深度学习文本分类在智慧政务系统上的应用

摘要：随着大数据、云计算、人工智能等技术的发展，文本信息中包含着大量有价值的知识，如何有效的管理这些文本信息创造出价值、实现变现的过程，成为了我们面临的巨大挑战。

现如今我国多措并举保障人民民意、诉求得以及时反映和妥善解决。随着网络问政平台的兴起，各类社情民意相关文本信息不断攀升。基于自然语言处理技术的智慧政务系统已经逐渐取代以人工来进行留言划分和热点整理的相关部门的工作。智慧政务系统是社会治理创新发展的新趋势，大大推动了政府的管理水平和施政效率。

针对问题一：建立关于留言内容的一级标签分类模型。实际上，这是一个文本分类（多分类）问题，文本分类中包含了大量的技术实现，根据附件 2 中的留言主题与留言详情提供的文本资料，采用 KNN 算法的文本分类模型，将留言划分为城乡建设、环境保护、党务政务等问题类型，并利用所给评价指标 F-Score 对分类模型进行评价。

针对问题二：根据某一时段内群众集中反映的某一问题可称为热点问题，及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。根据附件 3 将某一时段内反映特定地点或特定人群问题留言，采用 LDA 算法进行归类，定义合理的热度评价指标，并给出评价结果，按照格式给出排名前 5 的热点问题，还有具体留言信息。所以需要从附件 3 众多留言中识别相似的留言，并将留言进行归类，再对分类后的留言问题进行热度评价。

针对问题三：对相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。所以需要从不同角度分析答复意见，即需要进行层次分析，答复意见的内容是否与问题相关，是否满足某种规范，是否能够让民众理解等，建立评价模型。

关键词：文本挖掘； KNN算法；LDA算法；层次分析

Application of Deep Learning Text Classification in Intelligent Government System

Abstract

With the development of big data, cloud computing, artificial intelligence and other technologies, text information contains a lot of valuable knowledge. How to effectively manage these text information to create value and realize the process of realization has become a great challenge for us.

Nowadays, our country has taken many measures to ensure that people's public opinion and demands can be reflected and properly resolved in time. With the rise of the network political platform, all kinds of social sentiment and public opinion related text information is rising. The intelligent government system based on natural language processing technology has gradually replaced the work of the relevant departments to divide messages and organize hot spots manually. Intelligent government system is a new trend of social governance innovation and development, which greatly promotes the management level and governance efficiency of the government.

Question one: Establish a first-level label classification model about message content. this is actually a text classification (multi-classification) problem. the text classification contains a lot of technical implementation. according to the text information provided by the message subject and message details in annex 2, the text classification model of the KNN algorithm is used to divide the message into urban and rural construction, environmental protection, party affairs and other problem types, and the classification model is evaluated F-Score using the evaluation indicators given.

Question two: according to a certain period of time, a certain problem reflected by the masses can be called called a hot issue, timely discovery of hot issues, help the relevant departments to deal with targeted, improve service efficiency. According to annex 3, the paper classifies the message which reflects the specific place or the specific crowd in a certain period of time, uses the LDA algorithm to classify, defines the reasonable heat evaluation index, and gives the evaluation result, according to the format gives the top 5 hot issues, and the specific message information. So we need to

identify similar messages from many messages in Annex 3, and classify the messages, and then evaluate the heat of the classified messages.

Question three: to the relevant departments to reply to the comments, from the point of view of the relevance, integrity, interpretability of the response to the quality of comments to give a set of evaluation, and try to achieve. Therefore, it is necessary to analyze the response opinions from different angles, that is, to carry out a hierarchical analysis, whether the content of the response opinions is related to the problem, whether it meets certain norms, whether it can be understood by the public, and so on, and to establish an evaluation model.

Keywords: text mining; KNN algorithm; LDA algorithm; hierarchical analysis

1. 问题背景

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。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。自然语言处理[1]是计算机科学领域与人工智能领域中的一个重要方向。

所以基于自然语言处理和文本挖掘的方法，结合大量的智慧政务数据，能够及时发现热点问题，有助于相关部门进行有针对性地处理，从而提高政府部门的服务效率。

2. 总体流程与步骤

2.1. 总体流程

本文的总体架构及思路如下：

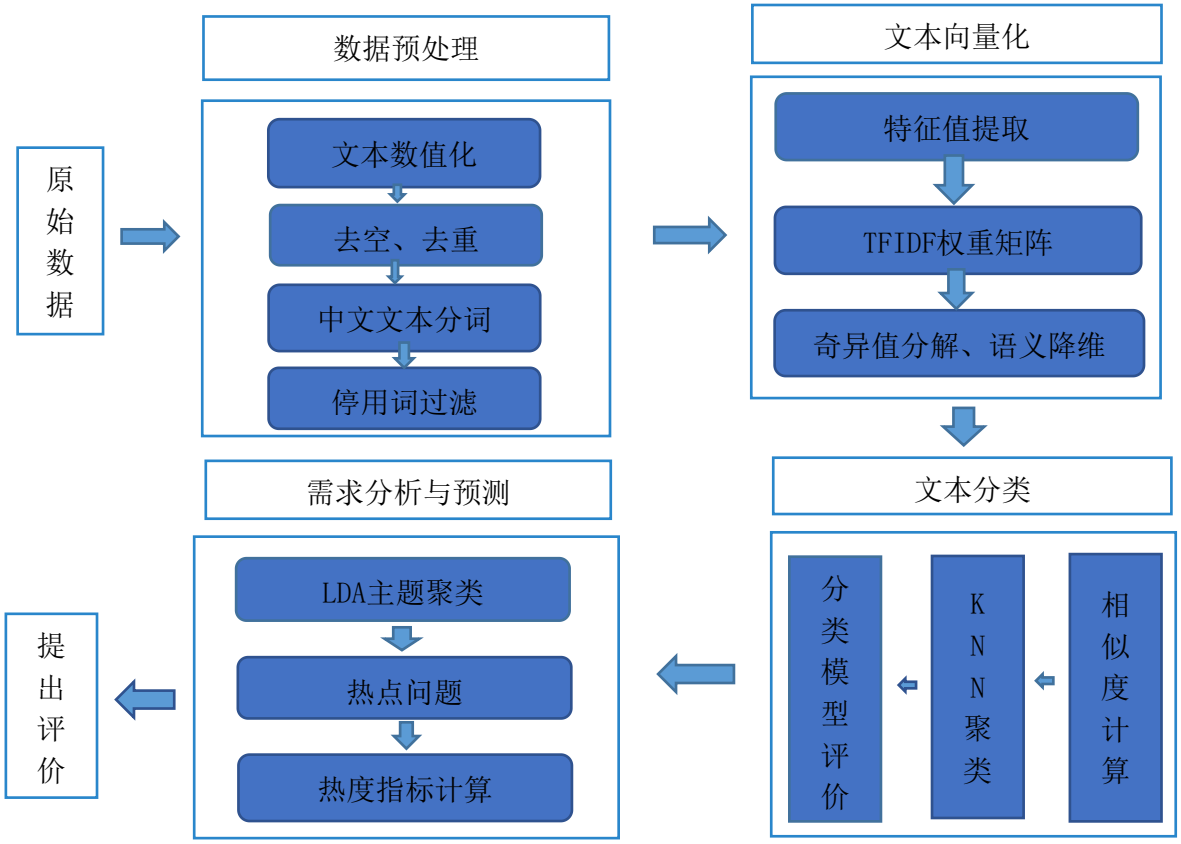

图1 整体框架

2.2. 步骤

步骤一：数据预处理，参考附件1提供的内容分类三级标签体系对留言进行分类，对附件2文本数据数值化处理，对附件3文本去除重复项及空行、中文文本分词、停用词过滤，以便后续分析；

步骤二：文本向量化，基于TFIDF权重法提取关键词，构造词汇-文本矩阵，进而利用奇异值分解算法进行语义空间降维，去除同义词的影响，简化计算。

步骤三：文本聚类，根据文本向量，计算文档间的欧式距离，再基于LDA模型对各个主题描述进行聚类。

3.模型简介

3.1. KNN算法

KNN算法又称为邻近算法[2]，或K最近邻(KNN，K-NearestNeighbor)分类算法，所谓K最近邻，就是k个最近的邻居，即每个样本都可以用它最接近的k个邻居来代表。KNN算法的原理是通过测量不同特征值之间的距离进行分类，其核心思路是一个样本在特征空间中的k个最相似的样本中的大多数属于某一个类别，则该样本也属于此类别，并具有这个类别上样本的特性。KNN算法中，所选择的邻居都是已经正确分类的对象。在KNN中，用于计算对象间距离的相似度指标一般使用欧氏距离:

$$d(x,y)=\sqrt{\sum_{k-1}^{n}(x_k-y_k)^2} \tag{1}$$

KNN算法的主要步骤为:

i计算测试数据与各个训练数据之间的距离;

ii.按照距离的递增关系进行排序;

iii.选取距离最小的k个点;

iv.确定前k个点所在类别的出现频率;

v.返回前k个点中出现频率最高的类别作为测试数据的预测分类。

KNN算法优点在于思维较为简单，是机器学习算法中最简单的算法之一，且理论成熟，准确度较高。随着技术不断革新，KNN算法的改进算法也不断出现，改进的方向主要有计算复杂度的降低、优化相似度度量方法、选取恰当的k值等。

3.2.LDA模型

LDA是一种非监督机器学习技术[3]，可以用来识别大规模文档集或语料库中潜藏的主题信息。它采用了词袋(bag ofwords)的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。

3.2.1 LDA生成过程

对于语料库中的每篇文档，LDA 定义了如下生成过程：

1.对每一篇文档，从主题分布中抽取一个主题；

2.从上述被抽到的主题所对应的单词分布中抽取一个单词；

3.重复上述过程直至遍历文档中的每一个单词。

语料库中的每一篇文档与 T（通过反复试验等方法事先给定）个主题的一个多项分布 （multinomialdistribution）相对应，将该多项分布记为 θ 。每个主题又与词汇表（vocabulary）中的 V 个单词的一个多项分布相对应，将这个多项分布记为 φ 。

3.2.2 LDA 学习过程

LDA 算法开始时，先随机地给θd 和φt 赋值（对所有的 d 和 t）。然后上述过程不断重复，最终收敛到的结果就是 LDA 的输出。再详细说一下这个迭代的学习过程：

1.针对一个特定的文档 ds 中的第 i 单词 wi，如果令该单词对应的 topic 为 tj，可以把上述公式改写为：

$$pj(wi|ds) = p(wi|tj) * p(tj|ds) \quad (2)$$

2.现在我们可以枚举 T 中的 topic,得到所有的$pj(wi|ds)$,其中 j 取值$1{\sim}k$。然后可以根据这些概率值结果为 ds 中的第 i 个单词 wi 选择一个 topic。最简单的想法是取令 $pj(wi|ds)$最大的 tj（注意，这个式子里只有 j 是变量），即 $argmax[j]pj(wi|ds)$；

3.然后，如果 ds 中的第 i 个单词 wi 在这里选择了一个与原先不同的主题，就会对θd 和φt 有影响了（根据前面提到过的这两个向量的计算公式可以很容易知道）。它们的影响又会反过来影响对上面提到的 $p(w|d)$的计算。对 D 中所有

的 d 中的所有 w 进行一次 $p(w|d)$的计算并重新选择主题看作一次迭代。这样进行 n 次循环迭代之后，就会收敛到 LDA 所需要的结果了。

表 1 LDA 算法的输入输出

算法输入：分词后的文章集(通常为一篇文章一行) 主题数K，超参数α和β
算法输出：1.每篇文章的各个词被指定的主题编号：tassign-model.txt 2.每篇文章的主题概率分布θ：theta-model.txt 3.每个主题下的词概率分布φ：phi-model.txt 4.程序中词语word的id映射表：wordmap.txt 5.每个主题下φ概率排序从高到低top n特征词：twords.txt

3.3. 层次分析法（AHP）

Analytic Hierarchy Process[4]，简称AHP。是一种定性与定量相结合确定因子权重的科学方法。AHP是将决策总是有关的元素分解成目标、准则、方案等层次，在此基础之上进行定性和定量分析的决策方法，基本思想如图3。运用AHP法进行决策时，需要经历以下4个步骤：

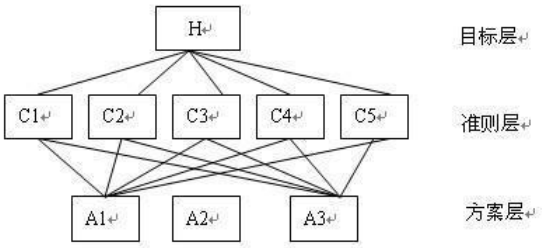

图2 AHP层次分析法基本思想

（1）分析系统中各因素之间的关系，建立系统的递阶层次结构；

（2）对同一层次的各元素关于上一层次中某一准则的重要性进行两两比较,构造两两比较判断矩阵；

（3）由判断矩阵计算被比较元素对于该准则的相对权重,并进行一致性检验；

（4）计算各层元素对系统目标的合成权重，并进行排序。

4.数据处理

4.1. 数据预处理

4.1.1 数据描述

根据所给数据观察,可以发现数据量较大,且附件中的字段大多为文本格式,需将其量化成数值形式才能对其进行分析,如果附件中有大量空行以及重复的情况，不做处理将会对后续分析造成影响，并且留言文本信息存在大量噪声特征，如果把这些数据也引入进行分词、词频统计乃至文本聚类等，则必然会对聚类结果的质量造成很大的影响，于是本文首先要对数据进行预处理。

4.1.2 文本预处理

我们把这些文本数据的预处理分为四个部分：

（一）文本数值化

为了让计算机能够理解词语,我们需要将词语的信息映射到一个数值化的语义空间中，这个语义空间我们可以称之为词向量空间。文本的数值化方式有：TF-IDF、BOW、One-Hot、分布式的表示方式（word2vec、Glove）等。

本文采用最常用的word2vec工具，它是一种无监督的学习模型，可以在一个语料集上，实现词汇信息到语义空间的映射，最终获得一个词向量模型。

以上操作完成之后先将文本分词做embeeding得到词向量, 将词向量经过一层卷积,一层max-pooling, 最后将输出外接softmax来做n分类。

（二）去重、去空

首先要对文档中的无用信息进行预处理，去除文档中的标点符号、数字、特殊字符以及格式等一些无关信息。

去除文档中的停止词，在文档中存在着很多无用的特征，计算机没有直接区分和预测能力，如冠词、助词、代词、介词、连词。去掉这些无用的特征，可以减少特征总数。只提取文本中的动名词作为特征词，去除文本中大量的虚词和部分实词。

（三）中文分词

由于中文文本的基本特点是词与词之间没有明显的界限,从文本中提取词语时需要分词，本文采用Python开发的一个中文分词模块——jieba分词[5]，对附件2及附件3中每一条留言主题描述进行中文分词，jieba分词用到的算法：

基于Trie树结构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)；

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jieba分词系统提供分词、词性标注、未登录词识别，支持用户自定义词典，关键词提取等功能。

（四）停用词过滤

为节省文本存储的空间以及提高搜索的效率，在处理文本之前会自动过滤掉某些表达无意义的字或词，这些字或词即被称为Stop Words（停用词）。停用词具有两个特征：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。

为了找出这些停用词，需要一些标准估计词的有效性。而高频词通常与高噪声值具有相关性。词语在各个文档中出现的频率可以作为一个噪声词的衡量标准，事实上一个只在少数文本中出现的高频词不应被看作是噪声词。因此用以下指标衡量词语的有效性：

1、词频(TF)

TF是一种简单的评估函数，其值为训练集合中此单词发生的词频数。TF评估函数的理论假设是当一个词在大量出现时，通常被认为是噪声词。

2、文档频数(DF)

DF同样是一种简单的评估函数，其值为训练集合中包含此单词的文本数。DF评估函数的理论假设是当一个词在大量文档中出现时，这个词通常被认为是嗓声词。本文选用DF方法筛选出如下停用词：我,有,的,了,是等。将筛选出的停用词加入停用词表，再利用停用词表过滤停用词，将分词结果与停用词表中的词语进行匹配，若匹配成功，则进行删除处理。

4.2．文本特征的抽取

经过上述文本预处理后，虽然已经去掉部分停用词，但还是包含大量词语，给文本向量化过程带来困难，所以特征抽取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，以此来降低向量空间维数，从而简化计算，提高文本处理的速度和效率。常用的方法有词频-逆向文档频率(TF-IDF)、互信

息、信息增益、X2统计等。

(一) 互信息[6](Mutual Information，MI)

在统计语言模型中，互信息用于表示两变个量间(表征 f 和类别c之间)的相关性。其互信息记作$MI(f,c)$可由下式计算：

$$MI(f,c)=\log\left(\frac{p(f,c)}{p(c)p(f)}\right) \quad (3)$$

由于互信息没有考虑单词发生的频度，这是互信息一个很大的缺点，它导致互信息评估函数经常倾向于选择稀有词。

（二）x^2统计[7](CHI)

x^2统计方法度量词条t和文档类别之间的相关程度，并假设t和c之间符合具有一阶自由度的x^2分布。令N表示训练语料中的文本总数，c为某一特定类别，t表示特定的词条，A表示属于c类且包含t的文档频数，B表示不属于c类但是包含t的文档频数，C表示属于c类但是不包含t的文档频数，D是既不属于c也不包含t的文档频数，则t对于c的x^2值由下式计算：

$$x^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)} \quad (4)$$

如果词条对于某类的x^2统计值越高,则表示它与该类之间的相关性越大，携带的类别信息也较多。

(三) 信息增益[8]

IG是一种在机器学习领域应用较为广泛的特征选择方法。它从信息论角度出发，以各特征取值情况来划分学习样本空间，根据所获信息增益的多少来筛选有效的特征。IG可以用下式表示：

$$IG(t)=p(t)\sum_{i=1}^{m}p(C_i|t)\,lg\frac{p(C_i|t)}{p(C_i)}+p(\bar{t})\sum_{i=1}^{m}p(C_i|\bar{t})l\,g\frac{p(C_i|\bar{t})}{p(C_i)}) \quad (5)$$

式中$p(C_i|t)$表示文本中出现词条t时文本属于C_i的概率，$p(C_i|\bar{t})$表示文本中不出现词条t时文本属于C_i的概率，$p(C_i)$表示类别出现的概率，$P(t)$表示语料中包含词条t的文本的频率。

(四) 词频-逆向文档频率[9](TF-IDF)

传统的 TF-IDF：

词频(Term Frequency，TF)是词语在文本中出现的频率，如果某一个词在一个文本中出现的越多，它的权重就越高，基本公式：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (6)$$

以上式子，$n_{k,j}$中是该词在文件d_j中的出现次数，而分母则是在文件中d_j所有字词的出现次数之和。

逆向文档频率(Inverse Documentation Frequency，IDF)是指在少数文本中出现的词的权重比在多数文本中出现的词的权重高，因为在聚类中这些词更具有区分能力。它的基本公式如下：

$$idf_i = \log \frac{N}{|\{j:t_i \in d_j\}|} \quad (7)$$

其中，N：语料库中的文件总数，$|\{j:t_i \in d_j\}|$包含词语t_i的文件数目($n_{i,j} \neq 0$的文件数目）如果该词语不在语料库中，就会导致被除数为零，因此一般情况下使用$1 + |\{j:t_i \in d_j\}|$。

在Shannon的信息论的解释中:如果特征项在所有文本中出现的频率越高，它所包含的信息熵越小；如果特征项集中在少数文本中，即在少数文本中出现频率较高，则它所具有的信息熵也较高。

最后可以得出：

$$w_{i,j} = tf_{ij} \times idf_i \quad (8)$$

这就是词的权重。上述方法各有利弊，IG计算量相对其它几种方法较大；对于MI方法，在相同的条件概率下，稀有名词会比一般词获得更高的得分；x^2方法基于x^2分布，如果这种分布被打破，则对低频词不可靠。因此本文采用目前公认的比较有效的TF-IDF 算法抽取特征词条，将权重按照从大到小的顺序排列，抽取权重最大的前 5000个特征词作为候选特征词。

4.3．文本的空间向量模型

计算机不能够直接处理文本信息，所以我们需要对文本进行处理，将文本表示成为计算机能够直接处理的形式，即文本数字化。文本表示[10]也称为文本特征表达，它不仅要求能够真实准确的反映文档的内容，而且要对不同的文档具有区分能力。目前常用的文本表示模型有：向量空间模型、布尔模型和概率模型等。

向量空间模型[11](Vector Space Model, VSM)最早是由 Salton 和 McGill 于 20世纪60年代末提出的，是目前在文本挖掘技术中最常用的表示模型。

其主要思想：将每一个文本表示为向量空间的一个向量，并以每一个不同的特征项(词条)对应为向量空间中的一个维度，而每一个维的值就是对应的特征项在文本中的权重，这里的权重可以由TF-IDF等算法得到。向量空间模型就是将文本表示成为一个特征向量：

$$V(d) = (t_1, w_1(d), \cdots\cdots t_n, w_n(d)) \quad (9)$$

其中，$t_i(i = 1,2\cdots\cdots n)$为文档d中的特征项，$w_i(i = 1,2\cdots\cdots n)$为特征项的权值，可由TF-IDF算法得出。

5.模型的建立与求解

5.1 本文选择使用Knn模型作为文本分类模型

本文采用了KNN模型对附件2中的共9210条数据进行文本分类，其中城乡建设2009条、环境保护938条、交通运输613条、教育文体1589条、劳动和社会保障1969条、商贸旅游1215条、卫生计生877条。由于我们的目的是将留言者的留言文本与上述一级分类标签进行整理归类，因此首先要创立训练集。

5.1.1 创建训练集、测试集

通常训练集样本量越大，其包含的文本特征越多，对于整体文本的代表性也就越强，反之则无法覆盖到所有文本特征；但训练集样本量越大，训练时间就越长，无法突显出文本自动分类的方便高效。所选出的训练集对于文本特征越具有代表性，其训练效果就越好，因此我们需要在训练集样本量和分类性能中寻找到一个最合适的平衡点。目前基于机器学习的分类算法通常遵循一个假设：训练集和测试集的分布是一致的，这样在训练集上训练出来的分类器应用于测试集时才会比较有效[2]。选取80%作为训练集，剩余作为测试集。

首先对数据集做监督学习，需要人工标注数据。在确定好训练集后，首先对训练集样本进行标注，运用 R 语言中的Rwordseg 包和 tm 包进行分词，数据包原始的语料库中包含 10 万条以上的中文词条，根据本次文本数据的实际内容，对语料库进行了补充，并用导入的停词表去除了留言中的虚词和语气词；最将处理后的训练集和测试集同时导入，使用 class 包中的 KNN 算法对测试集进行分类。

5.1.2 构建模型及模型评估

结果评估：在机器学习领域对于文本分类结果的评估有三个基本的指标：召回率（Recall Rate），指检索出的相关文档数和文档库中所有相关文档的比率，衡量的是检索系统的查全率。

召回率 = 系统检索到的相关文档数 / 系统所有相关文档的总数。

准确率（Precision），指检索出的相关文档数与检索出的文档总数的比率，衡量的是检索系统的准确率。准确率 = 系统检索到的相关文档数 / 系统检索到的所有文档总数

F-Score：机器学习中常用的评价标准。

5.2 本文选择LDA算法对留言进行文本聚类分析

LDA[3]（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。 我们采用 LDA 算法对留言主题进行聚类，根据留言主题的概率对主题进行归类。

5.2.1 优化主题数目

由于数目需要人为选取，会影响到主题的聚类效果，因此我 们对主题数目的选取进行了优化。选取主题数目的原则：类间距离越大、类内距离越小。 我们定义主题差异度：

$$D_w = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N-1} \left\| T_i - T_j \right\| \quad (10)$$

其中，T_i为LDA主题模型中主题i在t年的概率向量，N为t年所有主题的数量，T_i和T_j 之间的一范数代表两个不同主题i和j之间的距离。 由于不同主题之间的概率向量相差较远，相近主题的概率向量距离较近，因此主题差异度D_w 能够较好地反映出不同数目主题之间的总的差异度。

为了能够抑制专利主题数目对主题差异度的影响，我们定义了专利主题平均差异度 D_{ave}^{t}。

每两个主题之间的平均距离越大，聚类效果越好。根据这一原则，我们可以比较不同主题数的聚类结果。

5.2.2 主题强度

主题强度由隶属于该主题的留言数量决定。 由前述可知，最佳主题数量可

基于LDA模型确定，我们可以使用以下公式来界定隶属于某个主题的留言数量：

$$n_{i,p}^{t}=\begin{cases}1,Prob=max\left\{Prob_{j,p}^{t}\right\}\\0,其他\end{cases} \quad (11)$$

其中，$Prob$为基于LDA模型，在第t次留言专利p隶属于主题的概率；$n_{i,p}^{t}$是一个 01变量，表征在第t次留言，留言p是否隶属于主题i。

5.2.3 主题内容演变指标

主题内容演变指标可以通过两个主题之间的相似度来刻画，我们定义第t次主题i和第t＋k次主题j的相似度如下：

$$sim(T_i-T_j)=\frac{T_i\times T_j}{\|T_i\|\times\|T_j\|} \quad (12)$$

其中 T_i代表基于 LDA 模型第t个主题i的概率向量。第t个主题i和第t＋k个主题j的相似度可以用来刻画主题随时间的知识继承关系。如果相似度低于给定阈值，则认为第t＋k个主题j继承第t个主题i内容较少；如果相似度高于给定阈值，则可以认为第t＋k个主题j继承了第t个主题的知识。

对于第t个主题i来说，第t＋1个所有主题和第t个主题的最相关的主题j可以看作是主题i的继承者。这样就可以构建留言主题内容演变。

5.3 本文选择层次分析法对答复意见的质量进行评价

在深入分析实际问题的基础上，将有关的各个因素按照不同属性自上而下地分解成若干层次，构造层次结构模型。随后构造成对比较阵，计算权向量并做一致性检验。对于每一个成对比较阵计算最大特征根及对应特征向量，利用一致性指标、随机一致性指标和一致性比率做一致性检验。若检验通过，特征向量(归一化后)即为权向量：若不通过，需重新构造成对比较阵。

6.结论

6.1 问题一：

本文选取了KNN模型对附件2中的留言进行分类，并通过混淆矩阵、准确率、召回率以及F值对模型进行评价，结果较好。

6.2 问题二：

我们利用LDA算法对附件3中的留言文本进行文本聚类，对每一篇文档，从主

题分布中抽取一个主题，再从上述被抽到的主题所对应的单词分布中抽取一个单词；重复上述过程直至遍历文档中的每一个单词。

结合距离计算公式各类热点问题间的距离及每类热点问题各问题的距离，从而得到了6种热点问题，并通过建立热点评价指标，即热度指数=相应问题的留言条数，从而可以得到排名第一的热点话题为A2区丽发新城小区修建搅拌厂噪音扰民，污染环境问题。针对整理汇总得到的热点问题留言明细表，相关部门能够更好地更有针对性地解决居民提出的问题，帮助政府部门解决因传统人工分类带来的错判误判的概率。

表2 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	15	2019/07/21至2019/09/25	A5区魅力之城小区	小区临街餐馆油烟噪音扰民
2	2	13	2019/7/3至2020/01/26	A2区丽发新城小区	修建搅拌厂噪音扰民，污染环境
3	3	13	2019/7/18至2019/8/28	A市伊景园滨河苑	坑害购房者
4	4	10	2019/11/15至2020/1/9	丽发新城小区	附近搅拌站的一些问题
5	5	9	2019/01/15至2019/12/24	A市居民	咨询A市缴纳社保购房公积金等问题
6	6	8	2019/01/16至2019/12/02	A市居民	咨询A市人才购房补贴政策
7	7	8	2020/01/05至2019/07/07	A市居民	咨询A市军转安置问题
8	8	6	2019/1/至2019/7/9	A3区西湖街道茶场村五组	五组何时启动拆迁
9	9	6	2019/3/16至2019/10/21	A市高新区	道路交通扰民问题
10	10	5	2019/4/2至2019/11/22	A3区中海国际社区	三期四期中间空地夜间施工噪音扰民

6.3 问题三：

我们基于层次分析模型，建立三层的留言评价模型，从答复的相应性、完整性、可解释性等方面对附件4的留言评价，将留言评价分为四个等级，优秀、良好、及格、较差。根据所给留言，我们可以得到政府部门工作人员所给出的留言基本上都为优秀或良好等级的留言，在留言的完整性和相应性的分数较高，但是有些留言可解释性不强，市民没有办法很好地理解其中含义。希望今后根据此评

价体系，能够提高留言回复的质量。评价体系如下图3

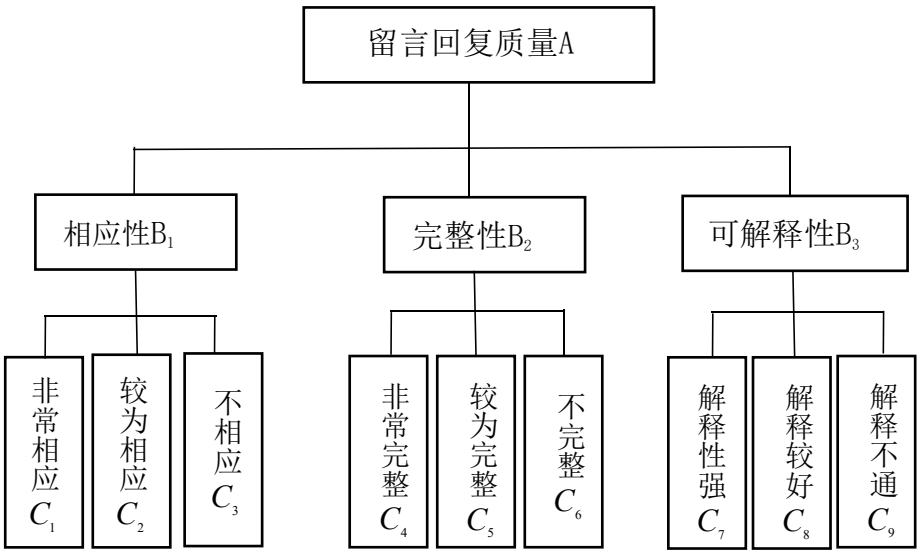

图3 留言回复评价体系

7.总结与展望

在未来的学习中，会对本文的工作进一步的研究。可以对深度学习的模型中的算法进行改进，尝试使用深度学习中其它模型如卷积循环神经网络（CRNN）模型对文本分类进行特征学习。总之，基于深度学习模型的文本分类，在未来的发展中会有很好的应用前景。随着深度学习理论的不断完善，对本文的进一步研究中会加入更多的深度学习新的研究成果，使得基于深度学习模型的文本分类器的性能有很好的提高。

参考文献：

[1]王灿辉,张敏,马少平.自然语言处理在信息检索中的应用综述[J].中文信息学报,2007(02):35-45.
[2]陈曦.文本挖掘技术在社情民意调查中的应用[J].中国统计,2019(06):27-29.
[3]王元波,骆浩楠,汪 峥.文本挖掘在主题发现和相关性评估中的应用[D].东南大学,2018
[4]Saaty,T.L.The analytic hierarchy process[J].McGraw-Hill, New York,1980
[5]陶伟.警务应用中基于双向最大匹配法的中文分词算法实现[J].电子技术与软件工程,2016(4).
[6]郭飞,张先君,叶俊.基于改进互信息的特征提取的文本分类系统[J].四川理工学院学报:自然科学版,2008,21(3):93-96.
[7]徐明,高翔,许志刚,等.基于改进卡方统计的微博特征提取方法[J].计算机工程与应用,2014,50(19):113-117.
[8]陈小莉.基于信息增益的中文特征提取算法研究[D].重庆大学,2008.
[9]王美方,刘培玉,朱振方.基于TFIDF的特征选择方法[J].计算机工程与设计,2007,28(23):5795-5796.
[10]周昭涛.文本聚类分析效果评价及文本表示研究[D].中国科学院研究生院(计算技术研究所),2005.
[11]胡学钢,董学春,谢飞.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.