“智慧政务”中的文本挖掘应用

摘　要：

近年来，随着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 1，通过去除停用词且进行去重、去空格、去*序列来进行数据清洗。利用 jieba 中文分词工具进行分词，并通过 TF-IDF 算法得到每个群众留言信息的 tf-idf 权值向量。再建立多项式朴素贝叶斯模型，将已经得到的 tf-idf 权值向量用于训练模型。使用训练完的模型进行留言分类，得出每条留言对应的类别。再构造混淆矩阵，进行 F-Score 方法评价分类模型，得到模型的平均分类准确度为 0.887191，F-Score 平均得分为 0.886354。

对于问题 2，首先对政务留言信息表的数据预处理，进行剔除留言详情中的停用词，利用 jieba 中文分词工具对留言详情的中的有效信息进行分词。再通过 TF-IDF 算法得到每个留言详情的 TF-IDF 权重向量，利用余弦相似度对 TF-IDF 权重向量进行两两相似度分析和匹配聚类，当相似度大于 0.6 的归结为一类。提取各类中的时间跨度（最晚留言时间-最早留言时间）以及各类的数量，并以单位时间中的留言数量（留言数量/时间跨度）作为热度评价指标对各个类别进行降序排序，并筛选出排名前五的 5 个类别，将各类中的留言信息保存为“热点问题明细表”。对于“热点问题明细表”，提取各类中的留言详情的关键信息，如：留言时间、地区、人群、问题描述关键词，加以语言组织并保存为“热点问题表”。

对于问题 3，从完整性，可解释性，相关性，答复时效 4 个不同角度通过层次分析法建立一套对答复意见的评价标准方案。利用 jieba 中文分词工具对数据进行分词，且以正则表达式为辅助，从答复意见信息中匹配关键词，从 4 个不同角度建立评价模型，再以斯坦福 NER 训练完毕的模型对每条答复意见信息进行实体识别，提取出实体识别结果中的机构实体和日期加强评价模型的效果。然后运用层次分析法，建立完整性、可解释性、相关性和答复时效性四个因素的两两对比判断矩阵，通过 matlab 求解，得到各自的权重，计算出答复意见的综合得分。答复意见综合得分的表达式：0.4949×s1＋0.2933×s2＋0.1375×s3＋0.0743×s4，其中 s1-s5 分别为相关性、完整性、可解释性和回复时效性的得分。

关键词：中文分词　聚类　TF-IDF　算法　余弦相似度　层次分析法

Abstract:

In recent years, as the online political platform has gradually become an important channel for the government to understand public opinions, gather people's wisdom and gather people's spirit, the amount of text data related to various social situations and public opinions has been increasing, which has brought great challenges to the work of relevant departments that used to mainly rely on manual workers to divide messages and sort out hot topics. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoting the management level and efficiency of the government.

For problem 1, data cleaning is carried out by removing stop words and performing de-duplication, de-space and de-* sequences. Jieba Chinese word segmentation tool is used for word segmentation, and TF-IDF algorithm is used to obtain the TF-IDF weight vector of each crowd message information. Then the polynomial naive bayesian model is established, and the TF-IDF weight vector obtained is used to train the model. The trained model was used to classify the messages, and the corresponding category of each message was obtained. To construct confusion matrix, F - Score method to evaluate classification model, the average classification accuracy of model is 0.887191, the F - Scored Scored an average of 0.886354.

For question 2, firstly, the data of the government message information table is preprocessed to remove the stop words in the message details, and the effective information in the message details is segmented with jieba Chinese word segmentation tool. Then, the TF-IDF weight vector of each message is obtained by TF-IDF algorithm, and the TF-IDF weight vector is k-means clustering and pairwise matching clustering by cosine similarity. When the similarity is greater than 0.7, it is reduced to a class. Extraction of various time span (the latest message time - the first message time) and the number of all kinds of, and with the number of messages in unit time (message quantity/time span) as evaluation index to descending order of each category, the heat and out the top five of the five categories, will all kinds of the message information is saved as a schedule of "hot spots". For the "list of hot issues", extract the key information of all kinds of message details, such as: message time, region, population, problem description keywords, organize the language and save as "hot issues list".

For question 3, a set of evaluation standard scheme for response comments is established through analytic hierarchy process (ahp) from

the perspectives of completeness, interpretability, relevance and response time. Jieba Chinese word segmentation tool is used to analyse the data points, and is assisted with regular expressions, matching keywords from the reply opinion information, establish evaluation model, from four different angles to Stanford NER finished training model for each entity recognition reply opinion information, date of the entity and entity recognition results are extracted to strengthen the effect of the evaluation model. Then, the analytic hierarchy process is used to establish a pairwise comparison judgment matrix of four factors: integrity, interpretability, relevance and timeliness of response. The respective weights are obtained through matlab solution, and the comprehensive score of response comments is calculated. The expression of the comprehensive score of replies:
0.4949×s1+0.2933×s2+0.1375×s3+0.0743×s4, where s1-s5 is the score of relevance, completeness, interpretability and timeliness of reply.

Keywords: Chinese word segmentation TF-IDF algorithm
cosine similarity analytic hierarchy process

目录

1.研究目标

本次建模目标是利用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等数据，利用 jieba 中文分词工具进行分词、gensim 训练聚类主题等技术，达到以下 三个目标：

1)利用多项式朴素贝叶斯的高斯模型对群众留言内容数据进行分类模型的建立，根据 F-Score 对分类方法进行评价，得到最终标签分类的评价

2)根据群众问政留言记录的数据，采用聚类分析方法，制定热度评价指标，得出相对应的各类问题的热度评价指标，进行排序得出排名前五的“热点问题表”以及相对应的“热点问题留言明细表”

3）根据相关部门对留言的答复意见数据，从答复意见的相关性、完整性、可解释性以及时间有效性等方面，制定关于答复意见的质量的评价方案

2.技术方案

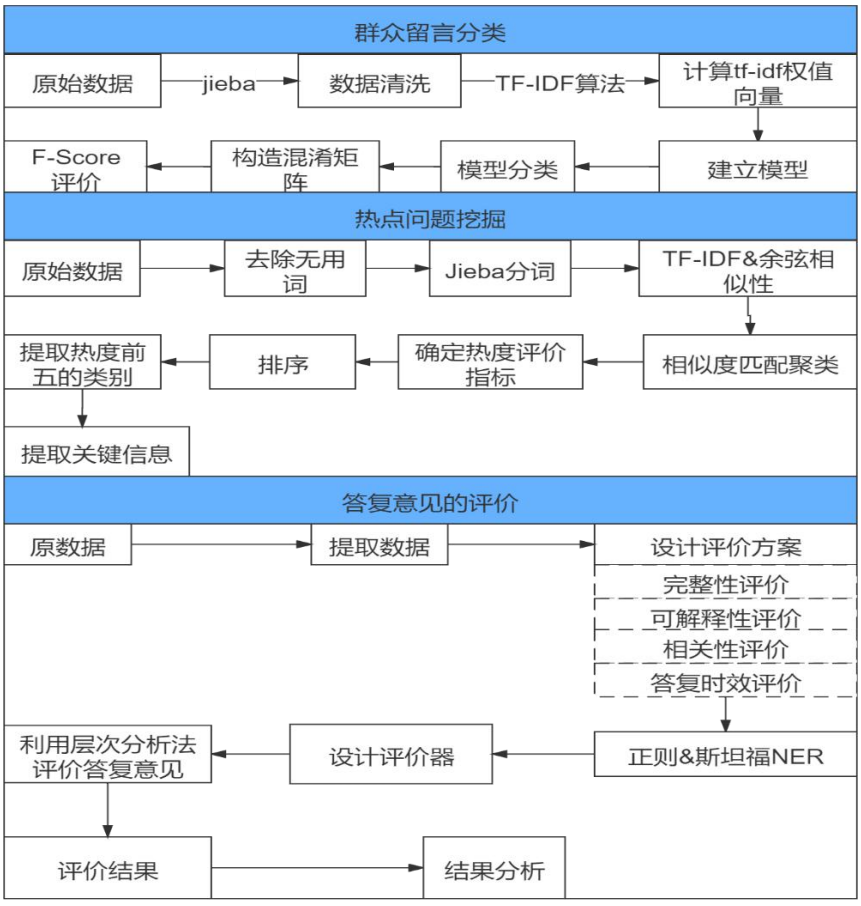

图 1 全文脉络图

3. 分析方法与过程

3.1 问题 1 分析方法与过程

3.1.1 流程图：

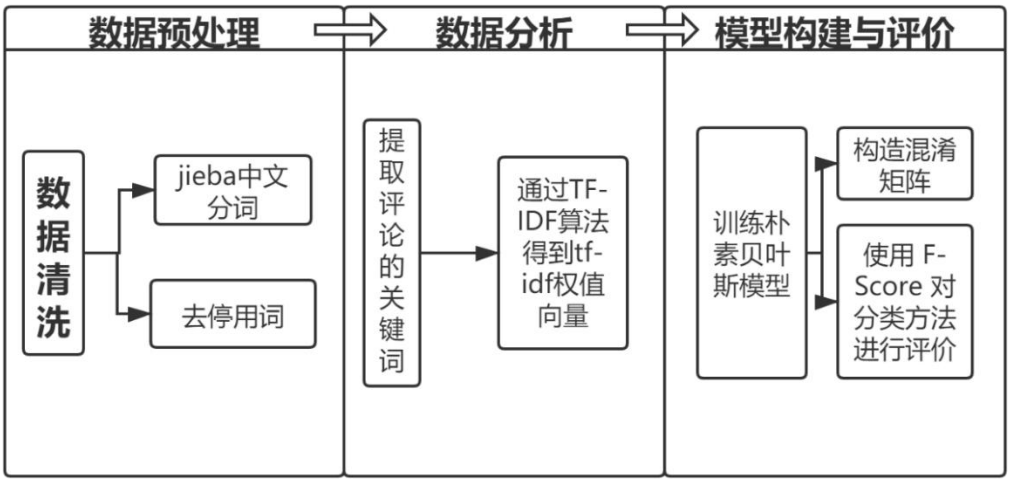

图 2 问题一流程图

3.1.2 数据预处理

3.1.2.1 去重、去空格、去*序列

在附件所给出的所有数据中，为避免数据的不规范和冗余，我们进行一定的数据预处理。首先，我们进行信息的去重。考虑到群众有可能进行重复留言，同一时间多次提交同样的留言或者数据整合是重复了，我们要为了减少工作量，通过对留言内容使用 drop_duplicates()进行去重；其次我们要去掉*序列。因为留言中有许多信息被使用*代替，我们通过 data_dup.apply(lambda x:re.sub('*',' ',x))在处理过程中要去掉。最后要去掉留言中的空格，因为留言是根据一定的模板书写的，会产生很多空格。去掉空格有利于接下来的数据分析。对留言信息的数据预处理的代码保存在 Data_process.py 中。

3.1.2.2 中文分词

对群众留言信息进行处理之前，我们要把非结构化的留言信息转化为能够被计算机识别的计算机结构化信息。所以，对于附件 2 中给出的群众留言信息，我们提取出留言主题和留言详情信息，并利用 python 的中文分词包 jieba 对其

进行中文分词[1]。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。同时，我们根据实际的群众留言信息，自定义词典，添加在留言信息中出现过词语但 jieba 词库中没有的词，提高分词的准确率。最后，为了得到有用信息，我们导入停用词库去掉没有实际含义的词，简化精确留言信息。

3.1.2.3 TF-IDF 算法

对于对群众留言进行分词之后，需要把分词后得到的词语转化为词向量矩阵进行挖掘分析。这里采用 TF-IDF 权重策略，将留言信息转化权重向量。TF-IDF 具体原理如下：

TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。

TF-IDF 的主要思想是：如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。

- 词频（TF）表示词条（关键字）在文本中出现的频率。

  因为留言信息的长度不统一，要对词频进行归一化，即词频除以文章总词数，以防止它偏向长的文件。

$$tf_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

即：
$$\mathrm{TF} = \frac{\text{某个词在文本中出现得次数}}{\text{文本的总词数}}$$

- 逆向文件频率（IDF）：即逆文档频率（Inverse Document Frequency），需要建立一个语料库(corpus 用来模拟语言的使用环境某一特定词语的 IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。IDF 越大，则说明词条具有很好的类别区分能力。

$$idf_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

即：
$$\mathrm{IDF} = \log\left(\frac{\text{语料库的文本总数}}{\text{包含该词的文本数}+1}\right)$$

- TF-IDF 实际上是：TF × IDF

  某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

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。

3.1.2.4 多项式朴素贝叶斯

朴素贝叶斯算法基本原理：
朴素贝叶斯法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。这里的朴素是指在计算似然估计时假定了条件独立。基本原理可以用下面的公式给出：

$$P(Y \mid X) = \frac{P(X)P(X \mid Y)}{P(X)}$$

$$P(X \mid Y) = P(X1, X2, X3, ...Xn \mid Y) = P(X1 \mid Y)P(X2 \mid Y)..P(Xn \mid Y)$$

其中，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。

多项式朴素贝叶斯[2]：

训练阶段：
先验概率：

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

条件概率：

$$P(wi \mid c) = \frac{\text{词}wi\text{在属于类}c\text{的所有文档中出现次数}}{\text{属于类}c\text{的所有文档中的词语总 数}}$$

先验概率和条件概率的计算都利用了最大似然估计。它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。

3.2 问题 2 分析方法与过程

3.2.1 流程图

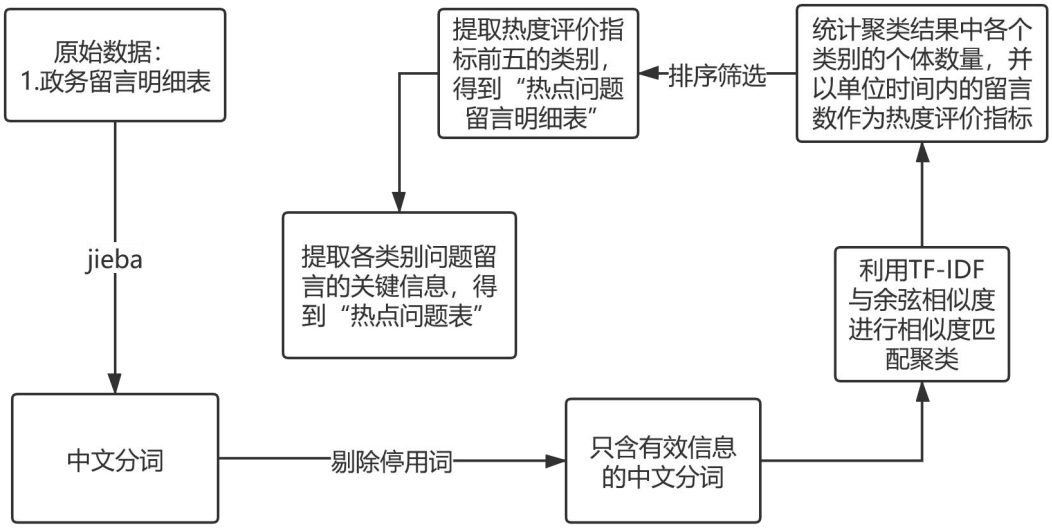

图 3 问题二流程图

3.2.2 数据预处理

3.2.2.1 中文分词

在对热点问题数据挖掘之前，要先对政务留言明细表中的留言详情进行去重工作，因为有些群众有可能重复留言，再进行中文分词工作。采用的是 python 中的中文分词包 jieba 进行分词。Jieba 是基于前缀词典实现高效的词图扫描，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，使用了 Viterbi 算法，得到相对应的中文分词词组。

3.2.2.2 剔除停用词

在对中文分词词组进行算法操作以及聚类操作前，要对这些中文分词词组进行去杂工作，即剔除停用词工作，这些停用词会降低中文分词词组聚类和相似度比较的效率。第一，这些停用词极其普遍。记录这些词在每一个分词词组中的数量需要很大的磁盘空间。第二，由于它们的普遍性和功能，这些词很少单独表达文档相关程度的信息。如果在检索过程中考虑每一个词而不是短语，这些停用词基本没有什么帮助。剔除停用词可减少索引量，增加了检索效率，

并且通常都会提高检索的效果。停用词主要包括英文字符、数字、数学字符、标点符号及使用频率特高的单汉字等。

3.2.2.3 TF-IDF 算法

在对留言详情信息分词后，需要把这些词语转换为向量，以便实现数据挖掘的目的。这里采用 TF-IDF 算法[3]，把留言详情信息转换为权重向量。TF-IDF 算法的具体步骤如下：
（1）第一步，计算词频（2）第二步，计算逆文档频率（3）第三步，计算 TF-IDF

3.2.2.4 生成 TF-IDF 向量

（1）建立关于留言详情信息的词典（2）基于词典，将分词列表集转换成稀疏向量集，称作语料库（3）将关键词转换为稀疏向量（4）创建 TF-IDF 模型，传入语料库训练（5）用训练好的 TF-IDF 模型处理被检索文本和搜索词（6）计算和生成政务留言明细表中的各留言详情的 TF-IDF 值，生成各留言详情的 TF-IDF 权重向量，计算公式如下：

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

3.2.3 留言详情分类

根据政务留言明细表中的留言信息生成的 TF-IDF 向量，使用 gensim 包中的 similarities 相似度（余弦相似度）计算，两两进行相似度计算，将相似度大于 0.6 的则归为同一类。

余弦相似度原理如下：
余弦距离，也称为余弦相似度，是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小的度量。

余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，这就叫"余弦相似性"。

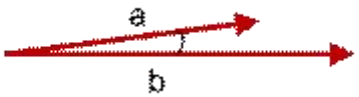

图 4 向量夹角小于 90 度图

图 4 中两个向量 a,b 的夹角很小可以说 a 向量和 b 向量有很高的的相似性。极端情况下，a 和 b 向量完全重合，可认为 a 和 b 向量是相等的，也即 a，b 向量代表的文本是完全相似的，或者说是相等的，如图 5：

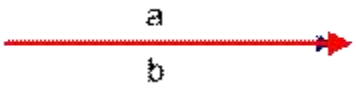

图 5 向量重合图

如果 a 和 b 向量夹角较大，或者反方向，两个向量 a,b 的夹角很大可以说 a 向量和 b 向量有很低的的相似性，或者说 a 和 b 向量代表的文本基本不相似。如图 6：

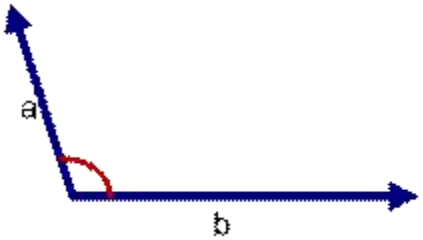

图 6 向量夹角大于 90 度图

若向量 a 和 b 是 n 维向量，余弦值的计算公式如下图：

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

$$= \frac{a \bullet b}{|a| \times |b|}$$

3.2.4 热度评价指标

根据政务留言明细表中的留言详情分类情况，统计各类中的留言详情数量，并计算各类中的留言详情时间跨度（即最晚时间-最早时间），制定留言详情热度评价指标：单位时间内的留言详情数量，计算公式如下：

单位时间留言详情数量=总留言详情数量/时间跨度

根据各类中的热度评价指标，进行排序，筛选出热度评价指标前五的类别，并保存各类相对应的留言信息为“热点问题留言明细表.xls”

3.2.5 提取“热点问题表”

根据“热点问题留言明细表.xls”各类的留言详情中的留言时间得热点问题的时间范围（即最早时间—最晚时间），根据留言详情、留言主题等信息，提取留言详情中的地点、人群以及关于问题描述的关键词，加以语言组织并保存为“热点问题表.xls”。其中共有前五的热点问题所涉及的有噪音问题、社会资源问题、人身安全问题等等

3.3 问题 3 分析方法与过程

3.3.1 分析方向

针对相关部门对留言的答复意见，从答复的完整性，可解释性，相关性，答复效率四个角度，按照一定的评价标准，对答复意见分进行评价。

3.3.2 问题 3 分析流程图

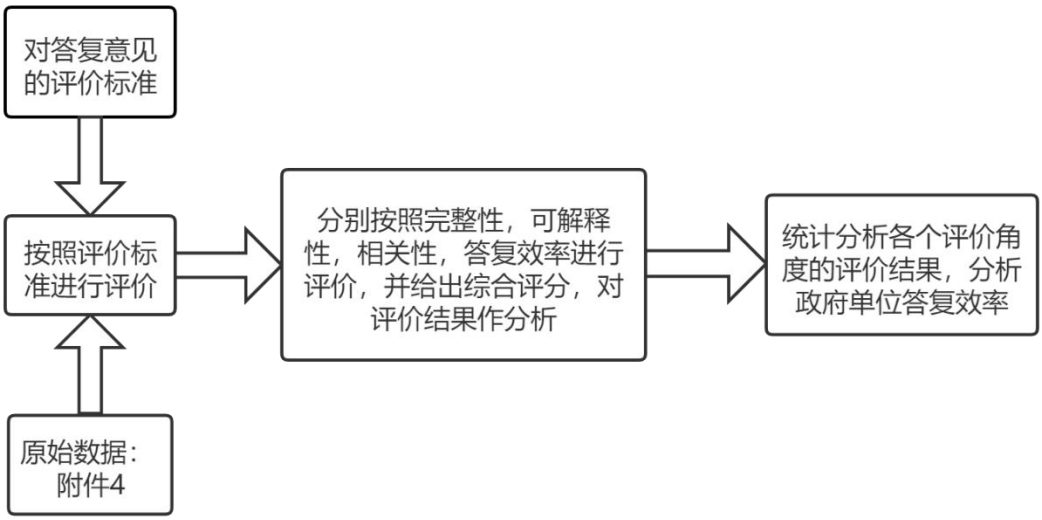

图 4 问题三流程图

3.3.3 评价方案与过程

3.3.3.1 完整性

1） 对于完整性的评价，主要是从礼貌角度与答复格式角度进行评价，再以答

复文本中是否含有咨询电话为准则辅助评价。

2) 评价流程图

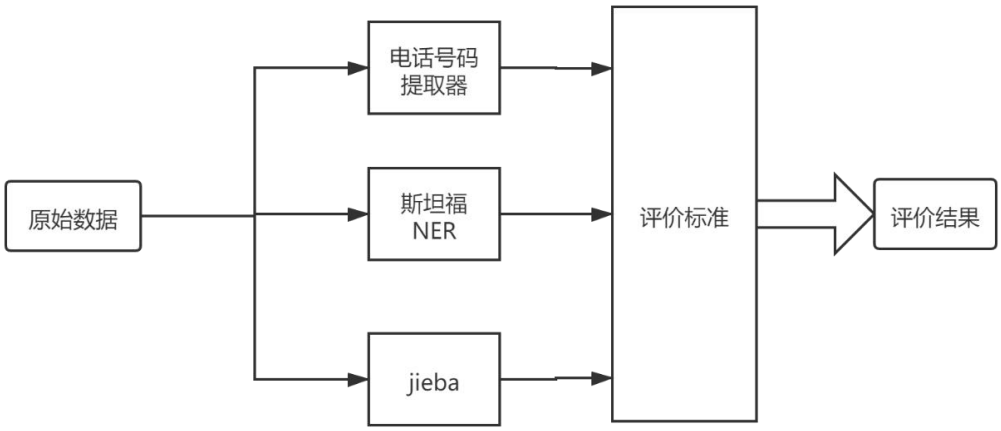

图 5 完整性评价流程图

3) 礼貌评价标准
   a) 首先，从网络上爬取日常生活中常用的礼貌用语（如您好，谢谢），将这些礼貌用语记为集合 P，将 P 保存在 polite.csv 文件中。
   b) 利用 python 中的 jieba 中文库进行分词，获得分词结果记为集合 F。将集合 F 与集合 P 进行相似度对照，得出的结果作为评价结果的一部分。
4) 包含咨询电话
   a) 采用 python 中的 re 正则模块，构建出电话提取器，对答复意见中的电话号码进行匹配提取。
5) 标准的答复格式评价标准
   a) 标准答复格式如下：

〈称呼〉：[正文]〈署名〉〈日期〉

   b) 技术介绍：斯坦福 NER(全称 Stanford Named Entity Recognizer) 是斯坦福大学自然语言研究小组发布的成果之一，它带有用于命名实体识别的精心设计的特征提取器，以及定义特征提取器的许多选项，接受 CoNLL，MUC-6，MUC-7 和 ACE 命名实体语料库的混合训练。斯坦福 NER 也被称为 CRF 分类器，提供了(任意阶)线性链条件随机场(CRF)序列模型的一般实现，同时它也支持中文。
   c) 采用斯坦福 NER，对答复意见进行处理。将标准答复格式分为称呼，正文，署名，日期四个部分进行匹配，按照匹配结果进行评价。
   d) 利用斯坦福 NER 识别答复意见中的实体，提取出出现在文末的署名和日期，经过正确性检验后对答复意见进行评价。

3.3.3.2 可解释性

对答复意见进行正则匹配，匹配提取书名号中的法规，章程，条文等信息，记为集合 Z。

对集合 Z 中的信息进行有效性判断，去除无效信息。 统计所有答复意见中的有效信息，将每条答复意见与总体可解释性水平进行对比，将对比结果作为

评价结果。

3.3.3.3 相关性

数据预处理：

1. 因数据信息中存在大量多余字符，这些无用字符可能降低评价效率与评价结果，对于这些多余字符，使用*号来代替：re.sub('*',' ',data)。

Jieba 分词：

1. jieba 装载自定义词典
2. 打开保存停用词的 stopword.txt 文件，将停用词读取记为集合 S，使用 S 分别对数据中的”留言详情”和”答复意见”进行分词，分词结果分别记为 R1，R2。

TF-IDF 模型：

1.建立词典，基于该词典将分词列表集 R1，R2 转换为稀疏向量集，记为语料库 Y。创建 TF-IDF 模型，使用语料库 Y 进行训练。

2.利用训练完毕的模型处理被检索文本与搜索词，从而得到相关性比较的结果作为评价结果。

3.3.3.4 答复效率

令从群众留言到政府答复的时间跨度为 K 天，将 K 的取值分为 5 个等级，按照对应等级对答复效率进行评价。

行标题	等级	分数
K≤3	1	5
K≤7	2	4
K≤14	3	3
K≤30	4	2
K≤90	5	1
K>90	6	0

图 6 答复效率等级得分表

3.3.4 层次分析法评价模型

3.3.4.1 建模思路：

利用层次分析法对准则层里的影响因素和方案层的选择进行权重分析，得到满足一致性检验的权重，对每个方案进行综合评价，得到综合得分来进行排

序。

(1) 建立递阶的层次结构：根据对问题的分析，缕清问题所包含的因素，确定出各个因素之间的关联和隶属关系，按这些因素的共同特性，将它们分为目标层、准则层、方案层等多个层次。

(2) 建立两两判断矩阵[6]：判断矩阵表示针对上一层次的某元素，本层次与它有关的元素之前相对重要性的比较。

A	B_1	B_2	……	B_n
B_1	b_{11}	b_{12}	……	b_{1n}
B_2	b_{21}	b_{22}	……	b_{2n}
……	……	……	……	……
B_n	b_{n1}	b_{n2}	……	b_{nn}

表 1 判断矩阵的一般形式

(3) 判断矩阵中的b_{ij}一般采用九分制标度法，根据资料数据、专家意见或者系统分析人员的经验，经过反复研究后确定。

标度	含义
1	表示两个因素相比，具有同样重要性
3	表示两个因素相比，一个因素比另一个因素稍微重要
5	表示两个因素相比，一个因素比另一个因素明显重要
7	表示两个因素相比，一个因素比另一个因素强烈重要
9	表示两个因素相比，一个因素比另一个因素极端重要
2, 4, 6, 8	上述两相邻判断的中值
倒数	A和B相比如果标度为3，那么B和A相比就是1/3

图 7 两两比较尺度表

(4) 计算各元素权重：通过对判断矩阵的运算，计算出本层所有元素对上一层相关元素的权重，再利用单层次权重的计算结果，进一步综合出对更上一层次元素的权重。通过权重排序，挑选出最优方案。

3.3.4.2 模型建立

层次分析法的原理：

层次分析法[7]，简称 AHP，是由美国匹兹堡大学教授 T.L.Satty 于 20 世纪 70 年代提出的一种多目标决策分析方法论 [1]。其原理是将与决策有关的因素分解成目标层、准则层、方案层等若干层次，通过对各因素的计算和比较，得出不同因素的权重，为决策者选择最优方案提供参考依据 [2]。

建立评价体系：

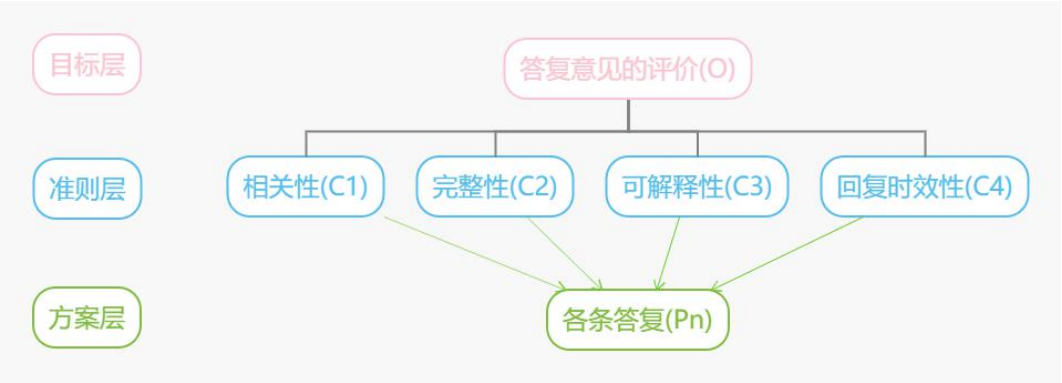

图 8 评价体系图

3.3.4.3 求解方法

先构建判断矩阵：

O	C1	C2	C3	C4
C1	1	2	4	5
C2	1/2	1	2	5
C3	1/4	1/2	1	2
C4	1/5	1/5	1/2	1

表 2 判断矩阵 O-C

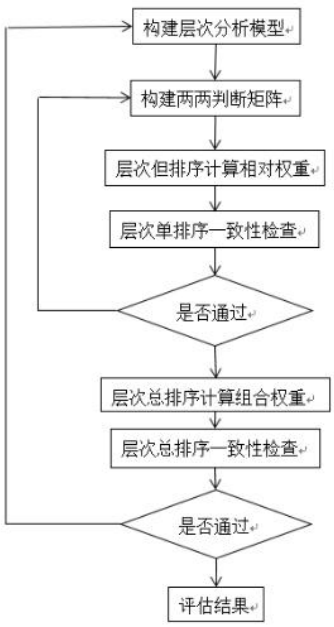

图 9 层次分析法流程图

第一步：计算一致性指标 CI

$$CI = \frac{\lambda_{\max} - n}{n-1}$$

$$CI = 0.0158$$

第二步：查找对应的平均随机一致性指标 RI

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
RI	0	0	0.52	0.89	1.12	1.26	1.36	1.41	1.46	1.49	1.52	1.54	1.56	1.58	1.59

$$CR = \frac{CI}{RI}$$

$$CR = 0.0177$$

因为 CR<0.10，所以该判断矩阵 A 的一致性可以接受。

4.实验结果

4.1 问题 1 实验分析

4.1.1 多项式朴素贝叶斯模型分类结果

通过对群众留言的文本进行数据预处理和中文文本分词，利用 IF-IDF 算法算出各条群众留言的 tf-idf 权值向量，并利用 0.8 的数据量训练模型和用 0.2 的数据量进行预测，得到了模型的分类准确度为 0.887191，F-Score 得分为 0.886354。

```
#模型训练
adata,data_after_stop,labels=data_process()
data_tr,data_te,labels_tr,labels_te=train_test_split(adata,labels,test_size=0.2)

coutVectorizer=CountVectorizer()
data_tr=coutVectorizer.fit_transform(data_tr)
X_tr=TfidfTransformer().fit_transform(data_tr.toarray()).toarray()

data_te=te=CountVectorizer(vocabulary=coutVectorizer.vocabulary_).fit_transform(data_te)
X_te=TfidfTransformer().fit_transform(data_te.toarray()).toarray()
model=MultinomialNB()
#model=GaussianNB()
model.fit(X_tr,labels_tr)
model.score(X_te,labels_te)
print('分类准确度：',sep="")
print(model.score(X_te,labels_te))
```

```
n Console × model × model (1) × model (2) × model (3) × model (4) × model (5) ×
分类准确度：
0.8707403055229143
利用f1_score函数算出来的F-Score：
0.8703973024379612
根据公式定义算出来的F-Score：
0.87039727

Process finished with exit code 0
```

图 10 多项式朴素贝叶斯模型分类结果

图 11 测试子集分类的混淆矩阵

4.2 问题 2 实验分析

通过对政务留言明细表进行聚类分析、制定热度评价指标以及排序工作，得到“热点问题留言明细表”和“热点问题表”。通过分析“热点问题表”可

见近年来人们对于政务问题大多聚焦于噪音问题、政策问题以及城市资源问题，如下表：

热度指数	时间范围	地点/人群	问题描述	数量
81	2019/11/16 至 2020/01/26	A 市 A2 区丽发新城小区	搅拌站噪音扰民，造成环境污染大	81
3.483333	2019/10/15 至 2020/01/05	A 市 A7 县	资源，政策规划以及道路安全问题	209
3.333333	2020/01/02 至 2020/01/06	A 市南站	候车室人员分配及管理问题，以及周边出租车辆非法载客问题	30
2	2019/09/17 至 2019/10/30	A 市 A4 区	教师待遇问题，医院以及“假冒期刊”问题	6
1.857143	2019/08/26 至 2019/12/04	A 市魅力之城	小区墙体开裂，商铺噪声严重，环境污染大	13

表 3 热点问题表

各问题数量占比图如下：

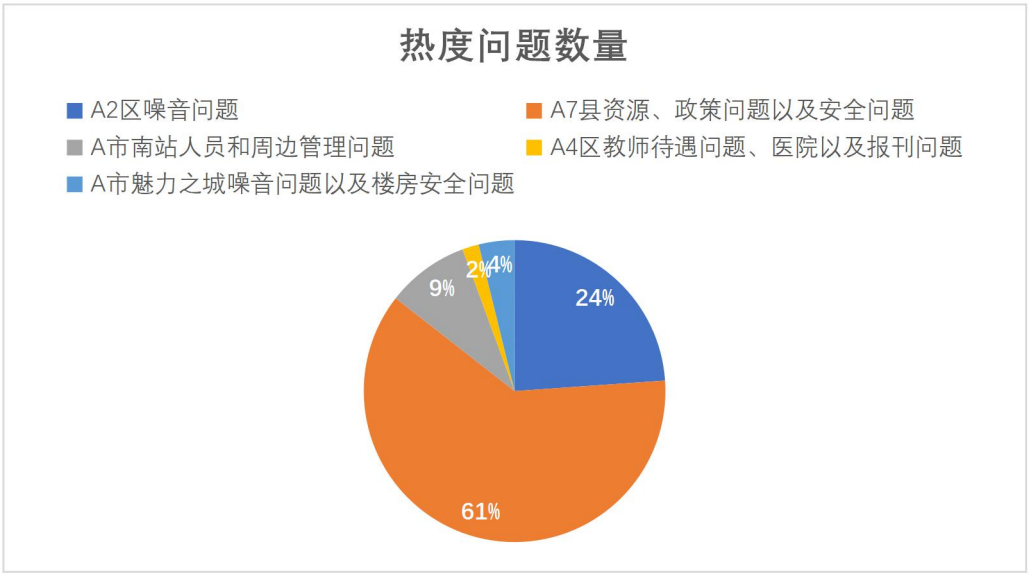

图 12 热点问题数量比重图

通过以上数据可得，人民群众在生活中的问题大多数任是环保问题、社会资源问题以及人身利益问题，政府以及相关机关部门可通过这些数据，更好的去实施相对应的措施，使民生问题更好的得到解决以及反馈。

4.3 问题 3 实验分析

4.3.1 预处理

令 AVER 为完整性，可解释性，相关性，答复时效性的评价结果中的平均数，L1，L2，L3，L4 为四个档次。

档次	表示范围
L1	L1<AVER/2
L2	L1<L2<AVER
L3	L2<L3<AVER*3/2
L4	L3<L4

表 4 评价因素的档次

4.3.2 结果分析

4.3.2.1 对完整性的分析：

如图，答复意见的完整性分布较为平均，呈”橄榄型”分布格局。

图 13 完整性档次分布

取出若干条完整性得分在平均分左右的答复意见数据，观察发现大部分答复内容有按照基本的答复格式进行答复，但仍然存在欠缺署名或是欠缺答复时间等问题。政府工作人员对于礼貌和按照规范答复还有提升空间。

4.3.2.2 对可解释性的分析

如图，有 30%的答复意见是引用了相关的法律条文或是章程。引经据典可以让答复更加有据可依，让答复更有信服力。图中仍有 70%的答复意见没有引用，因此政府工作人员应加强这方面的答复，让答复更有信服力。

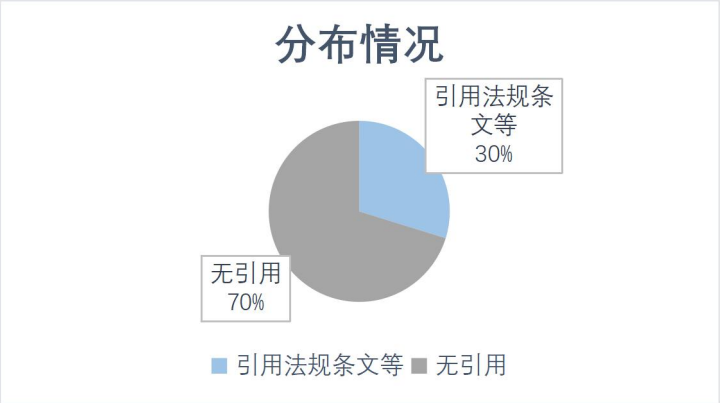

图 14 法规条文引用比重

4.2.2.3 对相关性的分析

如图，答复意见的相关性分布中只有少部分是相关性很差的，超过 80%的答复意见的相关性是可以接受的。在四个档次中取样观察，发现大部分答复意见的相关性是可以与下图情况相符合。由此可见政府工作人员对于留言的答复确实是在解决问题的。

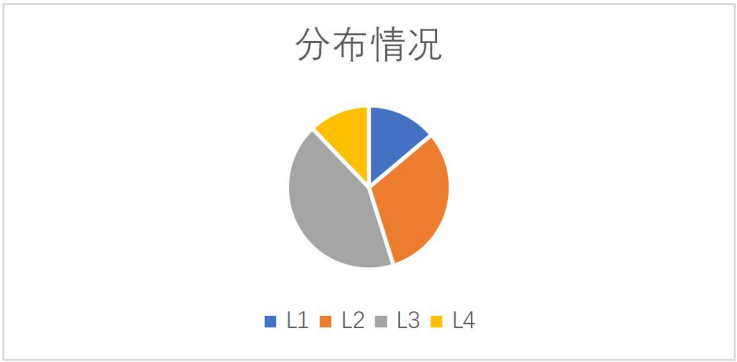

图 15 相关性档次分布比重

4.2.2.4 对答复效率的分析

如图，只有 39%的留言是在 14 天内被答复的，有 40%的留言超过 1 个月才会被答复，甚至有 21%的留言是 3 个月后才被答复。

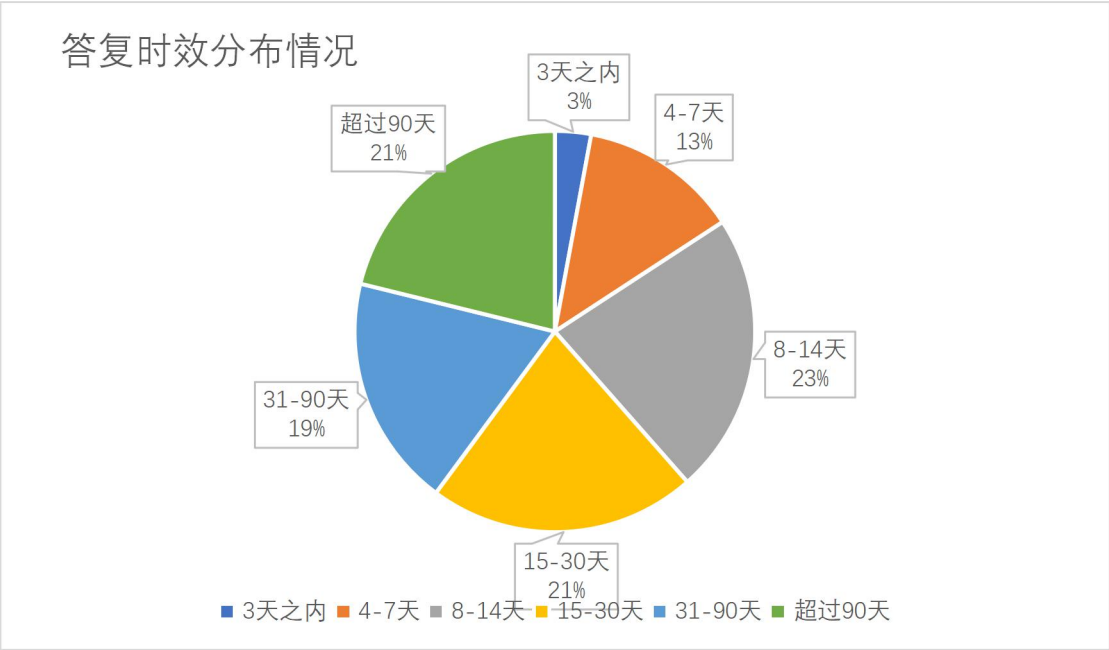

图 16 答复时效分布图

在这个瞬息万变的现代化社会，每隔一小段时间社会就可能发生天翻地覆的变化，因此对于民众积极反应的民生问题，答复的时效性就显得尤为重要，政府工作人员对于答复的效率仍需提高，为广大民众提供更及时的服务。

4.2.2.5 综合得分情况

计算每个因素的权重：

影响因素	算术平均法	几何平均法	特征值法
相关性	0.4928	0.4938	0.4949
完整性	0.2945	0.2936	0.2933
可解释性	0.1376	0.1388	0.1375
回复时效性	0.0751	0.0738	0.0743

表 5 各个因素权重表

最终结果表格：

	指标权重	各条答复得分
相关性	0.4949	s1
完整性	0.2933	s2
可解释性	0.1375	s3
回复时效性	0.0743	s4

表 6 指标权重排序表

答复意见质量评分=0.4949×s1＋0.2933×s2＋0.1375×s3＋0.0743×s4

图 17 指标权重排序图

求解结果：

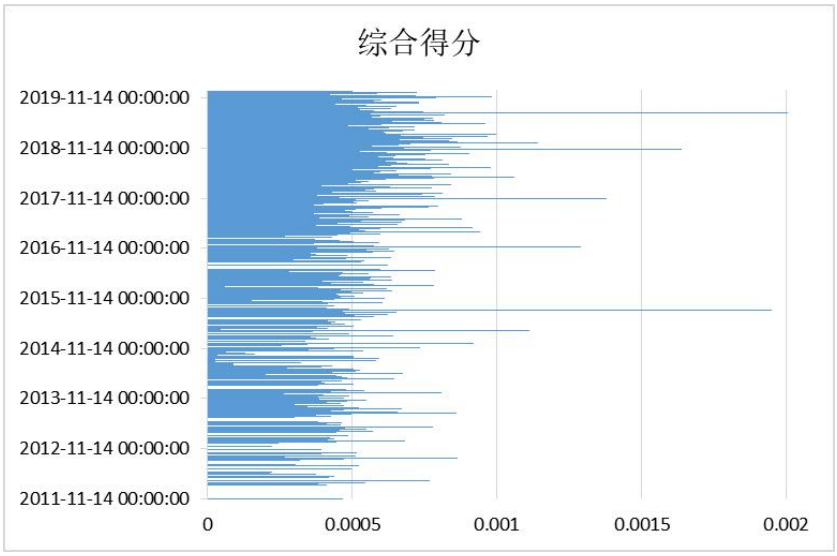

图 28 答复意见综合得分随时间变化图

根据完整性、相关性、可解释性和回复时效性的角度对各条答复进行评价，得到各条答复的综合得分。

如图，首先综合几个角度后，我们可以发现数据总体呈现”橄榄型”分布，答复效果极差或者极好的只占了一小部分，是一个相对比较健康的分布。
其次，答复意见中仍然存在答复效果极差的现象，政府应当积极消除这种现象，让人民的问题都能得到有效的解决。

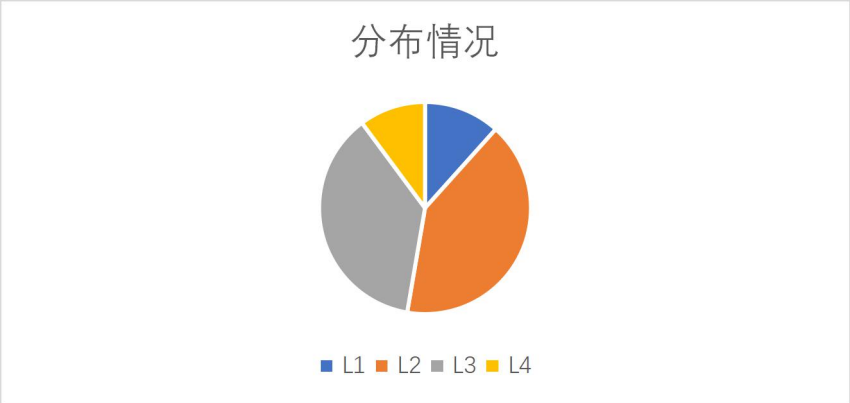

图 19 答复意见综合得分档次比重

4.2.2.6 层次分析法优缺点分析

优点：系统性的分析方法。层次分析法把研究对象作为一个系统，按照分解、比较判断、综合的思维方式进行决策，成为继机理分析、统计分析之后发展起来的系统分析的重要工具。系统的思想在于不割断各个因素对结果的影响，而层次分析法中每一层的权重设置最后都会直接或间接影响到结果，而且在每个层次中的每个因素对结果的影响程度都是量化的，非常清晰明确。这种方法尤其可用于对无结构特性的系统评价以及多目标、多准则、多时期等的系统评价。

缺点：特征值和特征向量的精确求法比较复杂在求判断矩阵的特征值和特征向量时，所用的方法和我们多元统计所用的方法是一样的。在二阶、三阶的时候，我们还比较容易处理，但随着指标的增加，阶数也随之增加，在计算上也变得越来越困难。不过幸运的是这个缺点比较好解决，我们有三种比较常用的近似计算方法。第一种就是和法，第二种是幂法，还有一种常用方法是根法。

5. 结论

对于网络问政平台文本挖掘分析研究，快速了解民情民意，对于政府提高工作效率和服务质量具有重要意义。同时，“智慧政务”也是文本分析的一个应用，传统的人工解读群众留言和毫无目的的低效率回复已经不能满足人民对于高质量服务的要求和政府服务的发展。本文采用多项式朴素贝叶斯分类模型对群众留言进行快速准确分类，统计每一类问题的留言，进而达到了提高工作效率的效果。

对待群众留言，要求政府有关部门要及时回复当前热门的问题，这对人工服务发出极大的挑战。通过 K-means 算法和中文文本相似度算法可以进行快速精确的聚类，得到群众留言的热点问题的数量和热度，让政府能够对热点问题进行及时的回复。可以发现搅拌站噪音扰民、环境污染、政策规划和道路安全问题是热点问题。

利用层次分析法从完整性、相关性、可解释性和回复时效性四个方面对政府答复意见进行评价，发现政府答复意见质量普遍较高，但有随时间整体提高

的趋势，可以看出政府的服务质量正在不断提高，向着新型服务型政府发展。

6.参考文献

[1]基于 Python 语言的中文分词技术的研究　祝永志；　荆静　通信技术 2019-07-10 期刊
[2]多项式朴素贝叶斯文本分类算法改进研究　张伦干　中国地质大学 2018-05-01 硕　士
[3]结合信息增益率和 K-means 聚类的协同训练算法　龚旭；吕佳；皮家甜　重庆师范大学学报(自然科学版) 2020-05-06 11:13 期刊
[4]武永亮,赵书良,李长镜,魏娜娣,王子晏.基于 TF-IDF 和余弦相似度的文本分类方法[J].中文信息学报,2017,31(05):138-145.
[5]正则表达式必知必会（修订版）[M].Ben Forta.人民邮电出版社.2015(1)
[6]基于层次分析法的教学质量评价模型的构建[J]. 吴结元,陈磊,雷大正.中小企业管理与科技(上旬刊). 2018(04)
[7]基于 AHP 法的教师教学质量评估改进模型[J]. 陈弘,李幽铮,郑钢.金陵科技学院学报. 2010(01)