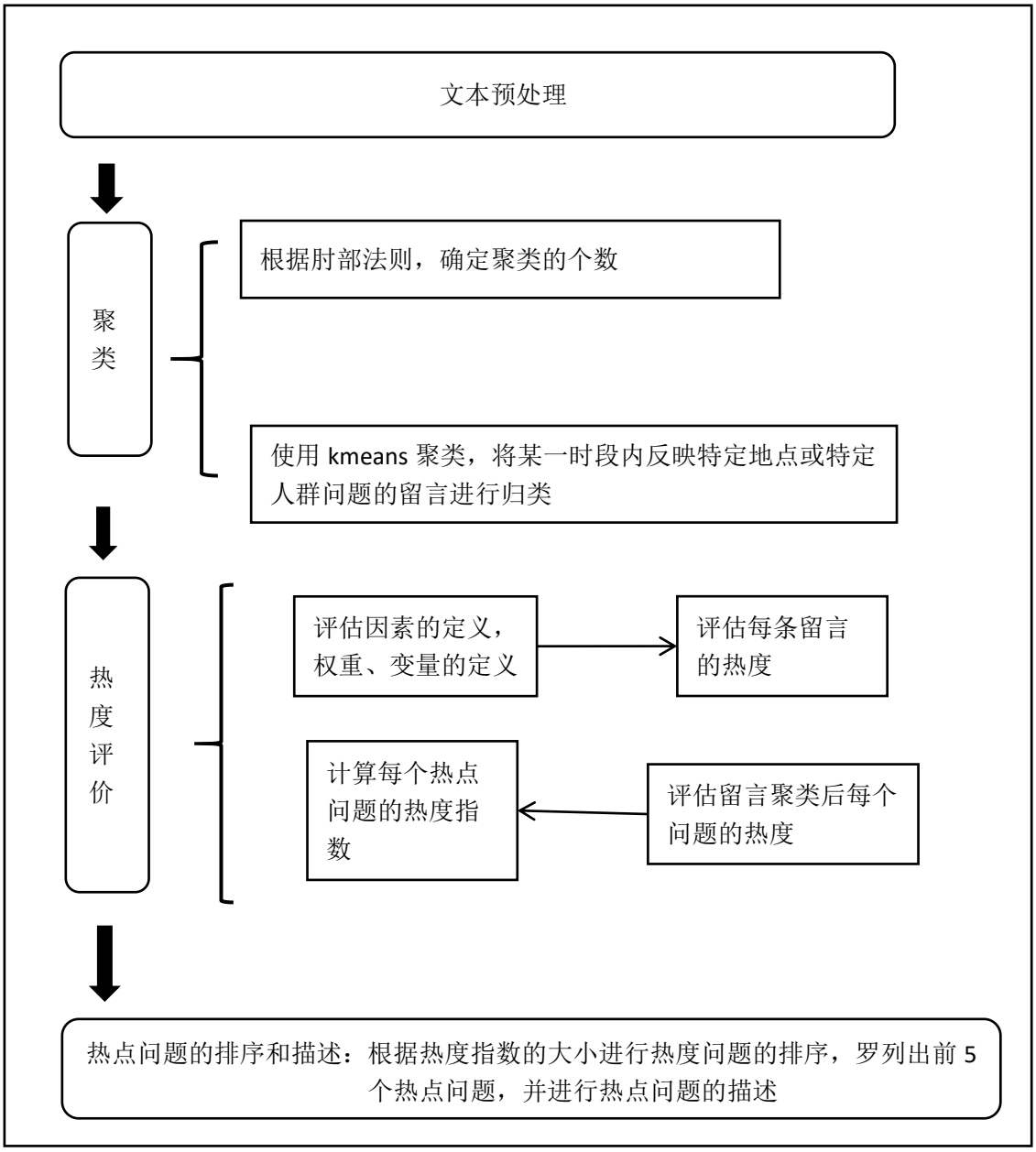


图 3：问题二流程图

2.2.3 肘部法则确定聚类数 K

经过文本预处理建立特征矩阵后，准备采用 K-means 算法将某一时段内反映特定地点或特定人群问题的留言进行归类。但在执行 K-means 算法时，需要选定数据空间中 K 个对象作为初始聚类中心。为提高聚类的准确性，使用肘部法则确定 K 个聚类中心。

具体算法如下：

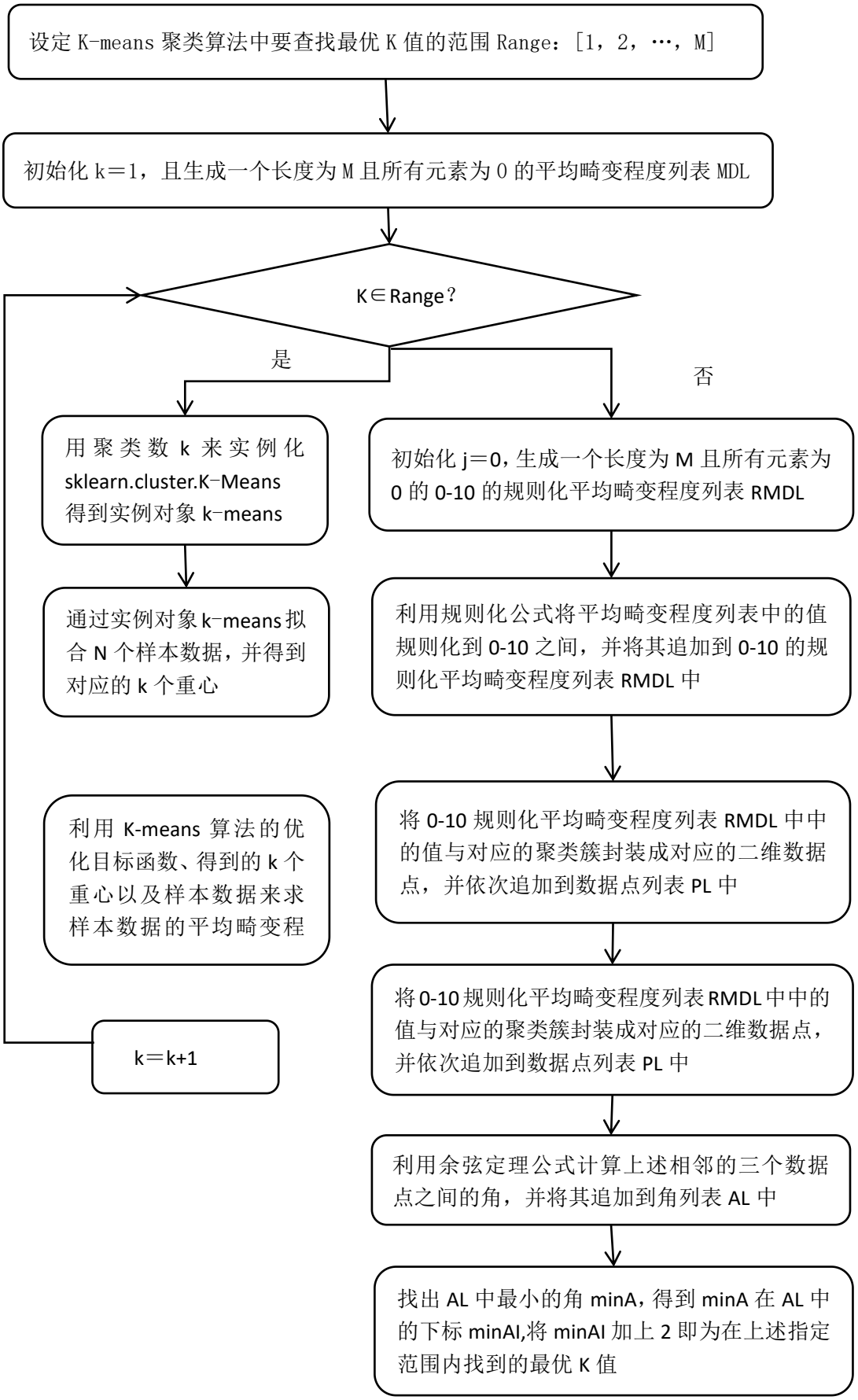


图 4：确定 K 个聚类中心的流程图

现有常用的肘部法则需要通过人工观察肘部方法画出的图进而识别最佳的聚类数 K。考虑到将本文中建立的模型运用到实际的问政平台中时，人工观察的不便性，留言数量多，更新快的特点。我们将现有的肘部法则改进，将不需要通过人工观察去识别，进而使其可以更好地应用到自动聚类的系统中。

2.2.4 K-means 聚类

聚类算法有很多种，K-means 在众多聚类算法中，具有简单、好理解、运算速度快的特点，在问政平台大量数据面前，K-means 聚类无疑是一个最佳的选择。

K-means 聚类的原理如下[4]：

随机选取 K 个对象作为初始的聚类中心，计算每个对象与各个聚类中心的距离，常见的有欧式距离、曼哈顿距离等。把每个对象分配给聚类中心，全部对象被分配后，每个聚类中心的所有对象被重新计算。不断重复上述过程，直至达到终止条件。在终止条件中，我们选择以没有（或者最小数目）聚类中心再发生变化作为终止条件。

在理解其原理、确定其终止条件后，我们做出如下步骤：

Step1:通过肘部法则将一个包含 d 维数据点的数据集 $X=\{X_1,X_2,\cdots X_i,\cdots,X_n\}$，组织为 K 个划分 $C=\{c_k,i=1,2,\cdots,k\}$。每一个划分代表一个类 c_k，每一个类 c_k 有一个类别中心 μ_i，即 μ_1，μ_2,…，$\mu_i\in R^n$。

Step2：选取欧式距离作为相似性和距离判断准则，计算该类各点到聚类中心 μ_i 的距离平方和。

$$J(C)=\sum_{X_i\in c_k}\|X_i-\mu_k\|^2 \tag{9}$$

该公式利用欧式距离计算出 Step1 划分后剩下元素到 K 个簇中心的相异度（距离代表相异程度），将这些元素分别划分到相异度最低的簇。

Step3：为准确地实现聚类，根据 Step2 的聚类结果，重新计算 K 个簇各自的中心，根据最小二乘法以及拉格朗日原理，聚类中心 μ_k 取 C_k 类各数据点的平均值，计算公式如下：

$$J(C)=\sum_{K=1}^{K}J(c_k)=\sum_{K=1}^{K}\sum_{X_i\in C_i}\|X_i-\mu_i\|^2=\sum_{K=1}^{K}\sum_{i=1}^{n}r_{ki}\|X_i-\mu_i\|^2 \tag{10}$$

其中，

$$r_{ki}=\begin{cases}1,若X_i\in C_i\\0,若X_i\notin C_i\end{cases} \tag{11}$$

即，r_{ki}表示数据点X_i被归类到C_i时为1，否则为0。

Step4：将X中全部元素按照新的中心重新聚类，不断重复，知道聚类结果不再发生变化后，将结果输出。

2.2.5 热度评价模型

Step1：计算每个留言的热度。将影响每一条留言热度评估的因素定义为反对数和点赞数。用P_i来表示第i个留言，$HD(P_i)$表示第i条留言P_i的热度；$AN(P_i)$表示第i条留言P_i的反对数；$GN(P_i)$表示第i条留言P_i的点赞数；a_1,a_2分别表示反对数、点赞数这两个影响因素的权值，且$\sum_{i=1}^{2}a_i=1$。那么，每一条留言的热度为：

$$HD(P_i)=a_1\times AN(P_i)+a_2\times GN(P_i) \tag{12}$$

Step2：计算每个热点问题的热度。将聚类结果的每一类定义为一个热点问题，即N条留言具有一个共同的热点问题T_j；该热点问题的热度为$HD(T_j)$；留言热度对该热点问题的热度影响记为$M(T_i)$则：

$$M(T_i)=\sum_{k=1}^{k}\omega_k\times HD(P_i) \tag{13}$$

ω_k为每个$HD(P_i)$权值，和整个热点问题内其他帖子的相似度越高，其权重越大，因此，我们采用Weibull分布，

$$f(x;\lambda,i)=\frac{i}{\lambda}\left(\frac{x}{\lambda}\right)^{i-1}e^{-\left(\frac{x}{\lambda}\right)^{i}} \tag{14}$$

取$\lambda=1$，$i=1.5$则：

$$\omega_k=\frac{1.5\sqrt{k}\times e^{-k^{1.5}}}{\sum_{k=1}^{k}1.5\sqrt{k}\times e^{-k^{1.5}}} \tag{15}$$

其中，$\sum_{k=1}^{k}\omega_k=1$。

Step3：b_1,b_2 分别表示留言热度和留言条数对该热点问题热度的影响因素权值。同样的，$\sum_{i=1}^{2}b_i=1$。那么，每一个热点问题的热度为：

$$HD（T_i）=b_1\times M(T_i)+b_2\times N \tag{16}$$

2.2.6 热点问题的排序和描述

通过 K-mean，知道可得到聚类后相同的结果，并将结果输出。结合热度留言评价模型，通过 Python，我们可以计算出每一个热点问题的热度，导出到 Excel 后，通过排序可得出各个热点问题的热度排名。将前五个热点问题列出后，从而达到热点问题排序和描述的目的。

2.3 问题 3 分析方法与过程

2.3.1 问题 3 流程图

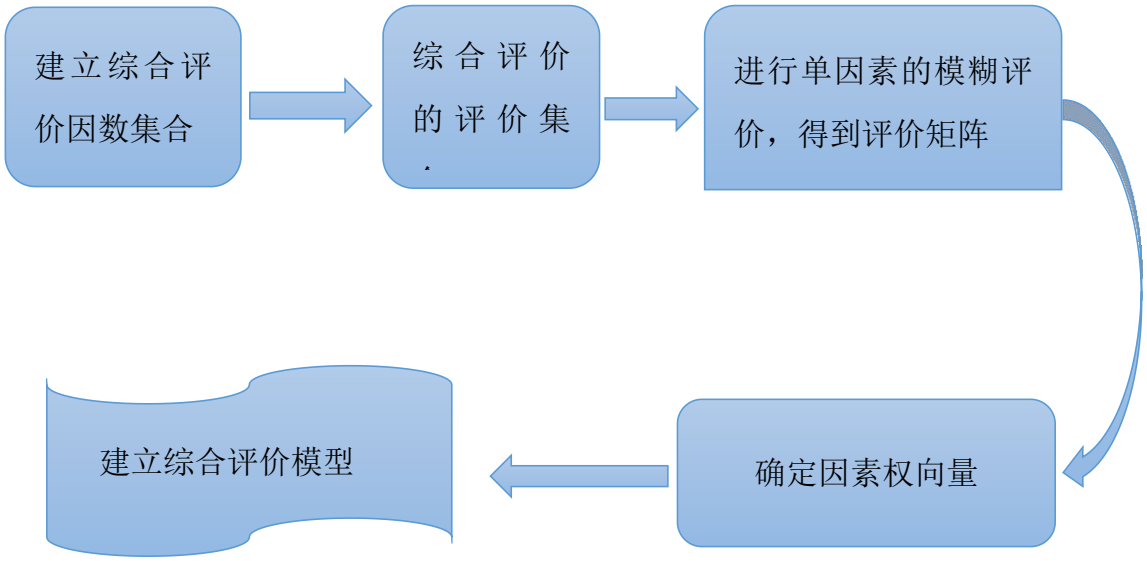


图 5：问题 3 流程图

模糊综合评价是对受多种因素影响的事物做出全面评价的一种十分有效的多因素决策方法，其特点是评价结果不是绝对地肯定或否定，而是以一个模糊集合来表示。这种评价方法非常适合对答复意见进行评价。

2.3.1 综合评价模型的建立步骤

- 第一步：建立综合评价的因素集

因素集是以影响评价对象的各种因素为元素所组成的一个普通集合，通常用 U 表示，$U=(u_1,u_2,\cdots,u_m)$，其中 U_i 元素代表影响评价对象的第 i 个因素。这些因素，通常都具有不同程度的模糊性[3]。

这里设评定留言答复意见质量等级的指标集 $U=(u_1,u_2,u_3,u_4)$，u_1 表示为答复时长，u_2 表示为答复有效程度，u_3 表示为答复字数，u_4 表示为礼貌程度

- 第二步：建立综合评价的评价集

评价集是评价者对评价对象可能做出的各种结果所组成的集合，通常用 V 表示，$V=(v_1,v_2,\cdots,v_n)$，其中元素 v_j 代表第 j 种评价结果，可以根据实际情况的需要，用不同的等级、评语或数字来表示。

这里设评定留言答复意见质量等级的评价集为 $V=(v_1,v_2,v_3,v_4)$，分别表示很好、较好、一般、不好。

- 第三步：进行单因素模糊评价，获得评价矩阵

若因素集 U 中第 i 个元素对评价集 V 中第 1 个元素的隶属度为 r_{i1}，则对第 i 个元素单因素评价的结果用模糊集合表示为：$R_i=(r_{i1},r_{i2},\cdots,r_{im})$，以 m 个单因素评价集 $R_1,R_2,\cdots,R_m$ 为行组成矩阵 R_{mn}，称为模糊综合评价矩阵。

- 第四步：确定因素权向量

评价工作中，各因素的重要程度有所不同，为此，给各因素 u_i 一个权重 a_i，各因素的权重集合的模糊集，用 A 表示: $A=(a_1,a_2,\cdots,a_m)$。

我们认为答复时长和答复有效程度同样重要，答复字数次之，礼貌程度再次之，所以在这里取权数分配 $A=(0.35，0.35，0.2，0.1)$

- 第五步：建立综合评价模型

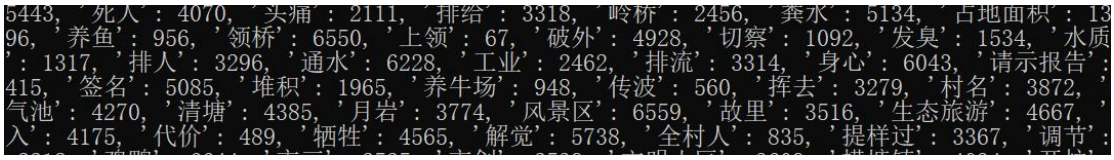
确定单因素评判矩阵 R 和因素权向量 A 之后，通过模糊变化将 U 上的模糊向量 A 变为 V 上的模糊向量 B，即 $B=A*R$

计算出模糊向量 B 后即可得到答复意见的综合评价。

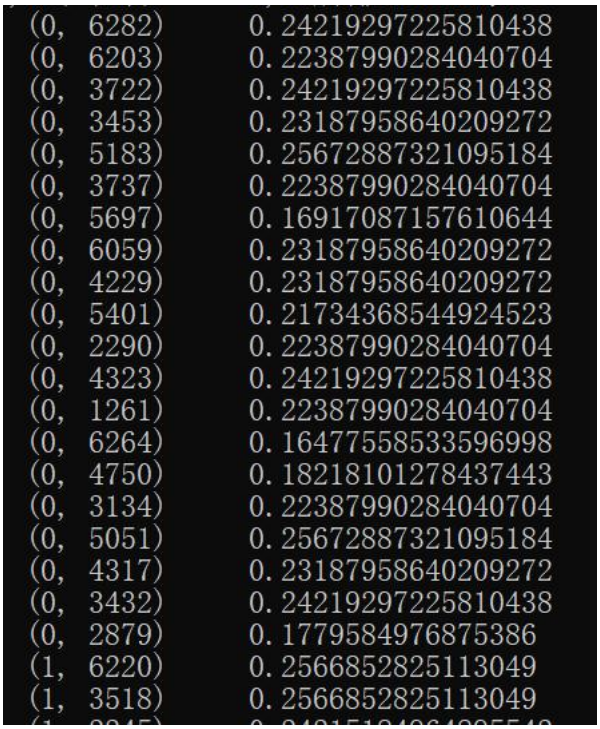
三、结果分析

3.1 问题 1 结果分析

3.1.1 词典和 TF-IDF 的结果



这里面是用 python 编的部分词典。我们通过上述的算法，通过编程，得到了 tf-idf 值。下面是截取部分数据。



3.1.2 分类评估的结果

通过我们所建立关于留言内容的一级标签分类模型，可以实现对留言主题的分类。如下图所示。

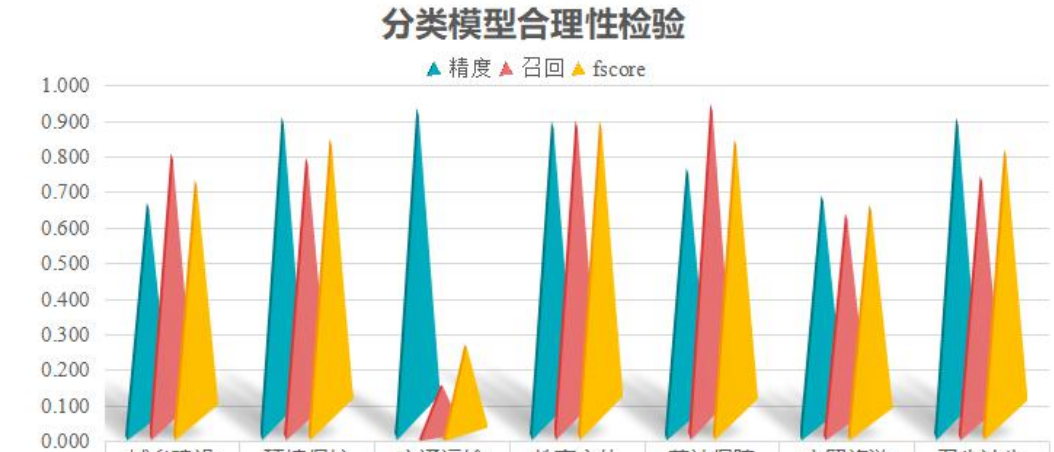
... ...

```
test_seg/环境保护/2943.txt : 实际类别:  环境保护 -->预测分类:  环境保护
test_seg/环境保护/2944.txt : 实际类别:  环境保护 -->预测分类:  环境保护
test_seg/环境保护/2945.txt : 实际类别:  环境保护 -->预测分类:  环境保护
test_seg/环境保护/2946.txt : 实际类别:  环境保护 -->预测分类:  教育文体
test_seg/环境保护/2947.txt : 实际类别:  环境保护 -->预测分类:  环境保护
```

为了可以有效的评价模型建立的合理性，我们通过计算 $F-Score$，然而准确率和召回率是存在一定的联系的，理想情况下是二者都高，但是一般情况下会成反例。即准确率高，召回率就低；召回率高，准确率就低。我们通过 python 可以得到下面关于“城乡建设”，“环境保护”，“交通运输”，“教育文体”，“劳社保障”，“商贸旅游”，“卫生计生”这 7 个一级留言主题的分类后的情况。

```
['城乡建设', '环境保护', '交通运输', '教育文体', '劳社保障', '商贸旅游', '卫生计生']
精度: ['0.670', '0.912', '0.937', '0.899', '0.768', '0.691', '0.909']
召回: ['0.810', '0.797', '0.158', '0.901', '0.949', '0.639', '0.746']
fscore: ['0.734', '0.851', '0.271', '0.900', '0.849', '0.664', '0.820']
平均fscore: 0.727
```

通过计算，我们可知最后的 $F-Score$ 的值分别为 0.734，0.851，0.271，0.9，0.849，0.664，0.82。



	城乡建设	环境保护	交通运输	教育文体	劳社保障	商贸旅游	卫生计生
精度	0.670	0.912	0.937	0.899	0.768	0.691	0.909
召回	0.810	0.797	0.158	0.901	0.949	0.639	0.746
fscore	0.734	0.851	0.271	0.900	0.849	0.664	0.820

因为一般情况下，查准率与查全率很难同时大，能得到上述结果，说明分类器已经达到很好的效果了，从而进一步说明一级标签分类模型适合度有不错的成效。

3.2 问题 2 结果分析

3.2.1 去重的结果分析

通过 excel，我们发现在同一类别中出现了相同 ID 给出的相同留言。找到后并去重。

例如：

A00077446	A1区A2区华庭地下车库要变成垃圾场了	2019/10/29 17:52:23
A00077446	A1区A2区华庭负一楼车库又成垃圾场了	2019/12/6 14:29:29

⋮

A000112212	A1区朝阳街道解放东路二里牌向韶村马路市场死灰复燃	2019/3/20 17:42:38
A000112212	A1区朝阳街道解放东路二里牌向韶村马路占道经营严重	2019/3/20 18:03:21

3.2.1 热点问题的排序的结果分析

热度排名	问题ID	热度	时间范围	地点/人群	问题描述
1	1	53.05	2019/4/10至2020/1/26	A2区丽发新城	附近修建搅拌厂噪音污染扰民
2	2	39.23	2018/1/12至2020/1/7	A市住宅区	麻将馆扰民
3	3	38.245	2019/7/7至2019/9/1	A市伊景园滨河苑	车位捆绑销售行为
4	4	36.18	2019/1/11至2019/7/8	西地省A市	车贷案件典型诈骗
5	5	28.02	2019/1/8至2019/12/4	A5区劳动东路魅力之城	餐馆油烟扰民

根据以上操作，我们得出最终的热点问题排名。

从图中我们可以知道，民众反应最强烈的问题分别为附近修建搅拌厂噪音污染扰民的问题、麻将馆扰民的问题、车位捆绑销售行为的问题、车贷案件典型诈骗的问题、餐馆油烟扰民的问题。它们的热度指数分别是 53.05、39.23、38.245、36.18、28.02。

部分热点问题的时间范围较短，例如：车位捆绑销售行为的时间范围仅两个月，说明该情况在出现两个月后，没有群众再次反映，该问题已经得到较好的改善。且部分热点问题（比如麻将馆扰民）的时间范围较长，说明该问题长期未被解决，但已打扰到许多人的生活，需要被重视起来。

五大热点问题大致可以分为民众与工业生产污染的矛盾、民众与小店面生产过程中不利行为的矛盾、消费者与生产者的矛盾以及诈骗行为。

排名第一的热点问题与其他四个问题的指数相差较大，说明该地修建搅拌厂已经打扰到许多民众的生活，迫切需要解决根据这五个热点问题，我们罗列出热点问题的明细，如下图所示：

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
1	188809	A909139	A市万家丽南路丽发新城居民区附近搅拌站扰民	2019/11/19 18:07:54	开发商在小区旁50米处建搅拌站，运渣车吵得人精神崩	0	1
1	189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	2019-11-13 11:20:21	小区不到百米的地方建搅拌站。可想而知，一个大型搅拌	0	0
1	190108	A909240	丽发新城小区旁边建搅拌站	2019-12-21 15:11:29	学，扬尘严重影响几千名学生的健康，很多业主反应	0	1
1	190523	A00072847	A市丽发新城违建搅拌站，彻夜施工扰民污染环境	2019/12/26 13:55:15	尘，噪音污染严重；3、搅拌站几百米外就是小学，扬尘	0	0
1	191943	A00038563	A市A2区丽发新城道路坑坑洼洼	2019/7/3 12:03:51	期与第二期中间的道路坑坑洼洼，下雨泥巴，天晴扬尘，	0	1
1	193091	A00097965	A市富绿物业丽发新城强行断业主家水	2019/6/19 23:28:27	公司发票，物业只提供地摊上买的收据，对于不交水费的	0	242
1	199379	A00092242	A2区丽发新城附近修建搅拌厂，严重污染环境	2019/11/25 10:17:56	影响，个别居民因此还得了疾病住院，该地区作为长株潭	0	0
1	203393	A00053065	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	2019/11/19 14:51:53	土搅拌带来了巨大的粉尘，严重影响居民健康；同时设	0	2
1	208714	A00042015	A2区丽发新城附近修建搅拌站，污染环境，影响生活	2020-01-02 00:00:00	题，小区内空气质量和声环境质量急剧下降，我们不敢打开	0	4
1	213464	A909233	投诉丽发新城小区附近违建搅拌站噪音扰民	2019-12-10 12:34:21	小区附近违建大型搅拌站。该搅拌站的设备太吵了，每天	0	0
1	213930	A909218	A2区丽发新城附近违规乱建混凝土搅拌站谁来监管？	2019-12-27 23:34:32	为此，全小区居民强烈呼吁政府和有关职能部门，加强	0	0
1	214282	A909209	A市丽发新城小区附近搅拌站噪音扰民和污染环境	2020-01-25 09:07:21	厂啊，天天吵天天吵，烦死了不仅吵还臭！说好的绿化	0	0

4	195090	A00039089	魅力之城小区临街门面油烟直排扰民	2019/09/03 12:29:01	烟烧烤熏死人。一天24小时都是烟。请政府关闭处理这个	0	3
4	360100	A324156	魅力之城小区临街门面油烟直排扰民	2019-09-05 12:29:01	烟烧烤熏死人。一天24小时都是烟。请政府关闭处理这个	3	0
4	245136	A909117	万科魅力之城小区底层门店深夜经营，各种噪音扰民	2019/09/04 21:00:18	耳边都充斥着吆喝声、拼酒声、炒菜烧烤的锅铲炭火声，	0	0
4	360109	A0080252	万科魅力之城小区底层门店深夜经营，各种噪音扰民	2019-09-04 21:00:18	耳边都充斥着吆喝声、拼酒声、炒菜烧烤的锅铲炭火声，	0	0
5	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	万分，急切盼望有案情消息总是失望，四处诉求也无效。	0	821
5	217032	A00056543	严惩A市58车贷特大集资诈骗案保护伞	2019/2/25 9:58:37	兑大股东苏纳和小股东、苏纳弟弟苏吕是挂名；说担保公	0	790
5	194343	A000106161	承办A市58车贷案警官应跟进关注留言	2019/3/1 22:12:30	。但是，A市A4区经侦并没有跟进市领导的留言，案件调	0	733
5	268251	A000106090	西地省58车贷立案近半年毫无进展，单位回复让人心寒	2019/2/2 15:03:05	就是不抓捕控制平台高管和资产，这种情况，要求还受害	0	25
5	240554	A00029163	A市58车贷老板跑路美国，经侦拖延办案	2019/2/10 20:58:40	不力，纵容嫌犯。这是涉嫌保护伞直观表现。邢速夫妇涉	0	6
5	226265	A000106448	恳请A市经侦公正办理58车贷案件，还我们受害人一个公道	2019/5/28 15:08:51	警官说，要相信经侦，但这样的经侦我们怎么相信呢？还	0	3
5	254532	A000106062	A市58车贷恶性退出立案近半年没有发过一次案情通报	2019/1/14 22:08:20	委会为其脱罪。于是，在58官网上挂出选举的公告和参选	0	3
5	272413	A000106062	西地省A市58车贷恶性退出，A4区立案已近半年毫无进展	2019/1/14 20:23:57	让广大出借人极为不满但我们出借人又没有任何办法，对	0	2
5	218132	A000106090	再次请求过问A市58车贷案件进展情况	2019/1/29 19:15:49	。当我们把此信息告知办案警官毛浚时，他说这事他知道	0	0
5	223787	A00034861	西地省58车贷案件创造全国典型诈骗案，立案至今无公告	2019/1/11 21:12:34	冻结，查出的相关资产4任由犯罪嫌疑人上蹿下跳，四处	0	0

从罗列出的明细中，我们发现大部分留言均贴近民众的生活，是生活中看似平常却很重要的事情。因此，问政平台应当多关注切近民众生活的小事，从小事逐渐改善大众的生活。

3.3 问题 3 结果分析

我们随机抽取了多条留言答复进行评价，验证了这套评价方案的可行性，同时发现了这套评价模型过于依赖人工建立的模糊综合评价矩阵，使得这套评价方案难以使用程序进行自动化的智能评价。

类似这这种属于答复不太完整，可解释性不强。没能给网友很好的答复。综合评价得数偏低。

2801	12387	UU0081236	洲新区高新安置	4/4/10 13:38	经有三年多了，	网友：您好！留言已收悉	2014/5/16 15:55:37
2802	12415	UU0081509	含浦镇白鹤社区的	4/3/27 17:16	这些难道都不	网友：您好！留言已收悉	2014/5/9 17:28:09
2803	12451	UU0081019	增加A市交通辅警	4/3/11 17:02	很低，住房公积	网友：您好！留言已收悉	2014/4/28 16:06:58
2804	12452	UU0081171	旅游发展的看法	4/3/11 11:56	尬现状也亟需集	网友：您好！留言已收悉	2014/4/28 16:05:42
2805	12458	UU00854	头镇连山村交通	14/3/7 22:39	断的扩宽提质，	网友：您好！留言已收悉	2014/4/14 12:13:50
2806	19602	UU008762	部门不作为，A1	14/7/1 22:42	领导或市领导或	网友：您好！留言已收悉	2014/7/14 11:08:55
2807	25918	UU0081182	段A1区南路路灯	14/4/21 1:40	A1区南路A市段	“UU0081182”	2014/5/28 15:11:52
2808	37482	A00062705	市规划一个校车接	8/12/7 18:48	多的交通事故！	2018年12月12日	2018/12/13 18:53:19
2809	88222	UU008133	农村购房补贴的	19/4/4 16:34	但是是公职人员	网友2019年4月8日	2019/4/8 14:45:38
2810	93271	UU0081810	市公租房分配问题	4/7/19 11:05	明是哪天，这样	网友：您好！留言已收悉	2014/7/24 15:56:02
2811	119660	UU0082261	7县龙凤巷乱摆摊	5/4/23 15:45	的安全隐患，严	网友：您好！留言已收悉	2015/4/30 17:14:22
2812	25431	UU0081037	镇能否根据房产	6/5/24 15:29	洛阳方面称无准	2016年6月12日	2016/6/12 10:51:48
2813	37459	A00039732	带小孩打疫苗要带	19/1/13 1:56	带什么证件呢？	2019年1月14日	2019/1/14 16:06:08
2814	116958	UU0081337	K市考取健康管理	19/8/7 16:02	是不知道向哪个	请咨询K市人社部门。	2019/8/12 8:14:45
2815	6556	UU0081320	狂犬役苗报销比例	8/3/20 15:19	打狂犬役苗报销	已收悉	2018/3/28 16:05:34
2816	30019	UU008151	全款购买二手房	6/11/3 10:00	易，房产局资金监	已收悉	2016/11/22 12:25:56
2817	114346	UU0081119	锦豪雍景园小区商	13/6/3 13:05	但购房补充合同	已收悉	2013/7/5 16:47:46

四、结论

本文应用 python，基于自然语言处理技术的智慧政务系统，建立关于留言内容的一级标签分类模型。方便后续工作人员将群众留言分派至相应的职能部门处理。得到的分类评价 $F-Score$ 为 72.7%。虽然存在一些误差，不过总体上已经达到很好的水平。

其次，采用 K-means 聚类方法，将相似的同一类留言归类在一起，最后我们得到了排名前 5 的热点问题，这种方式不仅有助于相关部门有针对性地处理问题，提升服务效率，而且可以得知哪些问题是群众急需解决的。网民留言处理是倾听民声、了解民情、维护人民利益、凝聚民心的重要渠道。也是党和政府联系群众的重要途径。做好网民留言的处理工作，对于保护人民群众的合法权益，维护社会的整体和谐稳定具有重要意义。

最后用模糊综合评价从答复的相关性、完整性、时间性、礼貌性等角度对答复意见的质量给出一套评价方案，并用附件 4 提供的数据进行测试，结果符合情况。

五、文献

[1]李舟军,范宇,吴贤杰.面向自然语言处理的预训练技术研究综述[J].计算机科学,2020,47(03):162-173.

[2]王艺颖.朴素贝叶斯方法在中文文本分类中的应用[J].中国高新科技,2019(07):57-60

[3]司守奎.数学建模算法与应用[M]出版社:国防工业出版社,2011.08.

[4]王俊丰,贾晓霞,李志强.基于 K-means 算法改进的短文本聚类研究与实现[J].信息技术,2019,43(12):76-80.