智慧政务留言的分析与挖掘

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。因此，运用网络文本分析和数据挖掘技术对群众留言信息的研究具有重大意义。

对于问题 1，首先对附件 2 进行去重，得到不重复的群众留言信息。接着对每类留言等比例抽取，利用 jieba 中文分词工具对留言详情进行分词、去停用词后将其划分为训练集和测试集，并通过 TF-IDF 算法得到词频矩阵。然后，采用高斯朴素贝叶斯、多项式朴素贝叶斯和支持向量机算法构建三个模型，比较得出最优模型。

对于问题 2，首先对附件 3 进数据预处理得到词频矩阵，采用 K-means 算法对词频矩阵进行聚类。然后，根据每类个数及其点赞数、反对数计算热度指数得到热度排名并得到热点问题留言明细表，对前五类的每类问题分别进行命名实体识别、提取每类摘要和最大时间与最小时间，然后构建热点问题表。

对于问题 3，首先对附件 4 的数据进行分词，并去除停用词，使用 TF-IDF 词袋模型对特征向量进行数字化映射后，对留言详情和答复意见使用余弦相似度做相关性矩阵，判断相关性。然后，判断答复意见是否满足答复规则确定完整性，采用 LDA 模型判断答复内容里是否有对应问题用来判断可解释性。最后，根据答复意见的相关性、完整性和可解释性对其评价。

关键词：TF-IDF，支持向量机，K-means，余弦相似度，LDA

Analysis And Excavation Of The Message Of Intelligent Government Affairs

Abstract

In recent years, with WeChat, such as weibo, mayor mailbox, sun hotline network asked ZhengPing stage gradually become the government understand the importance of public opinion, gathering intelligence, condensed bull channels, all kinds of public opinion related text data quantity rising, leave a message to past mainly depends on artificial to divide and hot spots of relevant departments work has brought great challenge.Therefore, the application of network text analysis and data mining technology to the study of message information has great significance.

For question 1, first of all, deduplication of attachment 2 is carried out to obtain the message message of the masses that is not repeated.Then, the proportion of each type of message was extracted, and jieba Chinese word segmentation tool was used to divide the message details into training set and test set, and the word frequency matrix was obtained by tf-idf algorithm.Then, three models are constructed by using gaussian naive bayes, polynomial naive bayes and support vector machine algorithm, and the optimal model is obtained by comparison.

For problem 2, the word frequency matrix was firstly obtained by preprocessing the data in annex 3, and k-means algorithm was used to cluster the word frequency matrix.Then, according to the number of each category and its thumb up number and opposition number, the heat index was calculated to get the heat ranking and the list of comments on hot issues. For each category of the first five categories, the named entity was identified, the abstracts of each category and the maximum and minimum time were extracted, and then the hot issues table was constructed.

For question 3, the data in attachment 4 is first segmented and stop words are removed. After digital mapping of feature vectors by using tf-idf word bag model, the correlation matrix is made by using cosine similarity for message details and response comments to judge the correlation.Then, it determines whether the reply satisfies the answer rule to determine the completeness, and USES the LDA model to determine whether there are corresponding questions in the reply content to determine the interpretability.Finally, the replies are evaluated according to their relevance, completeness and interpretability.

Key words: tf-idf, support vector machine, k-means, cosine similarity ,LDA

目 录

1 挖掘目标

本次建模目标是对所给的群众留言信息,利用 jieba 中文分词工具对留言信息进行分词、支持向量机算法、K-means 聚类的方法、余弦相似度及 LDA 主题模型，达到以下三个目标：

1） 利用文本分词和高斯朴素贝叶斯、多项式朴素贝叶斯和支持向量机算法构建三个一级标签分类模型，根据预测正确率选出最优模型。

2） 利用文本聚类的方法对留言信息聚类分析，根据聚类结果及每类的点赞数和反对数计算每类留言的热度指数，选出排名前五的热点问题。

3） 采用余弦相似度判断留言详情和答复意见的相关性，判断模板子串是否在答复意见字符串中得出答复的完整性，利用 LDA 主题模型得出答复意见的可解释性。根据相关性、完整性、可解释性制定出一套可行的评价方案。

2 分析方法与过程

总体流程

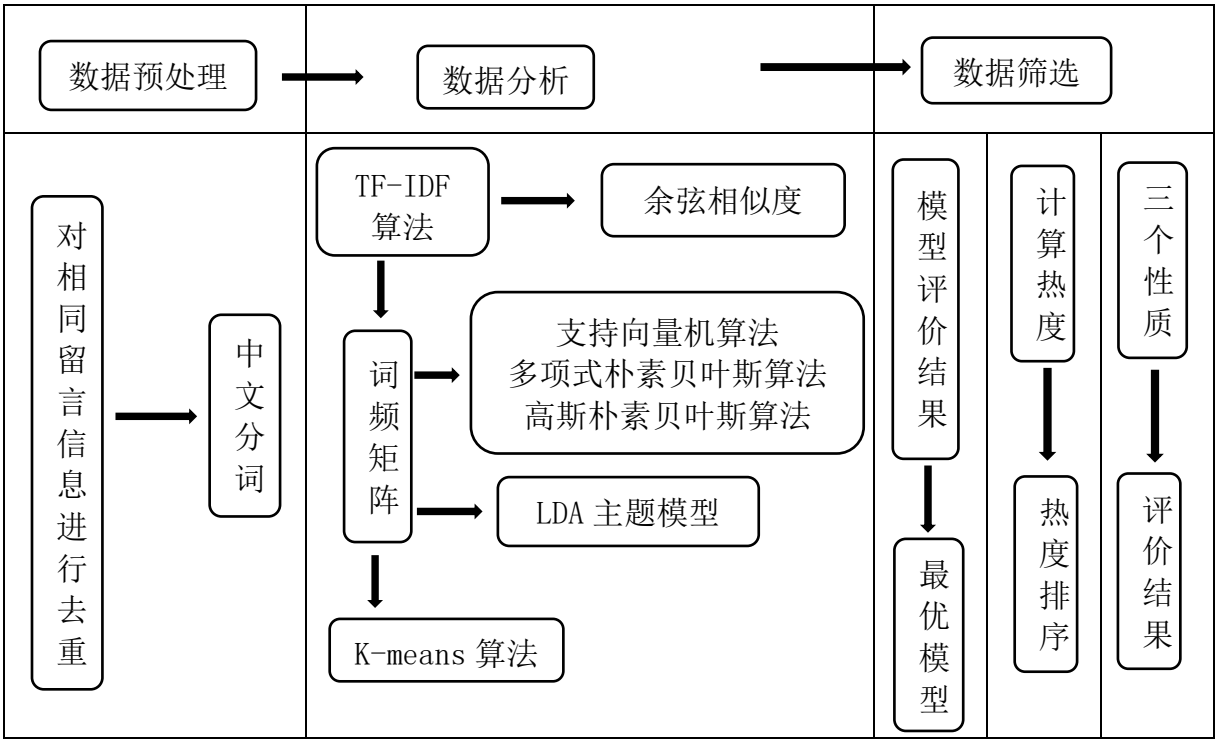

图 1：总体流程图

本用例主要包括如下步骤：

步骤一：数据预处理，去除所给留言信息中留言详情为空或重复的信息，然后进行分词并去除停用词。

步骤二：数据分析，在对留言信息分词后，采用 TF-IDF 算法把这些词转换为词频矩阵以供后续使用。对于问题一，采用支持向量机算法、多项式朴素贝叶斯算法及高斯朴素贝叶斯算法分别构建三种一级标签分类模型。对于问题二，采用 k-means 对留言进行聚类分析。对于问题三，利用余弦相似度判断相关性，字符串模式匹配判断完整性、LDA 主题模型判断可解释性。

步骤三：数据筛选，对于问题一，统计每种模型十次预测结果计算平均值得出最优模型。对于问题二，统计每类留言热度指数并降序排，选出排名前五的热点问题。对于问题三，根据已量化的相关性、完整性和可解释性对答复意见做出相关评价。

2.1 问题 1 分析方法与过程

2.1.1 流程图

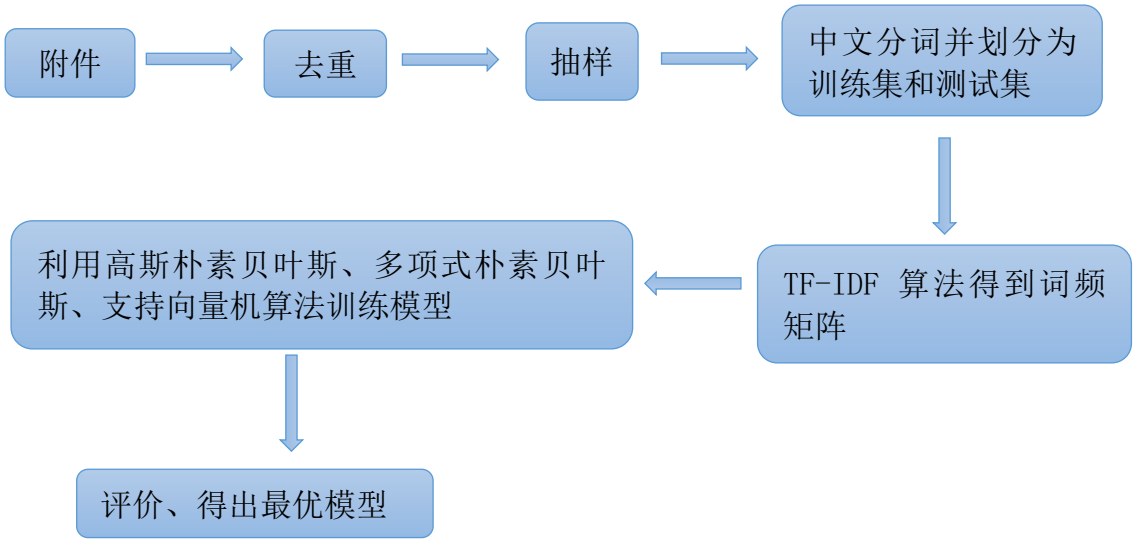

图 2：问题 1 流程图

2.1.2 数据预处理

2.1.2.1 去重、去空

对附件 2 中留言相同和留言为空的数据进行去除，以便抽取的数据不重复或为空。

2.1.2.2 抽样

附件 2 所给的留言中，每类留言的数量不均。因此，为了更好的训练模型需要等比例抽取每类留言的数量，进行纵向拼接形成新的数据集。

2.1.2.3 分词、去停用词

所给的留言信息为非结构化的文本信息，为了使计算机能够识别，在进行分析之前需要先转化为结构化的信息。在附件 2 中，以中文文本的方式给出了数据，为了便于转换，先将这些留言信息进行去除 X 序列:[\t,\n,\r,a-ZA-Z0-9,\u3000,\xa0]、空格、中文分词和去停用词。本文采用 python 的中文分词包 jieba 进行分词，导入停用词表和用户自定义词表去停用词。

Jieba 采用了基于前缀词典实现的高效此图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图，同时采用了动态规划查找最大概率路劲，找出了基于词频的最大切分组合，对于未登录词，加入了自定义词典，使得能够更好的实现中文分词效果。

2.1.2.4 划分数据集

在分词完毕后需要将数据划分为训练集和测试集方便后续使用，这里采用留出法对其划分(hold-out)，原理如下：

留出法把所有数据看成一个大的集合 A 然后划分为两个互斥的集合，一个集合为训练集 B 另一个则为测试集 C，即$A = B \cup C, B \cap C = \emptyset$。用训练集训练出模型后，用测试集来评估其测试误差，作为泛化误差的估计。训练集和测试集的划分要尽可能的保持数据分布的一致性，才能避免因数据划分过程引入额外的偏差而对最终结果产生影响。

留出法一般采用若干次随机划分，重复进行实验评估后取平均值作为留出法的评估结果。

2.1.2.5 TF-IDF 算法

在对留言信息分词后，采用 TF-IDF 把这些词转换为词频矩阵，以便后续使用。TF-IDF 算法的具体原理如下：

首先，计算词频（TF）

$$\text{词频(TF)} = \text{某个词在文档中出现的次数} \tag{1}$$

因为文本有长短之分，所以为了进行不同文本的比较要进行词频标准化，除以文本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。

计算方法为:

$$\text{词频} = \frac{\text{某个词在文章中出现的次数}}{\text{文章的总次数}} \tag{2}$$

或

$$\text{词频} = \frac{\text{某个词在文本中的出现次数}}{\text{该文本出现次数最多的词的出现次数}} \tag{3}$$

其次，计算逆文档频率（IDF），需要建立一个语料库，用来模拟语言的使用环境。IDF 越大，此特征性在文本中分布越集中，说明该次在区分该文本内容属性能力越强。

计算方法为：

$$\text{逆文档频率} = \log\left(\frac{\text{语料库的文本总数}}{\text{包含该词的文本数} + 1}\right) \tag{4}$$

最后，计算 TF-IDF 值

$$\text{TF} - \text{IDF} = \text{词频} \times \text{逆文档频率} \tag{5}$$

某个词在文本中越重要，TF-IDF 值越大。统计 TF_IDF 值有助于降低常见的词语特征

2.1.2.6 文本向量表示

生成词向量的具体步骤如下：

1）将训练集和测试集转换为一个稀疏矩阵并转化为数组

2）将数组进一步转换为 TF-IDF 权值的稀疏矩阵

2.1.3 构建模型

生成留言信息的词频矩阵后，将训练集用于模型的训练。这里采用高斯朴素贝叶斯算、多项式朴素贝叶斯算、支持向量机这三个算法分别构建三个模型。通过测试得到每个模型的 10 次预测结果，分别计算平均值选出最优模型。

2.1.3.1 朴素贝叶斯算法基本原理

朴素贝叶斯算法是利用贝叶斯定理首先求出联合概率分布，再求出条件概率分布。基本原理如下公式所示：

$$P(Y|X) = \frac{\mathrm{P(Y)P(X|Y)}}{\mathrm{P(X)}} \tag{1}$$

$$\mathrm{P(X|Y)} = P(\mathrm{X1, X2, \ldots Xn|Y}) = \mathrm{P(X1|Y)P(X2|Y) \ldots P(Xn|Y)} \tag{2}$$

P(Y|X)叫做后验概率，P(Y)叫做先验概率，P(X|Y)叫做似然概率，P(X)叫做证据。

2.1.3.2 高斯朴素贝叶斯算法

高斯朴素贝叶斯算法的原理如下：

高斯朴素贝叶斯：该算法可以根据样本文件进行预测，其算法本质是对某一件或某组事件及其结果为样本数据进行学习并根据学习结果进行预测。

在高斯朴素贝叶斯中，每个特征都是连续的，并都呈高斯分布，以均值为轴对称，如图所示：

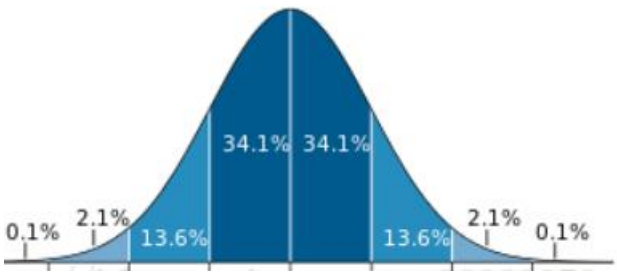

图 3：高斯分布图

GaussianNB 实现了运用于分类的高斯朴素贝叶斯算法。特征的可能性（即概率性）假设为高斯分布：

$$\mathrm{p}(\mathrm{x_i}|\mathrm{y}) = \frac{1}{\sqrt{2\pi\sigma_{\mathrm{y}}^2}}\exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right) \tag{3}$$

参数σ_y和μ_y使用最大似然法估计

2.1.3.3 多项式朴素贝叶斯算法

多项式朴素贝叶斯和贝叶斯的不同之处是，贝叶斯模型对某特征的不同取值代表不同的类别，而多项式朴素贝叶斯对某特征的取值代表着该特征决定该 label 类别的重要程度。

多项式朴素贝叶斯算法的原理如下：

先验概率

$$P(C=c) = \frac{\text{属于类 c 的文档数}}{\text{训练集文档总数}} \tag{4}$$

条件概率

$$P(w_i|\mathrm{c}) = \frac{\text{词}\mathrm{w_i}\text{在属于类 c 的所有文档中出现次数}}{\text{属于类 c 的所有文档中的词语总数}} \tag{5}$$

其中

1）条件概率$P(w_i|\mathrm{c})$表示词w_i在类别 c 中的权重

2）条件概率独立性假设，丢失了词语的位置信息，可以通过 ngram 的特征来减少损失。

3）先验概率和条件概率的计算都利用了最大似然估计。他们实际算出的是相对频率值，这些值能使训练数据的出现概率最大

拉普拉斯平滑

$$\mathrm{P}(w_i|\mathrm{c}) = \frac{\text{词}\mathrm{w_i}\text{在属于类 c 的所有文档中出现次数}+1}{\text{属于类 c 的所有文档中的词语总数}} \tag{6}$$

拉普拉斯平滑是采用均匀分布作为先验分布，即每次词项在每个类中出现一次，然后根据训练数据对得到的结果进行更新，即未登录词的估计值为 1/词汇表长度。

2.1.3.4 支持向量机算法

支持向量机是一个二分类模型，处理的是二分类问题。当用来处理多分类问题

时会转换为一对多的组合模式、一对一组合模式和 SVM 决策树；亦或是构建多个分类器这样才能够充分发挥模型的有点。

对于线性可分两类数据，支持向量机就条直线（对于高纬度数据就是一个超平面），两类数据点中的分割线有无数条，SMV 就是这无数条中最完美的一条，即这条线距离两类数据点越远，则当有新的数据点的时候使用这条线将其分类的结果也就越可信。

支持向量机方法基本步骤：

1）对样本空间利用核函数的方法转换到能线性可分的空间

2）利用最大化间隔的方法获取最优超平面，进而得出支持向量

3）利用最优超平面和支持向量，可以对新的样本进行分类预测。

线性可分支持向量机基本原理：

在 n 维的数据空间中找到一个超平面，这个超平面的方程可以表示为：

$$w^t x + b = 0 \tag{1}$$

分离超平面：

$$w^* * x + b^* = 0 \tag{2}$$

分类决策函数:

$$f(x) = \text{sign}(w^* x + b^*) => \begin{Bmatrix} H_1: w_0 + w_1 x_1 + w_2 x_2 \geq 1 \ \text{for}\, y_i = +1, \\ H_2: w_0 + w_1 x_1 + w_2 x_2 \leq 1 \ \text{for}\, y_i = -1, \end{Bmatrix} \tag{3}$$

推出

$$y_i(w_0 + w_1 x_1 + w_2 x_2) \geq 1 \tag{4}$$

函数间隔:

$$y = \min_{i=1\ldots N} y_i \tag{5}$$

对于样本空间点$(x_{i,}, y_i)$

$$y_i = y_i(w * x_i + b) \tag{6}$$

几何间隔：

$$y = \min_{i=1\ldots N} y_i = \frac{y}{\|w\|} \tag{7}$$

对于样本点(xi, y_i)

$$y_i = y_i\left(\frac{w}{\|w\|} * x_i + \frac{b}{\|w\|}\right) = \frac{y_i}{\|w\|} \tag{8}$$

欲想找到具有最大间隔的划分超平面，也就是 y 最大，即：

$$\max \frac{2}{\|w\|}$$

$$s.t. y_i(w^t * x_i + b) \geq 1, i = 1 \ldots m \tag{10}$$

最大间隔分类器就是求取的分类超平面，等于 max（最大间隔），而函数间隔假设为 1，就可得到最大间隔超平面：$max\frac{1}{\|w\|}$，而约束条件是因为函数间隔是所有样本点的间隔函数最小值。

2.1.4 模型的评价

将测试集的样本放入已训练好的模型进行测试，分别得到三个模型十次测试的结果，求取平均值得到最优模型。

此处采用 F-Score 对分类模型进行评价：

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

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率

$$\text{查准率} = \frac{\text{预测为正例且正确预测的样本数}}{\text{所有预测为正例的样本数}} \tag{2}$$

$$\text{查全率} = \frac{\text{预测为正例且正确预测样本数}}{\text{真实情况下所有为正例的样本数}} \tag{3}$$

一般情况下，查准率高时查全率低、查全率高时查准率低，更多的情况下希望同时参考查准率和查全率而不只是简单的计算一下准确率，这时可以用 F-Score 进行综合评价。

2.2 问题 2 的分析方法与过程

2.2.1 流程图

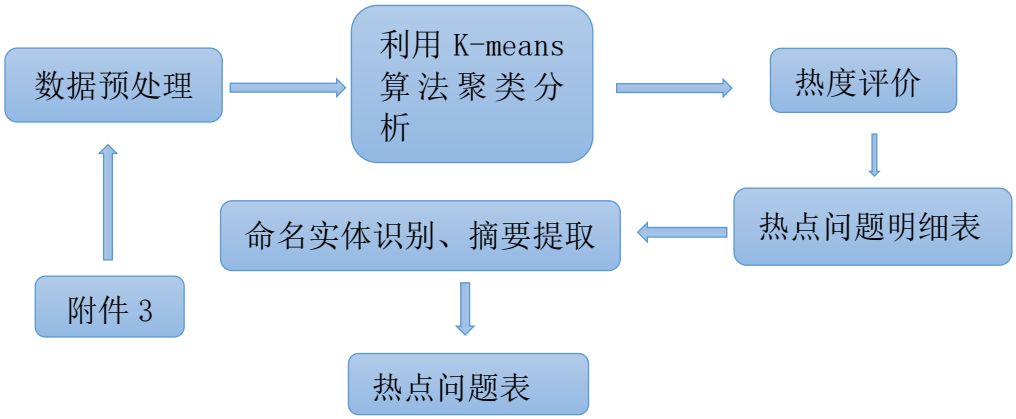

图 4：问题 2 流程图

2.2.2 数据预处理

数据预处理同问题 1 的方法一样：分词、去停用词、转换为词频矩阵

2.2.3 聚类分析

生成留言信息的词频矩阵后，对留言进行分类。这里采用 K-means 算法对其分类。

K-means 聚类的原理如下：

K-means 算法是迭代聚类算法，采用距离作为相似性判断，从而分出所给数据集的类别，并且类中心是所给数值的均值得到，每个类的中心用聚类中心来描述。要得到类别数，一般选取欧氏距离作为相似指标，聚类目标是使得各类总的距离平方和最小，即最小化：

$$J = \sum_{k=1}^{k} \sum_{i=1}^{n} \|x_i - u_k\|^2 \quad (1)$$

结合最小二乘法和拉格朗日原理，聚类中心为对应类别中各数据点的平均值，同时为了使算法收敛，在迭代过程中，应使得最终的聚类中心尽可能的不变。

K-means 聚类的算法步骤如下：

1）随机选取 K 个样本作为类中心

2）计算各样本与各类中心的距离

3）将样本归于最近的类中心

4）求各类的样本的均值，作为新的类中心

5）判定：若类中心不再发生变动或迭代达到次数，算法结束，否则回到第 2 步

6）将结果输出

K-means 聚类的算法流程图如下:

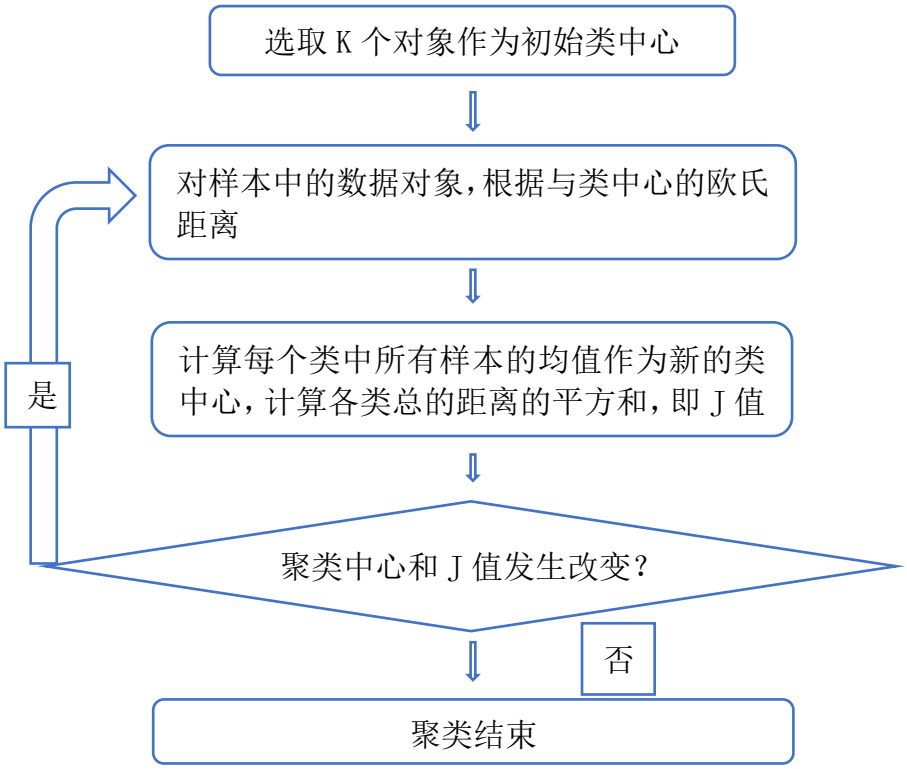

图 5：聚类算法流程图

将附件 3 的留言信息进行聚类后得到得到每条留言的聚类标签，将其以列表形式写入每条留言所对应的行。

2.2.4 热度统计

2.2.4.1 热度指数计算

通过聚类分析后，相似的留言得到同一标签，标记为同一个类别，结合每类留言总数、点赞数和反对数进行热度评价，具体方法如下：

$$\text{热度指数} = \frac{\text{每类留言的总点赞数和反对数之和}}{\text{所有的点赞数和反对数之和}} + \frac{\text{每类留言的总数}}{\text{总的留言总数}} \quad (1)$$

通过以上公式得出每类留言的热度指数

2.2.4.2 热度排序

将得到的热度指数写入每条留言所对应的行，用 excel 对其进行降序排序得出热点问题留言明细表。

2.2.5 热点问题表的构建

2.2.5.1 提取每类摘要

通过聚类分析得到热点问题留言明细表后，需要对每类留言问题描述进行总结，对此可以进行每类留言的摘要提取。这里采用的是 TextRanK 算法。

TextRanK 提取摘要的过程如下：

1）把文本切割成句子，以句子为结点构建图。

2）对句子进行分词、去停用词等，以便计算任意两个句子间的相似度。将相似度作为两个句子构成的边的权值。

3）根据公式，迭代传播权重计算各句子的得分。

4）将得到的句子进行倒序排序，抽取重要度最高的 N 个句子作为候选摘要。

5）根据字数或句子数要求，从候选句中抽取句子组成文摘。

2.2.5.2 命名实体识别

问题二中需要提取每类留言的地点/人群信息，对此可以采用命名实体识别的方法提取地点/人群信息，这里采用的是百度 LAC 进行命名实体识别。

首先从中选取该类中最具代表性的一个主题进行分词处理，然后进行词性标注后对人名、简单地名和简单的组织机构名进行识别，最后识别复合地名和复合组织机构名，最后将每类的地点/人群写进一个列表。

2.2.5.3 表格构建

首先，如下创建各列所需数据：

热度排名、问题 ID：利用列表推导式创建列表([1,2,3,4,5])

热度指数：把排名前五的热度评价指标写入列表

时间范围：提取每类留言的最小时间和最大时间组成列表

地点/人群：将命名实体识别出来的地点/人群写入列表

问题描述：将提取的每类留言的摘要写入列表

然后，将这些数据创建一个字典，将字典写入 excel 表格，从而的到热点问题表

2.3 问题 3 的分析方法与过程

2.3.1 流程图

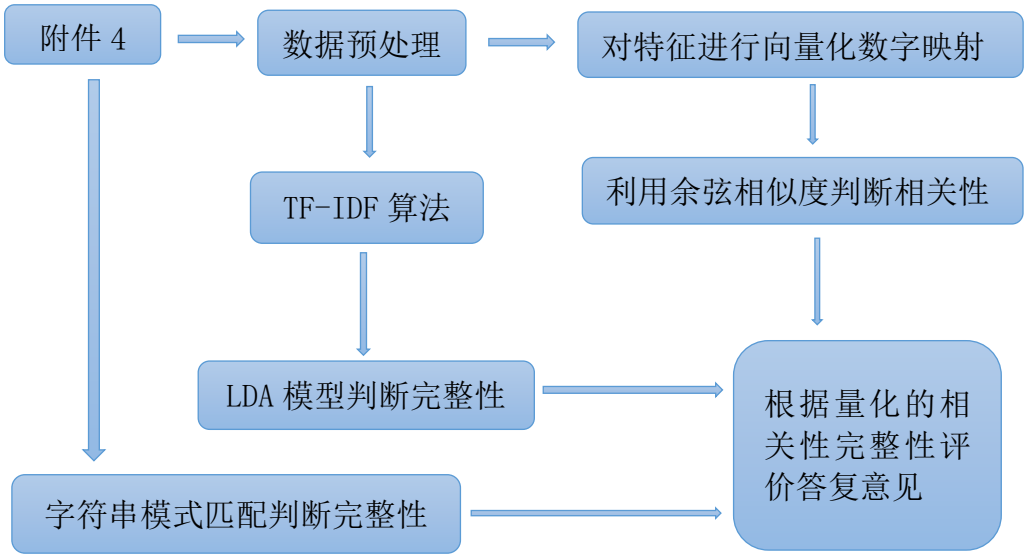

图 6：问题 3 流程图

2.3.2 数据预处理

对附件 4 中的留言详情和答复意见进行分词、去停用词。

2.3.3 答复意见的评价

对答复意见评价可以从答复的相关性、完整性及可解释性制定出一套评价方案。

2.3.3.1 相关性判断

若要判断答复意见和留言详情的相关性可以借助两者相似性的比较完成相关性判断。下图为判断相关性的模型：

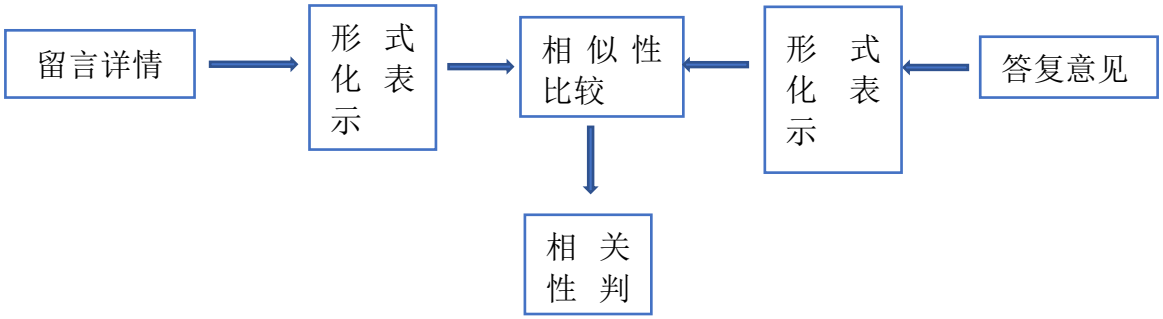

图 7：相关性判断模型

此处判断相似性采用余弦相似度算法，其算法原理如下：

首先将个体数据映射到向量空间，然后根据向量空间中两个向量夹角的余弦值判断两个样本之间的差异，余弦值越接近 1 即夹角越接近于 0 度，则两个向量越相似

两个向量的余弦值采用欧几里得点击公式计算：

$$a * b = |a| \times |b| \cos\theta, \qquad \theta \in [0,2\pi] \tag{1}$$

定义$\cos\theta$为两个向量之间的相似度，取值范围为[-1,1].两个个体的余弦相似度采用下面的公式计算得到

$$\cos\theta = \frac{Y_i * Y_s}{|Y_i||Y_s|} = \frac{(y_{i1}, y_{i2}, \ldots, y_{in}) * (y_{s1}, y_{s2}, \ldots, y_{sn})}{\sqrt{\sum_{j=1}^{n}(y_{ij})^2} \times \sqrt{\sum_{j=1}^{n}(y_{sj})^2}} \tag{2}$$

其中，Y_i和Y_s，分别代表第 i 和 s 个不同的向量。

本文通过计算留言详情和答复意见的相似性矩阵判断其相关性，具体方法如下：

1）对数据使用 DatsFrame 化，和数组化

2）进行分词、并去除停用词，连接成列表

3）使用 TF-IDF 词袋模型，对特征进行向量化数字映射

4）使用余弦相似度，对两两样本做相似性矩阵

5）判断相关性：相似度为 0 则不相关否则相关

2.3.3.2 完整性判断

判断答复的完整性可以通过设置特定的答复模板，判断答复意见是否满足模板格式从而判断完整性。所设计的答复模板如下：

条件一:【‘网友’或‘市民’或‘网名’或‘予以回复’】

条件二:【‘你好’或‘您好’】

条件三:【‘答复如下’或‘回复如下’或‘经查’】

条件四:【‘感谢您’或‘谢谢您’或‘特此回复’或‘感谢来信’或‘谢谢’或‘如有情况及时回复’或‘感谢网友’】

条件五:【‘日’】

若所给答复意见的字符串中同时存在所给的五个条件中的字符串，则判断该答复为完整答复否则为不完整答复。

2.3.3.3 可解释性判断

可解释性就是看答复的内容里是否有对应问题，有解释，有理论支撑。本文用LDA主题模型判断答复意见的可解释性。

LDA是一种概率主题模型，是一种词袋模型，她把文档看作是一系列词组成的集合，里面的元素并没有先后顺序。

LDA的生成过程如下：

1)对每个文本，从中选取一个主题

2)再从选取有的主题对应的词中任意抽取一个词

3)重复以上步骤直到整个文本的词都已经遍历完成

其实就是训练得出两个向量：

$$\text{文档第 i 个主题的概率} = \frac{\text{文档中有多少个词是第 i 个主题也有的}}{\text{文档中所有词的总数}} \quad (1)$$

$$\text{主题 t 生成 v 中第 i 个单词的概率} = \frac{\text{主题 t 对应到 v 中第 i 个单词出现的次数}}{\text{主题 t 下的所有单词总数}} \quad (2)$$

2.3.3.4 答复评价

通过量化得到的相关性、完整性、可解释性可以制定以下评价标准

表1：评价标准表

相关性	完整性	可解释性	评价结果
不相关	不完整	不可解释	D
不相关	不完整	可解释	C
不相关	完整	不可解释	C
相关	不完整	不可解释	C
相关	完整	不可解释	B
相关	不完整	可解释	B
不相关	完整	可解释	B
相关	完整	可解释	A

若答复意见同时满足相关性、完整性、可解释性评价为 A，满足其中两个性质评价为 B，满足其中一个评价为 C，都不满足评价为 D。

3 结果分析

3.1 问题 1 结果分析

通过去重后对文本进行分词、去停用词，利用 TF-IDF 得到词频矩阵后由高斯朴素贝叶斯，多项式朴素贝叶斯，支持向量机算法得到三种一级标签分类模型。将测试集放入模型测试十次的效果如图：

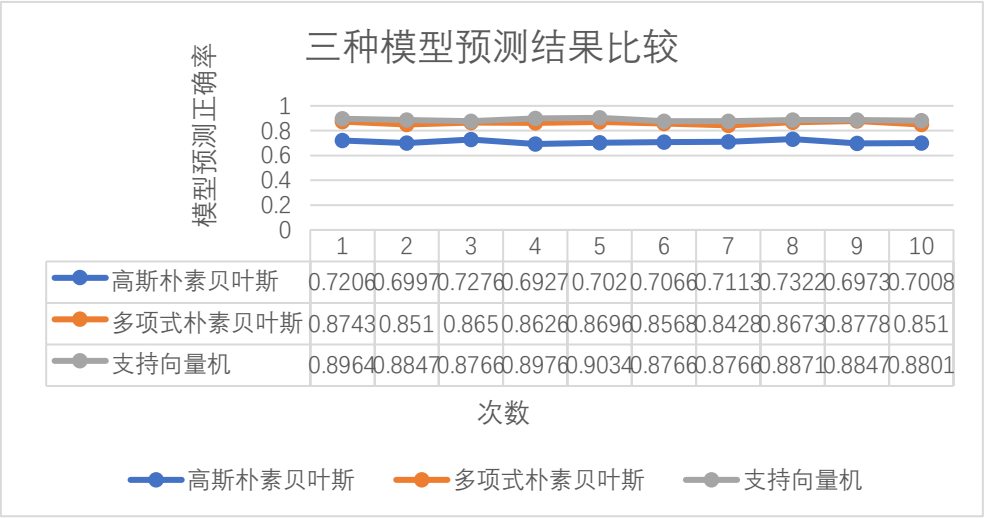

图 8：问题 1 模型测试结果图

从模型的测试结果可以看出：高斯朴素贝叶斯算法对模型的训练效果不是很好，多项式朴素贝叶斯算法和支持向量机算法对模型的预测想过相差不大，可以通过求取每类模型十次测试的平均值选出最优模型，如下图：

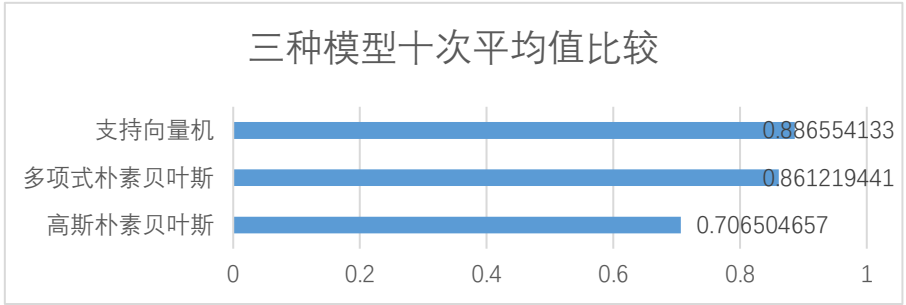

图 9：三种模型十次测试平均值

通过比较可知，最优模型为支持向量机模型，预测正确率为 88%左右。此外，还可以通过改善数据预处理环节和从外界获取更多数据来继续优化得到的模型。

3.2 问题 2 结果与分析

3.2.1 热点问题留言明细表

通过对附件 3 中的留言信息进行数据处理，采用 K-means 聚类分析根据聚类结果和每类点赞数反对数进行热度统计并排名得到结果如下：

表 2：热度排名明细表

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	194343	A0001061 61	承办 A 市 58 车贷案警官应跟进关注留言	2019/3/1 22:12:30	胡书记：您好！58 车贷案发……	733	0
1	217032	A0005654 3	严惩 A 市 58 车贷案特大集资诈骗案保护伞	2019/2/25 9:58:37	胡市长：您好！西地省……	790	0
…	…	…	…	…	…	…	…
5	188409	A0003274	A 市地铁 3 号线星沙大道站地铁出入口设置不合理！	2019/6/19 10:14:39	尊敬的领导您好，4……	4	0
5	211122	A0003550 1	投诉 A 市轨道 3 号线保洁服务项目招标	2019/9/27 16:32:20	我公司于 2019 年 9 月 3 日……	3	3

得到的该表为热点问题留言明细表，对其进行归纳总结可以得到热点问题表。

3.2.2 热点问题表

对热点问题明细表中的每类留言采用提取摘要的方法进行归类，利用百度 LAC 进行命名实体识别可以提取每类地点，通过一定的方法可以构建出热点问题表如下：

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	0.010652833	2019/01/11至2019/07/08	西地省	西地省 58 车贷案件创造全国典型诈骗案，立案至今无公告
2	2	0.010378721	2019/03/14至2019/12/31	五矿万境 K9 县	A 市五矿万境 K9 县房屋出现质量问题
3	3	0.007736191	2019/02/15至2019/04/11	市辰北三角洲幼儿园	A 市辰北三角洲幼儿园入园难、小学入学难
4	4	0.006507049	2019/07/07至2019/09/01	伊景园河滨苑	投诉 A 市伊景园滨河苑捆绑车位销售
5	5	0.005924918	2019/01/03至2020/01/06	地铁 3 号线	请问 A 市地铁 3 号线什么时候开通

通过以上表可以看出群众集中反应的热点问题为：西地省 58 车贷案立案至今无公告、五矿万境 K9 县房屋出现质量问题、市辰北三角洲幼儿园小学入学难、伊景园河滨苑捆绑车位销售 A 市地铁的一系列问题。有关部门可以通过得到的热点问题首先解决这类迫切问题，提高政务效率和让群众更满意。

3.3 问题 3 结果与分析

通过采用余弦相似度文本相关性矩阵、字符串模式匹配和 LDA 主题模型，可以对答复意见的相关性、完整性和可解释性进行量化得到每条答复的量化指标。根据量化指标对答复意见采用 A\B\C\D 四类评价指标对其评价,评价结果如过程数据里的附件 4-2。

从评价结果可以看出大多数答复满足了相关性和可解释性，对于留言回复的规范还有待提高，有关部门可以倡导相关人员进行规范性的答复。

4 总结与展望

对群众留言信息进行分析研究，了解群众群众所反映的热点问题，对相关部门提高政务处理效率有着重大意义。然而随着留言的不断攀升给以人工进行留言划分和热点问题整理的相关部门的工作带来了极大挑战。本文采用支持向量机、高斯朴素贝叶斯和多项式朴素贝叶斯进建立三个一级标签分类模型对留言进行分类处理，用 k-means 聚类方法和热度评价模型筛选出群众所反映的热点问题，根据量化的相关性、完整性和可解释性对回复意见进行评价。

由结果分析可以看出，支持向量机模型对群众留言分类效果最好，因此可以爬取更多的数据进一步优化支持向量机模型得到更优的一级标签分类模型。为了提高分类效率、解放人力可以用同样的方法构建二级标签分类模型和三级标签分类模型，从而建立一个分类系统来处理群众留言分类。对于已按三级标签分类的留言，可以用 k-means 聚类分析的方法让同一问题归为一类然后对其进行热度评价可以得到每类留言中群众所反映的热点问题，有关部门及时处理热点问题能够提升服务质量和效率，提高群众满意度。对已处理的问题，会做出相应的答复，可以建立一套答复意见评价标准来评价有关部门的答复是否解决实际问题同时能够反映出相关部门处理政务的质量。

通过以上方法，能够建立出一套基于自然语言处理技术的智慧政务服务系统，使用该系统进行政务处理能够大大提升政府的管理水平和施政效率。

5 参考文献

[1] 冀先朋. 多标签文本分类算法的研究与应用[D]. 2019.

[2] 张苗, 张德贤. 多类支持向量机文本分类方法[J]. 计算机技术与发展, 2008, 18(3):139-141.

[3] 余芳, 姜云飞. A Feature Selection Method for NB-based Classifier%一种基于朴素贝叶斯分类的特征选择方法[J]. 中山大学学报:自然科学版, 2004, 043(005):118-120.

[4] 王春龙, 张敬旭. 基于 LDA 的改进 K-means 算法在文本聚类中的应用[J]. 计算机应用, 2014(01):255-260.

[5] 郭学平, 牛志亮, 郝洪涛. 基于网络舆情的物流热度分析报告——2017 年 7-12 月[J]. 物流工程与管理, 2018, v.40；No.289(07):13-17.

[6] 林鸿飞, 高仁璟. 基于潜在语义索引的文本摘要方法[J]. 大连理工大学学报, 2001(06):117-121.

[7] 基于条件随机场的命名实体识别及实体关系识别的研究与应用[D]. 北京交通大学, 2015.
[8] 赵玉茗. 文本间语义相关性计算及其应用研究[D]. 哈尔滨工业大学.
[9] 陈大力, 沈岩涛, 谢槟竹,等. 基于余弦相似度模型的最佳教练遴选算法%A Measure Model of Similarity for Finding the Best Coach[J]. 东北大学学报(自然科学版), 2014, 035(012):1697-1700.