第八届泰迪杯数据挖掘

C 题：“智慧财务”中的文本挖掘应用

目录

摘要

随着互联网相关技术的广泛应用，政府获取民众意见、凝聚民气的渠道愈发多样，但是随之而来的是文本信息数量的几何倍数增加，依靠以往单单通过人工识别的分类方式越来越低效，数据的时效性不再匹配如今的大部分处理方式，如何在海量文本中快速、有效地找到关键可靠的信息成为了信息时代的重中之重。

本文针对比赛所提出的问题进行了文本挖掘与自然语言处理的模型建立，所使用的数据分析软件为 Python（PyCharm）。总体思路如下：

一、首先针对于未分类、存在标签的留言集合，运用监督学习中的多项式朴素贝叶斯算法（Multinomial Naive Bayes）结合词袋（doc2bow）与 TF-IDF（Term Frequency－Inverse Document Frequency）等方法进行分类，分类的准确度在 83.2%左右，召回率在 82.6%左右，最终 F1-score 得分为 0.823 分（满分 1 分）。

二、针对于热点问题挖掘方面，在考虑到其为无监督学习的性质后，本文采用了 LDA 主题模型（Latent Dirichlet Allocation）进行相关主题词的分类，在此之前通过计算困惑度（perplexity）对主题词数量进行确定，后制定评分系统将各个主题的文本数量、点赞数量等指标合成为综合评分，对热度进行加权综合排名，提取热点问题信息。

三、针对于答复意见的评价系统，我们通过从答复的相关性、完整性、内容充实度与文章的可解释性进行综合评分。对于相似度分析采用余弦相似度算法（Cosine Similarity）计算问答相似性；对于完整性运用词频占比率（Term Frequency Rate）进行分析；对于内容充实度运用文本长度转反正切函数（Inverse Tangent）进行评分；对于文本的丰富程度与可解释性而言，运用文本情感分析（Sentiment Analysis）进行质量评价。最后在计算综合得分之前，利用正态公式消除各个指标的分布不同的差异实现标准化，通过加权平均得到最终的答复质量评分。

关键词：LDA 主题模型 困惑度 TF-IDF 多项式朴素贝叶斯模型 文本情感分析

研究背景

随着第四次工业化革命的带来，在工业 4.0 时代大环境之下，社会逐渐朝着“信息化”所发展。在对信息的储存、运算、分析等方面，互联网有关技术的普及使得“大数据”应用变成了可能，而且不论是硬件还是软件都在这个信息化时代大放异彩：计算机硬件的快速更迭让数据的处理速度屡屡突破极限，云计算让海量数据的储存成本大大降低，人工智能算法让杂乱无章的数据变得井然有序.....

在如今这个属于大数据的时代，数据的种类多种多样，其中文本数据就占了半壁江山。我们每天在智能手机上使用微信、QQ 等软件所发送、收取的均为文本信息；企业日常记录财务、人事管理等方面所储存的数据格式是文本信息格式；政府有关部门在记录、分析、调差社会数据的时候，存储最多的数据格式也是文本数据类型。文本信息充斥着我们生活的方方面面，文本信息凭借着其独有的语法结构、编写顺序和时效性质让无数的数据分析爱好者为之痴迷，从机器学习，到深度学习甚至是人工智能领域，出现了无数的自然语言分析与文本挖掘算法，只为了能够在茫茫数据之中提取到最有价值的部分。但是文本数据的数量过于庞大，在进行文本分析的时候难免会觉得无从下手，再加上文本数据的时效性较短，导致了文本挖掘的难度进一步提高。

自十九大以来，坚持以“人民群众”为中心的新时代发展思想成为热议话题，如何让人民获得幸福感、安全感、更有保障，更可持续变成了政府最为关心的头等大事。习近平总书记在国家数据战略会议上提出：要实施国家大数据战略，加快建设数字中国的步伐。由此，政府微信平台，市长信箱，阳光热线等等的网络问政平台应运而生，民众不需要再经过一道道繁琐的程序表达自己的诉求与意见，政府也能够在第一时间获得有关当时当刻的民众议论重心，采取行之有效的应对措施，更好地服务人民。但是随着文本数据量的不断上升，处理文本的要求也在不断提高，仅仅通过人为提取信息的效率远远不足以支撑信息更新的速度，因此，“智慧政务”的网络问政想法孕育而生，问政与互联网结合，文本与算法相合并，对于政府的管理水平与施政效率有着很大的推动作用。

相关模型简介

1.1 朴素贝叶斯多项式分类模型

贝叶斯分类器具有三种不同的分类，分别是高斯朴素贝叶斯，多项式朴素贝叶斯以及伯努利朴素贝叶斯，其依据不同的性能有着截然不同的作用。本文使用文本分析最多的应用方法多项式朴素贝叶斯模型（Multinomial Naive Bayes）进行分类。贝叶斯分模型的基本原理是通过计算先验概率和调整因子对后验概率的求解。本质过程是计算先验概率和似然函数的过程，如下所示：

$$\hat{c} = \arg\max_{c \in C} P(c \mid d)$$

其中前提假设为现有存在类别 $C = \{c_1, c_2, c_3, ..., c_m\}$ 共计 m 个类别，且相互独立。上式所描述的为在所有 m 个类别之中，能够使得条件概率达到最大值时的类别。根据条件概率公式可以将上式进行变换，变换如下所示：$P(c \mid d)$

$$\hat{c} = \arg\max_{c \in C} P(c \mid d) = \arg\max_{c \in C} \frac{P(d \mid c)P(c)}{P(d)}$$

将公式进行转换以后可以发现，由于是每一次计算类别的时候，最大似然函数的分母所在的位置都是一致的，所以在进行概率大小的比较时候可以将其去除，方便比较。则可以转化为如下的形式：

$$\hat{c} = \arg\max_{c \in C} P(c \mid d) = \arg\max_{c \in C} P(d \mid c)P(c)$$

公式进行简化以后可以简化成两个部分，其中 $P(d \mid c)$ 称之为似然函数，$P(c)$ 称为先验概率。在生成文字向量的时候，本文使用了词袋的方法封装各个文本的向量、文本内容、标签等数据，这里假设词袋模型表示文档 d，此时的文档 d 的每一个特征表示为：$d = \{f_1, f_2, f_3, ..., f_n\}$，其中特征 f_i 表示的是每个单词 w_i 出现的频率（次数），因此可以对公式进行进一步转换：

$$\hat{c} = \arg\max_{c \in C} P(f_1, f_2, f_3, ..., f_n \mid c)P(c)$$

此时有一个较为重要的前提假设，即各个特征之间是相互独立的，由于在日常生活中相互独立的条件较为苛刻，因此形成模型的名字，也就是朴素贝叶斯中

“朴素”一词的由来。在假设独立的前提基础之下，可以得到 $p(f_1,f_2,...,f_n|c)=p(f_1|c)*p(f_2|c)*...*p(f_n|c)$，由此化简后得到：

$$c_{NB}=agr\max_{c\in C}P(c)\prod_{f\in F}P(f|c)$$

但是为了防止出现下溢（underflower）现象，即在每一个概率都很小的情况下，将极小概率直接相乘会导致所得的结果越来越小，无限趋近与 0，所以为了防止出现此类情况，在最后的表达式中加入 Log 函数，并且引入变量 w_i（词频），公式如下：

$$c_{NB}=\arg\max_{c\in C}P(c)\prod_{f\in F}P(f|c)$$

1.2 TF-IDF

TF-IDF 为词频-逆向文件频率，它是一种资讯检索与资讯探勘的加权技术，这种统计方法能够评价一个词或者在语料库中其中一份文件的重要程度是多少。TF-IDF 使用了词频数与逆向文件频率这两个指标，大体思想为短词的重要性随着其在同一篇文章中的出现频率的增加而增加，但是随着其在语料库中出现文章频率的增加而减少。

TF（term frequency）词频指的是一个给定的短词在指定文件中出现的次数，对于最后的 TF-IDF 呈正相关增长。公式如下（N(x)为指定短词数量，N 为指定文档短词总数）：

$$TF(x)=\frac{n_{x,j}}{\sum k\cdot n_{k,j}}$$

IDF（inverse document frequency）逆向文件频率是通过整个语料库的范围对一个短词的重要性判别度量，其主要思量不仅是考察单篇文章中短词的频率，而是将所有文章中的特定短词所出现的文章数频，其表达式如下：

$$IDF(x)=\log(\frac{N}{N(x)})$$

但是在一些比较特殊的情况下，会出现一些较为生僻的字词，这样一来会导致 IDF 分母出现 0，运算会出现错误。因此通常需要将 IDF 进行平滑，修改后的

公式如下所示：

$$IDF(x)=\log\frac{N+1}{N(x)+1}+1$$

在得到了 TF 与 IDF 后，则可以计算出 TF-IDF 的最终值：

$$TF-IDF=TF(x)\cdot IDF(x)$$

1.3 LDA 主题模型

LDA 主题模型（Latent Dirichlet Allocation）从结果上来说是一种文档主题提取生成模型，从结构上来说其是一种三层贝叶斯概率模型。其主要通过无监督的学习方法发现文本中隐藏的主题信息。与第一问使用方法的基本原理相类似，但主题模型包含了短词、主题与文档三个方面的层次结构，称其所谓的“生成模型”，是由于每一篇文章的每一个短词都是通过“用一定的概率选择了某个主题，并且从这个主题中以一定的概率选择这个词语”的过程，让文档到主题服从了多项式的分布，主题到词也是服从多项式分布。

LDA 主题模型的主要思想如下：

首先，使用随机分配的方法将文本的每一个单词分配主题 $z^{(0)}$，接着将每一个主题下出现的 term 数量与每个文档下主线的主题 z 中包含的短词数量进行统计。接着进行重新计算，即排除当前词的主题分配，依据其他所有词的主题分配估计当前主题词分配各个主体的概率。然后得到当前短词属于所有主题 z 的概率分布以后，根据这个概率分布为该词一个新的主题 $z^{(1)}$。然后依据相同的方法不断更新下一个短词的主题，直到发现每一个文档分布 $\vec{\theta}_m$ 和每一个主题下短词分布 $\vec{\phi}_k$ 收敛，此时算法停止，输出待估计参数 $\vec{\theta}_m$ 和 $\vec{\phi}_k$，最终每一个单词的主题 $z_{m,n}$。

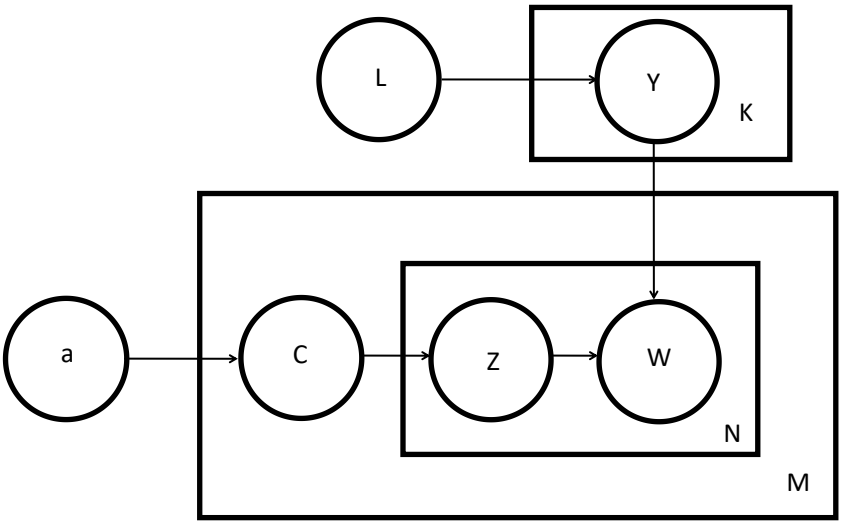

图 1 LDA 原理图

其中 LDA 主题模型步骤参数包括：K 为主题个数，M 为文档总数，N 是每个文档的单词总数，L 为每一个主题下词的多项式分布 Dirichlet 先验参数，a 为每一个文档下主题多项式 Dirichlet 先验参数,Z 为每个 m 个文档第 n 个主题词，W 是 m 个文档中的第 n 个词，隐含变量 C 与 Y 分别为 m 个文档下主题分布和第 k 个主题词的分布。

1.4 困惑度

困惑度(perplexity)是指度量一个概率分布或者概率模型预测样本的好坏，其使用方法针对于 LDA 主题模型配合使用。LDA 主题模型由于其无监督学习性质的特殊性，所以进行无监督学习之前需要进行分类数目的确定，因此需要使用困惑度进行主题个数的确定。困惑度也能够比较不同的概率分布或者概率模型。

用一个概率模型 q 去估计真实的概率分布 p，则可以通过测试集中的样本来定义这个概率模型的困惑度：

$$b^{-\frac{1}{N}\sum_{i=1}^{N}\log_b q(xi)}$$

其中测试样本 x1,x2,...,xN 是来自于真实概率分布 p 的观测值，b 的取值通常取 2，由此原因低的困惑度对说明 q 对于 p 的拟合越好，当模型 q“遇到”测试样本的时候，q 会不会感到“困惑”，可以看出指数部分为交叉熵。如下所

示：

$$H(\hat{p},q)=-\sum_{x}\hat{p}(x)\log_2 q(x)$$

其中 $\hat{p}$ 表示的是对于真实分布下样本点出现 x 的概率估计。

问题一：留言分类

2.1 分类思路

在处理网络问政平台的时候会遇到各种类的留言，但是政府部门的职能划分是十分明确的，每一个职能部门都有着其管辖的范围领域，此时如何进行分类就不能单单依靠人工经验，第一是因为数据量的庞大，一旦人工分类进行的时间达到人类的疲劳期，势必分类的准确度会大幅下降；第二数量巨大、重复且机械式的任务浪费了大量的时间，使得单位效益十分低下。由此，利用统计模型进行模型的建立会使得数据的处理效率大大提升。

对于题目要求所描述，本文对有关问题的原始数据结构分析后，选择文本挖掘中监督学习范围的多项式朴素贝叶斯算法（Multinomial Naive Bayes）进行分类。在进行模型的拟合以前，需要对数据进行处理：对于数据的系统存储使用 bunch 类，在将其持久化后，使用 TF-IDF 进行各个短词的加权处理，接着利用贝叶斯算法模型进行文档的分类，利用测试集检验分类的精确度与准确度，最后计算 F1-score 得分，读模型进行评价。主要流程图如下所示：

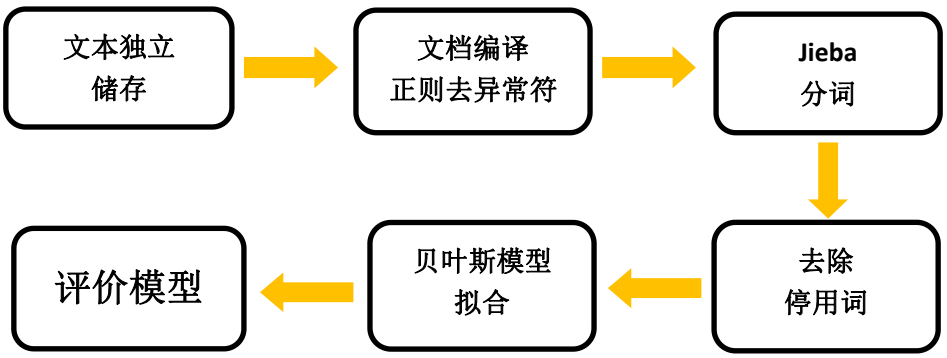

图 2 文本分类流程图

2.2 文本预处理

题目所给出的留言分类表格由以下部分构成：留言编号、留言用户、留言主题、留言时间、留言详情与一级标签构成，因为要依据一个标签进行分类，所以首先将文件按照一级标签-留言详情（0.txt，1.txt，2.txt...其中数字表示行

标）进行分类，以及标签包括：城乡建设，环境保护，运输交通，教育文体，劳动和社会保障，商贸旅游与卫生计生七个方面。分类结果如下图所示：

分类名称	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
分类数量	2009	938	613	1589	1969	1215	877

表 1 一级标签分类表

在进行监督学习之前，需要将原始数据划分为训练集与测试集，文本在划分数集时候，设置训练集与测试集的比例为 9:1：数据集的分类默认为前 90%数据自动划入训练集，后 10%的数据划入测试集，若划分数据数量出现浮点数时则向下取整。

在进行模型建立之前，需要将文本划分为单词组，所使用的工具为 jieba 中文分词组件，分词模式为默认模式（精确分词模式）。建立模型所使用的停用词集是从 GitHub 上进行下载，但是在建立分类模型的过程中，文本经过分词后暂时不进行删除停用词的单独步骤，由于本文在之后进行 TF-IDF 进行词空间向量建立的时候会设置停用词表，因此不单独进行分词处理的步骤。

本文在进行读取文本的时候，出现了一些编译解析错误的问题，在通过读问题文本的识别后，本文发现问题文章的格式多数为书信格式，其中由于书信格式的特殊性，使得其中出现了许多的全角空白符（\u3000）与空格（\t），因此本文对所有的文本进行 translate 后使用正则表达式匹配并且去除特殊符号。

2.3 建立模型

在建立模型之前，需要将所有的文本数据转化成为贝叶斯网络所能够识别的向量形式，在面对基本的数据转换中，python 提供了一个十分强大且有效的类型：“Bunch”。Bunch 类继承了 dict 字典的属性，因此其拥有了同字典一样的属性，可以在其构造器中设置任何我们需要的属性，可以对各个文本的属性进行动态设置。本文对 bunch 的设置如下所示（以“城乡建设”中的 0.txt 为例）：

属性名称	target_name	label	filenames	contents
数据类型	字符串	字符串	整型	字符串
含义	所有类别 标签	当前类别 标签	当前文件 名称	当前文件 内容
样例数据	交通运输，城乡建设等七个类名	城乡建设	0(.txt)	“A3 区大道西行便道，未管所路口至加……”

表 2　bunch 属性结构设置表

将所有分词完成后的文本数据进行存储以后，通过程序将文本 bunch 进行持久化存储，方便程序对其进行调用操作。而对于词汇的向量化表示，本文通过 TF-IDF（Term Frequency - Inverse Document Frequency）方法进行权重的计算有关。

在建立 TF-IDF 空间向量的时候，本文所采用依然是储存在 bunch 中，其 bunch 设置的具体参数如下所示：

属性名称	target_name	label	filenames	tdm	vocabulary
属性类型	字符串	字符串	浮点数	整型 （列表）	字符串、整型 （字典）
含义	所有类别标签	当前类别标签	当前文件名称	文件内容向量化	关键词及相对位置

表 3　TF-IDF 向量属性结构设置表

本文设置 TF-IDF 中的 max-df 参数为 0.5，即指定特征词出现在语料库中超过 50%以上的文档，则认为其实没有重要价值，进行剔除。

2.4 模型评价

将所有训练集进行训练后，本文将测试集数据标签剔除，使用训练集进行拟合，使用 metrics 方法中精确度、召回率与 f1-score 对模型进行评价。精确度与召回率均是通过混淆矩阵进行计算得来，混淆矩阵如下所示：

	Positive	Negative
Yes	True Positives	False positives
No	False Negative	True Negative

表 4　混淆矩阵表

此时的精确率、召回度与准确率的表达形式可以写为下列形式：

$$precision = \frac{TP}{TP + FP}, recall = \frac{TP}{P}, accuracy = \frac{TP + TN}{P + N}$$

精确率为正确预测为正项占全部预测为正的比例；召回率是指正确预测占全部正确样本的比例；准确率为预测正确占全部样本的比例。F1 得分是综合评价准确率与召回率的指标，其公式如下所示：

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

依据相关公式可以得到评价指标的有关数值，结果如下所示：

参考指标	精确度	召回率	F1-score
得分数值	83.2％	82.6％	0.823

表 5 指标得分表

由上述结果可见准确率与召回率均在 80％以上，分类结果较为理想。

问题二：热点问题挖掘

3.1 问题挖掘思路

依据题目所要求挖掘热点要求，本文在结合分析所给资料后，可以得出：所给留言均为无标签数据集，通时要求对文本进行分类，因此选用无监督学习方法LDA主题模型（Latent Dirichlet Allocation）进行文本的分类，但是LDA主题模型需要事先确定参数，即预分类的种类个数。在确定个数的方法上本文采用计算困惑度（perplexity），利用困惑度曲线的拐点确定最终分类个数，这样能够排除人为确定分类个数导致的主观性带来的问题，避免信息丢失和主题冗余。同样，在进行困惑度计算与LDA主题模型之前需要进行数据的预处理。

在将文本进行分类完成以后，本文对热点问题的评价指标时设立了评分机制，主要指标包括热点话题内所包含的留言个数（热点话题的数量与热度呈正比例增加），与对应留言中的点赞数与反对数（点赞数与反对数能够从侧面反映话题的热度），并且将话题进行加权平均，得到最后的综合评分，绘制热点问题表格。主要流程图如下所示：

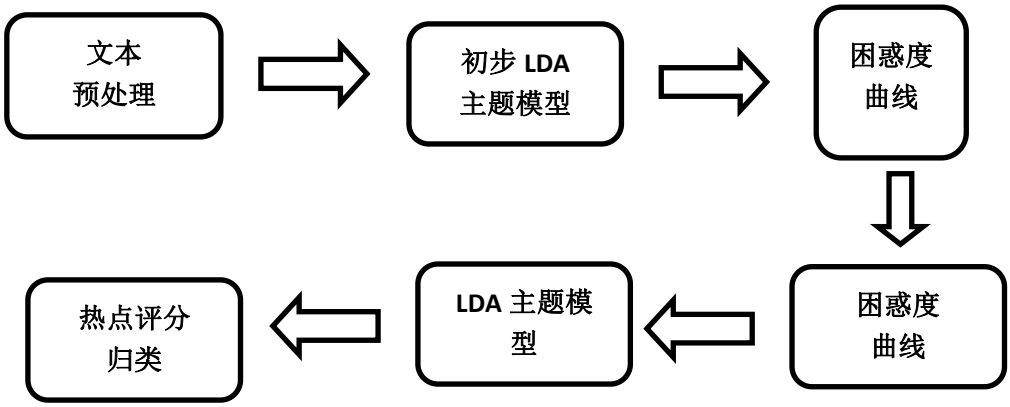

图 3 热点问题挖掘流程图

3.2 模型建立

本文对此题主要基于上个方面：于热点问题的挖掘与评分系统的建立。在挖掘有关热点问题时，本文首先随机确定模型主题个数，将其带入 LDA 主题模型进行运算，得到初步的训练模型，再依据困惑度生成的曲线拐点得到最终适宜分类的主题个数。

在对文本信息进行了分词、去除停用词的预处理操作后，文本首先进行 LDA 主题模型的建立，初步主题分类个数为 50 个（后期观察困惑度曲线确定主题词个数，范围在[1,50]之间进行选择）。本文使用了多核并行处理 LDA 主题模型，结合 doc2bow 词袋与 TF-IDF 统计方法进行模型的持久化。在函数的参数设置中，本文定义了 corpus 语料库与 dictionary 字典、迭代次数等参数，如下所示：

参数名称	num_topics	dictionary	corpus	iterations	passes
参数意义	设定主题个数	短词字典	语料库	迭代次数	全部语料训练次数

表 6　LDA 模型参数表

其中本文设置迭代次数为 4500 次，passes 训练次数为 5，使用多核运算核为 2 个，初步建立完成后得到语料模型与字典模型，同时生成主题词向量分布的 json 文件，方便后期困惑度的计算。本文在进行困惑度计算时，依据 perplexity 计算公式进行程序编写，最后得到困惑度在[1,50]的曲线图，如下所示：

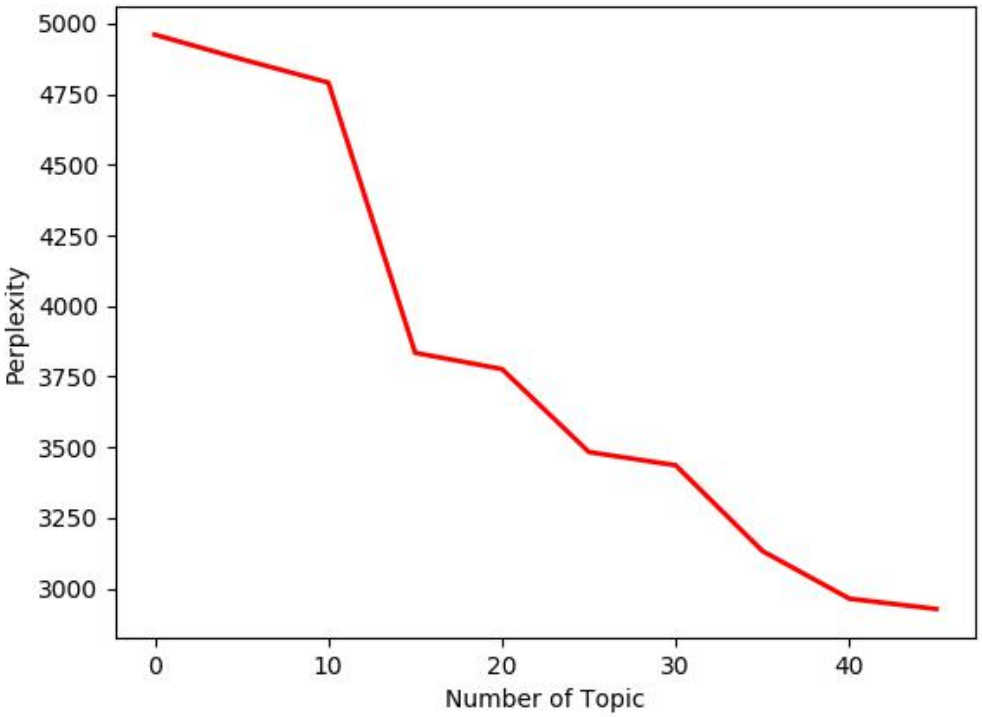

图 4 困惑度曲线表

由上所示的困惑度曲线图可以看出，曲线呈现出单调递减的下降走势，并且在大约 topic=15 时，曲线斜率发生了较大的变化。由此基本确定了 LDA 主题模型的主题个数为 15 个。再将主题个数带入到 LDA 模型中进行计算，得到了所有留言的主题概率以及所包含的主题词。

计算后得到的主题分类以及每一个分类对应的概率如下简表所示（话题热度与出现顺序无关，详见附录主题词与对应概率详细表）：

主题	包含主题词	主题词对应概率
主题 0:	扰民 施工 中学 噪音 汽车 号线 南站 工地 高速 宾馆	[0.16 0.12 0.10 0.08 0.04834193 0.048 0.041 0.0379 0.037 0.006]
主题 1：	地铁 拆迁 人行道 占用 楚江情况 黄兴 区 农民 北路	0.17 0.079 0.07 0.05 0.0481 0.038 0.032 0.027 0.016 0.005]
……	……	……
主题 14	新城 搅拌站 投诉 开发商 公寓 反对 小区 万科 魅力之城	[0.086 0.074 0.068 0.06 0.053 0.038 0.035 0.031 0.029 0.0238]

表 7 主题词与对应概率简表

对于热点问题的评分机制而言，本文认为能够影响热点问题的找方面包括两个：一、话题包含的留言个数。二、对应话题所包含的点赞与反对个数。但是这两个影响因素的重要程度存在差异，因此赋予不同的权值。如下所示：

	主要影响因素	辅助影响因素
影响因素名称	话题包含个数	点赞数与反对数
权重分配	90%	10%

表 8 影响因素与权重分配表

选择话题的包含数量主要是因为一个话题的热门程度，可以由其话题所包含的数量多少直接判断，一定时间内热度与谈论次数呈现单调递增的趋势。而点赞与反对数可以从侧面很好地反映一个话题的热门程度。

在经过对话题数量指标与点赞反对数指标共同标准化后赋予权重进行综合得分的计算。计算公式如下所示：

$$x_{summary}=\frac{x_i-\min(x_i,i\in M)}{\max(x_i,i\in M)-\min(x_i,i\in M)}\cdot 90+\frac{x_j-\min(x_j,j\in N)}{\max(x_j,j\in N)-\min(x_j,j\in N)}\cdot 10$$

得分的分值在[0,100]范围之内，分值越靠近 0 分说明话题的热度越低，分值越接近 100 分说明话题的热度越高。经过计算综合得分，并且通过编程实现了热点问题的时间提取与排序，本文得到了热点话题的排序，应题目要求提取排名前五位的热点话题。结果如下：

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	100	2019/07/21 至 2019/12/04	A 市万科魅力之城小区	小区临街餐饮店油烟噪音扰民
2	2	36.91	2019/07/11 至 2019/09/01	A 市广铁集团铁路职工	伊景园滨河苑商品房违规捆绑车位销售
3	3	31.9	2019/11/13 至 2020/01/25	A 市万家丽南路丽发新城居民区	小区旁边建搅拌站扰民
4	4	28.86	2017/06/08 至 2019/11/27	A 市经济学院学生	学校强制学生去定点企业实习
5	5	25.9	2019/05/05 至 2019/09/19	A 市五矿万境 K9 县	房屋出现质量问题

表 9 综合热点话题排名表

根据热度指数的综合排名，本文通过代码实现了热点话题的文本分类，最后得到了热点话题的有关分类，并且建立评价指标进行热点话题的量化，实现了题目要求内容。

问题三：答复意见的评价

4.1 评价意见思路

依据题目所给要求，在对相关部门的答复意见进行分析后，本文主要从留言答复的相关性，答复的完整性，答复内容充实程度与感情色彩和可解释性这几个方面进行留言质量的评价分析。

简单来说，相关性的判断是否为“答为所问”，不能“答不对问”；完整性是在有一定的相关性基础上，回答的范围是否完整，是否存在片面、缺失的问题；内容充实程度是在考察文段是否丰富，可解释性强弱与否；感情色彩指的是回答在情感、语句结构等方面的好坏。

评价指标使用的方法与原因如下：

对于相关性分析使用了文本挖掘中的余弦相似度（Cosine Similarity）算法。

本文采用此方法的原因是从文本的角度来说，回答是对提问的分析、解释、说明，所以回答与提问出现的关键词信息应该具有相似性；换而言之，如果是“文不对题”，则表述所用的关键词信息也会产生较大差别。

对于完整性采用计算 TF 词频占比（Term Frequency Rate）的方法。

基于文本相似性的基础之上，本文在研究文本完整性时发现，文本表述越是完整，则匹配的关键字数量会越多，也就是文本含有关键词信息与完整性呈正比例增加。

对于内容的充实度与可解释度，本文将处理后答复留言的文本长度进行反正切函数（Inverse Tangent）的拟合。

本文在对文本充实度与可解释性进行研究发现，文本的充实程度与可解释性与文本的相对长度有关。回复的充实程度与提问的文本长度具有一定的比例关系，而此种比例关系可以使用反正切函数的正向轴部分进行拟合，结果也是较为准确。

对于答复内容的感情色彩分析利用文本情感分析（Sentiment Analysis）进

行评价。

回复的情感色彩、关联词使用以及文段的结构等都能够对文本的可读性进行解释，因此使用文本情感分析能够在很大程度上反映文本的可解释性。

在计算完所有指标的对应得分后，考虑到所有指标的分布不同，所以最后利用标准正态公式消除各个指标之间的分布差异，再进行加权平均计算综合质量得分，对文本质量高低有一定的参考价值。

进行文本答复质量的评价流程如下所示：

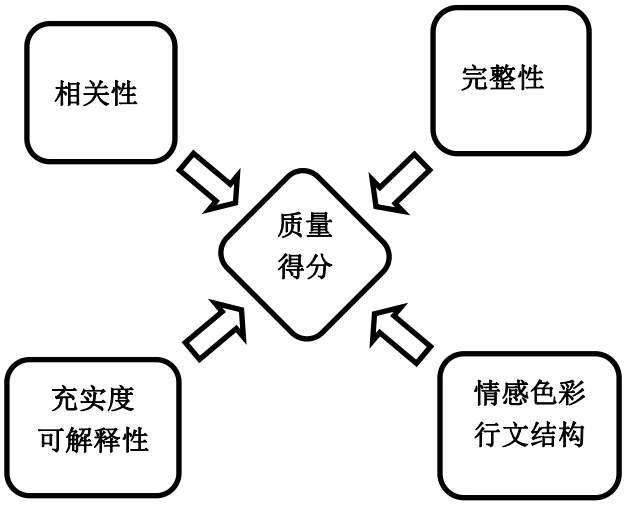

图 5 答复意见质量评价流程图

在进行得分计算的时候，由于部分得分的前提条件不允许文本为空值，但是第 1107 条内容回复出现异常，在进行仔细看到后发现其确实存在异常，所以在分值计算时候均记为 0 分。异常结果如下：

属性	提问详情	回复内容
异常文本	地点：A1 区南路暮云段 时间...A1 区南路马路边上 A7 县境内（伊莱克斯大道--莲湖安置小区段）强烈呼吁各位领导能够重视解决!万分感谢!	“UU0081182”

表 10 异常问答内容表

4.2 建立质量评价模型

4.2.1 文本相关性

对于不同文本的相似度而言，评价方法有很多种，例如 SVM、余弦相似度（Cosine Similarity）、BM25 算法等等，但是进过有关实验后发现，针对于短文本而言评价其相似度的最佳指标为文本余弦相似度，其他方法也可以实现相似度的计算，但效果欠佳。

余弦相似度是通过计算两个向量的夹角余弦值大小来评估向量之间的相似度，将向量写入二维空间或者是更高维度的空间进行库里的计算，向量之间对的夹角越小，则表示向量之间的相似度越高。下面以二维空间为例：

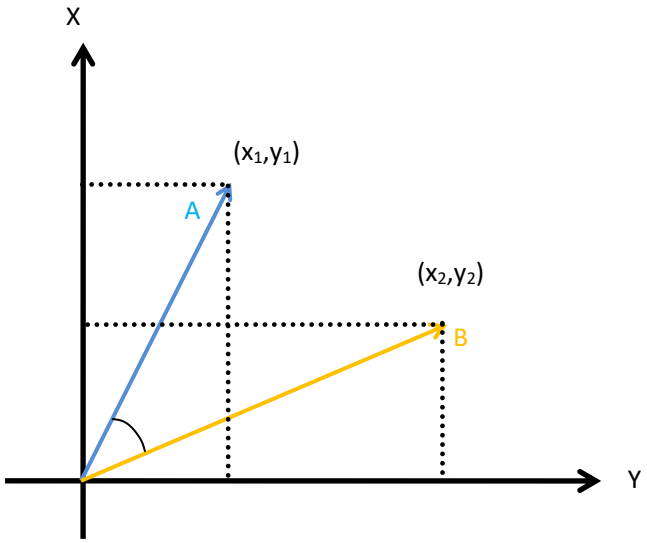

图 6 二维空间坐标图

在上图中，X 轴与 Y 轴组成了一个平面二维直角坐标系，向量 A 和 B 是两个特征向量，其夹角记为θ。在计算向量 A 与向量 B 之间的举例时使用到余弦坐标公式，如下所示：

$$\cos\theta = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}}$$

将余弦计算公式推广到多维向量的形式同样也适用。假设 A 与 B 均是 m 维的向量，记 A=[A1,A2,A3,...,Am]，B=[B1,B2,B3,...,Bm]，那么此时 A 与 B 之间的夹角余弦值计算公式如下所示：

$$\cos\theta = \frac{\sum_{i=1}^{n}(A_i \times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} = \frac{A \cdot B}{|A| \times |B|}$$

余弦相似度与欧氏距离相比，由于其形式的特殊性，所以让余弦得出的数值具有“方向性”，并不是强调意义上的长度与距离。

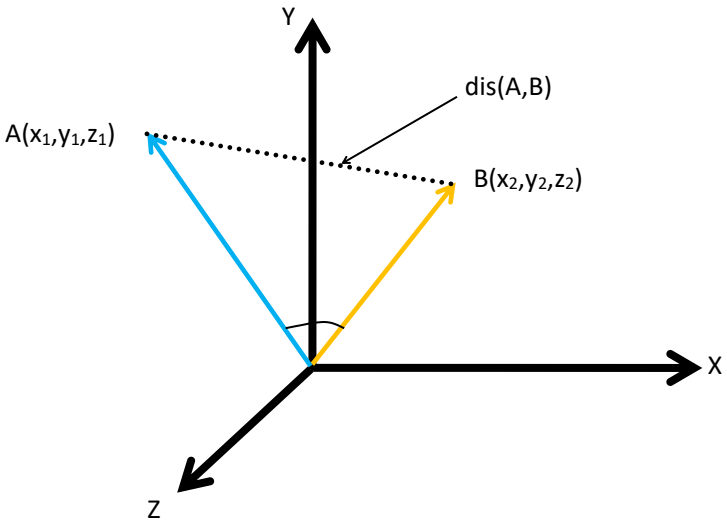

图 7 三位空间坐标图

依据文本余弦相似度的计算公式，本文首先将提问者的问题与有关部门的回答分别用进行 jieba 分词与删去停用词，分别保存成为提问与回答的文件。以一条提问问题与回答留言为例，如下所示：

	原文本	文本预处理
留言详情	潇楚南路从 2018 年开始修，到现在都快一年了，路挖得稀烂用围栏围起，一直不怎么动工，有时候今天来台挖机挖两几下.....	潇楚 南路 修 一年 路 挖 稀烂 围栏 围 动工 有时候 来台 挖机 挖 两 几下 几天 挖 几下 交通 生意.....
答复意见	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 区洋湖街道高度重视，立即组织精干力量调查处理，现回复如下：您反映的为潇楚大道西线道路工程项目，该项目位处于坪塘老集镇.....	区潇楚 南路 洋湖 段 修好 区 洋湖 街道 高度重视 现 回复 潇楚 大道 西线 道路 位 坪塘 集镇 土方 排水 施工 因该 次 大道 设计标准.....

表 11 文本样例数据表

经过文本预处理后将提问与答复的所有短词制成字典格式，字典中的键为出现的所有短词，而值为对应短词出现的频率：{’原’：0，’很大’：1，’土质’：2，

’挖’: 3, ’较差’: 4, ’天气’: 5, ’回复’: 6, ’污水’: 7, ’难度’: 8, ’较长’: 9, ’绕’: 10,... ’管线’: 81, ’管道’: 82, ’位’: 83}。

经过统计短词以及其词频数量后，本文将留言详情与答复意见进行编码，即用字典中的索引编号对每一句话进行编译。提问和答复的编译后成为列表形式，也就是向量化。编译结果如下所示：

	文本列表	文本编译向量
留言详情	['潇楚', '南路', '修', '一年', '路', '挖', '稀烂', '围栏','围','围', '动工', '有时候', '来台', '挖机', '挖', '两', ...,'街上', '老百姓', '出行']	[59, 31, 62, 20, 21, 14, 6, 30, 2, 22, 55, 65, 70, 14, 43, 24, 11, 14, 24, 81, 23, 60,..., 57, 72]
答复意见	['网友', '区潇楚', '南路', '洋湖', '段', '修好', '区洋湖', '街道', '高度重视', '现', '回复', '潇楚', '大道', '西线', ...,, '感谢您', '关心']	[74, 44, 31, 38, 46, 68, 77, 75, 39, 61, 32, 59, 27, 7,17, 54, 8, 25, 64, 56, 45, 41, 33, 27, 78,...,49, 19]

表 12 文本编译表

将文本表示成为向量形式以后便可以利用余弦计算公式计算留言与答复之间的相差距离。计算出最后得分，将部分得分与实际内容进行比较，发现得分能够较为直观地反映文本相关性，如下所示：

	留言提问	答复详情	余弦相似度
高相似度文本	经调查，反映我镇镇长张登武不作为的信访人为黄尊富..... 一、2011 年 12 月 22 日，经双方协商一致，黄刚明代表其全家与指挥部签订..... 二、由于黄刚明房屋与涵洞之间原有的道路是泥路..... 三、由于黄尊富常年在外，回家较少，黄尊富以其经常不在家且一次性协议.....	您好！来信收悉。现回复如下： 1、在浏醴高速建设期间，积极解决黄尊富一家的出行问题并给以补偿..... 2、黄尊富否认第一次协议效力，并以自己名义签订合约再次拿到补偿..... 3、江背镇政府积极召开信访答复会.....	0.88
低相似度文本	请问领导，农合费用增加了，打狂犬役苗报销比例是多少。盼回音。先谢了	已收悉	0.00

表 13 文本相似度对比表

上表中选取了相似度较高的提问与回答和相似度较低的提问与回答，对比后可发现，高相似度文本从回答方式、语句完整句等方面均十分契合提问问题，并

且每一条回复都是有理有据，并没有出现文不对题，或者拖泥带水的情况；反观相似度较低的答复与问题，可见回复者的回复内容对于提问而言有些敷衍，草草了事，相似程度较低。

4.2.2 文本完整性

本文在研究答复内容完整性时，发现完整度很高的文本均在相关度一定的前提条件下，对提问的内容解释越充实有效，以下提问与回复为例：

	留言问题	答复详情
完整性较高文本	建议将“白竹坡路口”更名为“马坡岭小学”，原“马坡岭小学”取消，保留“马坡岭”	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“白竹坡路口”更名为“马坡岭小学”，原“马坡岭小学”取消，保留“马坡岭”的问题。公交站点的设置需要方便周边的市民出行，现有公交线路均使用该三处公交站站名，市民均已熟知，因此不宜变更。感谢来信人对我市公共交通的支持与关心。
完整性较低文本	我是A5区桃花埠路上的居民。我们这里的居民小孩上学和日常出行都要穿过时代阳光大道。在省质监局站坐车……希望有关部门领导重视和马上解决，保障此处居民的安全！希望有关部门领导重视和马上解决，保障此处居民的安全！	您好！您的留言已收悉。现将有关情况回复如下：目前，我市正在开展2017年交通项目实施计划研究工作，对于您提出的意见建议，我们会予以认真考虑，进行研究。感谢您对我们工作的关心、监督与支持。

表 14 文本完整度对比表

由上表对比内容可以看出，文本完整度由两个方面进行描述：一、文本相似性。文本具有一定相似性的前提基础之下，才能够进一步判断文本的完整性。二、文本描述话题的完整程度。以上述为例，较为完整的文本是经过对提问的详细解释，精确进行答复，而完整性不高的文本在具有一定的相似度的情况下，只是在片面地、笼统地解释提问问题，并没有给出精确的答复内容，所以认为其回复的完整程度较低。

本文选用词频占比率（Term Frequency Rate）对文本的完整性进行解释，原理为将提问与答复分别建立语料库，通过计算答复词频与提问词频的契合度，

契合度越高，说明文本的完整性越高；反之而言，文本的契合度越低则说明文本的完整度就越低。如下所示：

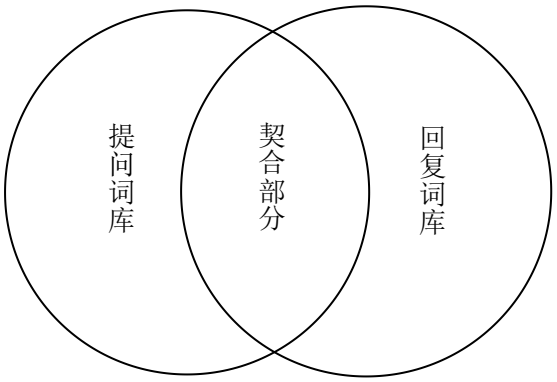

图 8 文本完整度图

上图所示契合部分能够较为直观地反映文本的完整程度。以上表为例，经过计算完整度得分后，高完整度文本最终得分为 98，低完整度文本得分只有 6 分，评价较为合理。

4.2.3 文本充实度（可解释性）

结合有关资料，本文在研究文本充实度时发现可解释性与充实程度与文段的相对长度有关。文段的相对长度在正向轴呈现单调递增的趋势，并且文本充实度在开始范围内随着范围的变大而逐渐增大，且增幅较为剧烈；但是在相对长度逐渐变大时随着范围的增大而逐渐趋于平缓，即一阶导单调递减。经过对现有函数进行拟合，发现反正切函数（Inverse Tangent）能够很好地拟合此种趋势，如下图所示：

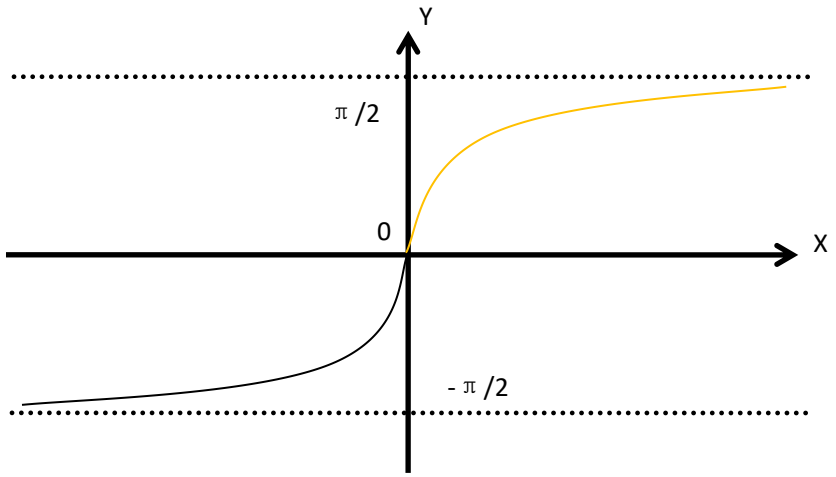

图 9 反正切函数图

文本相对长度的计算公式如下所示：

$$x_i = \frac{\arctan(abs(\frac{M_i}{N_i})) - \min(abs(\frac{M_i}{N_i}),\ i=[1,n])}{\max(abs(\frac{M_i}{N_i}),\ i=[1,n]) - \min(abs(\frac{M_i}{N_i}),\ i=[1,n])}$$

对每一对提问与回答进行得分计算后，得到所有问答项的充实度得分计算，取值范围为[0, 1]，数值越靠近 0 说明文本的充实度越低，越接近 1 说明文本的充实度越高。下面进行部分数据对比展示：

	留言问题	答复详情	充实度得分
高充实度文本	G7 县文盛小学引入特色班，每个学生必须参加，然后特色班老师就要学生购买相关的物品，一星期一节课，很多家长认为这所谓的特色班是没有必要的。(共计 67 字)	您好！获悉关于“对 G7 县文盛小学特色班的质疑”的网帖后，我局领导高度重视，并责成教育局基教股调查处理，现将调查处理情况回复如下：一、开展学生社团活动的意义…… 二是学校以学生社团活动为基础……（共计 618 字）	90.84
低充实度文本	刘市长：你好！我们知道你是一个勤政为民的人民好市长，。近几年，随着张界家旅游产业提质，…,新建便民加油站，望上级有部门领导重视,给予上报批准为谢(共计 504 字)	你好，你所反映的问题已转交相关部门调查处置。您好！也非常感谢您对家乡建设提出的宝贵建议…再次感谢您为家乡建设提出的宝贵建议，谢谢！2019 年 1 月 31 日(共 209 字)	24.44

表 15 文本充实度对比表

经过得分文本的充实度与可解释性得分计算后，可以看出得分能够较为直观反映出文本相对长度，即并不是回复内容越长充实度越高，而是要与提问者提问问题相结合进行考虑。

4.2.4 文本情感与结构

情感分析（Sentiment Analysis）是自然语言处理（NLP）领域的一类任务，又称倾向性分析、意见挖掘、主观分析等，它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。本题要求分析的答复意见内容里包括了程度副词、否定词以及一些具有情感倾向的词语、标点符号，这些属性都具有情感倾向。于是本文从这四个角度构建情感词典，基于情感词典匹配文本，最后对每个回复进行情感打分，得出该答复表达的是积极情绪还是消极情绪，进而属于主观还是客观描述，以此判断答复意见是否具有可解释性。

基于情感词典对文本进行情感分析，最关键在于情感词典的构建，它直接决定整个情感分析的准确性。本文在这里所使用的积极情感词表和消极情感词表均来源于 BosonNLP 数据下载的情感词典，否定词词典和程度副词词典均在《知网》情感分析用词语集（beta 版）中下载，中文停用词表则仍来自于在 GitHub 上下

载的哈工大停用词表。

实现情感分析的中心思想是通过逐个遍历分词后的语句中的词语，如果词语命中词典，则进行相应权重的处理。正面词权重为加法，负面词权重为减法，否定词权重取相反数，程度副词权重则和它修饰的词语权重相乘。最终，利用输出的权重值来挖掘正、负面信息，区分积极、消极情绪。其具体算法设计如下：

一、读取答复数据，对答复进行分句、分词以及去除停用词等数据清洗处理；

二、查找对每段话进行分词后的情感词，记录积极还是消极，锁定文本位置；

三、在情感词前查找程度词，如果找到则停止搜寻。若存在程度词则设权值，乘以情感值；

四、在情感词前查找否定词，完全部否定词，若数量为奇数，乘以-1，若为偶数，乘以 1；

五、判断分句结尾是否有感叹号，有叹号则往前寻找情感词，有则相应的情感值+2；

六、计算完每段回复所有分词的情感值，用数组（list）记录进行记录；

七、计算每段答复的积极情感均值，消极情感均值，积极情感方差，消极情感方差；

八、定义 result 函数，以积极情感均值减去消极情感均值作为每段答复最终的情感得分。

情感分析的主要流程图如下所示：

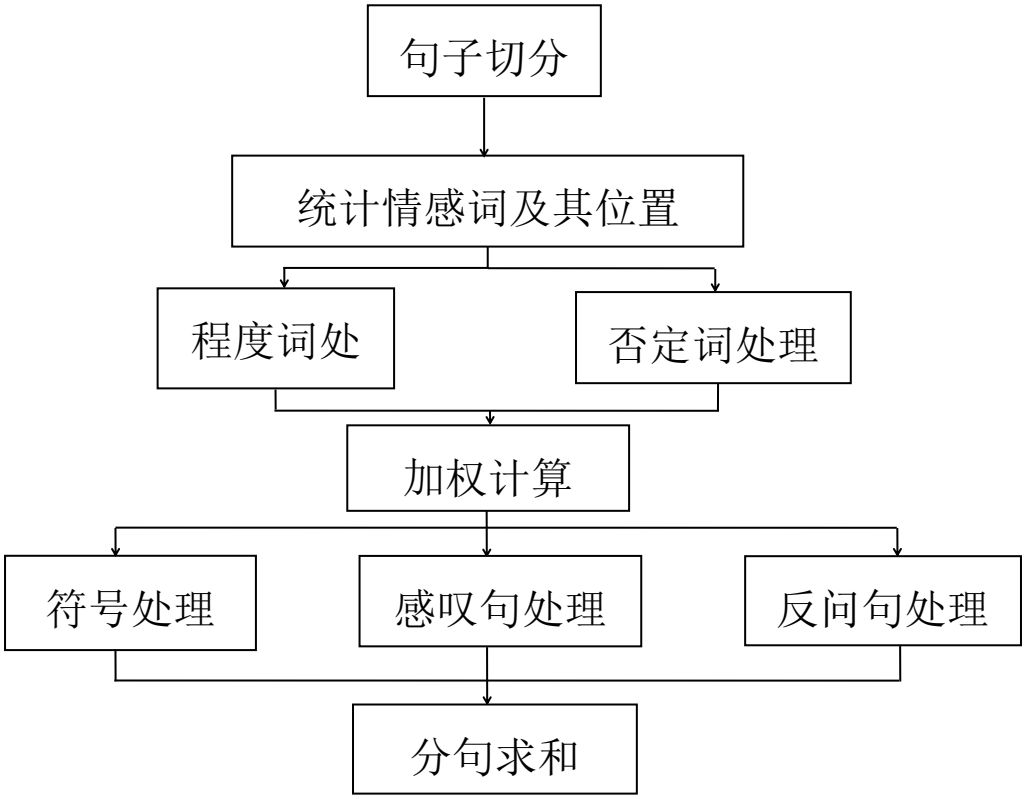

图 10 文本感情分析图

本文在计算得分时，对情感分析、语句结构进行了分别的评分计算，对于情感词、程度词与句式结构词进行定位分析，计算得分。对于情感积极向上，行文结构完整，用语规范的文本基于较高的评分；而对于情感色彩一般，行文结构有所欠缺的文本则给予较低的评分，最后最分数进行标准化，区间[0,100]，文本质量随着分数的上升而呈正比例增加。部分文本数据得分如下：

	留言内容	回复详情	得分
高分文本信息	G7 县文盛小学引入特色班，每个学生必须参加，然后特色班老师就要学生购买相关的物品，一星期一节课，很多家长认为这所谓的特色班是没有必要的。	获悉关于“对 G7 县文盛小学特色班的质疑”的网帖后，我局领导高度重视…一、开展学生社团活动的意义…学校把学生社团活动统一安排在学校课程中开展，符合国家、地方、学校三级课程管理要求。	96
低分文本信息	强烈反对 I 市 9 路公交车改线路获悉从…请政府、交通主管部门充分听取民众意见，合理规划，拿出切实有效的方案	您的留言已收悉。关于您反映的问题，已转市交通运输局调查处理。	18

表 16 文本对比表

4.2.5 文本综合得分

在对文本进行相关性分析、完整性分析、内容充实度分析和文本情感结构分析后，本文将把这四个指标进行综合评价。综合指标的评分方式为加权平均。在采用加权平均计算得分时，需要考虑两个问题：一、每一项得分的权重比例。二、每一项得分的分布存在差异。

对于指标的权重问题而言，本文将权重数值设置如下所示：

属性	相关性	完整性	充实程度与可解释性	情感与行文结构
权重	30%	25%	25%	20%

表 17 权重分配表

针对第一个权重非配问题，在进行权重分配时本文进行首先考虑到文本相似性的重要性，由于文本完整性的前提基础是建立在文本相似性之上，所以相关性的权重占比较高，设置为 30%。接着考虑到文本的完整性与内容充实程度的意义，这两者对于答复质量而言也占有较大的比重，因为答复的完整性直接决定了对提问问题是否回答完整，充实度与可解释性决定了文本内容的丰富程度。最后考虑到文本的情感色彩与行文结构因素，其是在文本表达正确、完整后对文章的润色程度、行文结构进行评价，因此给予 20%的权重。

对于第二个指标分布差异的问题，本文首先将各个指标按照公式对数据进行进行标准化，公式如下：

$$y_i = \frac{x_i - \min(x_i, i = 1,2,...,m)}{\max(x_i, i = 1,2,...,m) - \min(x_i, i = 1,2,...,m)}$$

进行标准化后的各项指标数据不能够直接进行加权运算，因为各个指标的分布不同，导致了可能同样一个数值得分所体现的价值截然不同。例如，第一条留言在计算文本相关性时与完整性时得分同样为 90 分（满分 100 分），但相关性的分数普遍分布在 80 左右，则 90 分对于相关性来说十分优秀；但是完整性的得分普遍均在 95 到 100 分分布，则 90 分对于完整性而言就略显逊色。如果此时将相关性与完整性的分数盲目相加，则依然会导致得分存在差异。如下图分布表示：

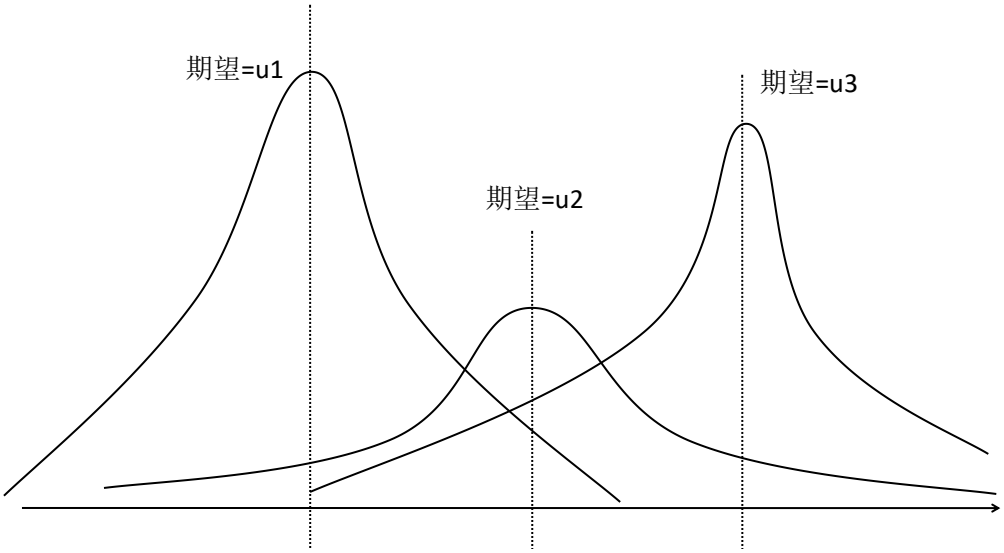

图 11 不同分布指标图

此时的各个指标分数存在差异，本文使用标准正态公式将各个指标的数据进行标准化，使的分布相同，在进行权值的分配。标准正态化公式如下所示：

$$x_z = \frac{x_i - \mu}{\sqrt{\delta^2}}$$

其中考虑到样本数数量均为大数据量，在考虑期望与方差的计算时，本文使用样本的均值与方差进行数据的标准正态化。经过标准正态化后的数据拥有同样的分布，在将期望值修改为权重，完成权值分配。

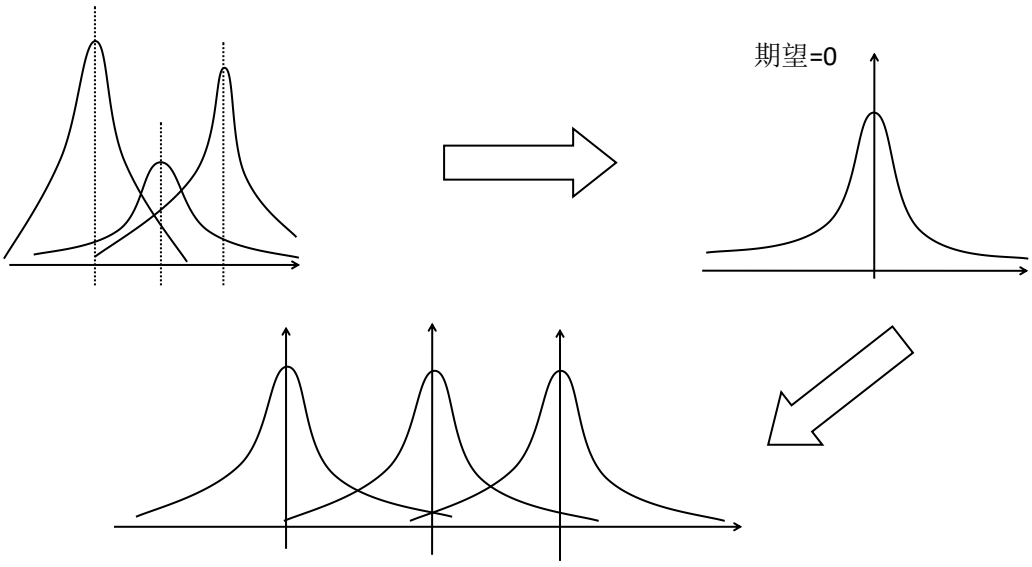

图 12 分布转化图

经过标准正态化后，对每一项指标进行正向轴的平移，即改变期望但是不改变分布形状。如下所示（x1 表示相关性，x2 表示完整性，x3 充实度与可解释性，x4 表示情感表达与行文结构）：

$$x_1 \sim N(30,1), x_2 \sim N(25,1), x_3 \sim N(25,1), x_4 \sim N(20,1)$$

在经过权重弄分配后，得到综合数据。再对得分进行标准化，设置分数为底分 70 分，满分为 100 分，在 70 分到 100 分之间按照比例分配分值。得到部分结果如下所示：

	文件名称	综合得分
具体数值	['0.txt', '1.txt', '2.txt', '3.txt', '4.txt', '5.txt', '6.txt', '7.txt', '8.txt',...,'2814.txt', '2815.txt']	[74.38,72.48,74.75,76.21,79.17,72.19,73.12, 74.22, 73.75,...,75.69, 74.19]

表 18 综合得分表

参考文献

[1]李春林,冯志骥.基于文本挖掘的新能源汽车用户评论研究[J].特区经济,2020(04):148-151.

[2]艾楚涵,姜迪,吴建德.基于主题模型和文本相似度计算的专利推荐研究[J].信息技术,2020,44(04):65-70.

[3]王影,库婷婷,许书萍,李伟强,袁博.敬畏感的情绪成分分析:基于社交网络的文本挖掘[J].心理技术与应用,2020,8(04):235-242.

[4]林毅焜.基于文本挖掘的典型旅游网站的旅游分享研究——以陕西省为例[J].价值工程,2020,39(08):243-247.

[5]姚潇,吴冬晓,庞守林.基于文本挖掘的管理层语调对公司债券信用利差的影响[J].经济理论与经济管理,2020(03):99-112.

[6]哈工大停用表(hit_stopwords.txt).https://github.com/goto456/stopwords/blob/master/hit_stopwords.txt,2019-12-18.

[7]BosonNLP情感词典.https://bosonnlp.com/dev/resource,2014-12-29.

[8]《知网》情感分析用词语集（beta版）.http://www.keenage.com/html/c_index.html,2007-10-22.

附录

1. 主题词与对应概率详细表：

主题	包含主题词	主题词对应概率
主题 0	扰民 施工 中学 噪音 汽车 号线 南站 工地 高速 宾馆	[0.16 0.12 0.10 0.08 0.0483 0.048 0.041 0.0379 0.037 0.006]
主题 1	地铁 拆迁 人行道 占用 楚江情况 黄兴 区 农民 北路	0.17 0.079 0.07 0.05 0.0481 0.038 0.032 0.027 0.016 0.005]
主题 2	企业 请问 教师 实习 整治 待遇 教育局 经济学院 小区 周边环境	[0.1068 0.0872 0.0868 0.046 0.0436 0.0344 0.028 0.0269 0.02459 0.0172]
主题 3	安全隐患 二期 道路 住房 区万境 湖湾 县泉塘 房屋 限购	[0.1366 0.0935 0.070 0.0678 0.0374 0.034 0.028 0.028 0.0263 0.0181]
主题 4	小区 新城 违建 搅拌站 区 请问 县星沙 请求 教师 扰民	[0.112 0.075 0.068 0.066 0.059 0.054 0.0505 0.043 0.0426 0.035]
主题 5	建设 安置 镇 人行天桥 村 请深业 睿 楼盘 东	[0.136 0.053 0.053 0.040 0.0396 0.0351 0.033 0.032 0.0285 0.0282]
主题 6	投诉 成 苑 滨河 景园 红绿灯乱 维权 商品房 出租车	[0.121 0.079 0.066 0.0625 0.0516 0.0401 0.0398 0.0340 0.0266 0.0227]
主题 7	西地省 违规 停车场 店 押金收取 驾校 拖欠工资 变 快递	[0.1953 0.176 0.082 0.0485 0.0305 0.0256 0.0221 0.014 0.0098 0.00414]
主题 8	搅拌站 万金 栋 诈骗 不合理区 区丽发 整治 街 城际	[0.0973 0.0705 0.0611 0.0551 0.054 0.039 0.0310 0.0292 0.0288 0.0279]
主题 9	地铁 拆迁 人行道 占用 楚江情况 黄兴 区 农民 北路	[0.177 0.0796 0.0739 0.0508 0.0489 0.0386 0.0323 0.0275 0.016 0.0053]
主题 10	请 拖欠 小孩 区 供水 民工天鑫城 市鑫 小区 包工头	[3.294e-01 9.3255e-02 2.3251e-02 2.241e-02 1.40716e-02 1.9250e-04 8.0294e-05 7.925e-05 7.253e-05 7.096e-05]
主题 11	建议 道 三期 山水 公交 火车站 区 平台 开工 派出所	[0.2228 0.0983 0.0516 0.051 0.0397 0.0262 0.0257 0.025 0.0237 0.0219]
主题 12	街道 社区 保利 建 区 公寓反对 县星 号 栋	[0.1468 0.106 0.0742 0.0554 0.0553 0.0540 0.050 0.0433

		0.035 0.0165]
主题 13	路 城 大道 东路 车 调整 路口 渣土 扰民 小区	[0.1916 0.1055 0.0674 0.0593 0.039 0.039 0.035 0.0354 0.0237 0.0187]
主题 14	新城 搅拌站 投诉 开发商 公寓 反对 小区 万科 魅力之城	[0.086 0.074 0.068 0.06 0.053 0.038 0.035 0.031 0.029 0.0238]