“智慧政务”中的文本挖掘应用

基于智慧政务中的数据挖掘文本分析

摘 要

随着社交网络、移动多媒体为代表的新兴技术的日益盛行，社会公共事务交流与传播愈发网络化，“网络问政”作为一种新型议政问政方式，使得与政务相关的以文本集为代表的非结构化数据呈现井喷式增长，为政府的管理带来极大的挑战与转型冲击。政府管理者能否及时的给予反馈回复，高效率、高准确率的处理这些信息显得尤为重要。这就与他们是否真正理解公众所表达的信息息息相关，其中关键就在于对反馈者的留言文本数据进行信息处理。本文将基于自然语言处理和文本挖掘的方法对市民的留言数据及政府部门对相关留言的评价进行潜在信息的提取、挖掘与分析。

在本次数据挖掘过程中，我们首先对获取到的数据进行分析，利用 python 进行数据预处理并实现对评论数据的优化，提升其可建模度。具体包括： 数据清洗（去重，去敏感词，特殊字符，正则表达式）、通过 jieba 库进行中文分词和词性标注、剔除停用词、利用 wordcloud 绘制词云图、文本特征选择以及文本表示。接着，采用多种方法来构建数据挖掘模型，为后面的数据分析奠定基础。为此我们使用基于 sklearn 的朴素贝叶斯分类器 MultinomialNB，实现留言内容的一级标签分类模型的构建，再利用 F1-Score 进行模型评估，并对其不断优化提高预测准确率。

再通过 K-means 聚类方法对相似留言进行相似度识别，并考虑出现频次、涉及时间、地点、范围等方面来聚类，定义合理的热度评价指标，并从中提取出排名前五的热点问题。最后，运用构造出来的多种数据挖掘模型的结果，对这些留言数据进行多方面多角度的文本分析，提取留言中隐藏的信息。对于答复意见，提取相关部门答复数据，确定一套答复的模版来保证其完整性，将留言的关键词与答复的关键词作对比，保障答复的相关性，并对答复的要点找出其理论法律依据确保可解释性，综合考虑三大方面将答复量化，构建一套完整的、可行的评价体系，并加以实现。

关键词：政务文本分析 自然语言处理 深度学习 文本聚类 朴素贝叶斯算法

Abstract

As social networks, mobile multimedia as representative's emerging technologies, the growing popularity of social public affairs exchanges and growing network of transmission, network asked administration as a new town asked administration way, make government affairs related to text sets on behalf of unstructured data presents the blowout growth, brings great challenges to the government's operation and management and transformation. Whether the government managers can give feedback and reply in time, it is particularly important to process these information efficiently and accurately. This is closely related to whether they truly understand the information expressed by the public, and the key is to process the message text data of the feedback. Based on the methods of natural language processing and text mining, this paper extracts, excavates and analyzes the potential information of the message data of citizens and the evaluation of relevant message by government departments.As social networks, mobile multimedia as representative's emerging technologies, the growing popularity of social public affairs exchanges and growing network of transmission, network asked administration as a new town asked administration way, make government affairs related to text sets on behalf of unstructured data presents the blowout growth, brings great challenges to the government's operation and management and transformation. Whether the government managers can give feedback and reply in time, it is particularly important to process these information efficiently and accurately. This is closely related to whether they truly understand the information expressed by the public, and the key is to process the message text data of the feedback. Based on the methods of natural language processing and text mining, this paper extracts, excavates and analyzes the potential information of the message data of citizens and the evaluation of relevant message by government departments.

In this data mining process, we first analyze the acquired data, and use Python to preprocess the data to optimize the comment data and improve its modelability. Specifically, it includes: data cleaning (de duplication, de sensitive words, special characters, regular expressions), Chinese word segmentation and part of speech tagging through Jieba library, removing stop words, drawing word cloud map with matploylib, text feature selection and text vectorization.Then, compare the advantages and disadvantages of various methods to build data mining model, select the best model, and lay the foundation for the later data analysis. For this reason, we use multi nomialnb, a naive Bayesian classifier based on sklearn, to build a multi classification model of the first level tags of the message content, and then use F1 score to evaluate the model, and optimize it to improve the prediction accuracy.

Then, K-means clustering method is used to analyze the similar messages, and the frequency, time, place and scope are considered to classify them. Then, a reasonable heat evaluation index is defined and the top five hot issues are extracted.

Finally, using the results of the constructed data mining model, the paper analyzes the message data from many aspects and angles, and extracts the hidden information in the message. For the reply opinions, we extract the reply data of relevant departments, determine a set of reply template to ensure its integrity, compare the key words of the

message with the key words of the reply, ensure the relevance of the reply, find out the theoretical and legal basis for the key points of the reply to ensure the interpretability, compare the speed of the reply time, consider its timeliness, and comprehensively consider the four aspects of the reply quality Quantitative evaluation is quantified, and a complete set of feasible evaluation system is constructed and realized.

Keywords: Government text analysis; natural language processing; naive Bayes; text clustering; similarity analysis

目录

1. 挖掘目标

本次建模针对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的文本评论数据，利用自然语言处理和文本多分类的方法对留言进行文本挖掘，通过对比文本相似度进行热点问题的挖掘以及运用文本挖掘方法建立关于评价答复意见的模型。

通过以上处理，对数据进行情感态度以及所隐藏信息的挖掘，来得到此次文本挖掘的目标。

2. 分析方法与过程

2.1. 总体流程

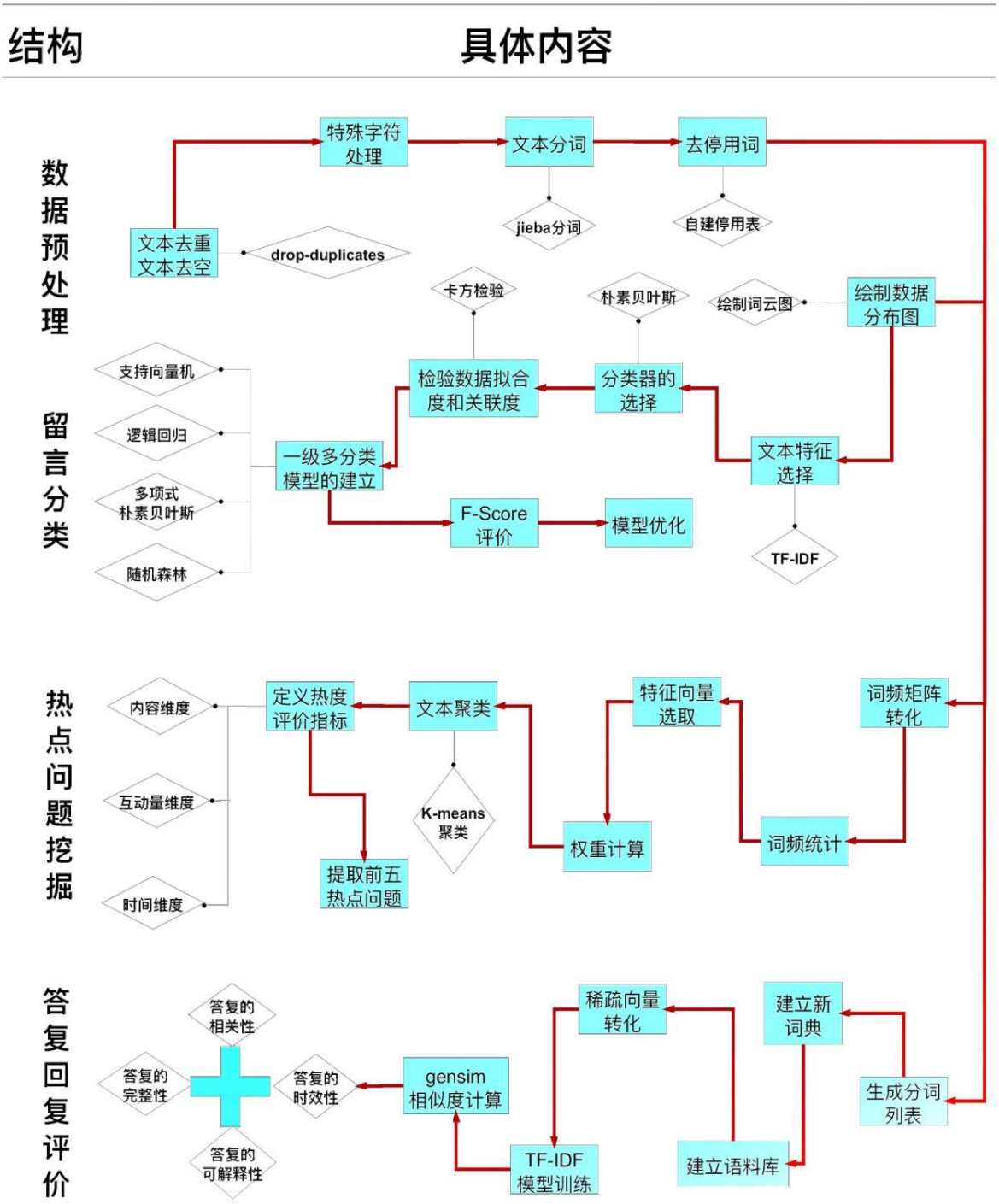

图 1 总体流程图

本论文的分析流程可大致分为以下四步：

第一步：获取分析用的原始数据，并对数据进行数据清洗、去停用词、分词等预处理操作。

第二步：基于 Sklearn 算法对留言建立一级标签多分类模型， 使用 F1-Score 对模型的好坏进行评价；

第三步：对某一时段内群众集中反映的热点问题,通过 K-Means 聚类算法将反映相同问题的留言进行归类，并定义合理的热度评价指标，按指定格式给出热点排名前 5 的相关热点问题及问题对应的留言信息；

第四步：针对相关部门对留言的答复意见的质量，采用 Gensim 相关算法从留言与回复的相关性、完整性、可解释性、时效性等多角度综合给出一套合理的评价方案。

2.2. 具体步骤

2.2.1 数据介绍

本文使用的实验数据包含三级标签分类的具体内容及涵盖城乡建设、环境保护、交通运输等多个方面的相关留言数据，以及相关部门对于居民留言问题给出的具体答复意见。经比较，可以发现数据类别比较均衡，无需再外部爬取，故所用全部数据从官方平台获取。

2.2.2 文本预处理

通过观察所给的数据，可以发现数据量比较大，且所给文件中字段大多为文本格式，需要将其转化为数值形式才能对其进行分析，除此之外，文本信息存在大量干扰信息，如果把这些数据也引进分词、词频统计、乃至文本聚类等，则必然会对结果造成很大的影响，使得结果与实际情况大相径庭，因此在使用前就必须要先进行文本预处理，把这些大量的价值含量很低甚至没有价值含量的条目去除。

运用 Python3.6 对这些文本数据的预处理主要由以下几个重要部分组成：

2.2.2.1 文本去重、去空

文本去重的基本解释、原因以及具体步骤：对于群众发表的留言意见，大多过程详细，叙述完整。而在实际分析中，有些较长的文本有大量的冗余词汇及句子，会给进一步的文本分析带来阻碍，因此精确、快速的筛选出文本留言中的关键词就显得十分重要。文本中出现的重复词汇及句子，干扰了问题的分析，采取 drop_duplicates 方法，将重复文本从整体数据中删除并将索引重置。同时检查文本数据是否存在空值，并将其去除。

2.2.2.2 文本分词

2.2.2.2.1 文本分词的基本解释与原因

中文分词是文本信息处理的基础与关键。在中文中，只有字、句和段落能够通过明显的分界符进行简单的划界，而对于“词”和 “词组”来说，它们的边界模糊，没有一个形式上的分界符。因此，进行中文文本挖掘时，首先需要文本分词，即将连续的字序列按照一定的规范重新组合成词序列的过程。分词结果的准确性对后续文本挖掘算法有着不可忽视的影响，如果分词效果不佳，即使后续算法优秀也无法实现理想的效果。例如在特征选择的过程中，不同的分词效果，将直接影响词语在文本中的重要性，从而影响特征的选择。

2.2.2.2.2 中文分词的原理及常见文本分词算法概述及缺陷

常见文本分词算法可分为三大类：基于字符串匹配的分词方法、基于理解的分词方法和基于统计的分词方法。

①基于字符串匹配的分词方法：这种方法又叫做机械分词方法，它是按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功（识别出一个词）。

②基于理解的分词方法：这种分词方法是通过让计算机模拟人对句子的理解，达到识别词的效果。其基本思想就是在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。它通常包括三个部分：分词子系统、句法语义子系统、总控部分。在总控部分的协调下，分词子系统可以获得有关词、句子等的句法和语义信息来对分词歧义进行判断，即它模拟了人对句子的理解过程。这种分词方法需要使用大量的语言知识和信息。由于汉语语言知识的笼统、复杂

性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统还处在试验阶段。

③基于统计的分词方法：给出大量已经分词的文本，利用统计机器学习模型学习词语切分的规律（称为训练），从而实现对未知文本的切分。例如最大概率分词方法和最大熵分词方法等。随着大规模语料库的建立，统计机器学习方法的研究和发展，基于统计的中文分词方法渐渐成为了主流方法。主要统计模型：N 元文法模型（N-gram），隐马尔可夫模型（Hidden Markov Model ，HMM），最大熵模型（ME），条件随机场模型（Conditional Random Fields，CRF）等。

2.2.2.2.3 中文分词选用的方法及原因

对于中文分词，选用 jieba 分词的原因在于它强大且高效的文本处理方法。jieba 分词支持三种分词模式： 精确模式、全模式、搜索引擎模式。除了上述的三种分词模式外，jieba 分词还自带有关键词摘取算法：基于 TF-IDF 算法的关键词抽取和基于 TextRank 的算法关键字抽取。

留言编号	[留言，主题]
360114	[A，市，经济，学院，体育，学院，变相，强制，实习]
289408	[在，A，市，人才，app，上，申请，购房，补贴，为什么，通不过]
336608	[希望，西地省，把，抗癌，药品，纳入，医保，范围]
360103	[A5，区，劳动，东路，魅力，之，城，小区，临街，门面，烧烤，夜宵，摊]
323149	[请，给，K3，县，乡村，医生，发，卫生室，执业，许可证]
360107	[A5，区，劳动，东路，魅力，之，城，小区，一楼，的，夜宵，摊，严重...
360108	[A5，区，劳动，东路，魅力，之，城，小区，一楼，的，夜宵，摊，严重...
343985	[A，市，能否，设立，南塘，城轨，公交站，？]
286572	[请求，A，市，地铁，2，#，线，在，梅，溪湖，CBD，处，增设，...
316619	[请问，A，市，什么，时候，能，普及，5G，网络，？]
360100	[魅力，之，城，小区，临街，门面，油烟，直排，扰民]
360101	[A5，区，劳动，东路，魅力，之，城，小区，油烟，扰民]
360110	[A，市，经济，学院，寒假，过年，期间，组织，学生，去，工厂，工作]
323034	[L，市，物业，服务，收费，标准，应，考虑，居民，的，经济，承受能力]
319659	[A，市，江山，帝景，新房，有，严重，安全隐患]
360106	[A，市，魅力，之，城，小区，底层，商铺，营业，到，凌晨，，，各种,...
313964	[12123，上，申请，驾驶证，期满，换证，，，一个，星期，了，都，无...
360105	[A5，区，魅力，之，城，小区，一楼，被，搞，成，商业，门面，，，...
337458	[能否，分层，单独，补交，超，面积，地款，？]
360104	[A，市，魅力，之，城，商铺，无，排烟，管道，，，小，区内，到处，...
360109	[万科，魅力，之，城，小区，底层，门店，深夜，经营，，，各种，噪音，扰民]
304503	[A，市，什么，时候，能，实行，独生子女，护理，假，？]
353426	[J4，县，供销，合作社，在，岗，失业，职工，追缴，社保]
321736	[A，市能，不能，提高，医疗，门诊，报销，范畴]
360112	[A，市，经济，学院，强制，学生，实习]
360102	[A5，区，劳动，东路，魅力，之，城，小区，底层，餐馆，油烟，扰民]
360113	[A，市，经济，学院，强制，学生，外出，实习]
360111	[A，市，经济，学院，组织，学生，外出，打工，合理，吗，？]
351074	[对，A，市，参保，记录，的，几点，疑问]
342119	[咨询，移动，通信，业务，问题]

图 2 jieba 分词结果

2.2.2.2.4 分词结果展示

分词结果是根据 jieba 库的分词词典进行分词，结果基本符合要求，但对于特定的文本，还需要根据情景对上述一些与预期结果有偏差的词语进行调整，通过加载自定义的分词词典，可以得到指定的词语，如对留言中的第一条数据进行处理，得到预期的词语‘A 市，经济学院，体育学院’。

表 1 加载自定义分词词典后分词结果

留言编号	留言主题
360114	[A 市，经济学院，体育学院，变相，强制，实习]

2.2.2.3 停用词过滤

上面的分词结果是尚未经过停用词过滤的结果，从中可以观察到，文本含有较多标点符号以及大量无意义的字词，若保留这些字词，则会对后面的文本分析过程造成较大影响，因此接下来将进行停用词过滤。

停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为 Stop Words（停用词）。这些停用词都是人工输入、非自动化生成的，生成后的停用词会形成一个停用词表。

停用词的两个特征为：一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。例如中文中的“的”、“了”、“地”、“啊”等，英文中的“is”、“are”、“the”、“that”等，在特征选取的过程中，停用词的介入可能会造成选出的特征几乎都是停用词，从而影响结果的分析。但是在停用词的去除中，应注意要保留其中的否定词，可以对停用词表进行人工筛选相结合的方式，再添加一些该场合经常用到的一些客套用语，对停用词进行处理。

2.2.3 留言分类

文本分类是数据挖掘领域中文本挖掘的一个重要研究方向。文本分类就是指把一组分类过的文本进行训练，对其进行分析后得出分类模式，用得出的分类模

式对待测试文本进行分类的过程。本题中需要对群众留言建立一级分类模型，以便后续分派给相应职能部门处理，提高工作效率。

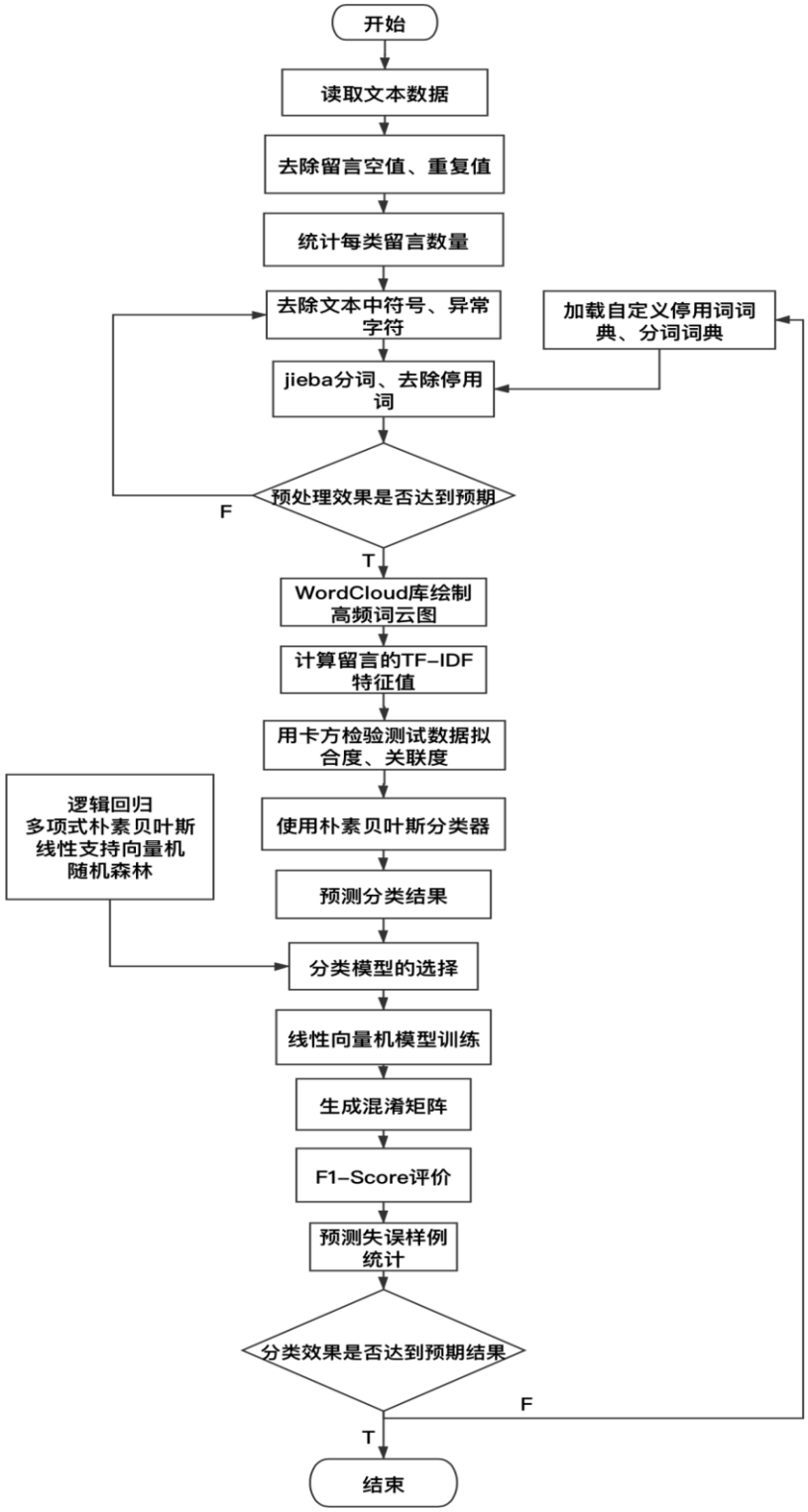

图 3 题目一流程图

2.2.3.1 一级分类模型

文本分类是个有监督的学习过程，其包括训练和分类两个过程。训练过程包括对分类过的文本进行中文分词，绘制数量分布图，特征选择，建立词频矩阵，将文本抽象成文本向量模型（VSM），最后用特定的分类算法对文本向量进行训练，提取分类信息和分类规则，建立分类模型。分类过程就是指用建立的分类模型对待测试文本进行分类。

文本分类包括以下几个关键技术: 中文分词、特征选择、向量空间模型建立、分类算法选取、分类器的性能评价。其中特征选择方法的选择和特征空间的高维性是文本分类的两个难点。

2.2.3.1.1 留言数据直观展示

为了方便模型的建立以及后续对数据的处理，需要先对数据进行形象化展示，使我们更加了解数据，便于模型的优化。具体操作如下：

2.2.3.1.2 绘制数量分布图

根据留言的一级标签，统计出每个标签中对应的文本数量，并绘制出分类数量分布图。如图 4，从图中可直观看出文本中包含的类别数量以及各类别对应的文本数量。

类别	数量
城乡建设	2009
劳动和社会保障	1969
教育文体	1589
商贸旅游	1215
环境保护	938
卫生计生	877
交通运输	613

图 4 文本类别数量图

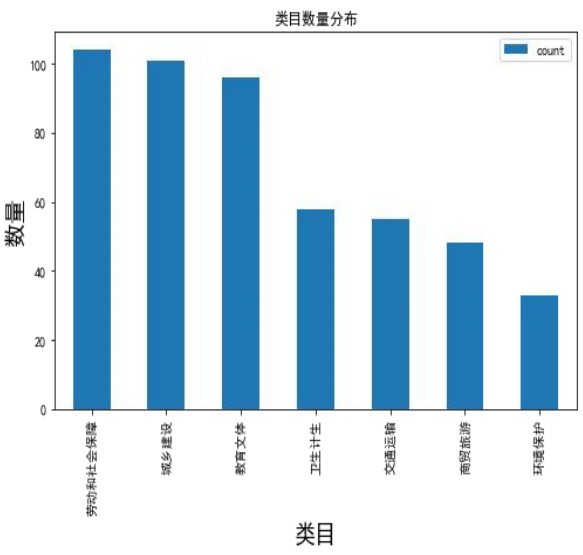

图 5 文本类别数量分布图

2.2.3.1.3 绘制词频图

“词云”就是通过形成“关键词云层”或“关键词渲染”，对留言文本中出现频率较高的“关键词”产生视觉上的突出。词云图过滤掉大量的文本信息，使查看者只要一眼扫过文本就可以领略文本的主旨。

当下，词云盛行，词云能够准确快速地筛选出重要的文本信息，把关键字以图片的形式展现出来，吸引人们的眼球，且词云的生成简单，制作工具容易获得。我们使用 wordcloud 根据词频来绘制关于各部分留言关键词的词云图，如图所示。

图 6 教育文体类高频词云图

图 7 环境保护类高频词云图

2.2.3.1.4 特征选择与权重计算

经过上述文本预处理后，虽然已经去掉部分停用词，但文本中仍包含大量词语，为文本的向量化带来困难，所以特征选择的主要目的是在不改变文本原有核心信息的情况下尽量减少要处理的词数，以此来降低文本的特征向量维数，从而简化计算，使模型泛化能力更强，同时具有区分能力的特征可以提高系统的效率与精度，所以文本分类处理中的特征选择准确与否至关重要。常用的方法有：文档频率(DF)、互信息(MI)、信息增益（IG）、X2 统计、期望交叉熵、文本证据权和优势率等。这些方法的基本思想都是对每一个特征（在此处是中文词），计算某种统计度量值，然后设定一个阈值 T，把度量值小于 T 的特征量过滤掉，剩下的即认为是有效特征。

权重计算是指经过文本特征选择后，计算并集中每个特征词在每篇文档中的权重，以形成词频矩阵，建立文本向量模型。

向量空间模型的基本思想是以向量来表示文本: $(W_1, W_2, \ldots, W_j, \ldots, W_n)$，其中 W_j 为第 j 个特征项的权重。本文采用 TF-IDF 来计算留言文本的特征值，TF-IDF（term frequency–inverse document frequency）是一种用于信息检索与数据挖掘

的常用加权技术。TF 意思是词频(Term Frequency)，IDF 意思是逆文本频率指数(Inverse Document Frequency)。

TF-IDF 是在单词计数的基础上，降低了常用高频词的权重，增加罕见词的权重。因为罕见词更能表达文本的主题思想,比如在一篇文本中出现了“中国”和“卷积神经网络” 两个词,那么后者将更能体现文本的主题思想,而前者是常见的高频词,它不能表达文本的主题思想。所以“卷积神经网络”的 TF-IDF 值要高于“中国”TF-IDF 值。这里使用 TF-IDF Vectorizer 方法来抽取文本的 TF-IDF 特征值。这里使用了参数 ngram_range=(1,2),这表示除了抽取评论中的每个词语外,还要抽取每个词相邻的词并组成一个“词语对”,如: 词 1，词 2，词 3，词 4，(词 1，词 2)，(词 2,词 3)，(词 3，词 4)。这样就扩展了我们特征集的数量,有了丰富的特征集才有可能提高分类文本的准确度。参数 norm='l2',是一种数据标准化处理的方式,可以将数据限制在一点的范围内，例如(-1,1)。如表 2，可以看到我们的 features 的维度是(9210,696045),这里的 9210 表示总共有 9210 条评价数据，696045 表示特征数量。其中包括全部评论中的所有词语数+词语对(相邻两个单词的组合)的总数。

表 2 文本特征值维度及权重表

特征值维度	特征值权重
(0,409269)	0.15103443199035319
(0,406440)	0.15103443199035319
(0,406440)	0.15103443199035319
(0,244201)	0.09658847972586007
(0,409197)	0.09580745273705077
(0,276822)	0.15103443199035319
(0,677160)	0.07257079096756588
(0,639173)	0.15103443199035319
:	:
(9209,146691)	0.12269604327605854
(9209,399273)	0.0706604018051018
(9209,249855)	0.053511789351971276

下面利用卡方检验的方法来找出每个分类中关联度最大的两个词语和两个词语对。

2.2.3.1.5 卡方检验

卡方检验是一种统计学的工具,用来检验数据的拟合度和关联度，就是统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，如果卡方值越大，二者偏差程度越大；反之，二者偏差越小；若两个值完全相等时，卡方值就为 0，表明理论值完全符合。

具体步骤：

①提出原假设：

H0：总体 X 的分布函数为 F(x).

如果总体分布为离散型，则假设具体为

H0：总体 X 的分布律为 $P\{X = x_i\} = p_i, i = 1,2,\ldots$

②将总体 X 的取值范围分成 k 个互不相交的小区间$A_1, A_2, A_3, \ldots, A_k$，如可取

$A_1 = (a_0, a_1], A_2 = (a_1, a_2], \ldots, A_k = (a_{k-1}, a_k]$其中$a_0$ 可取$-\infty$，a_k可取$+\infty$，区间的划分视具体情况而定，但要使每个小区间所含的样本值个数不小于 5，而区间个数 k 不要太大也不要太小。

③把落入第 i 个小区间的 A_i的样本值的个数记作f_i，称为组频数（真实值），所有组频数之和$f_1 + f_2 + \ldots + f_k$等于样本容量n。

④当 H_0为真时，根据所假设的总体理论分布，可算出总体 X的值落入第 i个小区间A_i的概率p_i，于是，np_i 就是落入第i 个小区间 A_i的样本值的理论频数（理论值）。

⑤当H_0为真时，n 次试验中样本值落入第i 个小区间A_i 的频率 f_i/n与概率p_i应很接近，当H_0不真时，则f_i/n 与p_i 相差很大。基于这种思想，皮尔逊引进如下检验统计量

$$x^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i}$$

在 0 假设成立的情况下服从自由度为$k - 1$ 的卡方分布。

在这里使用 sklearn 中的 chi2 方法后，我们可以看到已经找出了每个分类中关联度最强的两个词和两个词语对。这些词和词语对能很好的反映出分类的主题。

表 3 分类主题

类别 (class)	关联度最强的词 (Most correlated unigrams)	关联度最强的词语对 (Most correlatedbigrams)
交通运输	租车 .出租车	[租车 出租车].[出租 租车]
劳动和社会保障	劳动 . 社保	[养老 养老保险].[劳动 合同]
卫生计生	医生 . 医院	[生子 子女]. [独生 生子]
商贸旅游	传销 . 电梯	[传销 组织].[屠宰 屠宰场]
城乡建设	开发商 . 业主	[积金 公积金]. [开发 开发商]
教育文体	学校 . 教育	[主任 班主任]. [教育 教育局]
环境保护	污染 . 环保	[环保 环保部]. [环保 环保局]

2.2.3.1.6 使用朴素贝叶斯分类器

当有了词向量以后就可以开始训练我们的分类器。分类器训练完成后,即可对没有见过的文本进行预测。目前常用的文本分类的有:朴素贝叶斯(Naïve Bayes)、支持向量机(Supporting Vector Machine，SVM)、KNN(K-Nearest Neighbor) 、神经网络(Neural Net)、决策树(Decision Trees) 等。其中朴素贝叶斯分类算法是一种简单易行的文本分类算法，它的核心是后验概率算法的实现和对特征词进行有效的降维。

朴素贝叶斯分类器最适合用于基于词频的高维数据分类器，最典型的应用如垃圾邮件分类器，准确率可以高达 95%以上。这里使用的是 sklearn 的朴素贝叶斯分类器 MultinomialNB，首先将文本转换成词频向量,然后将词频向量再转换成 TF-IDF 向量，还有一种简化的方式是直接使用 TF-IDF Vectorizer 来生成 TF-IDF 向量(正如前面生成 features 的过程)，这里还是按照一般的方式将生成 TF-IDF 向量分成两个步骤：1、生成词频向量；2、生成 TF-IDF 向量。最后开始训练 MultinomialNB 分类器。

贝叶斯基本理论：假设有类别集合C和待分类样本空间S，其中集合C表示为$C = \{c_1, c_2, c_3, \dots, c_n\}$，样本空间S中包含任意事件 e，事件可表示为 $e = \{e_1, e_2, e_3, \dots, e_n\}$，进而对于任意事件e都有$P(e) > 0$，则有公式:

$$P(c_i|e) = \frac{P(e|c_i)P(c_i)}{\sum_{j=1}^{n} P(e|c_j)P(c_j)} (i = 1,2, \dots, n)$$

基于统计的贝叶斯理论基本步骤是

(1)求得类别集合 c 中的先验概率$P(c_i)$。

(2)根据先验概率$P(c_i)$，运用贝叶斯公式得出其后验概率

$P(c_i|e)$。

(3)根据所得后验概率，将待分类样本归为概率值最大的类别。

朴素贝叶斯分类器：假设某文本可以用一组具有代表性的词来表示，该组词可以称为特征词把特征词定义为代表文本的特征向量，用V（$v_1, v_2, v_3, \dots, v_n$）来表示。同时在已知训练文本集中有 k 个不同类别，用C（$C_1, C_2, C_3, \dots, C_k$）来表示。如果要确定具有特征向量 V 的文本所属的类别，需要求后验概率$P(c_i|v)$，进行比较后选出最大的概率，即 得出 V 所属类别。贝叶斯理论有一个重要前提，要求特征向量 n 之间是相互独立互不联系的，然而在实际中，这种情况是不存在的．在求其结果时，只需得知后验概率最大值即可确定类别，则有朴素贝叶斯分类公式：

$$max_{j=1}^{k}\big(P(C_i|V)\big) = max_{i=1}^{k}\left(P(C_i)\prod_{j=1}^{N}P\big(v_j|C_i\big)\right), (i = 1,2, \dots, n)$$

通过预测函数可看出朴素贝叶斯分类器分类效果与预期结果表 4 分类结果一致。

表 4 基于预测函数的分类结果

结果分类	留言文本分类结果预测
城乡建设	这条铁路怎么还没有修好，周围的房子都拆了。

2.2.3.1.7 模型选择

通过尝试不同的模型，并评估其准确率，将不同模型的准确率作对比，可得到适合选题的最佳模型。我们将尝试以下四种模型：Logistic Regression(逻辑回归)， Multinomial Naive Bayes(多项式朴素贝叶斯)， Linear Support Vector Machine(线性支持向量机)，Random Forest(随机森林)。

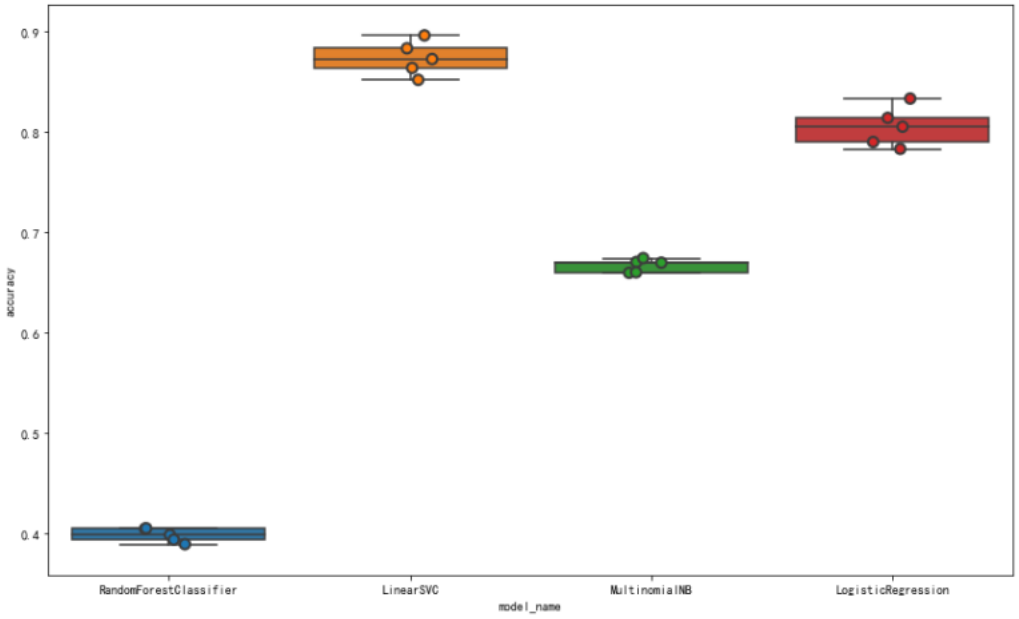

图 8 模型精度对比箱体图

从模型精度对比箱体图上可以看出随机森林分类器的准确率最低，因为随机森林属于集成分类器(由若干个子分类器组合而成)，一般来说集成分类器不适合处理高维数据(如文本数据),因为文本数据有太多的特征值,使得集成分类器难以应付,其中线性支持向量机的准确率最高。通过对文本预处理部分的优化，进一步提高了分类的准确性。通过对样例数据的测试，并绘制预测结果的混淆矩阵图，可直观看出分类的结果。

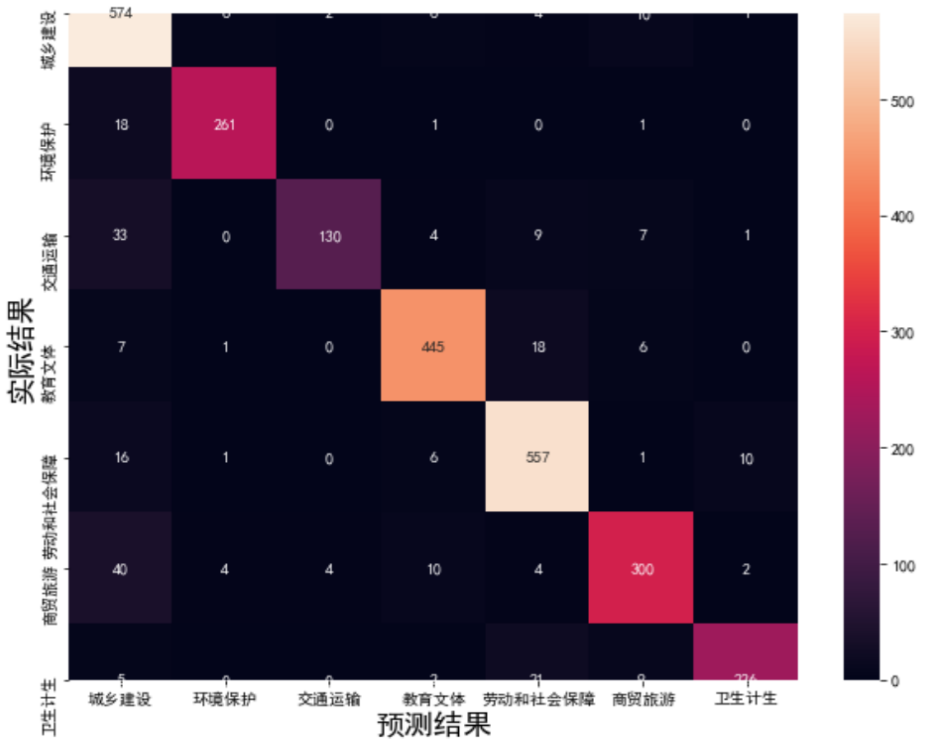

图 9 示例数据预测结果

混淆矩阵的主对角线表示预测正确的数量,除主对角线外其余都是预测错误的数量.从上面的混淆矩阵可以看出"交通运输"类预测最准确,只有一例预测错误。“城乡建设”预测的错误数量较多。

多分类模型一般不使用准确率(accuracy)来评估模型的质量,因为 accuracy 不能反应出每一个分类的准确性,当训练数据不平衡(有的类数据很多,有的类数据很少)时， accuracy 不能反映出模型的实际预测精度,因此我们通过使用 F1-Score 方法对分类结果进行评价。

2.2.3.2 使用 F1-Score 对分类方法进行评价

分类模型的评估方法 F1 分数(F1-ScoreLogistic Regression(逻辑回归))基本原理与解释：

精确率(Precision)和召回率(Recall)评估指标,理想情况下做到两个指标都高当然最好，但一般情况下，Precision 高，Recall 就低，Recall 高，Precision 就低。所以在实际中常常需要根据具体情况做出取舍，例如一般的搜索情况，在保证召回率的条件下，尽量提升精确率。而像癌症检测、地震检测、金融欺诈等，则在保证精确率的条件下，尽量提升召回率。

引出了一个新的指标 F1-score,综合考虑 Precision 和 Recall 的调和值

$$\mathrm{F1}-\mathrm{Score}=(1+\beta^2)\cdot\frac{Precision\cdot Recall}{\beta^2\cdot Precision+Recall}$$

当β=1 时，称为 F1-score 或者 F1-Measure，这时，精确率和召回率都很重要，权重相同。

当有些情况下，我们认为精确率更重要些，那就调整β的值小于 1，

如果我们认为召回率更重要些，那就调整β的值大于 1。

F1 指标(F1-score): F1-score 表示的是 precision 和 recall 的调和平均评估指标。

$$\mathrm{F1}=2\cdot\frac{Precision\cdot Recall}{Precision+Recall}$$

通过 F1—Score 对模型评价的结果如下表：

表 5　F1-Score 评价结果

	精确度	召回率	F1-Score	测试数量
城乡建设	0.83	0.94	0.88	603
环境保护	0.94	0.92	0.93	281
交通运输	0.92	0.73	0.81	184
教育文体	0.95	0.95	0.95	477
劳动和社会保障	0.94	0.94	0.94	591
商贸旅游	0.88	0.84	0.86	364
卫生计生	0.92	0.89	0.90	263
准确率			0.91	2763

2.2.4 热点问题挖掘

针对问题二热点问题挖掘，需要将反应特定地点或特定人群问题的留言进行归类。所谓留言归类实际上就是文本聚类，即将无类别标记的文本信息根据不同的特征，将有着各自特征的文本进行分类，使用相似度计算将具有相同属性或者相似属性的文本聚类在一起。这样就可以通过文本聚类的方法根据不同的留言内容，对相似的留言归类后再进行分类。通过 K-Means 聚类方法，可以得到相似问题的集合，并从中统计出热点问题将之汇总。具体方法如下：

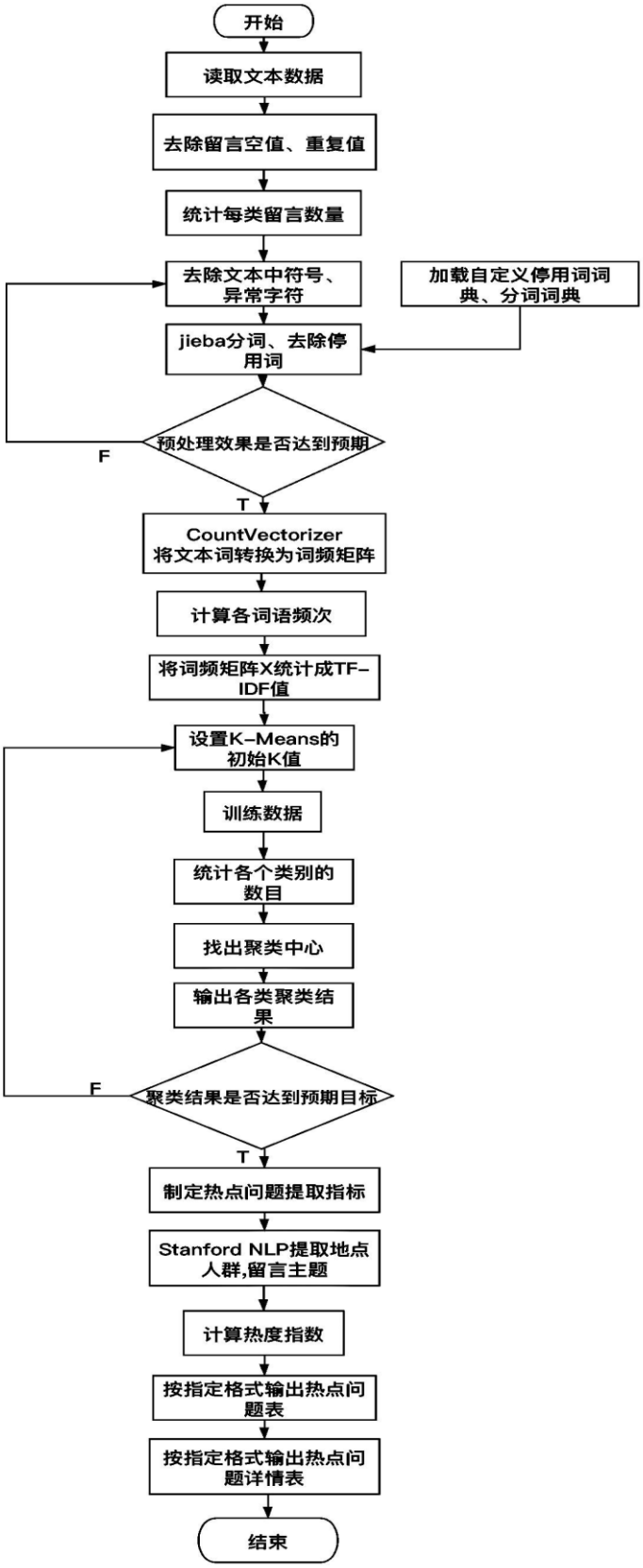
开始
读取文本数据
去除留言空值、重复值
统计每类留言数量
去除文本中符号、异常字符
加载自定义停用词词典、分词词典
jieba分词、去除停用词
预处理效果是否达到预期
F
T
CountVectorizer将文本词转换为词频矩阵
计算各词语频次
将词频矩阵X统计成TF-IDF值
设置K-Means的初始K值
训练数据
统计各个类别的数目
找出聚类中心
输出各类聚类结果
聚类结果是否达到预期目标
F
T
制定热点问题提取指标
Stanford NLP提取地点人群,留言主题
计算热度指数
按指定格式输出热点问题表
按指定格式输出热点问题详情表
结束

图 10 题目二流程图

2.2.4.1 文本相似度计算

相似度是用来衡量文本间相似程度的一个标准。在文本聚类中，需要研究文本个体间的差异大小，也就是需要对文本信息进行相似度计算，根据相似特性将信息进行聚类。目前相似度计算方法分为距离度量和相似度度量。本文采用的是基于距离度量的欧几里德距离计算留言文本见差异。当然这里也可以扩展用其他距离进行计算，如曼哈顿距离，标准欧氏距离等。传递距离原理对于相似性度量具有较高的扩展性。

欧氏距离定义：令 i = （x_1，x_2，…，x_p）和j = （y_1，y_2，…，y_p）是两个被 p 个数值属性标记的对象，则对象 i 和 j 之间的欧式距离为：

$$\mathrm{dis}(i,j) = \sqrt{\sum_{k=1}^{p}(x_k - y_k)^2}$$

2.2.4.2 文本聚类

文本聚类实际上是将一组没有给定的任何标签、标记的样本划分至多组或者类别的过程。它是研究随机样本分类问题的一种多元统计方法，包括谱系聚类、K-means 聚类及模糊减法聚类等，各聚类分析的思想共同点是将样本按相似度大小逐一归类,关系密切的聚集到较小一类,关系疏远的聚集到较大一类,直到所有样本分类完毕。

对于给定的数据样本,谱系聚类定能将样本分为一类，分类结果唯一，但最大缺点是样本一旦被分到某类就不能再改变,且样本量较大时，计算较较慢； K-means 聚类分析法能克服以上缺点,但运用其进行分析前，需对样本分类数 K 进行初步预估,然后在 K 类中便有 K 个凝聚点,所以需寻找凝聚中心，所以本题采用的 K-means 聚类分析法。

（1）聚类原理介绍

K-Means 聚类算法也称为 K-平均或 K-均值，是一种非监督学习的硬聚类算法，是典型的基于原型的目标函数聚类方法的代表，且是一种使用最广泛的聚类算法。它是以数据点到原型的某种距离作为优化的目标函数,利用函数求极值的方法得到迭代运算的调整规则,主要采用误差平方和准则函数作为聚类准则函数,以欧式距离作为相似度测度。算法的主要思想是通过迭代过程把数据集划分为不同

类别，使评价聚类性能的准则函数达到最优，从而使生成的每个聚类类内紧凑，类间独立。具有计算速度快。操作简单，时间复杂度近似线性的特点,适合挖掘大规模数据集，且对大数据集分析有较高效率以及可伸缩性。

（2）K - means 聚类分析步骤

设 n 个样本的 P 元观测指标数据矩阵为:

$$\mathrm{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

（1）式中，X_{ij}表示第 i 个样本的第 j 项指标($\mathrm{i}=1,2,\ldots,\mathrm{n};\ \mathrm{j}=1,2,\ldots,\mathrm{p}$) . 任意两个数据样本间距离采用欧氏距离，其表达式为:

$$\mathrm{d}(x_i,x_j)=[\sum_{k=1}^{p}(x_{ik}-x_{jk})^2]^{1/2}$$

（2）假设样本中 k 个凝聚点的集合是:

$$\mathrm{L}^{(0)}=\{x_1^0,x_2^0,\ldots,x_k^0\}$$

并引入

$$G_i^0=\left\{x: d\left(x,x_i^{(0)}\right)\le d\left(x,x_j^{(0)}\right)\right\}, i,j=1,2,\ldots,k, i\ne j$$

于是将样品分为不相交的 K 类:

$$\mathrm{G}^{(0)}=\{G_1^0,G_2^0,\ldots,G_k^0\}$$

现在从 G_0开始，计算新的聚点集合:

$$L^1x_i^{(1)}=1/\mathrm{n\_i}\sum_{x^i\in G_i^0}x_i, i=1,2,\ldots,k$$

其中，n_i是 $\mathrm{G}_{(0)}$样品中的样品个数，得到一个新的集合:

$$\mathrm{L}^{(1)}=\{x_1^1,x_2^1,\ldots,x_k^1\}$$

现又从 L 开始分类，将样本记为新的分类有:

$$G_i^{(1)}=\left\{x: d\left(x,x_i^{(1)}\right)\le d\left(x,x_j^{(1)}\right)\right\}$$

取值同上，得到新的分类: $G_i^{(1)} = \{G_1^{(1)}, G_2^{(1)}, \ldots, G_k^{(1)}\}$. 依次重复计算下去，上述步骤重复 m 次，可得 $G^{(m)} = \{G_1^{(m)}, G_2^{(m)}, \ldots, G_k^{(m)}\}$

其中，x_{mi} 是 G_{m-1i} 类的重心，当 m 逐渐增大时，分类也会逐渐趋于稳定，同时 x_{mi} 也可近似看为 G_{mi} 的重心，此时可得: $x_{mi} \approx x_{m+1i}$，$G_{mi} \approx G_{m+1i}$，计算结束. 在实际计算过程中对于一个样本进行 m 次的分类，当满足

$$G_i^{(m)} = \left\{G_1^{(m)}, G_2^{(m)}, \ldots, G_k^{(m)}\right\} = \{G_1^{(m+1)}, G_2^{(m+1)}, \ldots, G_k^{(m+1)}\}$$

时，计算可以结束，分类完成。

表 5 K-Means 聚类结果表

	留言一	留言二	留言三	留言四	留言五
簇 1	-5.421011e-20	0.000000e+00	1.355253e-20	-1.355253e-20	1.355253e-20
簇 2	-8.131516e-20	6.776264e-20	5.421011e-20	6.776264e-20	5.421011e-20
簇 3	-8.131516e-20	6.776264e-20	5.421011e-20	2.710505e-20	5.421011e-20
簇 4	-8.131516e-20	6.776264e-20	5.421011e-20	6.776264e-20	5.421011e-20
簇 5	-5.421011e-20	6.776264e-20	6.776264e-20	-2.710505e-20	6.776264e-20
簇 6	-8.131516e-20	5.421011e-20	-1.355253e-20	0.000000e+00	-1.355253e-20
簇 7	-5.421011e-20	2.710505e-20	-1.355253e-20	-2.710505e-20	-1.355253e-20
簇 8	-8.131516e-20	-2.710505e-20	1.084202e-19	-1.355253e-20	1.084202e-19
簇 9	4.065758e-19	4.595475e-04	4.187239e-04	5.014435e-19	4.187239e-04
簇 10	-8.131516e-20	8.131516e-20	9.486769e-20	-8.131516e-20	9.486769e-20
簇 11	-8.131516e-20	6.776264e-20	5.421011e-20	4.065758e-20	5.421011e-20
簇 12	-5.421011e-20	-1.897354e-19	1.084202e-19	1.219727e-19	1.084202e-19
簇 13	-5.421011e-20	6.776264e-20	8.131516e-20	-4.065758e-20	8.131516e-20
簇 14	-5.421011e-20	6.776264e-20	8.131516e-20	-5.421011e-20	8.131516e-20
簇 15	-5.421011e-20	5.421011e-20	2.710505e-20	4.065758e-20	2.710505e-20
簇 16	-8.131516e-20	6.776264e-20	4.065758e-20	5.421011e-20	4.065758e-20
簇 17	-8.131516e-20	5.421011e-20	-1.355253e-20	0.000000e+00	-1.355253e-20
簇 18	-5.421011e-20	6.776264e-20	8.131516e-20	-6.776264e-20	8.131516e-20
簇 19	-8.131516e-20	-6.776264e-20	1.084202e-19	2.710505e-20	1.084202e-19
簇 20	-8.131516e-20	5.421011e-20	0.000000e+00	1.355253e-20	0.000000e+00
簇 21	-8.131516e-20	6.776264e-20	5.421011e-20	4.065758e-20	5.421011e-20

（2）K-Means 聚类结果

K-Means 算法要求在计算之前给定 k 值。根据对题二所给的文本数据进行分析可得知，文本中存在大量属于不同类别的留言数据。为了减少表达相似主题但实际并不为同一问题的留言数据对聚类结果的干扰，在对数据初步处理后，将整个文本数据分为 40 簇，并以此作为 k 的值，令 k=40，即总计分为 40 个类别，分类结果汇总后如图所示

第3类:
投诉 滨河 滨河苑 广铁 职工 购房 霸王
景园 伊景园 滨河 滨河苑 捆绑 销售 车位 维权 投诉
A市 景园 伊景园 滨河 滨河苑 协商 购房 购买 车位
车位 捆绑 违规 销售
广铁 集团 广铁集团 铁路 职工 铁路职工 定向 商品 商品房 景园 伊景园 滨河 滨河苑 项目
投诉 A市 景园 伊景园 滨河 滨河苑 捆绑 车位 销售
A市 武广 新城 违法 捆绑 销售 车位 投诉
景园 伊景园 滨河 滨河苑 捆绑 车位 销售 合法
反对 景园 伊景园 滨河 滨河苑 强制 捆绑 销售 车位
景园 伊景园 滨河 滨河苑 强行 捆绑 车位 销售 业主
景园 伊景园 滨河 滨河苑 项目 绑定 车位 出售 合法 合规
投诉 A市 景园 伊景园 滨河 滨河苑 定向 限价 商品 商品房 违规 涨价
房 景园 伊景园 滨河 滨河苑 销售 若干 投诉
A市 景园 伊景园 滨河 滨河苑 捆绑 销售 车位
A市 景园 伊景园 滨河 滨河苑 欺诈 消费 消费者
投诉 武广 新城 景园 伊景园 滨河 滨河苑 广铁 集团 广铁集团 定向 商品 商品房
A市 景园 伊景园 滨河 滨河苑 定向 限价 商品 商品房 项目 违规 捆绑 销售 车位
投诉 A市 景园 伊景园 滨河 滨河苑 捆绑 销售 车位
景园 伊景园 滨河 滨河苑 车位 捆绑 销售 广铁 集团 广铁集团 做
无视 消费 消费者 权益 A市 景园 伊景园 滨河 滨河苑 车位 捆绑 销售
和谐 社会 和谐社会 背景 A市 景园 伊景园 滨河 滨河苑 车位 捆绑 销售
维护 铁路 职工 铁路职工 权益 取消 景园 伊景园 滨河 滨河苑 捆绑 销售 车位
景园 伊景园 滨河 滨河苑 捆绑 销售 车位 投诉
景园 伊景园 滨河 滨河苑 开发 开发商 强买强卖
举报 广铁 集团 广铁集团 景园 伊景园 滨河 滨河苑 项目 非法 绑定 车位 出售
A市 景园 伊景园 滨河 滨河苑 诈骗 钱财
投诉 景园 伊景园 滨河 滨河苑 项目 违法 捆绑 车位 销售
违反 自由 买卖 A市 景园 伊景园 滨河 滨河苑 车位 捆绑 销售
投诉 景园 伊景园 滨河 滨河苑 捆绑 销售 车位
A市 景园 伊景园 滨河 滨河苑 欺压 百姓 欺压百姓
投诉 A市 景园 伊景园 滨河 滨河苑 开发 开发商 违法 捆绑 销售 产权 车位
惊 A市 景园 伊景园 滨河 滨河苑 商品 商品房 捆绑 销售 车位
A市 景园 伊景园 滨河 滨河苑 坑害 购房 购房者

图 12　聚类结果由聚类结果分析得到具体热点问题及问题详情。

	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
83	1	190213	A0003161	请A市加快	2019/1/1(	地处时代	0	0
151	1	191872	A0003161	请A市加快	2019/3/1	地处中部	2	9
219	1	193514	A0003161	请加快A市	2019/3/2(	地处月亮	0	4
301	1	195905	A0003161	A市加快打	2019/7/2	在国家大	0	0
343	1	196908	A0003161	请加快A市	2019/1/4	地处银杉	0	9
485	1	200013	A0003161	请加快A市	2019/2/2	地处师大	0	1
495	1	200220	A0005160	请加快对	2019/7/1	胡书记：	1	0
517	1	200611	A0003161	请A市加快	2019/2/1	A市在加快	0	3
644	1	203779	A0003161	加快A市经	2019/12/	地处西地	0	2
754	1	206587	A0003161	请加快省	2019/2/1	在建设文	0	0
869	1	209202	A0003161	希望加快	2019/12/	地处长株	0	0
905	1	210000	A0003161	请加快创	2019/1/1(	地处时代	0	0
937	1	210896	A0003161	请加快A市	2019/1/2	地处月亮	0	1
943	1	210961	A0003161	请A市加快	2019/3/1	在国家大	0	0
1159	1	215805	A0003161	请加快体	2019/2/1	规划多年	0	0
1176	1	216151	A0003161	请A市加快	2019/2/2	在国家大	0	1
1292	1	218732	A0003161	请A市加快	2019/6/5	地处中部	0	0
1413	1	221652	A0003161	请加快A市	2019/3/1	A市就是L	0	1
1560	1	225471	A0003161	请A市加快	2019/1/3(	A市每到	0	0
1566	1	225574	A0003161	请加快A(	2019/1/2	地处月亮	0	2

图 13 问题聚类详情

2.2.4.3 热度评价指标

经过 K-Means 聚类方法对文本数据处理后，可得到反映群众较多的热点问题。我们使用自定义的热度评价指标对上述数据进行热点问题筛选，得到具体的热度排名前五的热点问题。

综合分析各问题的具体内容可得知，在内容维度方面，当问题讨论的用户量越多，表明该问题受影响的群众基数越大，待解决程度加深；从互动量维度来看，所反映问题得到群众支持或反对数越大，表明问题被关注越多，需要得到进一步解决；从时间维度讨论，被讨论问题时间跨度长、与当前时间接近的问题更需得到解决。由此，我们将从上述三个维度展开，定义合理的热度评价指标，提取热点问题。

1．内容维度

同一类问题中留言频数作为热点问题热度指数的一项重要判断指标。留言数越多，说明该问题涉及范围越广，涉及人群越大，受关注度越高，越需要相关部门给予反馈，及时解决。

具体计算方法如下：将每一类问题的留言数据 n 进行统计后，除以总的留言数据 N，以 Score_content 为内容维度结果值，即：

$$\text{Score\_content} = n/N$$

2．互动量维度

在对留言定义合理的热度评价指标中，我们选用留言的点赞数与反对数作为评价指标之一。对于留言点赞数与反对数，我们参考 Reddit 基于用户投票的排名算法进行处理，具体定义如下：

（1）点赞数与反对数的差 x

x = 点赞数 - 反对数

（2）留言的受肯定(否定)的程度 z

$$z = \begin{cases} |x|, if\ x \neq 0 \\ 1, if\ x = 0 \end{cases}$$

z 表示点赞数与反对数之间差额的绝对值。如果对某个留言的评价，越是一边倒，z 就越大，那么该留言所受群众关注度就越高。如果点赞数等于反对数，z 就等于 1，以Score_inter作为结果值。

Reddit 的最终得分计算公式如下：

$$\text{Score\_interact} = \log_{10} z$$

$\log_{10} z$这个部分表示，点赞数与反对数的差额 z 越大，得分越大。

需要注意的是，这里用的是以 10 为底的对数，意味着 z=10 可以得到 1 分，z=100 可以得到 2 分。也就是说，前 10 个评价与后 90 个(乃至再后面 900 个)的评价权重是一样的，即如果一个留言特别受到关注，那么越到后面进行点赞，对得分越不会产生影响。

当点赞数等于反对数，z=1，因此这个部分等于 0，也就是不产生得分。

3．时间维度

由于我们的定位是针对群众留言问题的热度分析，所以选取了问题的反映时间跨度和最后一条留言距离当下时间的多少作为时间维度。具体意义上表现在：一方面是时间跨度较大的问题，实际上说明该问题存在时间已久，但仍未解决，需要引起相关部门注意，所以可以获得较高的热度，促使有关部门尽快解决；另一方面是距离当下时间很短的问题，实际上考虑了问题的时效性，越是最新发生的问题，那么相关部门就可以给予较高关注度，结合当下情形尽快解决问题，不仅提高了工作效率，还可以回馈公民一个满意的答复。

在这里采用计算公式：$\text{Score\_time} = （T_f - T_l）/360 + 30/（T_n - T_l）$。其中$T_f$是该类问题的第一条留言，$T_l$即该类问题的最近一条留言的时间，$T_n$是当前的时间，时间跨度和时效性的单位都是天。考虑到数据中时间跨度大的特点，时间跨度要除以一年 360 天，而时效性的距离时间越小，则应给予较高重视，并且距离时长大概在一年以内，大多为几个月，所以用月数 30 除以距离时长。如此便可度量问题的时间维度热度。

总的热度计算公式为：$\text{Score} = \text{Score\_content} + \text{Score\_inter} + \text{Score\_time}$

2.2.5 答复意见评价

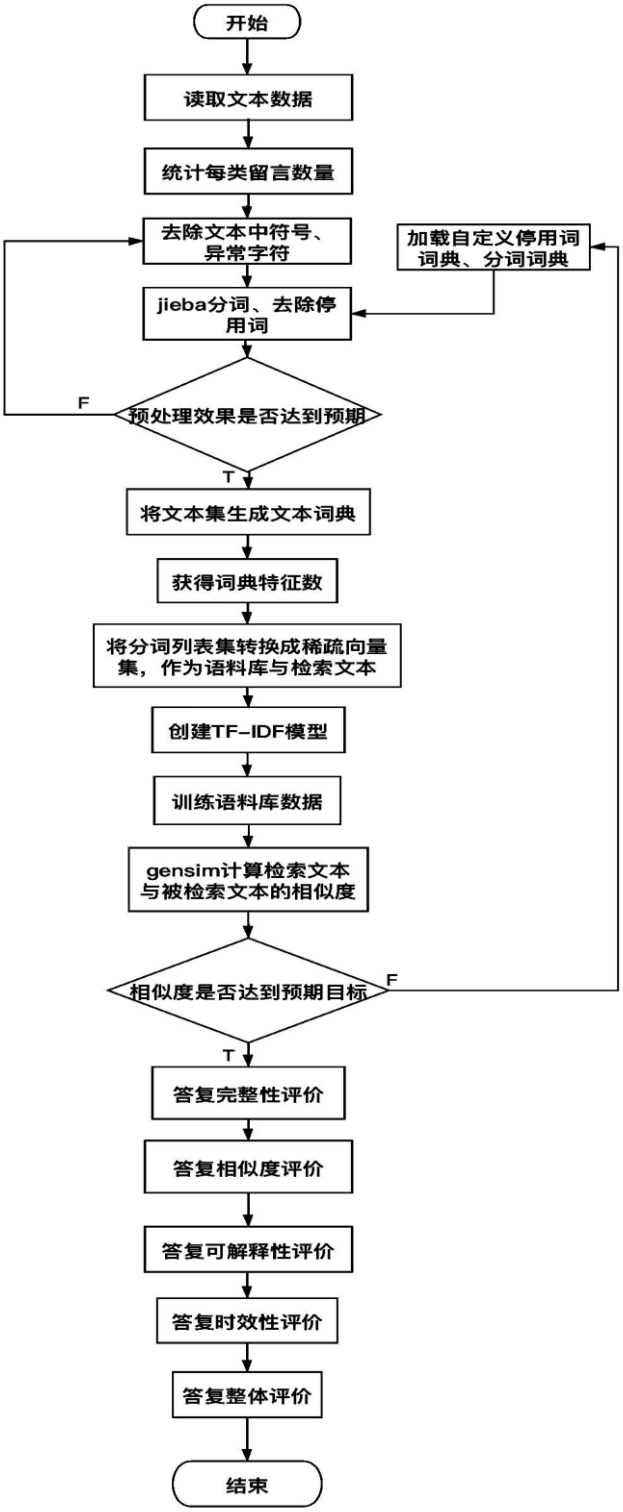

图 13　问题三流程图

1、从答复的相关性评价

在文本处理中，比如相关政府部门答复意见文本挖掘，有时需要了解政府每条答复意见和留言问题的描述之间相似度，以此衡量答复意见的质量。答复意见和问题描述的相似度越高，说明留言的答复联系到了问题的实际，正面或侧面回答了市民的问题，那么这答复意见的质量就可以得到肯定。

实现相似度的测量主要用到的是 python 的程序包：gensim，接下来主要说一下针对留言和答复意见之间的相似度，怎么使用 gensim 来计算。

（1）基本原理

①jieba 分词过滤无用词

对文本集中的文本进行中文分词，并返回分词列表

②基于文本建立词典，获取特征数

将用户留言在已分好词的基础上，建立文本词典，并获取其中所包含的特征数。

将与用户留言相对应的回复作为关键词文本，建立关键词词典，获得其词典特征数。

③稀疏向量的转化,基于稀疏向量集建立语料库

将文本词典中所有词语取集合，并为其中每个词语分配一个号码,将其转换成稀疏向量，作为语料库。并将关键词词典转化为稀疏向量矩阵,作为检索文本。

（2）处理关键词文本

根据建立好的关键词文本词典，计算其中每个词的 TF-IDF 值。

其主要思想是：如果某个词或短语在一篇文章中出现的频率高，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。

第一步：统计文本总词数M，统计第一个留言文本的词数 N，计算留言文本中的第一个词在该文本中出现的次数 n，再找出该词在所有文档中出现的次数 m。则该词的 TF-IDF 值为：$\frac{n}{N} * 1/\frac{m}{M}$，（还有其它的归一化公式，这里是最基本最直观的公式）

第二步：重复第一步，计算出单个留言文本中所有词的 TF-IDF 值。

第三步：重复第二步，计算出所有留言文本中每个词的 tf-idf 值。

（3）相似度的计算

将得到的语料库作为被检索文本，与检索文本的 TF-IDF 模型的稀疏向量集进行计算，将其中具有相同号码的词文本做乘积运算，将每个词文本的向量值做和运算得到最终的留言与评价的文本相似度。

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	答复相关性质量
2549	A00045581	A2区景荟华苑物业管理有问题	2019/4/25 9:32:09	\n\t\t\t\t\t\n\t\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景荟花苑物业管理有问题”的调查核...	2019/5/10 14:56:53	0.689939
2554	A00023583	A3区潇楚南路洋湖段怎么还没修好?	2019/4/24 16:03:40	\n\t\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	2019/5/9 9:49:10	0.665252
2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	\n\t\t\t\t\t\n\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来...	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14	0.726757
2557	A000110735	在A市买公寓能享受人才新政购房补贴吗?	2019/4/24 15:07:30	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的书记：您好！我研究生毕业后根据人才新政...	网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	2019/5/9 9:49:42	0.597687
2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	\n\t\t\t\t\t\n\t\t\t\t\t建议将“白竹坡路口”更名为“马坡岭小学”，原...	网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡...	2019/5/9 9:51:30	0.666723
...	...	...	...	...	...	...	...
181267	UU008766	汽车北站进站口附近居民强烈反对建设I市平康肾病医院!	2018/12/12 15:20:46	\n\t\t\t\t\t\n\t\t\t\t\t我们是I市汽车北站进站口的周围居民。在这里的...	您的留言已收悉。关于您反映的问题，已转I1区委、区人民政府调查处理。	2019/1/8 16:54:53	0.570991

图 14 答复相关性计算结果

2、从答复的完整性评价

因为考虑到每一条答复意见是从政府的角度出发的，那么答复意见的规范性，完整性就值得考量。在相关政府部门答复意见的质量评价中，首先考虑的就是政府每条答复意见的完整性，也就是是否有一个比较合理完整的答复框架，以此衡量答复意见的质量。答复意见的完整性越高，说明留言的答复意见越符合规范。

完整性主要由以下四个部分组成，首先是明确答复意见的对象，表明你是对哪个人做的答复，其次是对答复意见的对象做一个肯定的回应，表明你已经了解到了该条留言所反映的问题，再者就是对该问题的具体答复与解决方案，具体可见相关性，紧接着的是感谢留言用户对相关部门工作的监督与建议，欢迎他们继续关注和提议，最后便是具体答复日期的书写。具体实施方法如下：

（1）稀疏向量的转化,基于稀疏向量集建立语料库

将经过分词处理得到的留言评价文本转化为稀疏向量，作为检索文本。将判断评论完整性所需的相关文本词转换为稀疏矩阵，作为语料库。

（2）用语料库训练 TF-IDF 模型

统计语料库总词数M，统计第一个文本的词数 N，计算文本中的第一个词在该文本中出现的次数 n，再找出该词在所有文档中出现的次数 m。则该词的 TF-IDF 值为：$\frac{n}{N} * 1/\frac{m}{M}$，综合计算得到语料库中每个词的 TF-IDF 值。

（3）相似度的计算

将得到的语料库作为被检索文本，与检索文本的 TF-IDF 模型的稀疏向量集进行计算，将其中具有相同号码的词文本做乘积运算，将每个词文本的向量值做和运算得到检索文本与被检索文本的文本相似度。

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	答复完整性质量
2549	A00045581	A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	\n\t\t\t\t\t\n\t\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核...	2019/5/10 14:56:53	0.528346
2554	A00023583	A3区潇楚南路洋湖段怎么还没修好?	2019/4/24 16:03:40	\n\t\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	2019/5/9 9:49:10	0.389813
2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	\n\t\t\t\t\t\n\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来...	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14	0.427121
2557	A000110735	在A市买公寓能享受人才新政购房补贴吗?	2019/4/24 15:07:30	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的书记：您好！我研究生毕业后根据人才新政...	网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	2019/5/9 9:49:42	0.442259
2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	\n\t\t\t\t\t\n\t\t\t\t\t建议将“白竹坡路口”更名为“马坡岭小学”，原...	网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡...	2019/5/9 9:51:30	0.496919
...	...	...	...	...	...	...	...
181267	UU008766	汽车北站进站口附近居民强烈反对建设l市平康肾病医院!	2018/12/12 15:20:46	\n\t\t\t\t\t\n\t\t\t\t\t我们是l市汽车北站进站口的周围居民。在这里的...	您的留言已收悉。关于您反映的问题，已转l1区委、区人民政府调查处理。	2019/1/8 16:54:53	0.720082

图 15 答复完整性计算结果

3、从答复的可解释性评价

可解释性主要由以下三个部分组分组成，首先是对答复意见进行整体分析，提取出各个答复中所涉及的相关法律、法规、事实依据作为可解释性的被检索文本，依次将各答复文本与被检索文本进行相似度的计算。具体实施方法如下：

（1）基于稀疏向量集建立语料库

将经过分词处理得到的留言评价文本转化为稀疏向量，作为检索文本。将判断评论可解释性性所需的相关文本词转换为稀疏矩阵，作为语料库。

（2）训练 TF-IDF 模型

统计语料库总词数 M，统计第一个回复文本的词数 N，计算文本中的第一个词在该文本中出现的次数 n，再找出该词在所有文档中出现的次数 m。则该词的 TF-IDF 值为：$\frac{n}{N} * 1/\frac{m}{M}$，综合计算得到语料库中每个词的 TF-IDF 值。

（3）相似度的计算

将得到的语料库作为被检索文本，与检索文本的 TF-IDF 模型的稀疏向量集进行计算，将其中具有相同号码的词文本做乘积运算，将每个词文本的向量值做和运算得到检索文本与被检索文本的文本相似度

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	答复可解释性质量
2549	A00045581	A2区景蓉华苑物业管理有问题	2019/4/25 9:32:09	\n\t\t\t\t\t\n\t\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核...	2019/5/10 14:56:53	1.0
2554	A00023583	A3区潇楚南路洋湖段怎么还没修好?	2019/4/24 16:03:40	\n\t\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	2019/5/9 9:49:10	1.0
2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	\n\t\t\t\t\t\n\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来...	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14	1.0
2557	A000110735	在A市买公寓能享受人才新政购房补贴吗?	2019/4/24 15:07:30	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的书记：您好！我研究生毕业后根据人才新政...	网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	2019/5/9 9:49:42	1.0
2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	\n\t\t\t\t\t\n\t\t\t\t\t建议将“白竹坡路口”更名为“马坡岭小学”，原...	网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡...	2019/5/9 9:51:30	0.0
...	...	...	...	...	...	...	...
181267	UU008766	汽车北站进站口附近居民强烈反对建设l市平康肾病医院!	2018/12/12 15:20:46	\n\t\t\t\t\t\n\t\t\t\t\t我们是l市汽车北站进站口的周围居民。在这里的...	您的留言已收悉。关于您反映的问题，已转l1区委、区人民政府调查处理。	2019/1/8 16:54:53	1.0

图 16 答复可解释性计算结果

4、从答复的时效性评价

政府部门对公众留言及时回应，有助于提升政府形象和公信力，通过对附件 4 所给文本数据进行分析，利用答复时间和留言时间两列作为对答复及时性分析的主要数据。进行留言数据的时效性评价。

通过计算答复时间与留言时间的相隔大小并进行比较，找出相隔时间较短的答复，即具有答复的时效性。

利用 datetime 模块计算两个时间差，python 中通过 datetime 模块可以很方便的计算两个时间的差, datetime 的时间差单位可以是天、小时、秒, 甚至是微秒。

评价标准：每条留言的最终比值=最小时间差/每条留言的时间差

2.3 结果分析

2.3.1 问题一结果分析

1．对留言内容应用一级标签分类模型结果展示：

通过抽取部分数据对分类模型进行测试，可得到如下图所示结果：

表 6 留言文本分类结果

结果分类	留言文本分类结果预测
城乡建设	这条铁路怎么还没有修好，周围的房子都拆了。(自定义文本)
教育文体	学校怎么还不开学，都快到放暑假的时间了。（自定义文本）
卫生计生	请求严惩西地省儿童医院不负责的医护人员。(留言主题)
劳动和社会保障	投诉 C 市威胜电气有限公司强迫员工无偿加班。（留言主题）
环境保护	E8 县环保局、又兰镇政府与又兰坳子上砖厂勾结。（留言主题）

从结果可看出该文本分类模型通过训练后，对于文本分类有着较好的效果。无论是自定义文本、留言主题或是具体的留言详情，该模型都能够精确的依据文本特征词将其划分到正确的文本类别中。

2．对模型使用 F1—Score 进行评价：

表 7　F1—Score 评价结果

	精确度	召回率	F1-Score	测试数量
城乡建设	0.83	0.94	0.88	603
环境保护	0.94	0.92	0.93	281
交通运输	0.92	0.73	0.81	184
教育文体	0.95	0.95	0.95	477
劳动和社会保障	0.94	0.94	0.94	591
商贸旅游	0.88	0.84	0.86	364
卫生计生	0.92	0.89	0.90	263
accuracy			0.91	2763
macro avg	0.91	0.89	0.90	2763
weighted avg	0.91	0.91	0.91	2763

为了进一步确定模型的精确度，通过 F1-Score 方法对模型进行整体评价，从测试数据的评价结果中可看出，一级标签的七项子类的平均准确率为 91%，教育文体类准确率最高，可达 95%。

通过上述对留言数据进行处理,可将居民发布的问题反馈信息按照不同的留言内容分类。按照划分的交通运输、环境保护、劳动与社会保障等日常生活类别，可将归于不同类别后的数据统一分发给相关部门进行处理，这将大幅度的提升行政机构的服务效率，能够及时、有效的通过网络平台处理公众请求。

2.3.2 问题二结果分析

1、留言聚类结果展示：

留言数据经过聚类模型处理后可得到如图 17 所示结果：

第20类:
A市 万家 丽 南路 新城 丽发新城 居民 居民区 搅拌 搅拌站 扰民
投诉 A2区 新城 丽发新城 建 搅拌 搅拌站 噪音 扰民
新城 丽发新城 小区 旁边 建 搅拌 搅拌站
A市 新城 丽发新城 违建 搅拌 搅拌站 施工 扰民 污染 环境 污染环境
A市 丽发 小区 建 搅拌 搅拌站 噪音 污染
A市 A2区 新城 丽发新城 道路 坑坑 坑洼 坑坑洼洼
A2区 新城 丽发新城 修建 搅拌 厂 污染 环境 污染环境
A市 新城 丽发新城 小区 侧面 建设 混 泥土 搅拌 搅拌站 粉尘 噪音 污染
投诉 小区 搅拌 搅拌站 噪音 扰民
A2区 新城 丽发新城 修建 搅拌 搅拌站 污染 环境 污染环境 影响 生活
投诉 新城 丽发新城 小区 违建 搅拌 搅拌站 噪音 扰民
A2区 新城 丽发新城 违规 乱建 混凝土 搅拌 搅拌站 监管
A市 新城 丽发新城 小区 搅拌 搅拌站 噪音 扰民 污染 环境 污染环境
A2区 新城 丽发新城 小区 旁边 搅拌 厂 合法 经营 合法经营
A2区 新城 丽发新城 小区 太吵
搅拌 搅拌站 加工 砂石 料 噪音 污水 影响 新城 丽发新城 小区 环境
新城 丽发新城 小区 旁 搅拌 搅拌站 影响 生活
A2区 新城 丽发新城 区内 垃圾 垃圾站 散发 臭味
噪音 灰尘 污染 A2区 新城 丽发新城 环保 部门 环保部 环保部门
A2区 新城 丽发新城 修建 搅拌 厂 影响 睡眠
新城 丽发新城 小区 旁建 搅拌 厂 扰民
天暗 暗地 昏天暗地 噪声 新城 丽发新城 小区 A2区
暮云 街道 新城 丽发新城 小区 水泥 搅拌 搅拌站 非法 经营 非法经营 何时休
新城 丽发新城 小区 建 搅拌 搅拌站
A市 A2区 新城 丽发新城 小区 遭 搅拌 搅拌站 污染
A市 A2区 新城 丽发新城 小区 搅拌 搅拌站 明目 明目张胆 污染 环境 污染环境
新城 丽发新城 小区 搅拌 搅拌站 噪音 扰民

图 17　部分留言聚类结果

2、热度评价结果展示:

由聚类结果进一步分析并通过定义的热度评价指标可得到具体热点问题：

	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
0	1	1	83.65	2019/01/02至2020/01/06	A市	国家中心城市建设刻不容缓
1	2	2	75.30	2019/06/25至2020/01/26	A市	丽南路丽发新城居民区附近搅拌站扰民
2	3	3	70.36	2018/11/15至2020/01/06	A市	新政补贴最近两个月的怎么还没发?
3	4	4	69.38	2019/07/07至2019/09/01	A市伊景园滨河苑	捆绑车位销售
4	5	5	51.31	2019/01/02至2019/12/26	A市	住房公积金贷款的相关问题

图 18 经聚类结果分析所得热点问题图

由图 18，分析出热点问题排名前五的问题均与 A 市民生相关，为更加详细的了解各类问题反映的具体信息，可按问题 ID 整理成如图 22 所示的热点问题详情表。

	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
83	1	190213	A00031618	请A市加快自来水深度净化改造力度	2019/1/16 12:53:26	\n\t\t\t\t\t\n\t\t\t\t\t\t地处时代倾城小区的自来水烧过开水后电水壶就有...	0	0
151	1	191872	A00031618	请A市加快轨道交通建设力度	2019/3/1 15:19:28	\n\t\t\t\t\t\n\t\t\t\t\t\t地处中部中心城市的A市在高铁和地铁建设极为落...	2	9
219	1	193514	A00031618	请加快A市月亮岛片区公共服务力度	2019/3/20 16:39:22	\n\t\t\t\t\t\n\t\t\t\t\t\t地处月亮岛片区近年人口迅猛增长，成了全区人口...	0	4
301	1	195905	A00031618	A市加快招商引资方面有何具体行动?	2019/7/24 13:48:16	\n\t\t\t\t\t\n\t\t\t\t\t\t在国家大力倡导加快制造业转型升级的大背景下，...	0	0
343	1	196908	A00031618	请加快A市汉王陵考古遗址公园建设力度	2019/1/4 15:00:57	\n\t\t\t\t\t\n\t\t\t\t\t\t地处银杉路旁边的国家级汉王陵考古遗址公园已规...	0	9
...	...	...	...	...	...	...	...	...
3551	5	273765	A000106426	强烈建议提高A市直单位公积金缴存比例	2019/5/22 12:18:32	\n\t\t\t\t\t\n\t\t\t\t\t\t市直单位职工一直执行最低档的公积金缴费比例，...	0	3
3748	5	278622	A00022513	投诉A市楚熙水郡拖延办理公积金贷款	2019/7/9 17:06:21	\n\t\t\t\t\t\n\t\t\t\t\t\t我是楚熙水郡三期（西地省A市A3区洋湖垸潇楚...	0	0

图 19　热点问题详情图

由图 19 可看出各用户所反映的具体信息，可根据问题的具体描述，将上述信息按照所反映具体内容类别移交至相关的部门，由部门工作人员进一步通过对留言进行热点问题挖掘，有利于相关部门有针对性地及时处理问题，提升服务效率。

2.3.3 结果三结果分析

通过从相关性、完整性、时效性、可解释性主要四个方面进行处理评价，得出如下图所示的评价结果，我们利用星级对评价结果进行直观的表示，等级从低到高为一星至五星。

答复意见质量评价结果如图所示：

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	答复相关性质量	答复完整性质量	答复可解释性质量	答复综合质量	答复评价分值
2549	A00045581	A2区景荟华苑物业管理有问题	2019/4/25 9:32:09	\n\t\t\t\t\t\n\t\t\t\t\t2019年4月以来，位于A市A2区桂花坪街道...	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景荟花苑物业管理有问题”的调查核...	2019/5/10 14:56:53	0.689939	0.528346	1.0	0.671735	★★★★★
2554	A00023583	A3区潇楚南路洋湖段怎么还没修好?	2019/4/24 16:03:40	\n\t\t\t\t\t\n\t\t\t\t\t潇楚南路从2018年开始修，到现在都快一年了...	网友“A00023583”：您好！针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	2019/5/9 9:49:10	0.665252	0.389813	1.0	0.623186	★★★★★
2555	A00031618	请加快提高A市民营幼儿园老师的待遇	2019/4/24 15:40:04	\n\t\t\t\t\t\n\t\t\t\t\t地处省会A市民营幼儿园众多，小孩是祖国的未来...	市民同志：你好！您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下：为了改善...	2019/5/9 9:49:14	0.726757	0.427121	1.0	0.652830	★★★★★
2557	A000110735	在A市买公寓能享受人才新政购房补贴吗?	2019/4/24 15:07:30	\n\t\t\t\t\t\n\t\t\t\t\t尊敬的书记：您好！我研究生毕业后根据人才新政...	网友“A000110735”：您好！您在平台《问政西地省》上的留言已收悉，市住建局及时将您反...	2019/5/9 9:49:42	0.597687	0.442259	1.0	0.618650	★★★★★
2574	A0009233	关于A市公交站点名称变更的建议	2019/4/23 17:03:19	\n\t\t\t\t\t\n\t\t\t\t\t建议将“白竹坡路口”更名为“马坡岭小学”，原...	网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡...	2019/5/9 9:51:30	0.666723	0.496919	0.0	0.355343	★★★
...	...	...	...	...	...	...	...	...	...	...	...
181267	UU008766	汽车北站进站口附近居民强烈反对建设I市平康肾病医	2018/12/12 15:20:46	\n\t\t\t\t\t\n\t\t\t\t\t我们是I市汽车北站进站口的周围居民。在这里的...	您的留言已收悉。关于您反映的问题，已转I1区委、区人民政府调查处理。	2019/1/8 16:54:53	0.570991	0.720082	1.0	0.690893	★★★★★

图 20　答复意见质量评价结果

若答复意见与留言相关性高，同时回复格式完整，解释性强，同时回复时间与留言时间相距短，则留言回复的评价等级越高，即星级越高。

3. 结论

伴随着中国进入大数据时代，网民持续增长，网络空间的政治互动日趋重要，直接改造了政府和公民之间的互动模式。政府能有效回应公民诉求来实现良性政治互动，是理解任何政治制度之运行逻辑的关键。

本文通过对公民与政府行为记录的大数据分析，要是采取文本分析和统计分析考察了网络政治互动的两个维度:公民诉求表达和政府回应。探讨了现阶段中国网络政治互动的发展状况。基于网民留言数据和相关部门的回应资料，了解了当下时代令群众困扰的问题存在于哪些方面，这对相关部门的工作有重大意义。

基于朴素贝叶斯算法对群众留言数据进行建立分类模型，根据 k-means 聚类方法挖掘热点问题，能够更加直观地了解民众所关心反应的问题。从公民留言的所属类别和排名前五的热点问题中可以发现，城乡建设类、劳动和社会保障类的关注度居高，教育文体类议题次之，而有关商贸旅游、环境保护、卫生计生，交通运输的问题相对较少。随着时代的变迁，可以发现公众关注的重点由传统的吃饱喝足到更广更具体的领域延伸，人们的生活越来越好，越注重于生活的享受，关注脚下的土地，家园的保护，随之有关这类的留言建议也越来越多。

对相关部门的留言回复数据多角度进行评价，有利于更好地评估和提升回复的质量，也更加激发了公民诉求表达。对留言的及时准确回复，回复是否可以有效地解决问题，定然会影响着政府形象和公信力。所以政府应大力强化网络空间的治理能力建设，以有效回应公民诉求来实现良性政治互动。其次，留言问题归属的影响反映了网络政治互动中政府行为者的策略。政府回应的留言问题选择可以由不同留言问题的诉求主体和议题复杂度来解释，相对强势的诉求主体和较低复杂度的留言问题更易得到政府回应，反之则难以获得政府回应。环境保护的留言问题由于太复杂而回应性低，就业和农业农村等留言问题则由于诉求主体弱势而获得较低回应，拆迁征地不仅留言问题复杂且诉求主体弱势而获得很少政府回应;反之，城市建设、交通、企业事务、医疗卫生 和社会治安等留言问题在操作上复杂度较低，诉求主体也大多为城市居民、企业主等优势群体而获得较多回应;贪污腐败留言问题尽管其诉求主体相对比较弱势，从这个意义上，强化政府回应性的政治平等和回应能力建设至关重要。

另一方面政务部门要随时关注社交平台舆情变化，反应民生问题热点 在信息爆炸的时代，任何个人每天都面对着海量信息，但是真正与老百姓相关的民生问题，始终会吸引人们的关注，政策的制定与政府服务都应围绕民生问题进行。官方沉默或者延后反馈会大大降低其在民众心中的公信力，不利于后续的政策推广。在社交平台上民众看到的信息是已经经过多次传播的信息，很容易出现不同的舆论导向，官方应及时回应，避免舆论导向错误。如在热点问题‘加快 A 市城市建设’的相关问题上，应加强与民众的沟通，允许民众合理地表达不满，及时披露信息，并对政府所采取的应对措施以及具体成效进行推广，可以得到群众广泛的认可，有效引导舆论。为了及时响应，政务大数据舆情研究更加重要，需要进一步提升政府治理问题与安抚群众的能力。

总而言之，政府首先要学会合理利用网络问政平台，引导一些议题，倾听大众的声音，形成一个正反馈。这样，公众才会觉得自己的意见是有用的，从而大家讨论的意见变成决策时的重要依据，那么就更体现了人民是国家的主人的含义。

同时政府要把网络当成改善治理的重要渠道，上行下效，并将网络问政纳入政府考核体系之中，这样才能有效果。

4.参 考 文 献

[1]李少温. 基于网络问政平台大数据挖掘的公众参与和政府回应问题研究[D].华中科技大学,2019.

[2]潘亚星.基于 Python 的词云生成研究——以柴静的《看见》为例[J].电脑知识与技术,2019,15(24):8-10.

[3]张学新,贾园园,饶希,蔡黎.网络招聘信息的分析与挖掘[J].长春师范大学学报（自然科学版）.2017, 36(5):28-36.

[4]刘佩鑫,于洪志,徐涛.基于朴素贝叶斯的档案分类研究[J].河北大学学报(自然科学版),2018,38(05):549-554.

[5]袁文生,王晓峰.基于朴素贝叶斯的中文海事文本多分类器研究[J].计算机与现代化 JISUANJI YU XIANDAIHUA.2011,(5):150-153.

[6]孟天广,李 锋.网络空间的政治互动:公民诉求与政府回应性———基于全国性网络问政平台的大数据分析[J].清华大学学报(哲学社会科学版).2015,30(3):17-29.

[7]阮永芬,魏德永,高 骏,吴 龙,丁海涛,刘克文.K-means 聚类分析高原湖相沉积软土参数[J].昆明理工大学学报(自然科学版) Journal of Kunming University of Science and Technology(Natural Science).2020,45(1)：85-91.

[8]王侠林,贺建峰.基于 K-Means 聚类的微生物群落结构研究[J]软件导刊 Software Guide.2018,17(1):146-148+151.

[9]戴天辰. 基于传递距离的度量学习和聚类算法研究[D].扬州大学,2018.