“智慧政务”中的文本挖掘应用

摘 要：近年来，随着互联网的飞快发展，各大网络问政平台的社情民意相关的文本数据量不断攀升，因此，为构建更加高效的智慧政务系统，运用自然语言处理和文本挖掘技术对各类群众问政留言及相应的答复意见进行研究，具有十分重要的研究意义。

对于问题一，首先基于附件 2 和附件 3 的留言主题，补充分词词典 appendix，结合 jieba 分词自带的词典，对附件 2 留言主题进行文本分词。其次网络搜索成熟的停用词词库 stoplist，再进行词频统计，从词频大于 50 的词中选择停用词，构建停用词变量 stop_add，利用两类停用词过滤，并进行文本向量化表示。最后，利用朴素贝叶斯分类法、KNN 分类法、SVM 分类法进行七个一级标签分类，比较 F_1 得分。最后得出相对最佳分类模型—SVM，准确率为 83.26%。

对于问题二，首先对附件 3 中的数据进行清洗，去除了重复值、异常值，转换了日期格式。其次利用 appendix 字典，stoplist 以及词频大于 20 中选择停用词构建停用词变量 add，进行分词、去停用词，计算 TF-IDF 值构建词频矩阵。其次用 LSA 模型、PCA 模型进行降维，层次聚类法进行聚类，得到 1981 类。最后，运用 AHP 构权法，基于留言条数、点赞数、反对数三个热度评价指标计算类热度指数，排序得到前 5 热点问题。

对于问题三，对附件 4 的留言详情、答复意见做数据清洗后，采用正则表达式的方法计算出本文定义的完整性、相关性、可解释性等指标。对指标进行同量纲处理，采用层次分析法进行综合赋权，对答复意见进行评价。

关键词：SVM；LSA；层次聚类；层次分析；主成分分析；

Application of text mining in “Intelligent Government”

Abstract： In recent years, with the rapid development of the Internet the amount of text data related to social situation and public opinion of major online political platforms has been increasing. Therefore, in order to build a more efficient intelligent government system, it is of great significance to use natural language processing and text mining technology to study all kinds of political messages and corresponding replies.

For the first question, firstly based on the message subjects of Annex 2 and Annex 3, and combined the defined word segmentation dictionary named appendix with the dictionary of Jieba , the paper split the texts in the message subjects of Annex 2. Secondly, the paper searched many mature stopwords named stoplist from internet, and then counted word frequency. Selected some stopwords named stop_add from words that frequency are more than 50. Then used the two kinds of stopwords to filter text, and carried on the text vectorization representation. Next, the paper used Naive Bayes classification, KNN classification and SVM classification to classify seven first level tags, and compared F1 scores. Finally, the paper concluded that SVM model is the best classification model which accuracy is 83.26%.

For the second question, firstly the paper cleaned the data in the Annex 3, removed the duplicate values and abnormal values, and transformed the date format. Then with the appendix dictionary, the stoplist and stopwords named add selected from words that frequency are more than 20, the paper split and flitter the texts. Next calculated the TF-IDF value to build the word frequency matrix. Following the LSA model and PCA model were used to reduce the dimension, and the hierarchical clustering method were used to cluster which obtained 1981 classes. Finally based on the indexes of message’s number, approval number and oppose number, the paper used AHP method to obtain

weights and got the top five hot issues by calculating and sorted the heat indexes.

For the third problem, cleaning the data in message details and reply opinions from Annex 4, the paper used the regular expression method to calculate the indexes of integrity, correlation and interpretability. Then the paper eliminated the influence of dimension, and used AHP for comprehensive weighting. Finally, the paper evaluated the responses comprehensively.

Key words: Support Vector Machine；Latent Semantic Analysis；Hierarchical Clustering；Analytic Hierarchy Process; Principal Component Analysis；

目 录

1. 挖掘目标

本文基于自然语言处理和文本挖掘技术，对各类群众问政留言及相应的答复意见进行研究，主要解决留言分类、热点挖掘、回复评价三个方面的问题。

第一，对群众留言进行分类。利用文本分析和文本分类的方法对群众问政平台的 9210 条留言进行分类，建立关于“城乡建设”、“卫生计生”等 7 个一级标签的分类模型，并对各类分类方法进行评价，选出最优分类模型。

第二，对群众反映的热点问题进行挖掘。基于 4326 条群众留言，对某一时间段反映特定地点或特定人群问题的留言进行聚类，采用留言数、点赞数、反对数定义合理的热度指标，根据计算结果找出排名前 5 的热点问题，以及相应的具体留言信息，进而挖掘出群众关注的热点信息。

第三，对相关部门对留言的答复意见进行评价。基于 2816 条关于群众留言及对应的部门答复意见的文本记录，定义、量化答复的相关性、完整性、可解释性这三个指标，最后对答复意见的质量进行综合评价分析。

2. 总体步骤

本文基于 Python 软件，进行文本数据的分析与挖掘[6]。首先进行文本分词，去除停用词，其次进行文本向量化表示，最后构建相应模型，得到结论。

2.1 文本向量化表示

对于文本数据，数据清洗后需进行分词、停用词过滤处理。基于附件 2 和附件 3 的留言主题，补充分词词典 appendix，结合 jieba 分词自带的词典，对附件 2、附件 3 的留言主题，以及附件 4 的答复意见进行文本分词。

网络搜索成熟的停用词词库 stoplist，再进行词频统计，从词频大于 20 或 50 的词中选择停用词，构建停用词变量，利用两类停用词过滤，并进行文本向量化表示。

2.2 留言分类模型构建

对于问题 1 分类问题，根据七个一级标签，对附件 2 中的群众留言进行分

类。在建立关于留言主题的文本向量后，运用朴素贝叶斯模型、KNN 模型、SVM 模型进行文本分类，并计算相应的 F_1 值，进行分类准确率比较。

2.3 留言聚类模型构建

对于问题 2 中聚类问题，首先，对文本向量化后的矩阵，计算 TF-IDF 值，构建词频矩阵。其次，为了解决同义词问题，采用 LSA 算法进行语义降维。接着使用 PCA 算法进一步降维。最后对降维后的矩阵，使用层次聚类法进行文本聚类[9]。

2.4 热度指数计算

对于问题 2 的热点指数计算问题，基于留言条数、点赞数、反对数三个热度评价指标进行计算。由于留言条数相对于点赞数和反对数的重要性不相同，因此选择层次分析法，进行指标重要性的两两比较，进而确定指标权重，计算各类热度指数，最终得到前 5 个热点问题。

2.5 答复质量评价指标量化

对于问题 3 的部门答复意见质量评价问题，从完整性、可解释性、相关性三个角度计算综合评价值。

对于完整性的指标值，定义为包含称呼、招呼、收阅状态、处理状态、感谢表示、回复时间 6 个结构的完整性的综合评价值。首先，由于处理状态结构反映了回复的完整性，且回复表达差异化的存在，故对原始文本，进行了关键字匹配统计后，对部分未匹配的文本计算其有效长度，综合衡量处理状态结构完整性。接着，对其余 5 个结构直接进行关键字匹配统计。最后，依据 5 个部分的重要性不同，进行 AHP 构权，加权计算得到完整性指标值。

对于可解释性，定义为有效文本长度值。对原始文本进行分词、去停用词，得到有效文本，对文本计算有效长度。

对于相关性的指标值，定义为附件 4 中留言详情和答复意见文本的相似度数值，故对向量化文本，基于余弦距离计算文本的相似度。

2.6 答复质量综合评价

将答复指标量化和进行归一化处理后，对相关性、完整性、可解释性等三个指标进行两两比较，依据层次分析对这三个指标赋权，得得到答复质量综合性得分，进而对答复质量进行评价。

3. 群众留言分类

3.1 问题 1 流程图

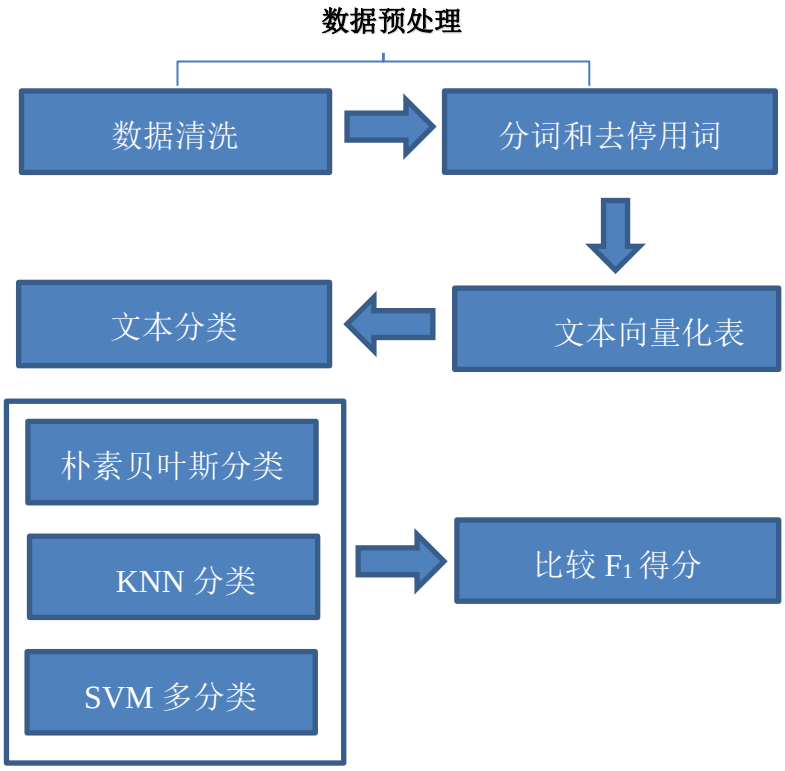

图 3-1 问题 1 流程图

3.2 数据预处理

3.2.1 数据清洗

对于附件 2 中的列“留言时间”，使用时间转换函数将其转化为时间格式。查询重复值，输出结果表明无重复值，因此无需进行去重处理。

3.2.2 文本分词

1、jieba 工具包简介

在对留言主题进行分类之前，需要将计算机不可识别的文本信息转化为可以识别的结构化信息。在附件 2 中，留言主题以及一级标签都是以文本形式给出，为了便于后续分析，首先对留言主题进行分词[1]。这里采用了 python 的工具包 jieba。jieba 分词算法使用了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生成词情况所构成的有向无环图(DAG),然后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，使用了 Viterbi 算法。

2、补充词典 appendix 构建

在尝试进行初步分词以后，观察结果发现有些实体地名分词错误，如长株潭分为了长株和潭。为使这类专有名词的分词结果更加准确，在 jieba 自带的词典的基础上，添加了自定义的词典 appendix。

由于群众留言分类问题和热点挖掘问题均涉及到群众留言，且留言主题中实体地点具有较多的重合，因此综合浏览了附件 2 和附件 3 中的“留言主题”列，选出行政级的地点（市、县、区、镇等）、居民区的地点（小区、街道、苑等）、公共区域（交通道路、交通线、学校、医院、广场、公园、大厦、其他设施等）、企业（集团、公司等）、特定事件名词（58 车贷、潇楚卡等）、一些由于长度过长或算法原因可能会分错的民生词（老虎机、垃圾站、渣土车、搅拌站、夜宵摊等），以及一些其他补充词，进而构建了自定义的分词词典 appendix。

3、分词结果展示

结果保存在 data_cut 数据集，部分分词结果如图 3-2 所示。

0	['A市', '西湖建筑集团', '占', '道', '施工', '有', '安全隐患']
1	['A市', '在水一方大厦', '人为', '烂尾', '多年', ',', '安全隐患', '严重']
2	['投诉', 'A市', 'A1区', '苑', '物业', '违规', '收', '停车费']
3	['A1区', '蔡锷', '南路', 'A2区', '华庭', '楼顶', '水箱', '长年', '不洗']
4	['A1区', 'A2区', '华庭自来水', '好大', '一股', '霉味']
5	['投诉', 'A市', '盛世耀凯', '小区', '物业', '无故', '停水']
6	['咨询', 'A市', '楼盘', '集中', '供暖', '一事']
7	['A3区', '桐梓坡', '西路', '可可小城', '长期', '停水', '得不到', '解决']

图 3-2 部分分词结果

3.2.3 停用词过滤

为了提高工作效率，分词之后需要去除留言主题中表达无意义的字符、词汇，如的、了、呀、此外、而且等等，这些字词就被称为停用词。这些词出现的频率高，且对于文本分类无作用。因此基于停用词表，进行留言主题中停用词的过滤。

1、停用词表构建

本文对于停用词的词库构建分为两部分：

一是搜集已经比较成熟的停用词，构建停用词表 stoplist。

二是结合本题分类要求，根据词频 TF 补充定义停用词，在 Python 中构建停用词变量 stop_add。

TF 是一种简单的评估函数，其值为留言主题中单词发生的频数，其主要原理为假设一个词语在大量的出现时，则认为为噪声词。因此，本文从词频大于 50 的词中，根据实际分类需求选出部分停用词（如社保局、社保卡之类词虽然词频高，但具有实际的意义，对于留言类型的区分具有标识意义，因此不选择加入停用词表），选择如对于分类无帮助的行政类地点（市县区），以及其他可能的停用词，放进停用词表中。

2、去停用词结果展示

基于停用词表 stoplist 和停用词列表 stop_add，得到去除停用词后的部分结果如图 3-3 所示，结果保存为 data_after 数据集。

索引	留言主题
0	['西湖建筑集团', '占', '施工', '安全隐患']
1	['在水一方大厦', '人为', '烂尾', '多年', '安全隐患', '严重']
2	['投诉', 'A1区', '物业', '违规', '收', '停车费']
3	['A1区', '蔡锷', '南路', 'A2区', '华庭', '楼顶', '水箱', '长年', '不洗']
4	['A1区', 'A2区', '华庭自来水', '好大', '一股', '霉味']
5	['投诉', '盛世耀凯', '物业', '无故', '停水']
6	['咨询', '楼盘', '集中', '供暖', '一事']
7	['桐梓坡', '西路', '可可小城', '长期', '停水', '得不到', '解决']
8	['收取', '城市', '垃圾处理', '费', '平等']
9	['魏家坡', '脏乱差']
10	['魏家坡', '脏乱差']

图 3-3 部分去停用词后的分词结果

3.3 文本向量化表示

对文本进行分词以及停用词过滤以后，构建一个空间向量矩阵，即采用 one-hot 编码将文本进行向量化表示，即每一个文本表示一个向量，并且以每一个不同的特征项（分词词语）表示向量空间中的每一个维度，每一维的值表示对应的特征项在文本中出现的频率，即出现表示分词词语则记为 1，否则记为 0。对于一级标签，用数字进行表示，即 0~6 分别表示城乡建设、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。最终形成了一个 9210*13553 的矩阵。

3.4 文本分类

3.4.1 分类方法简介[5]

1. 朴素贝叶斯分类

朴素贝叶斯分类是一种十分简单且比较常见的分类方法。它的主要思想为：对于给出的未被分类的样本，计算得到未被分类的样本在各种类别中出现的概率，比较概率大小，将未分类项划分给概率最大的类别。其流程见图 3-4 所示。由流程图可知，朴素贝叶斯主要分为三个阶段：

第一阶段为准备阶段。在附件 2 中的数据中已经给出了每个留言主题的具体分类，无需人工确定特征属性并且加以分类。

第二阶段为分类器训练阶段。在这个部分主要是为了得到分类模型，即计算每一个类别在训练集中出现频率，以及每个特征属性对每个类别的条件概率估计，并且保存结果。即输入的为训练集以及训练集分类结果，输出为分类器。

第三个阶段为应用阶段。这个过程主要是采用测试集对分类器进行测试，并且采用 f1 得分计算分类器的正确性。

2. KNN 分类

KNN 算法[4]也是一种比较常见且常用的一种分类算法。其主要思想为“近朱者赤，近墨者黑”，即一个样本在特征空间中最邻近的 K 个样本中大多数都属于第 i 个类别，那么该样本也属于第 i 个类别。其算法描述如下：

（1）计算测试数据与各个训练集数据之间的距离；

（2）按照距离的递增关系进行排序，选取距离最近的 K 个点；

（3）确定这 K 个点所在的类别出现的概率；

（4）得到 K 个点中出现频率最高的类作为测试数据的预测分类；

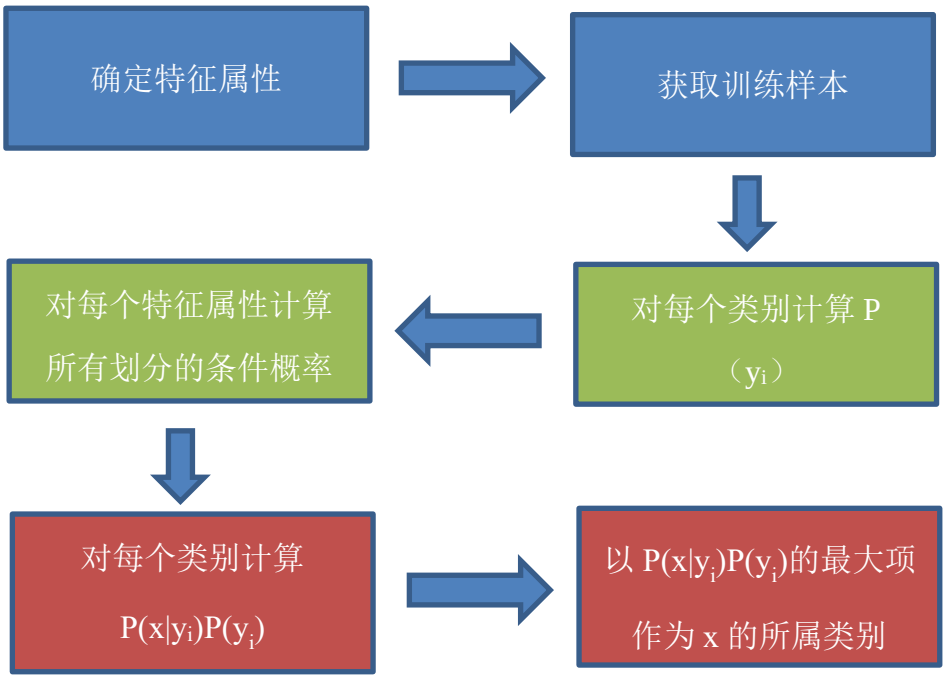

图 3-4 朴素贝叶斯流程图

3. SVM 多分类

SVM 算法[2]最初是为二值分类问题设计的，它主要是为了建立一个超平面作为决策曲面，使得两类之间的隔离边缘被最大化，如图 3-5 所示，中间的直线所表示的分类面，表示两个点集到该分类面的最小距离最大，其边缘点到该分类面的距离最大，这即为最优分类面。SVM 是一种典型的二分类器，但实际解决的问题往往是多分类问题，因此需要构造合适的多类分类器。构造多分类支持向量机分类器主要有两种方法，一种为同时考虑所有类的方法，但是这种方法实现比较困难；另一种为通过组合多个二分类器实现对多分类器的构造。

假设一共有 n 类，在训练时，一次将第 i 个类的样本归为一类，剩下的 n-1 个类归为另一类，这时候 n 个类别的样本就需要构造 n 个 SVM，对于未知样本分类时，将其划分为具有最大函数值的那一类，这种方法被称为一对多法，但是因为训练数据并非全部数据，这种方法中容易形成较大的偏差。为此在仍以两类样本之间构建一个 SVM，进行分类，一共构建了 n*(n-1)/2 个 SVM 分类器，将

未知样本归属于得票最多的那一类，这种方法即为一对一法，也是 python 中 svm 多类分类方法的实现。

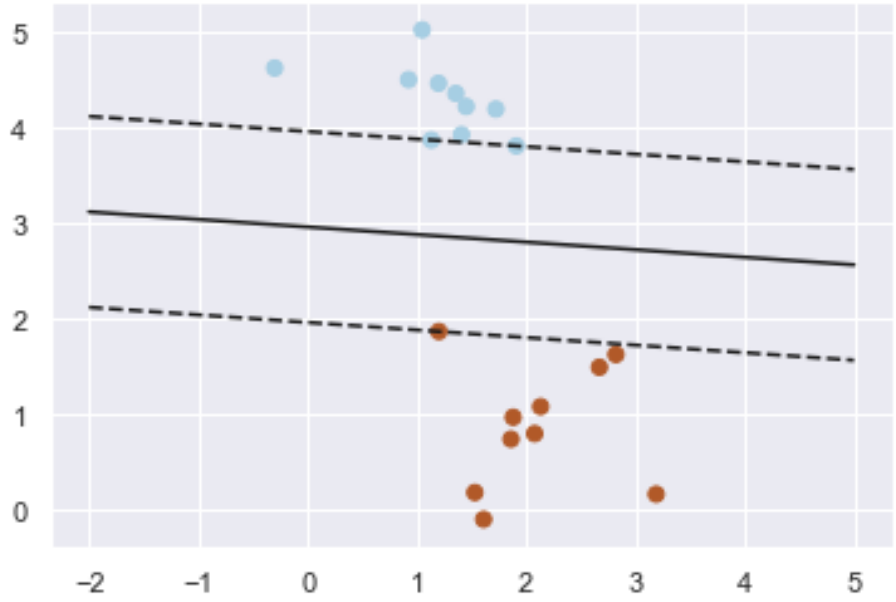

图 3-5　SVM 分类原理图解

对于简单的二分类问题，为了寻找最优分类面，定义点 x 到超平面的距离公式为：

$$d = \frac{|w^T x+b|}{||w||} \qquad (3\text{-}4\text{-}1)$$

其中，直线上方的类别点有$w^T x + b > 0$，下方的类别点则有$w^T x + b < 0$。

由于支持向量机的目标是使得两类数据最边缘的数据点（支撑向量）之间的距离（定义为ρ）越大越好，因此则使得下式最大：

$$\rho = 2\frac{|w^T x_{support}+b|}{||w||} = 2\frac{y_{support}(x_{support}+b)}{||w||} \qquad (3\text{-}4\text{-}2)$$

其中，直线上方的点$y_i = 1, w^T x_i + b \geq \frac{\rho}{2}$，下方的点$y_i = -1, w^T x_i + b \leq -\frac{\rho}{2}$，即等价于$y_i(w^T x_i + b) \geq \frac{\rho}{2}$。

为了求出支持向量，并且使得ρ取得极大值，则上面的问题可以转化为下面的简化的目标函数求解问题：

$$\arg\max_{w,b} \frac{2}{||w||}$$
$$\text{s.t. } y_i(w^T x_i + b) \geq 1 \; for \; i = 1,2,3 \ldots, n \qquad (3\text{-}4\text{-}3)$$

对于此目标函数的求解，转化为对偶问题的凸优化问题：

$$\arg\min_{w,b} \frac{||w||^2}{2}$$

$$s.t.\ \ y_i(w^T x_i + b) \geq 1\ for\ i = 1,2,3\ldots,n \qquad (3\text{-}4\text{-}4)$$

接着运用拉格朗日乘子进行求解，有

$$L(w,b,\alpha) = \frac{||w||^2}{2} - \sum_{i=0}^{n} \alpha_i (y_i(w^T x_i + b) - 1), \alpha_i \geq 0 \qquad (3\text{-}4\text{-}5)$$

可以转化为先求关于 w 和 b 的极小值问题，再求关于α的极大值问题。

至此讨论的是线性可分的 SVM 分类模型。对于模型的推广，一方面考虑模型的泛化能力，因此引入松弛变量ξ，对一些分类不好的点进行选择性忽略。另一方面，为了解决线性不可分的问题，引入核函数对数据进行高维变换，期望在高维中将数据点分开。常用的核函数有：多项式核函数，RBF 径向基核函数，sigmoid 核函数等。本文使用了线性核函数（即一次核函数）。

3.4.2 留言分类

在进行分类之前，将数据拆分为了训练集和测试集，即 80%的训练集用来拟合模型，20%的数据作为测试集用于检验模型，并比较 F_1 得分。这里引入了 python 的 sklearn 工具包。三种分类方法的 F_1 得分整理见表 3-1。

表 3-1 F_1 得分

分类方法	F_1 得分
朴素贝叶斯分类	0.8274
KNN 分类	0.4913
SVM 多分类	0.8326

综合比较三种分类方法，可以看出 SVM 支持向量机分类法效果相对较好，准确率较其他两种方法高。因此构建基于线性核函数的 SVM 分类模型，对群众留言进行一级标签的分类。

4. 热点问题挖掘

4.1 问题 2 流程图

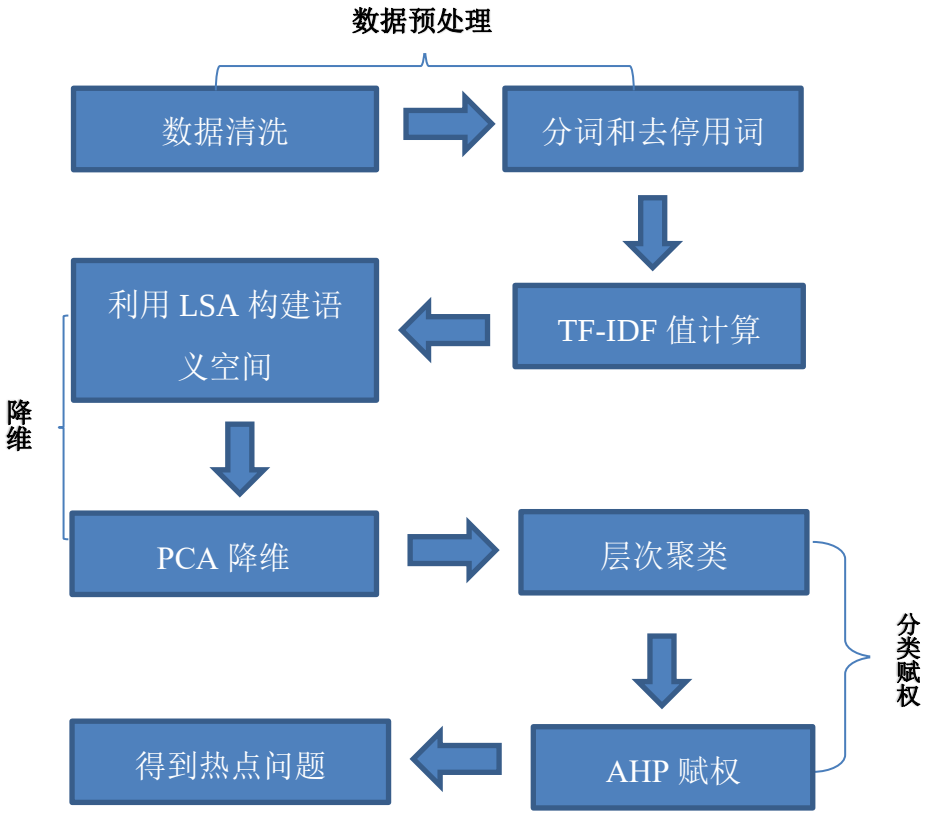

图 4-1 问题 2 流程图

4.2 相关算法简介

4.2.1 TF-IDF 算法

TF-IDF 算法[7]是一种信息检索与文本挖掘中常用的加权技术，用于评估一个字或词在一个文本集或者语料库中的相对重要程度，进而根据计算结果进行排序，提取出关键词。

其主要思想是字词的重要性随着它在文本中出现次数的增加而增加，随着在语料库汇总出现的频率的增加而减少，即如果某词在一篇文章中出现的频率 TF 高，而在其他的文章中出现的频率低，那么就认为这个字或者词具有很好的类别区分效果。本文对分词后的字或词语，运用 TF-IDF 提取关键词，具体计算过程

如下：

第一步，计算词频，即 TF 权重。表示某个词在文本中出现的频率，计算公式为：

$$\mathrm{TF} = \frac{\text{某个词在文本中出现的次数}}{\text{文本中的总词数}} \quad (4\text{-}2\text{-}1)$$

第二步，计算逆向文档频率，即 IDF 权重。计算公式为：

$$\mathrm{IDF} = \log\left(\frac{\text{语料库中文档总数}}{\text{包含某个词的文档数}+1}\right) \quad (4\text{-}2\text{-}2)$$

这里分母加 1 是为了避免分母为 0，导致分母无意义。

第三步，计算 TF-IDF 值。

$$\mathrm{TF}-\mathrm{IDF} = \mathrm{TF} * \mathrm{IDF} \quad (4\text{-}2\text{-}3)$$

某一文档内的高词语频率，以及该词语在整个文档集合（语料库）中的低文档频率，就可以产生较大的 TF-IDF 值，即较大的权重，至此构建了 TF-IDF 词频矩阵，开始进行后续的降维分析。

4.2.2 LSA 算法

LSA 算法[8]通过对大量的文本集进行统计分析，从中提取出词语的上下文使用含义。技术上采用 SVD 分解等处理，消除了同义词、多义词的影响，提高了后续处理的精度。算法流程如下：

1. 分析文档集合，建立词汇-文本矩阵 A
2. 对词汇-文本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

4.2.3 PCA 算法

PCA，即主成分分析，是一种常用的无监督学习方法。这一方法利用正交变换，将高维的线性相关变量进行降维，将原数据转换为少数几个由线性无关变量表示的数据，线性无关的变量即称为主成分，其中数据方差在第一主成分上方向上最大，其次是第二主成分方向上最大，依此类推。这一方法能够在保留大部分信息的基础上实现降维效果，因此适合应用于本文的高维文本向量分析。本文 LSA 去噪后的文本，运用 PCA 算法进一步进行降维。一般 PCA 具体计算步骤如下：

1. 数据标准化处理

对于 n 个样本（本文是留言记录），有 p 维特征来描述。现将数据进行标准化，去除量纲对数据的影响，得到新矩阵 X.

2. 计算特征值特征向量

对矩阵 X（n*p）计算协方差矩阵Σ（p*p），并求出协方差矩阵Σ对应的特征值λ和特征向量 u。其中$\Sigma = U \wedge U^T$，对角阵对角线元素即为特征值$\lambda_i, i = 1,2 \cdots P$。

3. 输出正交矩阵 U

按照特征值从大到小的顺序，将单位特征向量排列成矩阵，得到正交矩阵 U（p*p 维），由 UX 得到主成分矩阵.

4. 确定主成分个数

依据特征值占总特征值之和，计算方差贡献率和累计方差贡献率。一般取累计方差贡献率超过 85%的前 k 个主成分，或者根据具体的降维要求，直接选取前 k 个主成分。

5. 输出降维后的数据

依据选取的 k 个主成分，选取 U 中前 k 列个单位正交向量，构成的转换矩阵 P（p*k）。则可得到降维后的数据$Y = Xp$，此时数据是 n*k 的，即由 p 维降到了 k 维，且各主成分之间不相关。

本文中设定方差贡献率至少为 90%，进行文本向量降维。

4.2.4 层次聚类法

层次聚类法[3]假设类别之间存在层次结构，将样本聚到层次化的类中，具体地分为聚合聚类，分裂聚类，这里采用聚合聚类法。聚合聚类法开始将每个样本各自分到一个类，之后按照一定的规则，将满足规则条件的两个类合并，建立一个新的类，重复此操作直到满足停止条件，得到层次化的类别。因此，聚合聚类要先确定这三个要素：距离或相似度；合并规则；停止条件。

1. 距离或相似度

常用的距离测度有闵可夫斯基距离、马哈拉诺比斯距离，相似度测量有相关系数、夹角余弦。给定样本集合 X，X 是 m 维实数向量空间中点的集合，其中x_i，$x_j \in X, x_i = (x_{1i}, x_{2i}, \dots, x_{mi})^T, x_j = (x_{1j}, x_{2j}, \dots, x_{mj})^T$.

（1）闵可夫斯基距离

定义样本x_i与x_j的闵可夫斯基距离为：

$$d_{ij} = \left(\sum_{k=1}^{m}\left|x_{ki} - x_{kj}\right|^{p}\right)^{\frac{1}{p}} \quad (4\text{-}2\text{-}4)$$

当 p=1 时，即为曼哈顿距离。当 p=2 时，即为常用的欧氏距离：

$$d_{ij} = \left(\sum_{k=1}^{m}\left|x_{ki} - x_{kj}\right|^{2}\right)^{\frac{1}{2}} \quad (4\text{-}2\text{-}5)$$

当距离越小时，文本之间的相似度越高，反之则越低。本文用欧式距离来建模。

（2）夹角余弦

定义样本x_i与x_j的夹角余弦为：

$$s_{ij} = \frac{\sum_{k=1}^{m} x_{ki}x_{kj}}{\left(\sum_{k=1}^{m} {x_{ki}}^{2} \sum_{k=1}^{m} {x_{kj}}^{2}\right)^{\frac{1}{2}}} \quad (4\text{-}2\text{-}6)$$

余弦相似度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。夹角余弦越接近于 1，表示文本之间的相似度越高，越接近于 0，表示文本之间的相似度越低。

2. 合并规则

合并规则一般是类间距离最小。常用的类间距离，可以是最短距离、最长距离、中心距离、平均距离等等。

3. 停止条件

停止条件一般是类的个数达到阈值（极端情况类的个数是 1）、类的直径超过阈值。

本文对主成分降维后的文本向量，进行聚合聚类。具体的聚合聚类算法步骤如下：

（1）计算 n 条留言主题两两之间的欧式距离$\{d_{ij}\}$，记做矩阵D $= d_{ij}{}_{n\times n}$。

（2）构造 n 个类，每个类中只把包含一个样本。

（3）合并类间距离最小的两类，本文类间距离为平均距离，构造一个新类。

（4）重新计算新类与当前各类之间的距离，若类的个数为 1，终止计算，否则回到步骤三。

4.2.5 层次分析法

在得到分类结果后，热度指数采用留言条数、点赞数、反对数这三个指标反映，其中由于点赞数与反对数的数值过于庞大，为弱化其形成的影响，将二者的数值缩小 10 倍，且三个指标的权重由层次分析法确定。

层次分析法是一种比较常见的多目标决策方法，它主要是按照层次的结构计算加权和的赋权方法。其主要流程见图 4-2。

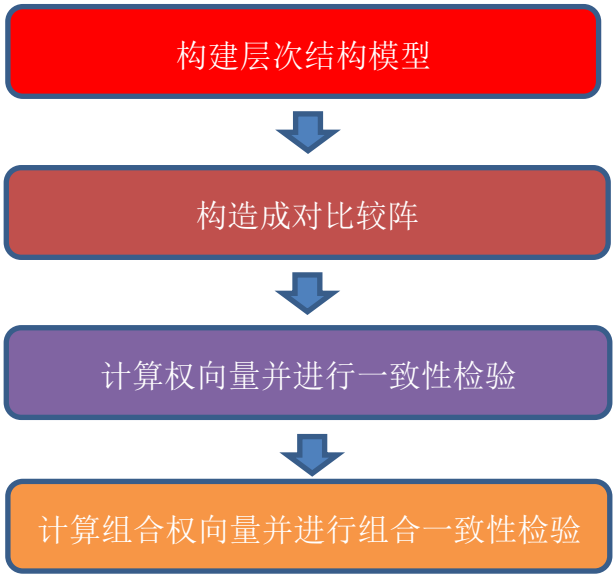

图 4-2 AHP 流程图

依据留言条数、点赞数、反对数这三个指标构建了三层层次结构。在构建成对比较矩阵中第 i 个元素与第 j 个元素相对于上一层某个因素的重要性的时候，采用数量化的相对权重 a_{ij} 进行表示。

4.3 热点问题挖掘

4.3.1 数据预处理

1. 重复值处理

考虑存在用户在一天之内多次留言，且留言内容一致，将其定义为重复留言，予以剔除。附件 3 原数据 4326 条，去除重复留言后，有效留言为 4280 条。

2. 异常值处理

考虑存在空值等异常留言，将留言按照文本长度进行升序排序，发现存在留

言为：“）”，即一个半括号，故予以剔除，得到有效留言 4279 条。

3. 数据格式转换

将留言时间转换为日期格式，便于后续分析。

4. 文本分词

同问题 1，对留言主题进行 jieba 分词，分词词典为 appendix 和 jieba 自带的词典。

5. 停用词过滤

运用词频分析法，从词频大于 20 的词里面挑出部分词，作为 stoplist 停用词表的补充，进而运用新的停用词表进行停用词过滤。

4.3.2 词频矩阵构建

1. 文本向量化表示

构建一个空间向量矩阵，即采用 one-hot 编码将文本进行向量化表示。最终形成 4279*7631 的矩阵，元素为 0 或 1。

2. TF-IDF 值计算

为方便后面聚类，计算 TF-IDF 值，得到关于每条记录每个词的频率，构建新的词频矩阵。

4.3.3 矩阵降维

利用 LSA 算法，处理同义词，进行语义降维，得到语义空间。运用 PCA 算法，对语义空间进一步降维，其中主成分分析中设置方差提取度为 90%。最终将列向量的维度降到 2147，得到 4279*2147 的词频矩阵。

4.3.4 文本聚类

运用层次聚类法，基于欧式距离和平均法，将文本分为了 1981 类。

4.3.5 热度评价指标定义

从留言条数、点赞数、反对数这三个指标的角度出发，运用加权法计算各类热度值。从实际出发，在留言条数、点赞数、反对数这三个指标中留言条数相对于点赞数、反对数要重要许多，因此运用层次分析法确定指标权重。

在层次分析法中，构建的对比较矩阵为$\begin{bmatrix} 1 & 9 & 9 \\ 1/9 & 1 & 3 \\ 1/9 & 1/3 & 1 \end{bmatrix}$。通过一致性检验后得到的 0.8082179、0.12951688 和 0.06226522 分别表示留言条数、点赞数、反对数的权重。故第 i 类热度指数：$z_i = 第i类留言总数 * 0.8082179 + 第i类点赞数 *$

$0.12951688 + 第i类反对数 * 0.06226522$。

4.3.6 热点问题展示

前五类热点问题词云图见图 4-3。可以看出，A 市、A7 县、K9 县、万矿万镜、泉塘等特定地点是热点问题出现的集中地点；虚假宣传、举报、投诉等是部分热点问题的反映。

图 4-3 热点问题关键词词云图

表 4-1 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	32.93	2019/05/05 至 2019/09/19	A 市五矿万境 K9 县	住宅房存在多种质量问题
2	2	26.89	2019/02/27 至 2019/07/22	A 市市民	咨询子女入学事宜
3	3	15.93	2019/01/21 至 2019/11/20	A 市房地产开发商	虚假宣传，欺骗消费者
4	4	14.08	2019/03/05 至 2020/01/07	A 市 A7 县泉塘街道	泉塘街道民生问题
5	5	12.25	2019/02/21 至 2019/08/02	A 市 A4 区	公布 58 车贷案件进展

计算热度指数，得到前 5 名的热点问题，相应问题见表 4-1。可以看出，其中：

排名第一的为 2019 年 5 月 5 日至 2019 年 9 月 19 日期间，群众集中反映 A 市五矿万境 K9 县的住宅房存在质量问题，该问题获得大量的点赞数，表明大部分群众对这一事实比较认可，问题热度高。

排名第二的为 2019 年 2 月 27 日至 2019 年 7 月 12 日期间，群众反映 A 市市民子女入学的相关问题，需咨询具体的入学相关事宜，尤其是金毛湾配套入学。

排名第三的主要为在 2019 年 1 月 21 日至 2019 年 11 月 20 日期间，群众反映 A 市的房地产开发商虚假宣传，欺骗广大消费者现象，对该问题进行举报、投诉。

排名第四的主要反映了在 2019 年 3 月 5 日至 2020 年 1 月 7 日期间，泉塘街道存在的一系列民生问题，包括道路建设、娱乐设施、装修施工等问题。

排名第五的主要反映了在 2019 年 2 月 21 日至 2019 年 8 月 2 日期间，群众对 A 市 A4 区“58 车贷”案件进展关注的相关问题。

5 答复意见评价

5.1 问题 3 流程图

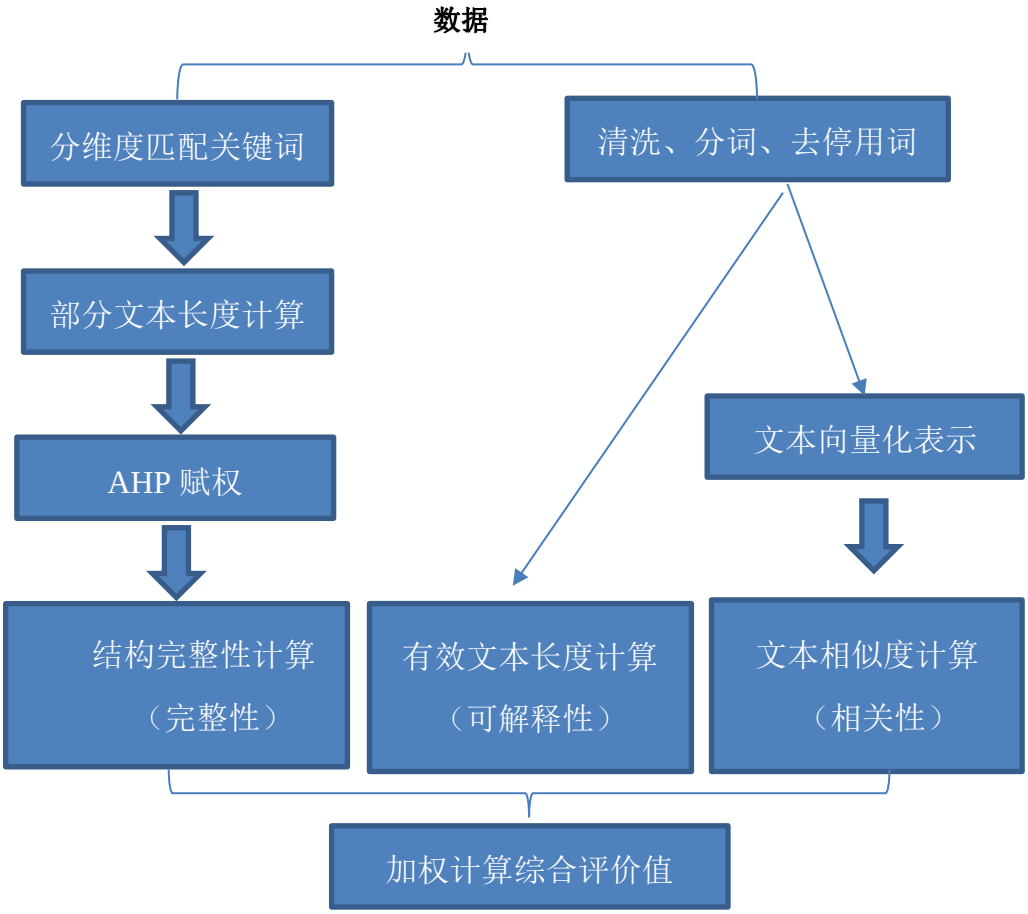

图 5-1 问题 3 流程图

5.2 数据预处理

5.2.1 去空白字符、去重

附件 4 中的“留言详情”和“答复意见”列包含了许多空白字符，这些空白字符会影响后续的数据清理，因此利用正则表达式剔除所有空白字符。

附件 4 中包含了部分留言用户对同一问题重复留言的情况。对完全相同的留言主题和留言详情，政府的答复意见差别很小，对后续的文本挖掘无意义。考虑到对同一用户的相同留言，答复意见中会出现类似“请勿重复提交”的文本，或出现回复意见更简短的情况，因此保留留言时间最早的重复留言。将文本按“留言时间”的升序排列，用 python 中的 drop_duplicates()函数剔除重复值，默认保

留第一个，即可剔除重复留言。

5.2.2 异常值处理

对“答复意见”列按字符长度升序进行排列后，发现“答复意见”列中存在仅回复日期、留言用户名称（“UU0081182”）、一串数字（0000-0000000020163）和单位名称（国网西地省 J9 县供电公司）的情况，这些回复没有具体意义，都属于异常回复，需要利用正则表达式将其剔除。

5.3 答复质量指标定义

5.3.1 完整性

完整性，本文理解为答复意见格式的规范性，即结构的完整性。因此在查阅了其他官方问政答复内容的规范格式，以及浏览了本文答复意见格式的基础上，总结出最基本、经典规范的答复结构框架：

网友***您好！您的留言已收悉。现将有关情况回复如下：***感谢您对我们工作的支持、理解与监督！*年*月*日。

观察答复结构可以看出，答复可以拆分成称呼、招呼、收阅状态、处理状态、感谢表示、回复时间 6 个部分。由于具体表达内容的差异，和时间表示格式上的差异，在观察、总结答复意见的基础上，总结出了下面这些同一结构，不同表达的词语，共 6 类词及结构。

1、称呼：网友、网名（UU0082054）、***同志、市民等；

2、招呼：您好；你好；您们好；

3、收阅状态：收悉；获悉；知悉；已阅

4、处理状态：答复；回复；反馈；已转

5、致谢：感谢；谢谢

6、回复时间：*-*-*-；*.*.*.；*年*月*日；*/*/*

此外，处理状态结构是主要是度量是否进行了有效回复，即回复的完整性。但观察留言发现，有的是直接进行回复的，无法匹配到上述词语。因此，对于此类留言，考虑有效文本长度，对于达到一定长度的留言视为处理状态结构完整。

故在进行完整性量化时，具体步骤如下：

1. 完整性结构统计

针对 1、2、3、5、6 类的结构部分，对于所有的答复意见记录，依次采用正

则表达式搜索匹配不同结构包含的关键词，若包含该结构任意一个关键词，就认为这条答复意见该结构部分视为完整，记为 1，否则记为 0。即若某一答复意见包含 1、2、5 类结构的关键词“网名”、“您好”、“感谢”，则认为该答复意见具有称呼、招呼、致谢结构的完整。

对于 4 中的处理状态的结构部分，由于有些留言并未使用“回复”等关键词，而是直接进行回复的。因此针对其它类结构部分使用的搜索匹配关键词的方式并不适用。对“答复意见”列按文本长度升序排列后，发现含有具体回复（即处理状态完整）的“答复意见”的内容的文本长度较长，而不包含具体回复的答复意见有两种情况：一是文本较短的以“***已收悉。”、“感谢***支持。”、“你好！*年*月*日”为结构的答复意见，二是答复意见文本较长，但在“回复如下：”、“链接：”后无内容。利用正则表达式筛选出上述两种情况的无具体回复的答复意见，其它答复意见则视为处理状态完整。

2. 完整性指标值计算

对于 6 个完整性构成部分，重要性不同，因此采用加权法计算权重。这里仍使用 AHP 层次分析法进行权重的确定，得到 1-6 类的权重分别 0.0554, 0.1046, 0.1204, 0.4229, 0.2693, 0.0274。对所有的答复意见在 1-6 类结构的取值（0 或 1）分别相应的权重相乘后加总，得到每一个答复意见的完整性指标值。

5.3.2 可解释性

对于可解释性，用剔除对留言详情无解释作用的文本后的文本长度表示。有效文本越长，表明有效的答复意见越详细，对群众的问题解释性越好。

无解释作用的文本主要分为：日期、致谢、符号、具体回复前面的内容（留言问题再描述、问候、留言收阅状态等）、回复内容中夹杂称呼（UU0082054）。

对日期、符号、致谢、称呼类，采用 re 包中的 sub()函数直接匹配删除。对具体回复前面的内容，通常以“回复如下”、“收悉”、“你好”等结尾，但是由于部分留言中不止出现一次上述词语，首先用 re 包中的 search()函数分别找到第一次出现上述词语的文本位置，位置靠前说明词语后为具体回复的起始，则剔除词语位置前（包括词语）的内容。需要注意的是：位置非常靠后说明词语后非具体回复的起始，这一类则需要保留。在对答复意见做分词后，将“尊敬”、“网友”、“目前”、“请”、“回复”、“你好”、“收悉”等词语作为停用词剔除后再合并成句

子。最后计算进行清洗后的答复意见的文本长度作为可解释性指标值。

5.3.3 相关性

相关性，即指对于某条问政记录，答复意见中逐条回复的主要内容，是否与群众留言反映的一系列问题相关，即官方答复是否是针对对应的群众留言中提出的某个问题或者一系列问题进行解答，而不是答非所问。

在剔除异常值后的 2275 条记录中，选取“留言详情”列和“答复意见”列对答复的相关性进行评价。在计算可解释性时已经对“答复意见”列剔除了符号、日期、感谢、问候等无意义的语句，因此只需对“留言详情”列、“答复意见”列进行分词、去停用词后分别计算它们的向量化文本。将“留言详情”列、“答复意见”列的文本向量化矩阵纵向拼接，计算 TF-IDF 值，构建新的词频矩阵后将拆分得到“留言详情”列、“答复意见”列的词频矩阵。

进行文本相似度的计算。衡量文本相似度的方法有很多，本文基于余弦距离计算文本的相似度，计算相关性指标。

5.4 文本相似度简介

文本的相似度，即指语义的相似度。常包括词与词、句与句、段落与段落、篇章与篇章，以及词与句、句与段落等之类的相似度问题。这里要解决的是段落与段落之间的相似度问题，即留言详情段落和答复意见段落之间的相似度问题。具体计算步骤如下：

1 文本向量化表示

同问题 1，分别对留言详情和答复意见，采用 jieba 分词，去除停用词，进行文本向量化表示。

2. 余弦相似度[10]计算

余弦值越接近 1，就表明夹角越接近 0 度，两个文本向量越相似。

相比距离度量，余弦相似度更加注重两个向量在方向上的差异，而非距离或长度上的差异，因此适合本题的文本向量。

5.5 答复质量综合评价

对 2775 条答复意见，分别计算出它们的完整性、相关性、可解释性指标值，然后对每一条答复进行质量评价。完整性、相关性、可解释性的对答复意见评价

重要性不同，因此采用加权法计算权重。仍使用 AHP 层次分析法确定权重：W= 0.1562, 0.1852, 0.6586。由于三类指标的量纲不同，需要进行同量纲处理：对每一类指标，每一条答复意见的指标值都减去该类指标的最小值，再用差值除以该类指标的极差，得到每一条答复同量纲下的该指标值。

对所有答复意见同量纲处理后的完整性、相关性、可解释性指标值，分别与相应的权重相乘后加总，得到每一个答复意见的答复质量综合评价。答复质量最高的前十名答复意见见图 5-2。

	回复评价	答复意见
1919	0.770245	首先感谢您对J9县扶贫工作的支持与关注！针对您咨询的关于加快推进原生态有机农产品品牌建设的决...
2038	0.753568	“UU0082390”首先感谢您对J9县扶贫工作的支持与关注！针对您咨询的关于加快推进原生态...
230	0.743697	网友“UU0081211”您好！您的留言已收悉。现将有关情况回复如下：根据《关于实施差别化购...
420	0.738215	网友“UU008424”您好！您的留言已收悉。现将有关情况回复如下：您好！我街道接到反映后，...
322	0.730297	网友“UU0081059”您好！您的留言已收悉。现将有关情况回复如下：1.关于地铁4号线罐子...
1605	0.714345	“A00049261”网友：您好！感谢您对我们人口计生工作的关注和支持！根据《西地省人口计生...
664	0.710559	网友“UU0081065”您好！来信收悉。现回复如下：该路段房屋拆迁业主为A7县途通公司，黄...
770	0.687129	网友“UU008835”您好！来信收悉。现回复如下：经镇综治办调查核实，信访人黄尊富，系我镇...
427	0.672239	网友“UU008144”您好！您的留言已收悉。经A5区教育局调查了解，现将有关情况回复如下：...
2048	0.671363	网友：您好！您于2019年11月14日12点53分发帖，反映有关“K市通中速递有限公司位于K...

图 5-2 前十高评价回复

6. 结论

本文基于 Python 软件，进行文本数据的分析与挖掘。首先结合补充的词典 appendix，利用 jieba 工具进行文本分词，其次综合成熟的停用词词库 stoplist 与补充的停用词词表，对停用词进行过滤，接着进行文本向量化表示，最后构建相应模型，得到结论。

为了对群众留言进行分类，对向量化文本，依次使用朴素贝叶斯分类法、KNN 分类法、SVM 分类法进行七个一级标签分类，并比较 F_1 得分。最后得出相对最佳分类模型为 SVM 模型，准确率为 83.26%。

为了对群众留言反映的热点问题进行挖掘，首先对数据进行清洗、分词、去停用词，得到向量化文本，计算 TF-IDF 值构建词频矩阵；其次，用 LSA 模型、PCA 模型进行降维，层次聚类法进行聚类，得到聚类值；最后，用 AHP 构权法确定留言条数、点赞数、反对数三个热度评价指标的权重，计算类热度指数，排序得到前 5 热点问题。

为了对答复意见质量进行综合评价，首先依据附件 4 的文本形式，合理定义了完整性、可解释性以及相关性等指标；其次，对这三个指标进行量化，且由于指标量纲不同，进行同量纲处理；最后采用 AHP 构权法赋权，得到综合评价得分，对答复意见进行评价。

总结本文的方法，问题一中尝试了 3 种分类方法，发现分类正确率有进一步提高的空间。问题二中在尝试了多种聚类方法后，发现层次聚类法比较合适，但还可以进一步优化此算法改进减少类的数目。问题三作为开放题，对于可解释性定义及量化，还可以进一步挖掘研究。

7. 参考文献

[1] 曹卫峰.中文分词关键技术研究.南京理工大学[D].2009.

[2] https://blog.csdn.net/xuxuxuxuri/article/details/81570576

[3] 李航.统计学习方法[M].北京:清华大学出版社，2012.

[4] 周志华.机器学习[M]. 北京:清华大学出版社，2016.

[5] 胡学钢，董学春.基于词向量空间模型的中文文本分类方法[J].合肥工业大学学报:自然科学版，2007,30(10):1261-1264

[6] Wes Mckinney.利用 Python 进行数据分析[M].北京:机械工业出版社，2018.

[7] 王美方等.基于 TFIDF 的特征选择方法[J].计算机工程与设计,2007,28(23):5795-5796.

[8] https://www.jianshu.com/p/9fe0a7004560

[9] 邬启为.基于向量空间的文本聚类方法与实现[D].北京交通大学,2014.

[10] 张振亚,王进,,程红梅. 基于余弦相似度的文本空间索引方法研究[J]. 计算机科学, 2005, 32(9):160-163.