C 题：“智慧政务”中的文本挖掘应用

摘要

近年来，随着政府对网络问政越来越重视，网络问政平台已成为政府与民众沟通的重要渠道。本文旨在利用自然语言理解、数据挖掘和机器学习的方法对群众问政留言进行留言分类、热点挖掘以及对政府的意见反馈进行评估，帮助政府提升管理水平和施政效率。

对于问题一：针对群众留言，基于自然语言理解，首先对数据进行预处理，进行中文分词、去除停用词等，然后使用 TF-IDF（Term Frequency－Inverse Document Frequency）算法提取关键词和生成权重向量，将文本数据转化为向量数据。接下来，基于朴素贝叶斯分类模型，在进行模型训练学习之后，对测试数据根据留言所属的管理部门进行分类。最后使用 F1 对分类方法进行评价，最后 F1 值达到 51.95%。

对于问题二：为提高热点挖掘的准确性，首先在数据预处理阶段，本文构造了专用分词表和停用词表。其次，基于主题生成模型（Latent Dirichlet Allocation，LDA）进行主题挖掘，生成主题向量。最后，基于 K 均值聚类算法（k-means clustering algorithm，K-means）进行聚类，找出样本数最多的前五簇的留言，分别提取时间段，地点和人群以及热点描述。实验结果显示，前 5 簇都是热点问题， 每类热点问题平均有 6 条留言，所提取的时间段，地点和人群以及描述均能较好的从时间、地点、人群和内容上反映热点。

对于问题三：基于模糊综合评价算法对政府相关部门的意见反馈进行综合评价。评价依据主要为两个指标：（1）群众留言与答复意见的相关性；（2）留言时间与答复时间的反馈时长。针对第一个评价指标，。本文基于潜在语义索引模型（Latent Semantic Indexing，LSI）通过余弦相似度计算群众留言与答复意见的相关性；生成相似度矩阵，从而得到留言与答复的相关性。针对第二个评价指标，计算答复时间与留言时间的时长。最后通过模糊评价算法，定义指标权重，通过这两个评价指标对每条答复进行评价，得出“差”、“一般”、“好”、“很好”四个等级。实验结果显示，每个等级大约覆盖 30 条左右的留言，评价等级较好的反映了政府部门的服务质量。

关键词：文本分类　朴素贝叶斯模型　LDA 模型　LSI 模型　模糊评价算法

目录

1.挖掘目标

本次文本挖掘的目标是利用网络问政平台的的群众问政留言记录，采用jieba分词工具、TF-IDF算法、朴素贝叶斯分类模型、LDA算法、模糊评价算法等解决以下三个问题：

（1）利用文本分词和TF-IDF算法提取关键词，并使用朴素贝叶斯算法按照附件一的标准把留言分到各自对应的一级类别，使用F-Score方法对分类方法进行评价。

（2)利用附件三的数据，找出某段时间内特定地点或特定人群反应的热点问题，列出排名前5的热点问题和热点问题对应的留言信息。

（3）从答复的相关性、完整性、可解释性等角度，利用模糊评价模型对政府相关部门对群众的答复意见质量建立一套评价体系。

2.分析方法与过程

本文通过对所给数据进行自然语言处理和文本挖掘，要实现对所给留言进行分类、提取热点问题以及对政府部门回应留言的答复意见进行评价。各问题主要步骤如下：

问题一：对留言进行自然语言处理，利用TF-IDF、朴素贝叶斯算法将留言进行分类。并代入测试数据检验模型精确度。使用F-Score方法来评价分类模型，以调整和优化模型。

问题二：将经过自然语言处理的留言数据带入LDA模型提取关键词和生成主题向量，再采用K-means算法对留言进行聚类，挖掘前五类热点主题词。

问题三：基于留言与答复意见的相关性、留言时间与答复时间的长短这两个指标建立模糊综合评价模型，以对政府部门的答复意见做出评价。

2.1 问题一分析方法与过程

为提取到更精确的关键词，本文对每条留言的留言主题和留言详情合在一起进行分词、去停用词处理，并根据附件一的分类为每条留言加上标签。采用 TF-IDF 算法，将处理好的留言信息转化成权重向量，再把他们代入朴素贝叶斯模型，把他们对应相关部门职能分成七类。带入测试数据对模型精确度进行检验，以及使用 F-Score 对模型进行评价。

2.1.1 流程图

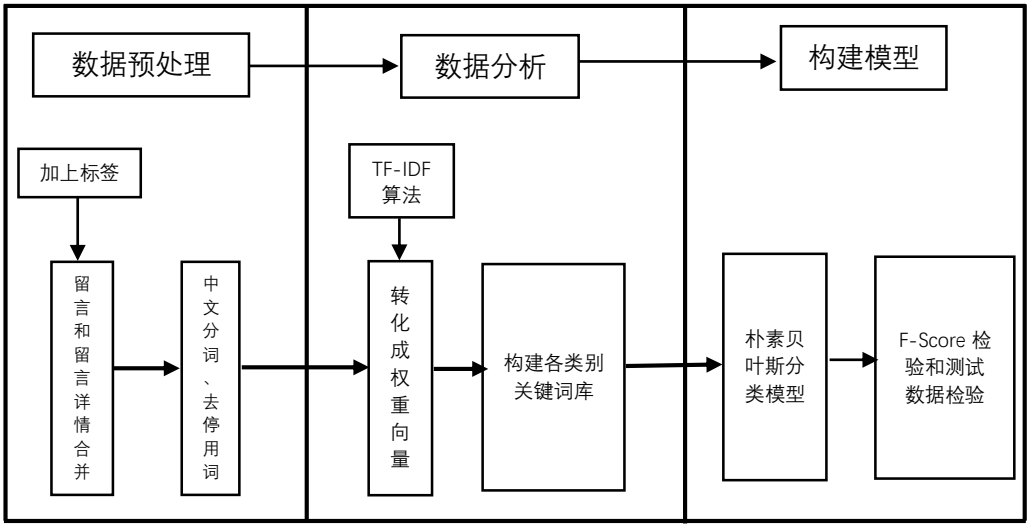

图 1 问题一工作流程图

2.1.2 数据预处理

计算机只能识别和处理数值化的信息，而非文本型。为了能进行数据转换，我们需要对文本数据先进行处理。调用 python 的 jieba 中文分词工具对文本进行分词，并且使用停用词表对留言去除无意义的词语。把处理好的数据封装成函数供后面操作调用。

2.1.2.3 文本向量化

采用 TF-IDF 算法，把处理好的留言信息转化为权重向量，实现文本向量化。TF-IDF 的计算原理如下：

第一步，计算词频 TF。（TF 表示单个文本中某个词的词频）

$$\text{词频}（TF）=\frac{\text{某个词在文章中出现的次数}}{\text{文章的总词数}} \quad （1）$$

第二步，计算逆文档频率 IDF。

$$\text{逆文档频率}（IDF）=\log（\frac{\text{语料库的文档总数}}{\text{包含该词的文档数+1}}） \quad （2）$$

这里分母加 1，是为了避免分母为 0（即所有文档都不包含该词）。

第三步，计算 TF-IDF。（TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。）

$$TF-IDF=\text{词频（TF）}\times\text{逆文档频率}（IDF） \quad （3）$$

计算得出的某个词的 TF-IDF 越大，说明这个词越重要，取前面 TF-IDF 值大的词语当作关键词，建立关键词语料库。代码中我们调用 sklearn 中的 CountVectorizer 函数，它可以把文本转化为词频矩阵，再调用 fit_transformer()函数将文本转化为权重向量。

2.1.3 留言类型的分类

生成 TF-IDF 权重向量以后，根据每条留言的权重向量进行分类，我们采用朴素贝叶斯算法把留言分成 7 类。

表 1　七类留言类别

类 别	类别 1	类别 2	类别 3	类别 4	类别 5	类别 6	类别 7
名 称	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	卫生计生

朴素贝叶斯分类算法是基于贝叶斯定理和特征条件独立假设的分类方法。它通过预测一个对象属于某个类别的概率，从而预测其所属类别。朴素贝叶斯的原

理[1]如下

设有样本数据集D={$d_1, d_2, d_3 \cdots, d_n$},对应样本数据的特征属性集为X={$x_1$，$x_2$，$x_3 \cdots$，$x_d$}，类变量为Y={$y_1, y_2, y_3 \cdots$，$y_m$}，即D可以分$y_m$个类别。其中$x_1$，$x_2$，$x_3 \cdots$，$x_d$相互独立且随机，则Y的先验概率$P_{prior}$=P(Y)，Y的后验概率$P_{post}$=P(Y|X),有朴素贝叶斯算法可得，后验概率可以由先验概率P_{prior}=P(Y)、证据P(x)、类条件P(Y|X)计算得出:

$$P(Y|X)=\frac{P(Y)P(X|Y)}{P(X)} \tag{4}$$

朴素贝叶斯基于各特征之间相互独立，在给定类别为y的情况下，上式可以进一步表示下式：

$$P(X|Y=y)=\prod_{i=1}^{d}P(x_i|Y=y) \tag{5}$$

由上式两式可以计算出后验概率为：

$$P_{post}=P(Y|X)=\frac{P(Y)\prod_{i=1}^{d}P(x_i|y)}{P(X)} \tag{6}$$

其中P(X)不变,所以一个样本数据属于某个类别的朴素贝叶斯计算公式为：

$$P(y_i|x_1,x_2,...,x_d)=\frac{P(y_i)\prod_{j=1}^{d}P(x_j|y_j)}{\prod_{j=2}^{d}P(X_j)} \tag{7}$$

具体算法步骤如下：

1.将附件二的数据经过自然语言处理后得到的每条留言的向量（$w_1, w_2, \ldots w_i$）代入朴素贝叶斯模型进行训练。

2.调用sklearn包，实现朴素贝叶斯模型的训练。

3. 代入附件二的测试数据，计算它与训练样本中的每个文本的相似程度。使用朴素贝叶斯模型自带的 score 函数测试相似度，找出相似概率最大的那个，确定所属类别。

4. 利用 F-Score 对朴素贝叶斯分类模型进行评价。

朴素贝叶斯分类算法流程图如下：

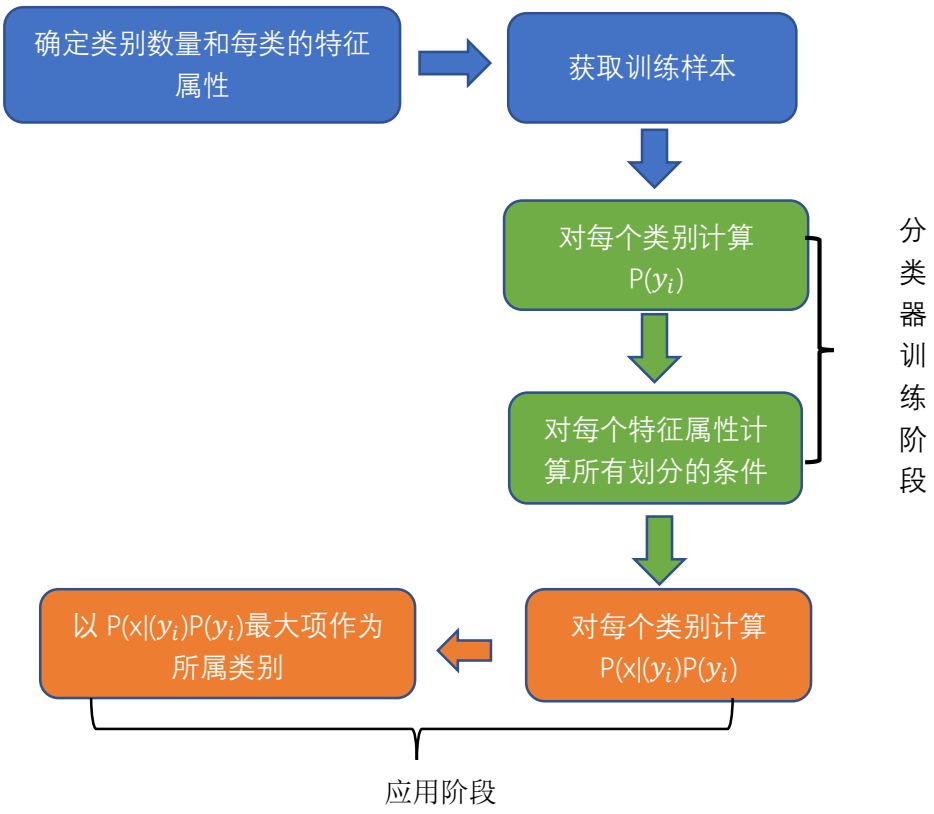

图 2　朴素贝叶斯分类算法实现过程图

2.2 问题二分析方法与过程

为了找到留言中的热点问题，本问进行主题挖掘。在自然语言处理后，采用 LDA 模型来提取每条留言的主题词，生成文档词频向量。然后进行聚类。使用肘部法则来决定聚成几类，再利用 K-means 聚类算法来对留言进行聚类，选取样本数最多的前五簇作为排名前五类的热点关键词。

2.2.1 流程图

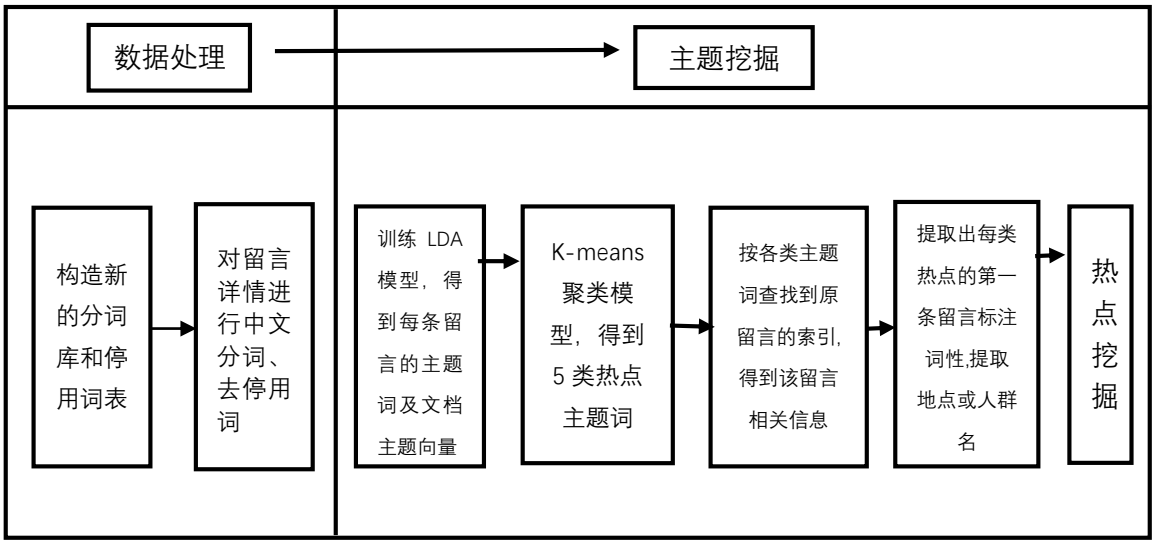

图 3 问题二工作流程图

2.2.2 数据预处理

经分析，本题需要找出热点问题和特定人群地点的关键词，所以需要对分词和去除无用词有更高的要求。所以我们在收集了留言中的专有名词和地点名词后，把它们加载到分词工具中。并且把分词后对句子影响不大的词语加到了停用词表。用新的分词工具和停用词表对留言详情内容进行去分词和去停用词，把数据保存到 text3.txt 中。

2.2.3 LDA 主题模型——提取主题词，生成文档词频向量

在对数据分词、去停用词后，我们需要在处理好的词语中提取与留言主题相关的关键词，使这些词能反应留言的主题。这里使用 LDA 主题模型提取主题词，并生成文档的词频向量，以供后面数据的挖掘。LDA 主题模型的原理如下：

假设有两个集合：文档集合 D，主题（topic）集合 T。D 中每个文档 d 都看作一个词语序列<w1，w2，…，wn>，d 有 n 个词语。D 中涉及的所有不同的词语组成一个大集合 VOC，LDA 以文档集合 D 作为输入，希望训练出的两个结果向量（设聚成 k 个 topic，VOC 中共包含 m 个词）：

对每个 D 的文档 d，对应到不同 topic 的概率 θ_d <pt_1，pt_{i2}，…，pt_k>. 其中，

pt_i=nt_i/n，其中pt_i表示 d 对应 T 中第 i 个 topic 的概率，nt_i表示 d 中对应第 i 个 topic 的词的数目，n 是 d 中所有词的总数。

对每个 T 中的 topict，生成不同词语的概率φ_t<pw1,…,pwm>，其中，pwi 表示 t 生成 VOC 中第 i 个单词的概率。计算方法同样很直观，pwi=Nwi/N，其中 Nwi 表示对应到 topict 的 VOC 中第 i 个单词的数目，N 表示所有对应到 topict 的单词总数。

LDA 的核心公式如下：

$$p(w|d)=p(w|t)\times p(t|d) \tag{8}$$

LDA 模型实现步骤具体如下:

1)从先验概率 p(d)选择文档 d;

2)在超参数为 a 的 Dirichlet 分布中取样生成文档 d 的主题分布α;

3)从主题分布α中采样生成文档中第 n 个词的主题β;

4)在超参数为 β 的 Dirichlet 分布中取样生成主题 α下的词分布 ϕ;

5)从词分布 ψ 下最终采样生成词 W。

LDA 算法流程图如下：

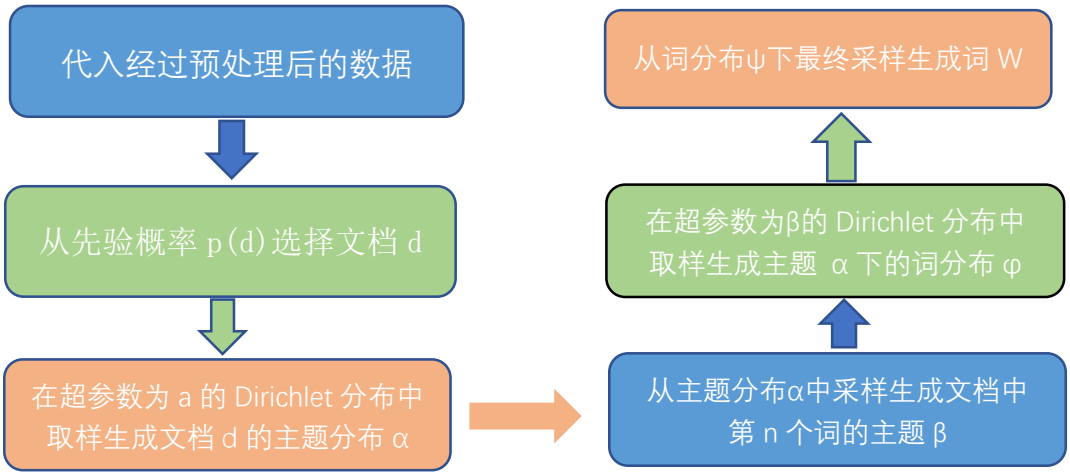

图 4 LDA 算法实现过程图

由此我们得到每条留言的主题词和文档主题向量。主题词近似的代表整条留言。对于一些留言提取主题词后，主题词数小于 2，我们认为它缺失信息严重，不能代表一条留言。后续处理将不带它。选取含有关键词大于 2 并且与聚类中心点余弦距离大于 0.9 的留言并在原数据的位置记录下来，便于以后的处理。

2.2.4 K-means 聚类模型

计算出来向量以后，我们就可以通过量化计算得出热点问题。我们想通过三个指标来确定热点问题，分别是一个问题的留言条数、点赞数和反对数。我们把留言的主题可以用向量表示，通过相似的主题累加计数来计算相似留言条数。通过将赞同数和反对数归一化，加到主题向量后，用此向量来衡量热度排名。对于 2.2.3 最终的数据，我们通过 K-means 聚类算法把它们分成 5 类。K-means 聚类算法原理如下：

假设有一个包含 n 个 d 维数据点的数据集 $X=\{X_1, X_2, \cdots, X_i \cdots, X_n\}$，其中 $X_i \in R^d$，K-means 聚类将数据集，组织为 K 个划分 C = {c2,i=1,2....K}。每个划分代表一个类 C_k，每个类 C_k,有一个类别中心 U_i..选取欧式距离作为相似性和距离判断准则，计算该类内点到聚类中心 u_i,的距离平方和

$$J(c_k)=\sum_{x_i \in x} \| x_i - u_i \|^2 \tag{9}$$

聚类目标是使各类总的距离平方和 J(C)= $\sum_{k=1}^{k} J(c)$最小

$$J(C)=\sum_{k=1}^{k} J(c_k)=\sum_{k=1}^{k}\sum_{x_i \in C_i} \|x_i - u_i\|^2=\sum_{k=1}^{k}\sum_{i=1}^{n} d_{ki} \|x_i - u_i\|^2 \tag{10}$$

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 U_k 应该取为类别 c_k 类各数据点的平均值。

K-means 具体步骤：

1. 从 x 中随机取 K 个元素，作为 K 个簇的各自的中心。

2、分别计算剩下的元素到 K 个簇中心的相异度，将这些元素分别划归到相异度最低的簇。

3、根据聚类结果，重新计算 K 个簇各自的中心，计算方法是取族中所有元素各自维度的算术平均数。

4、将 X 中全部元素按照新的中心重新聚类。

5、重复第 4 步，直到聚类结果不再变化。

6、将结果输出。

K-means 具体实现流程图：

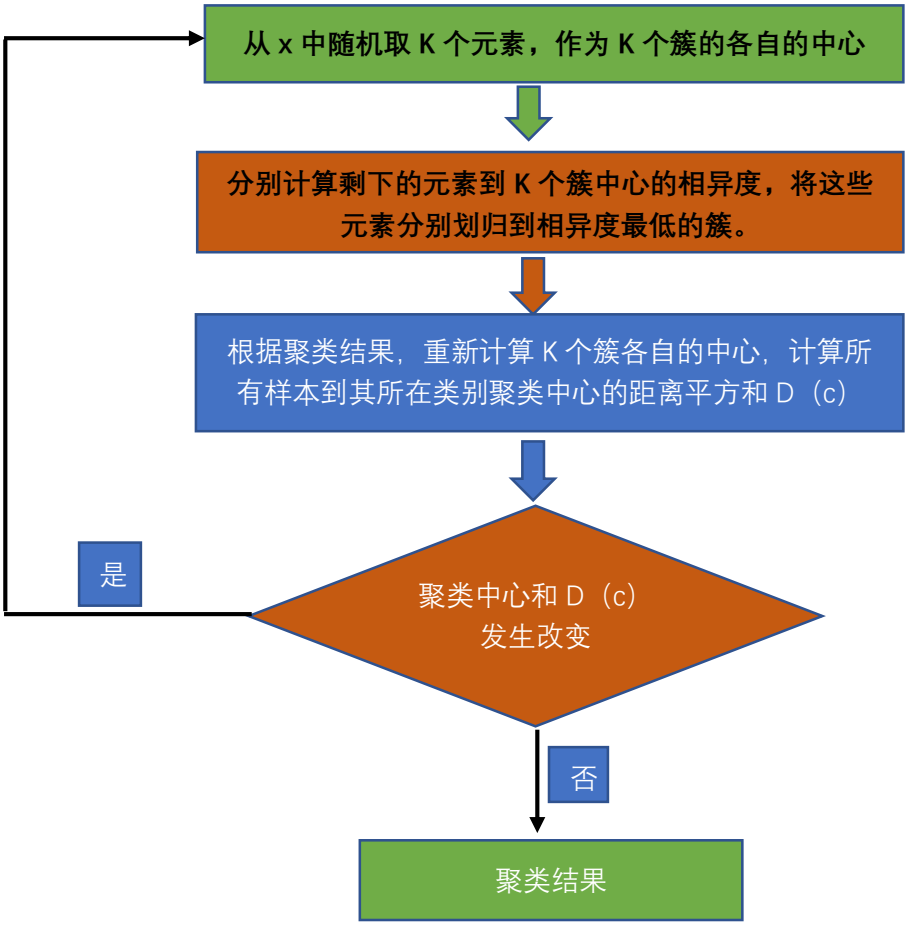

图 5 K-means 算法实现过程图

K-means 模型建立分成 5 类后，找出每类热点问题的索引值，进而找到完整的留言，记录下来留言编号、留言时间、留言主题。这就是排前五的热点问题。

2.2.5 热点问题细节的挖掘

得到排名前五的热点问题后，需要对热度进行具体排名，用热度指数来衡量。热度指数表示此类型留言在总留言中所占比例。本文找到所有类型留言各自的数目，经计算得出热度问题排名。因为每个热点问题虽然说法不一模一样，但是本文认为同类型的所有留言说的是同一个问题，问题描述取每类热点问题找到的第一条留言。对于热点问题的留言人群或地点提取，本文则对提取的热点问题描述进行词性标注。构造新的词性表：jieba 词性表+自定义词性表，自定义词性 x 标注分词中特殊的地点名、人群名、专有名词等。对每类中的第一条留言的留言主题提取出名词或词性为 x 的字符长度大于 3 的词语，把每个主题的前三个提取词作为地点人群。

2.3 问题三分析方法和过程

基于模糊综合评价模型建立政府相关部门对留言的答复意见评价模型，选取了两个指标：时间指标、留言与答复相关性指标，依据这两个指标建立模型。

2.3.1 流程图

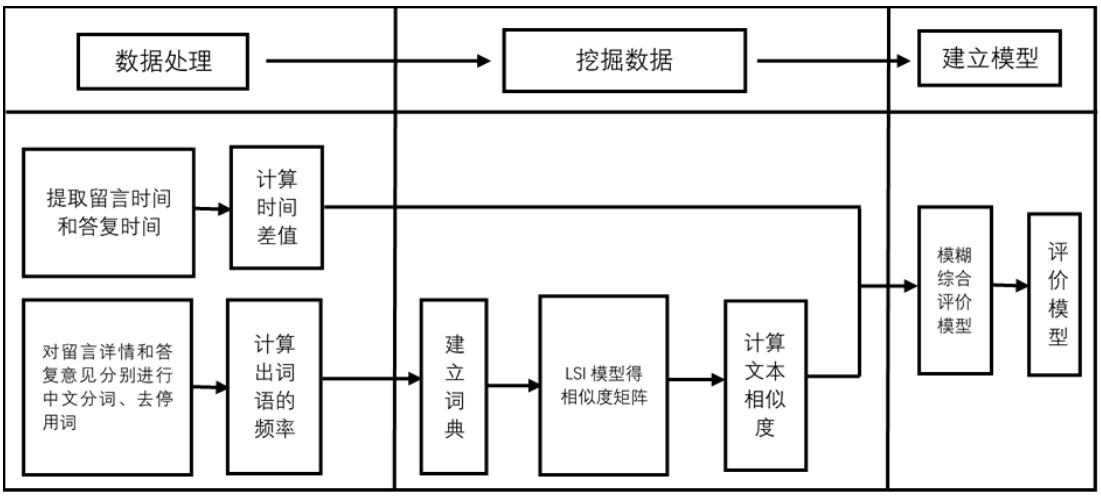

图 6 问题三工作流程图

2.3.2 时间转换与求差值

pandas 中的 to_datetime( )和 datetime( )有类似的功能。获取指定的时间和日期。datetime(%Y,%m,%d,%H,%M,%S)datetime 共有 6 个参数，分别代表的是年月日时分秒。其中年月日是必须要传入的参数，时分秒可以不传入，默认全为零。datetime 是模块，datetime 模块还包含一个 datetime 类，通过 from datetime import datetime 导入的才是 datetime 这个类。如果仅导入 import datetime，则必须引用全名 datetime.datetime。Python time strptime() 函数根据指定的格式把一个时间字符串解析为时间元组。我们可以利用它对时间进行一个处理，使其成为年月日时分秒的格式，当出现某些时间数据时分秒丢失的情况时，可以自动补全成为 00:00:00.

sim[i][i]为一个矩阵，可以将时间第 i+1 行第 i+1 列读取出来，将答复时间减去留言时间就可得到留言时间与答复时间的长短。

使用第二问的分词工具和停用词表对附件四的留言详情和答复意见分别进行中文分词和去停用词，存储到语料库 message.txt 和 reply.txt 里。

2.3.3 LSI 模型

对 message.txt 和 reply.txt 分别建立各自的词典，调用 doc2bow 生成词频矩阵，利用 tf-idf 计算出权重向量。在文本数据转换为数值型数据后，我们可以将数值数据带入模型计算文本相似度。这里我们采用潜在语义索引（LSI）模型可以进行特征提取和生成文本向量，从而计算文本相似度。LSI 模型的原理如下：

分析文档集合，建立词汇-文本矩阵 A。LSA(LSI)使用 SVD 来对词汇-文档矩阵进行分解，即将一个矩阵用其他几个矩阵的乘积来表示，SVD 是对矩阵进行分解的一种方法。假设有 m×n 的矩阵 A，那么 SVD 就是要找到一个分解，将 A 分解为 3 个矩阵的乘积。

$$A_{m\times n} = U_{m\times k} \times \sum_{k\times k} \times V_{k\times n} \tag{11}$$

其中，词汇---文本矩阵 A 是一个稀疏矩阵，其行代表词语，其列代表文档。一般情况下，词---文档矩阵的元素是该词在文档中的出现次数，也可以是是该词语的 tf-idf 权重向量。小矩阵 X 是对词进行分类的一个结果，它的每一行表示一个词，每一列表示一个语义相近的词类，这一行中每个非零元素表示每个词在每个语义类中的相关性。

在构建好词-文档矩阵之后，LSA 将对该矩阵进行降维，来找到词-文档矩阵的一个低阶近似，构建潜在语义空间。

LSI 模型的具体步骤为：

（1） 分析文档集合，建立词汇-文档矩阵

（2） 对词汇---文档矩阵进行奇异值分解

（3） 对 SVD 分解后的矩阵进行降维

（4） 使用降维后的矩阵构建潜在语义空间

（5） 计算向量的余弦相似度

2.3.4 模糊综合评价法

模糊综合评价法是一种基于模糊数学的综合评价方法。该综合评价法根据模糊数学的隶属度理论把定性评价转化为定量评价，即用模糊数学对受到多种因素制约的事物或对象做出一个总体的评价。它具有结果清晰，系统性强的特点，能较好地解决模糊的、难以量化的问题，适合各种非确定性问题的解决。

1. 评价因素：根据题目要求，我们选择了留言与答复意见的相关性以及留言时间与答复时间的长短两方面来设置评价指标。
2. 权重集：为反映各指标因素的重要程度，对各因素 ui 赋予一相应的权数 ai，分别将留言与答复意见的相关性与留言时间与答复时间的长短分别设为 0.65, 0.35。
3. 评价集——是评价者对评判对象可能作出的各种总的评判结果所组成的集合，一般写成：V={v1, v2, …, vn}， vj(j=1,2, …,n)代表各种可能的评判结果。我们将评价结果分为 “差”、“一般”、“好”、“很好” 四个等级，

设对评价对象的 ui 因素进行评价，对评价集中第 j 个元素 vj 的隶属程度为 rij，则按 ui 评判的结果为一模糊集，

记为：

$$R=(r_{i1},r_{i2},...,r_{in}) \quad , \tag{12}$$

或记为：

$$R=(r_{i1}/v_1,r_{i2}/v_2,...,r_{in}/v_n)(i=1,2,...,m) \tag{13}$$

从 m 个因素入手，得到单因素评判矩阵

$$R=[R_i]=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix} \tag{14}$$

4. 数据带入模糊隶属度函数中，得到隶属度矩阵

a_m1 = 1 - (abs(num - a1) / max_min) (15)

a_m2 = 1 - (abs(num - a2) / max_min) (16)

a_m3 = 1 - (abs(num - a3) / max_min) (17)

a_m4 = 1 - (abs(num - a4) / max_min) (18)

模糊综合评价法，关系到外界多种因素，因此要考虑到多方面的因素，另外得出的评价结果不会是绝对的好与坏，而是 对一个模糊集合进行解读。由评价指标构成有限集合 U，由等级评语构成有限集合 V，用矩阵 R 表示 U 和 V 的关系。由指标权重构成权重集 W=（W1，W2，W3 ... Wm）。综上，模糊综合评价法 就是，由指标权重构成权重集 W 构成的集合，与数量关系矩阵 R，二者的乘积形成的一个模糊集合 B。因此，模糊综合评价的数学模型如下：

$$B=W\bullet R=(W1\ \ W2\ \ W3\ \cdots\ \ Wm)\bullet\begin{bmatrix} r11 & r12 & r13 & \cdots & r1n \\ r21 & r22 & r23 & \cdots & r2n \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ rn1 & rn2 & rn3 & \cdots & rnn \end{bmatrix} \tag{19}$$

模糊综合评价算法流程图：

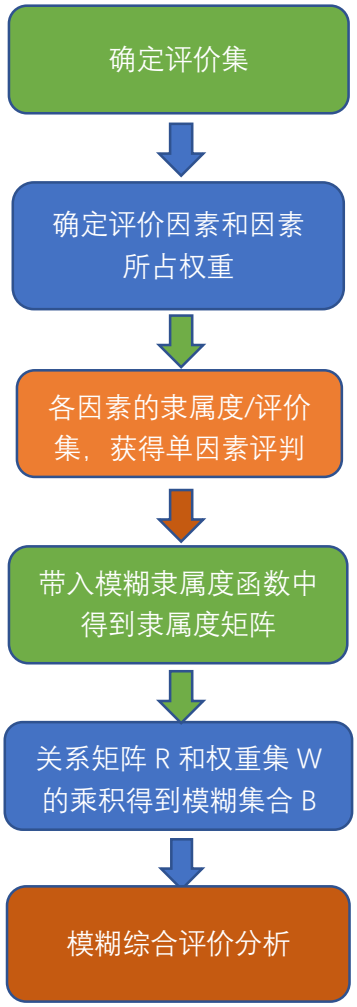

图 7　模糊综合评价算法过程图

3.结果分析

3.1 问题一结果分析

把每条的留言详情与留言主题加上标签，进行分词、去除停用词后，采用 TF-IDF 算法得到每条留言关键词的权重向量。把向量带入朴素贝叶斯模型训练后，得到了留言分类模型（见作品附件）。

表 2　部分留言分类：

留言编号	留言主题	留言详情	一级标签
118617	反映 K6 县公交车监控的有关问题	K6 县公交车破旧……	交通运输
130815	泸阳镇下坪村与壮稻村采石场严重破坏生态环境	泸阳镇下坪村与壮稻村，采石场，生态……	环境保护
168670	G2 区体育馆卫生状况堪忧	G2 区体育馆是 G1 区镇市民健身休闲的……	环境保护
15041	A 市新增 500 台出租车经营权竞标方案量身定制、指标内定	A 市新增 500 台出租车经营权竞标方案……	交通运输
8648	A 市民没有社保只有城镇居民医疗保险怎么办理社保卡	家人没有社保只有城镇居民医疗保险……	劳动和社会保障

由此可以看出，上述表格中编号为 118617 和 168670（标红的）的留言没有对应正确的类别，它们分别应该对应的正确类别是城乡建设和教育文体。经分析，留言分类出现差异的原因可能是因为所给的数据太少，留言主题中主题词存在交叉问题，即某些主题词出现在多个类别中。本文采用朴素贝叶斯模型自带的 Score 函数对模型评价，评估达到了 60%。又使用题目中所给的 F-score 对朴素贝叶斯模型进行评价，发现模型远达不到预期，只有 47.95%。经过调整模型参数，发现模型的评估达到 51.95%。

在此基础上，本文打算使用 LDA 模型再次进行生成含有语义分析的主题，重新进行分类，预计本次分类精确度会更高。

3.2 问题二结果分析

把经过分词和去停用词后的数据，建立词频矩阵和语料库，使用 LDA 模型，训练得到关于所有留言的 5 个潜在主题的特征词，然后生成了文档主题向量和代表每条留言的主题词。提取主题词大于 2 个的留言，记录下这些留言在原始数据中的索引。采用肘部法则和 K-means 算法将这些句子聚类，提取样本数最高的前五类，每一类即为一个热点问题。取每个类别的第一条留言当作这类热点问题的描述，发现它们基本上就是热点问题。生成的“热点问题表.xls”如下，“热点问题留言明细表.xls”见作品附件。

表 3　热点问题表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	0.384615	2017-06-08 17:31:20 至 2019-10-31 21:19:59	A5 区劳动东路魅力之城小区	A5 区劳动东路魅力之城小区临街门面烧烤夜宵摊
2	2	0.192308	2017-06-08 17:31:20 至 2019-09-25 00:31:33	城轨公交站	A 市能否设立南塘城轨公交站？
3	3	0.192308	2017-06-08 17:31:20 至 2019-09-25 00:31:33	人才 app 购房	在 A 市人才 app 上申请购房补贴为什么通不过
4	4	0.153846	2017-06-08 17:31:20 至 2019-09-25 00:31:33	经济学院体育学院变相	A 市经济学院体育学院变相强制实习
5	5	0.076923	2017-06-08 17:31:20 至 2018-11-15 16:07:12	物业服务收费标准居民	L 市物业服务收费标准应考虑居民的经济承受能力

表 4　部分热点问题留言明细表

问题 ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	360114	A0182491	A5 区劳动东路魅力之城小区临街门面烧烤夜宵摊	2019-09-25 00:31:33	A5 区劳动东路魅力之城小区临街夜宵摊、烧烤摊 24 小时营，油烟扰民。……	1	0
1	289408	A0012413	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气	2019-07-21 10:29:36	局长：你好，A5 区劳动东路魅力之城小区的一楼，开了几家夜宵快餐店，厨房内多个灶台油烟随意……	3	0
1	336608	A0005623	A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气，急需处理！	2019-08-01 16:20:02	局长：你好，A5 区劳动东路魅力之城小区的一楼，开了几家夜宵快餐店，厨房内多个灶台油烟随意排放，有的店灶台……	6	0
1	360103	A0012425	魅力之城小区临街门面油烟直排扰民	2019-09-05 12:29:01	魅力之城小区楼下烧烤摊、快餐店无证经营，长期油烟……	3	0

1	323149	A1241141	A5 区劳动东路魅力之城小区油烟扰民	2019-07-28 12:49:18	尊敬的政府：A5 区劳动东路魅力之城小区临街门面长期油……	4	0
1	360107	A0283523	A 市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦	2019-08-26 01:50:38	2019 年 5 月起，小区楼下商铺越发嚣张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音严重影响……	0	0

通过表 3 和表 4 可以看出，此方法找出的都是热点问题，描述的热点问题也比较清楚。上述明细表中的都是属于第一个热点问题对应的留言信息，说明找到的热点问题是可靠准确以及较完整的。

3.3 问题三结果分析

通过问题三建立模型，将留言与答复意见的相关性以及留言时间与答复时间的长短作为评价因素，经过分析得到了“差”、“一般”、“好”、“很好”四个等级。在大多数情况下都能通过评价指标得出与实际相符的等级评价。下表列出了 5 个评价结果。

表 5 评价等级分析

留言编号	相似度	时间差值（天）	评价等级
2549	0.97626287	15	很好
2557	0.51505435	14	一般
33978	0.87128717	8	好
34252	0.9748116	7	很好
37459	0.03459506	1	差

评价等级为“很好”的答复一般都有着满意程度高，时间差值短的特点。满意程度越低，时间差值越长，评价等级也就会越低。下表中最后一个数据，虽然它的时间差值很小，但是满意程度却很低，所以分析得出该答复等级为“差”，通过对原有数据进行比对，发现该条答复意见的确与留言没有恰当的关联。证实了此模型的准确度。

该方法有效解决了相关部门对答复意见的评价问题，便于更好地为人民服务。

4.结论

近年来随着网络的发展，网络问政平台以其独特的优势被政府愈发重视。对群众意见及时进行分类，找出热点问题，是政府在治理国家，了解群众，维护社会稳定所必须经历的步骤。对意见进行分类整理，并找出热点问题，可以让相关部门尽快地整理出群众的意见以及迫切需要解决的问题，并在第一时间解决问题。对相关部门的答复进行评价，使之能够及时得到反馈，高效地解决问题。智慧政务的发展使得政府管理水平和施政效率得到提升。

本文采用 TF-IDF、LDA、K-means、朴素贝叶斯分类等算法模型严谨地、系统地对群众问政留言进行留言分类、热点挖掘。这些模型的优点在于有着坚实的数学基础，可以简化问题的复杂性，模型结果稳定，算法更灵活。 但是不足之处在于分类时忽略了词语之间的关联，不能很好的挖掘出词语与主题之间的关系，导致分类有着一定的误差，数据量不够大，训练模型的准确度还有待提高。运用的模糊综合评价模型，通过精确的数字量化处理模糊的评价对象，作出比较科学、合理、贴近实际的量化评价，比较准确的刻画被评价对象。但是不足之处在于计算复杂，对指标权重矢量的确定主观性较强。针对这些不足之处，将改进完善算法模型方案，把分类以及评价的问题做的更精确，为相关部门提供更好的处理体系，提高服务效率。

5.参考文献

[1] 薛 彬，陶海军，王加强. 针对民生热线文本的热点挖掘系统设计.中国计量大学

[2] 张 聪，易秀双，朱明浩，王兴伟. 基于 Spark 的学术研究热点挖掘方法.东北大学.2019

[3] 谭章禄，彭胜男，王兆刚. 基于聚类分析的国内文本挖掘热点与趋势研究.中国矿业大学。2019

[4] 李肖肖，韩婧，刘邦凡. 基于模糊综合评价法的电子政务网公众满意度研究

[5] 许业超，基于文本挖掘的管理科学热点识别与演化分析. 哈尔滨工业大学.2019