基于自然语言处理的智慧政务系统构建

摘要

本项目旨在利用自然语言处理和文本挖掘的方法构建智慧政务系统的功能，包括建立中文文本多分类模型对群众留言进行分类处理、挖掘和整理热点问题和制定针对答复意见的质量评价方案，解决人工划分留言和整理热点时工作量大、效率低等问题，对提升政府管理水平和施政效率具有重大的意义。

针对问题 1，建立基于 BERT（Bidirectional Encoder Representation from Transformers）模型的分类模型，对群众留言进行分类。首先，使用 BERT 模型训练短文本数据，再通过 Softmax 分类器对训练得到的词向量进行分类处理，最终对模型性能进行评价，得到测试的 F1-Score 为 **0.959**。

针对问题 2，定义热度评价指标，挖掘热点问题。首先，对群众留言进行数据清洗、中文分词和特征提取，其次，计算每两则留言内容之间的文本相似度，并进行 DBSCAN 聚类，最终，**定义热度评价指标**为留言数量、点赞数和反对数的线性组合，系数分别取 0.4、0.6，计算热度指数，得到排名前 5 的热点问题，且热度指数分别为 **3908.4、1065.4、417、236.8 和 123**。

针对问题 3，通过定义关于答复意见的质量评价指标，制定评价方案。首先，对于答复意见进行数据预处理，提取文本特征后，通过 **TF-IDF 算法**计算文本相似度，其次，定义相关性指标由 **Jaccard 相似系数**表征，完整性指标由文本长度表征，时效性指标由留言时间和答复时间的间隔表征，最终，通过 **SVM 分类器**划分高质量与低质量的答复意见，并且得到三个评价指标的权重分别为 **0.16、0.67 和-0.12**，偏置项系数为 0.42，即为针对答复意见的质量评价模型。

关键词：BERT 模型；Softmax 分类器；热度指数；DBSCAN 聚类；质量评价模型

Abstract

The purpose of this project is to use natural language processing and text mining methods to build the functions of the intelligent government system, including the establishment of Chinese text multi classification model to classify and deal with the public message, mining and sorting out the hot issues and formulating the quality evaluation scheme for the reply opinions, solving the problems of heavy workload and low efficiency when manually dividing the message and sorting out the hot issues, so as to improve the government management Management level and governance efficiency are of great significance.

Aiming at problem 1, a classification model based on Bert (bidirectional encoder representation from transformers) model is established to classify the public comments. First of all, use the Bert model to train the short text data, and then use the softmax classifier to classify the word vectors, and finally evaluate the performance of the model. The F1 score of the test is **0.959**.

For problem 2, define the heat evaluation index and mine hot issues. First of all, data cleaning, Chinese word segmentation and feature extraction are carried out for the mass message. Secondly, text similarity between each two message contents is calculated, and DBSCAN clustering is carried out. Finally, **the heat evaluation index is defined** as the linear combination of the number of messages, the number of likes and dislikes. The coefficient is 0.4 and 0.6 respectively, and the heat index is calculated to get the top 5 hot issues, and the heat index is obtained They were **3908.4,1065.4,417,236.8 and 123** respectively.

In view of question 3, the evaluation scheme is formulated by defining the quality evaluation indexes of the reply opinions. First of all, we preprocess the response data, extract the text features, and calculate the text similarity through **TF-IDF algorithm**. Secondly, we define the correlation index as represented by **Jaccard similarity coefficient**, integrity index as represented by text length, timeliness index as represented by message time and response time interval. Finally, we use **SVM classifier** to divide high-quality and low-quality response opinions The weight of the three evaluation indexes is **0.16, 0.67, - 0.12**, and the coefficient of bias term is 0.42, which is the quality evaluation model for the reply.

Key word : Bert model ; Softmax classifier ; Heat index ; DBSCAN clustering ; Quality evaluation model

目录

1 问题重述

1.1 背景概述

随着互联网的快速发展和普及，网络已成为党和政府了解民情、听取民声、汇聚民智、凝聚民气，科学决策、民主决策的重要渠道，同时也是公众行使知情权、参与权、表达权和监督权的重要途径。微信、微博、市长信箱、阳光热线等网络问政平台汇集有关社情民意的信息量也日益庞大，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。因而，利用云计算、人工智能、数据挖掘等，开发基于自然语言处理技术的智慧政务系统，以提高相关机构办公、监管、服务、决策，已成为社会治理创新发展的新趋势。

1.2 问题提要

1.2.1 群众留言分类

根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型，并使用 F-Score 对模型进行评价。

1.2.2 热点问题挖掘

根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按表 1.1 的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 1.2 的格式给出相应热点问题对应的留言信息，并保存为“热点问题留言明细表.xls”。

表 1.1 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	…	2019/08/18 至 2019/09/04	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
2	2	…	2017/06/08 至 2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
…	…	…	…	…	…

表 1.2 热点问题留言明细表

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
1	360104	A012417	A 市魅力之城商铺无排烟管道，小区内到处油烟味	2019/08/18 14:44:00	A 市魅力之城小区自打交房入住后，底层商铺无排烟管道，经营餐馆导致大量油烟排入小区内，每天到凌晨还在营业……	0	0
1	360105	A120356	A5 区魅力之城小区一楼被搞成商业门面，噪音扰民严重	2019/08/26 08:33:03	我们是魅力之城小区居民，小区朝北大门两侧的楼栋下面一楼，本来应是架空层，现搞成商业门面，噪声严重扰民，有很大的油烟味往楼上窜，没办法居住……	1	0
1	360106	A235367	A 市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦	2019/08/26 01:50:38	2019 年 5 月起，小区楼下商铺越发嚣张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音严重影响了小区居民休息……	0	0
…	…	…	…	…	…	…	…
1	360109	A0080252	魅力之城小区底层门店深夜经营，各种噪音扰民	2019/09/04 21:00:18	您好：我是魅力之城小区的业主，小区临街的一楼是商铺，尤其是餐馆夜宵摊等，每到凌晨都还在营业，每到晚上睡觉耳边都充斥着吆喝……	0	0
2	360110	A110021	A 市经济学院寒假过年期间组织学生去工厂工作	2019/11/22 14:42:14	西地省 A 市经济学院寒假过年期间组织学生去工厂工作，过年本该是家人团聚的时光，很多家长一年回来一次，也就过年和自己孩子见一次面，可是这样搞……	0	0
2	360111	A1204455	A 市经济学院组织学生外出打工合理吗？	2019/11/5 10:31:38	学校组织我们学生在外边打工，在东莞做流水线工作，还要倒白夜班。本来都在学校好好上课，十月底突然说组织到外省打工……	1	0
…	…	…	…	…	…	…	…
2	360114	A0182491	A 市经济学院变相强制实习	2017/06/08 17:31:20	系里要求我们在实习前分别去指定的不同公司实训，我这的公司的工作内容和老师之前介绍以及我们专业几乎不对口，不做满 6 个月不给实训分，不能毕业……	9	0

1.2.3 答复意见的评价

针对附件 4 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。

2 问题 1：群众留言分类

2.1 问题分析

针对问题 1，根据附件 2 给出的数据，由于留言内容具有长度短、特征稀疏及上下文依赖性等短文本数据的特点，采用朴素贝叶斯（Nave Bayes，NB）和支持向量机（Support Vector Machines，SVM）等传统的文本分类方法直接进行分类的效果不佳，因此，本项目使用基于双向 Transformer 大规模预训练语言模型（Bidirectional Encoder Representation from Transformers，BERT）的分类模型，对留言内容进行分类，并通过 F1-Score 对模型性能进行评价，结果显示，F1-Score 高达 0.959。

2.2 数据准备

提取附件 2 数据中的留言主题、留言详情及一级分类标签，并将留言主题和留言详情的文本内容合并，切分训练集、验证集和测试集，完成数据准备工作，过程在文件 processing.py 中，接着将 训练集、验证集和测试集放在 THUCnews 文件夹的子文件夹 data 中。

2.3 模型建立

2.3.1 算法简介

BERT 模型基于 Transformer，是一种新的语言表示模型，它摒弃了常用的卷积神经网络（Convolutional Neural Networks，CNN）或者循环神经网络（Recurrent Neural Network，RNN）模型，采用 Encoder-Decoder 架构。其本质就是一个预训练结构，先通过利用大量原始的语料库训练，得到一个泛化能力很强的模型，再进行微调参数训练，将模型应用到任意的特定任务中[1]。

BERT模型基本结构[2]如图1.1所示，主要分为两部分，其中，$E_1,E_2,...,E_n$表示输入层，Transformer表示双向编码层，即在处理一个词时，可以根据上下文的语义关系，表征字在上下文中的具体语义。另外，$T_1,T_2,...,T_n$表示输出层。

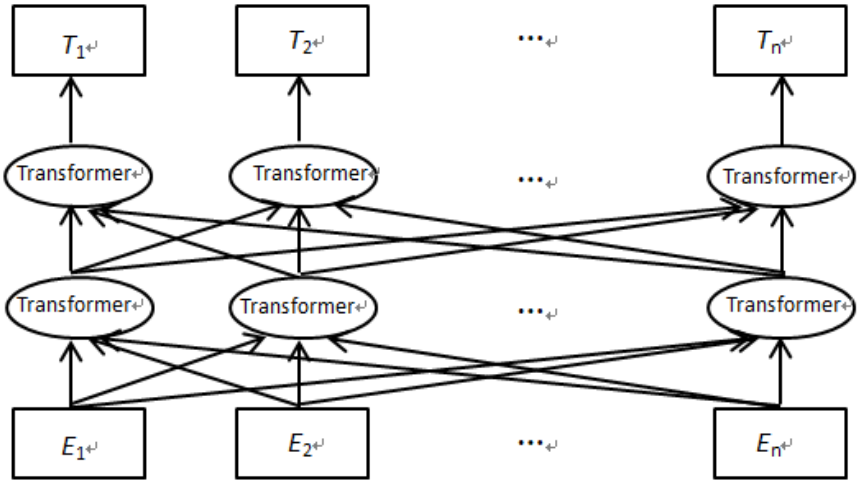


图1.1 BERT模型基本结构图

预训练是BERT模型的一个重要阶段，通过对海量语料的训练，使得单词学习到很好的特征表示。通过BERT模型训练得到文本的向量表示W：

$$W^{(i)}=\{w_1^{(i)},w_2^{(i)},...,w_n^{(i)}\}$$

其中，W(i)表示第i则文本的向量矩阵，w(i)表示单个字的表征向量，n表示最大句子长度（max_seq_length）。

2.3.2 分类模型

针对问题 1，本项目建立的分类模型是基于 BERT 模型，分类器的结构如图 1.2 所示，主要分为两部分：先通过 BERT 模型训练获取每则文本的语义表示，再将文本中每个字的向量表示输入到 Softmax 层进行分类处理，并输出文本标签。

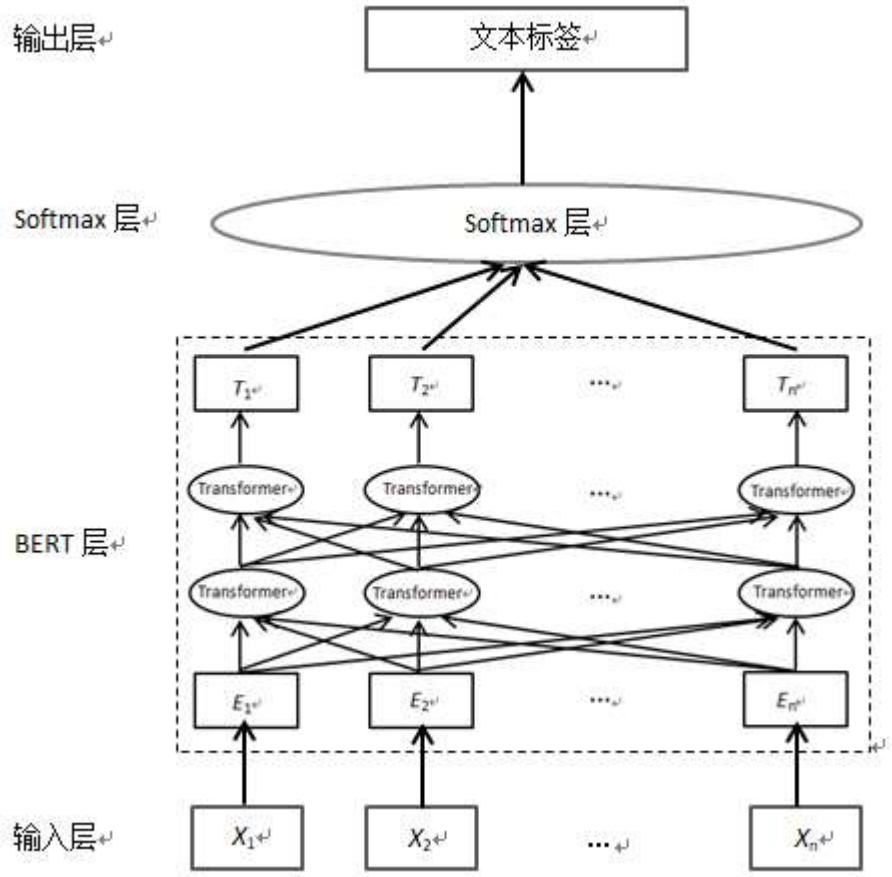


图 1.2 基于 BERT 模型的分类器结构图

2.4 模型求解

2.4.1 实验环境

本项目采用的环境如表 1.1 所示：

表 1.1 实验环境

名称	值
CPU	Intel(R) Core(TM) i5-6360U CPU @ 2.00GHz
内存/GB	16.0
编程语言	Python 3.6

2.4.2 数据说明

本项目使用的数据集是来自互联网公开来源的群众问政留言记录，针对问题 1，即

将按照一级分类体系，对群众留言内容进行多分类训练。

2.4.3 预训练模型说明

本项目涉及一些参数设置，在 BERT 层采用了预训练好的 bert_Chinese 模型，下载地址为 https://s3.amazonaws.com/models.huggingface.co/bert/bert-base-chinese.tar.gz，将模型载入后，可以直接输出训练好的字向量或句向量，本项目使用该模型获取句向量，并将其作为后续网络模型的输入，另外，对应的词表下载地址为 https://s3.amazonaws.com/models.huggingface.co/bert/bert-base-chinese-vocab.txt。

2.4.4 模型训练

由于 BERT 模型要求处理每条数据的长度，使其均等于数据中最长文本的长度，将留言主题与留言详情合并之后的文本数据长度较长，容易导致单词索引大于 vocab_size，综合考虑文本数据长度范围，选取 335 作为短填长切的长度。另外，当 batch_size 取 32 时，由于 pad_size 取 300 的值较大，导致程序在运行时出现内存不够的现象，因此，batch_size 只能默认取 16。

经过调参，测试集的查准率、查全率及 F1-Score 如表 1.2 所示，结果显示，当 batch_size 取 16，Learning rate 为 3e-5 时，测试集的 F1-Score 值最高，近似达到 0.959。

表 1.2 查准率、查全率及 F1-Score 汇总表

batch_size	Learning rate	Precision	Recall	F1-Score
16	2e-5	0.929	0.918	0.918
16	3e-5	0.961	0.959	0.959
16	5e-5	0.186	0.233	0.178

2.5 模型评价

2.5.1 评价指标的定义

关于分类模型性能的评估指标，有准确率（Accuracy）、查准率（Precision）、查全率（Recall）、F1-Score 值及 AUC 等，前四个均可以通过混淆矩阵（confusion matrix）直接计算得到，本项目主要考虑 F1-Score，其中，有关 Precision 与 Recall 的定义公式如下：

$$\begin{cases} \mathrm{Precision} = \dfrac{true\ positives}{num\,of\ \ predicted\ positive} \\ \mathrm{Recall} = \dfrac{true\ positives}{num\,of\ actual\ positive} \end{cases}$$

查准率表示在预测为 positive 的样本中真实类别为 positive 的样本所占比例，召回率表示在真实为 positive 的样本中模型成功预测出的样本所占比例[3]，并且，由查准率与查全率可计算得到 F1-Score 值，公式如下：

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

其中，P_i 表示第 i 类的查准率，R_i 表示第 i 类的查全率，$i=1,2,\ldots,15$，分别代表关于群众留言内容一级分类的 15 个类标签。

2.5.2 总结

综上所述，基于 BERT 模型的文本分类器实践过程，在调整 batch_size、Learning rate 两个参数之后，结果显示，当 batch_size 取 16，Learning rate 取 3e-5 时，多分类模型的性能最高，并且，测试集的查准率及查全率分别为 0.000 和 0.000，F1-Score 高达 0.959。

3 问题 2：热点问题挖掘

3.1 问题分析

针对问题 2，根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，在对留言进行预处理之后，进行文本特征提取，计算留言两两之间的文本相似度，选择基于密度聚类的 DBSCAN 聚类算法对留言进行归类。

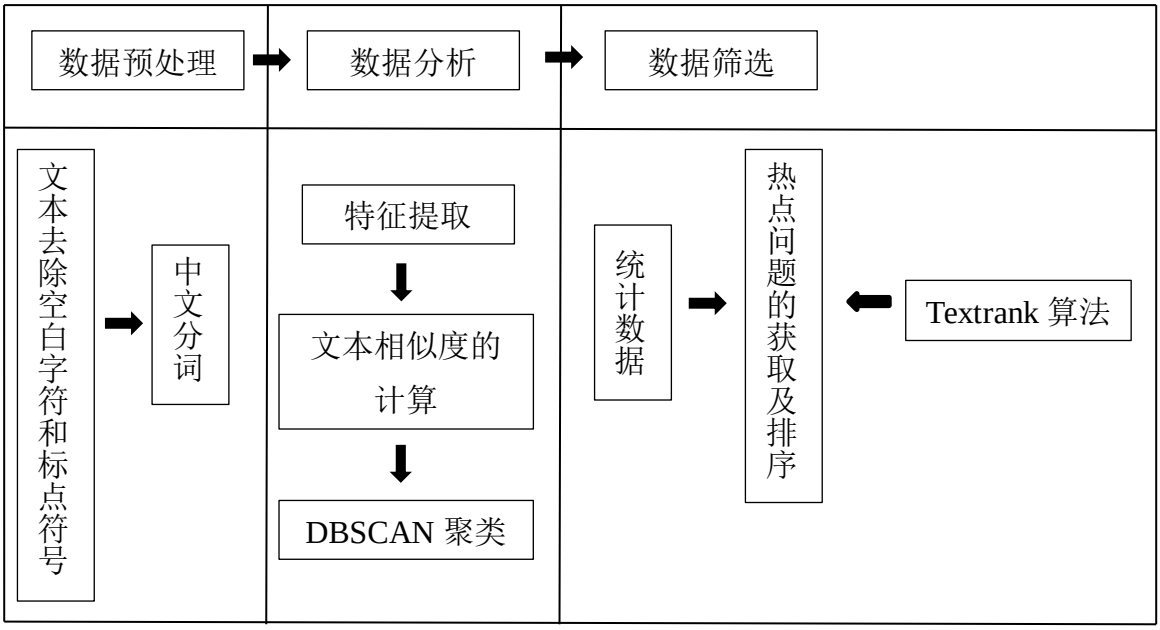
接着，根据热度评价指标的定义计算热度指数，根据热度指数的排序，得到排名前 5 的 5 个簇，即前 5 个热点问题的所有留言信息，获取各个簇中的留言主题，利用 TextRank 算法得到每个簇中影响最大的一个留言主题作为问题描述，接着利用基于 jieba 的 TextRank 算法对问题描述进行以人名、地名词性筛选、关键词提取，完成实体识别，进而得到相关的人群或者地名，最后给出排名前 5 的热点问题，保存为文件“热点问题表.xls”，给出相应热点问题对应的留言信息，并保存为‘热点问题留言明细表.xls’。

3.2 模型建立

热度问题是指某一时间段内群众集中反映的某一问题。定义合适的热度评价指标对热点问题的重要性进行排序，热度指标的定义如公式 3.2-1 所示。

热度评价指标 = 0.4 * 留言数量 + 0.6 *（点赞数 + 反对数）　　（公式3.2-1）

3.3 总体流程图



3.4 模型求解

3.4.1 数据预处理

3.4.1.1 数据清洗

读取群众留言的所有信息后，提取留言主题和留言详情结合的文本作为聚类的样本。查看文本可以发现，文本中存在很多空白字符，而且多数标点符号也会对聚类的效果产生影响，所以，需要将文本中的空白字符和标点符号进行清除。

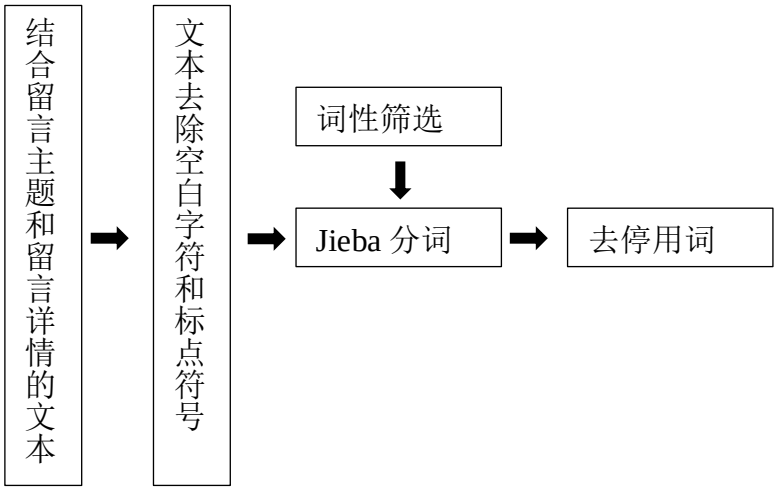
3.4.1.2 中文分词

文本相对计算机而言是无法识别的，因此，要把文本转换成计算机能够识别的结构化信息。为了计算机便于转换信息，首先，要对文本进行中文分词。中文分词是指将汉字序列按照一定规则、逐个切分为单独的词序列。Jieba 是目前最好的 Python 用于中文分词的组件。jiaba 分词是基于前缀字典实现词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图，同时利用动态规划查找最大概率路径，并找出基于词频的最大切分组合。对于未登录词，jiaba 工具采用了基于汉字成词功能的 HMM 模型，使得中文分词效果更好[4]。

对文本进行中文分词时，为了能够准确分辨文本、正确分词，指定了名词、动词、人名、地名、处所词、方位词等词性进行筛选。

为了更好地进行分词，对 jieba 分词后得到的关键词进行去停用词，删除一些不是特别重要的信息。

3.4.1.3 预处理流程图



3.4.2 文本特征提取

通过文本预处理得到文本的关键词之后，需要对他们进行特征提取，即将文本转换为计算机能够识别的信息。主要的过程有：首先，使用 CountVectorizer()函数，将文本的关键词转换词频矩阵，矩阵元素 a[i][j]，表示 j 词在 i 类文本下的词频；接着使用 TfidfTransformer()函数统计每个关键词的 tf-idf 权值，把 tf-idf 矩阵提取出来，矩阵中元素 w[i][j]表示 j 词在 i 类文本中的 tf-idf 权重，以下是 TfidfTransformer()函数的相关介绍。

TfidfTransformer()函数基于 tf-idf 算法。tf-idf 算法是一种用于文本挖掘的常用加权技术，用以评估一个词语对于一个语料库的其中某一篇文章的重要程度。若某个词语在一篇文章中出现的频率高，并且在其他文章中极少出现的话，则说明该词具有很好的类别区分能力。tf-idf 算法包括两个部分——tf 和 idf，由两者相乘即可得到 tf-idf 算法[5]。

词频(tf)表示词语在语料库中出现的频率，词频(tf)的计算公式如公式 3.4.1-1 所示。

$$词频(tf) = 某个词语在文章中的出现次数 \quad (公式3.4.1-1)$$

若用词频决定一个词语的重要程度，通常情况下，同一个词语在长文本的词频会比在短文本的词频高，因此，需要将词频标准化，如公式 3.4.2-2 所示。

$$词频(tf) = \frac{某个词语在文章中的出现次数}{文章的总词数} \quad (公式3.4.1-2)$$

逆向文档频率(idf)表示用于调整词频的权重系数，逆向文档频率(idf)的计算公式如公式 3.4.2-3 所示。

$$逆向文档频率(idf) = \log(\frac{语料库的文章总数}{包含该词的文章数+1}) \quad (公式3.4.1-3)$$

分母之所以用包含该词的文章数+1 代替是为了避免某词可能从来没有在语料库中出现，而导致分母为 0。从公式 3.4.2-3 可以看出，若一个词语越常见，分母就会越大，逆向文档频率(idf)就会越小，越接近于 0。

最后计算出每个词语的 tf-idf 值，如公式 3.4.2-4 所示。

$$tf\text{-}idf = tf \cdot idf \quad (公式3.4.1-4)$$

3.4.3 文本相似度的计算

通过文本特征提取之后，可以得到 tf-idf 矩阵，计算 tf-idf 矩阵的余弦相似度，得到的结果就是留言之间的文本相似度，更有利于进行 DBSCAN 文本聚类。

3.4.4 基于 DBSCAN 聚类算法的留言聚类

随着网络的不断发达，网络问政平台给广大人民提供了一个表情达意的机会，然而从众多的留言记录中，想要人工确定聚类中心是相当难的，因此，选用基于密度的 DBSCAN 算法[6]进行留言聚类是一种较为可行的方法，这种算法的优点在于可以不用设置聚类中心的个数，而且聚类速度快、能够有效处理噪声点以及发现任意形状的簇，同时，还可以过滤样本中的离群点，离群点不会被分在任意一个簇中，减少聚类的偏差。

3.4.4.1 DBSCAN 聚类算法的主要参数

ε：参数 ε 是由用户指定的每个对象的领域半径值，$\varepsilon > 0$ 。

MinPts：每个簇的最少含量。

3.4.4.2 DBSCAN 算法流程

(1)输入待聚类的数据集、ε、MinPts；

(2)标记所有样本为 unvisited；

(3)随机选取一个 unvisited 样本 p；
(4)标记 p 为 visited；
(5)如果 p 的 ε -领域至少有 MinPts 个样本
(6)创建一个新簇 C，并把 p 添加到 C；
(7)令 N 为 p 的 ε -领域中的样本集合
(8)循环 N 中的每个样本 p
(9)如果 p 是 unvisited；
(10)标记 p 为 visited；
(11)如果 p 的 ε -领域至少有 MinPts 个样本，把这些样本添加到 N；
(12)如果 p 还不是任何簇的成员，把 p 添加到 C；
(13)结束循环；
(14)输出 C；
(15)否则，标记 p 为噪点；
(16)直到没有标记为 unvisited 的样本。

3.4.4.3 DBSCAN 聚类部分结果

对留言内容进行 DBSCAN 聚类，部分结果如图 3.4.4-1 所示。

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	label
360114	A0182491	A市经济学院体育学院变相强制实…	2017-06-08 17:31:20	…	9	0	0
289408	A0012413	在A市人才app上申请购房补贴为…	2018-11-15 16:07:12	…	0	0	1
336608	A0005623	希望西地省把抗癌药品纳入医保…	2019-09-08 21:01:59	…	0	0	-1
360103	A0012425	A5区劳动东路魅力之城小区临街…	2019-09-25 00:31:33	A5区劳动东路魅力之城小区临街…	1	0	2
323149	A1241141	请给K3县乡村医生发卫生室执业…	2019-06-20 20:38:47	…	0	0	-1
360107	A0283523	A5区劳动东路魅力之城小区一楼…	2019-07-21 10:29:36	局长：　你好，A5区劳动…	3	0	3
360108	A0283523	A5区劳动东路魅力之城小区一楼…	2019-08-01 16:20:02	局长：　你好，A5区劳动…	6	0	3
343985	A108051	A市能否设立南塘城轨公交站？	2019-10-31 21:19:59	…	0	0	-1
286572	A23525	请求A市地铁2#线在梅溪湖CBD处…	2018-10-27 15:13:26	…	3	0	-1
316619	A235259	请问A市什么时候能普及5G网络？	2019-05-14 11:22:13	…	0	0	-1
360100	A324156	魅力之城小区临街门面油烟直排…	2019-09-05 12:29:01	魅力之城小区楼下烧烤摊、快餐…	3	0	2
360101	A324156	A5区劳动东路魅力之城小区油烟…	2019-07-28 12:49:18	尊敬的政府：A5区劳动东路魅力…	4	0	2
360110	A110021	A市经济学院寒假过年期间组织学…	2019-11-22 14:42:14	…	0	0	0
323034	A012414	L市物业服务收费标准应考虑居民…	2019-06-19 17:46:24	…	0	0	4

图 3.4.4-1

从图 3.4.4-1 可以看出，每条留言都会得到各自的标签，标签为-1 的留言属于离群点，说明与该留言相似的留言太少，以至于无法构成热点问题。

完成对留言的 DBSCAN 聚类后，根据热度评价指标计算热度指数，根据热度指数的排序，得到前 5 个热点问题的所有留言信息，并整合到热点留言问题明细表.xls 中，部分结果如图 3.4.4-2 所示。

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
1	188006	A0001029	A3区一米	2019/2/28	座落在A市	0	0
1	188007	A0007479	咨询A6区	2019/2/14	A市A6区追	0	1
1	188031	A0004006	反映A7县	2019/7/19	本人系春华	0	1
1	188119	A0003502	对A市地铁	2019/5/27	我是一名在	0	0
1	188170	A8801132	A市6路公	2019/12/23	12月21日下	0	0
1	188249	A0008408	A3区保利	2019/9/17	保利麓谷林	0	0
1	188251	A0001309	A7县特立	2019/10/19	近来，下午	0	0
1	188399	A0009793	A市利保壹	2019/7/3 6	您好，我想	0	0
1	188455	A0003590	咨询异地办	2019/5/16	书记您好！	0	0
1	188467	A0005018	投诉A市温	2019/3/28	退费之日起	0	1

图 3.4.4-2

3.4.5 使用 TextRank 算法为热点问题生成摘要和关键词

TextRank 算法[7]基于 Fgoogle 的 PageRank 算法，将主要用于为文本生成权重最高的摘要和关键词。

TextRank 算法自动生成文本摘要的基本过程是：将文本中的每个句子看成 PageRank 图中的一个节点，计算两个节点之间的相似度，若相似度大于指定的阈值，说明这两个句子是有关联的。相似度为两个节点之间无向边的权值，利用 PageRank 算法即可得到句子的权重，把权重最高的句子作为文本的摘要。

TextRank 算法自动生成关键词和生成文本摘要的过程类似，只是关键词要把句子先分词和词性筛选才能计算两个节点之间的相似度。

完成对留言的 DBSCAN 聚类后，根据热度评价指标计算热度指数，根据热度指数的排序，得到前 5 个热点问题的所有留言信息。接着利用 TextRank 算法生成摘要和关键词作为热点问题表的问题描述和相关的人群或地点，完善并得到热点问题表如表 3.4.5-1 所示。

表 3.4.5-1

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	3908.4	2020/1/8 9:32:33 至 2017/06/08 17:31:20	问题社保卡办理	咨询 A 市办理社保卡问题
2	2	1065.4	2019/4/7 18:25:52 至 2019/3/12 11:19:22	问题御府子女	请 A3 区协调解决旭辉御府业主子女入学问题
3	3	417	2019/9/6 18:36:16 至 2019/8/23 14:21:38	距长小区高铁	A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到，合理吗？
4	4	236.8	2019/9/30 14:38:58 至 2019/1/14 20:03:10	物业河畔小区	A1 区 A9 市河畔小区物业很差

5	5	123	2020/1/3 16:29:11 至 2019/1/3 10:20:02	月亮岛路	关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉

4 问题 3：答复意见的评价

4.1 问题分析

答复意见的评价方案主要是通过对所给的留言信息和答复意见建立模型，通过模型评估给出质量评价。

通过附件 4 提取留言意见与答复文本的文本相似度，答复文本的长度、留言与答复的间隔天数，接着将他们标准化，人工为附件 4 加上标签，构建分类器，获取权重 w 和截距 b 完善评价模型。

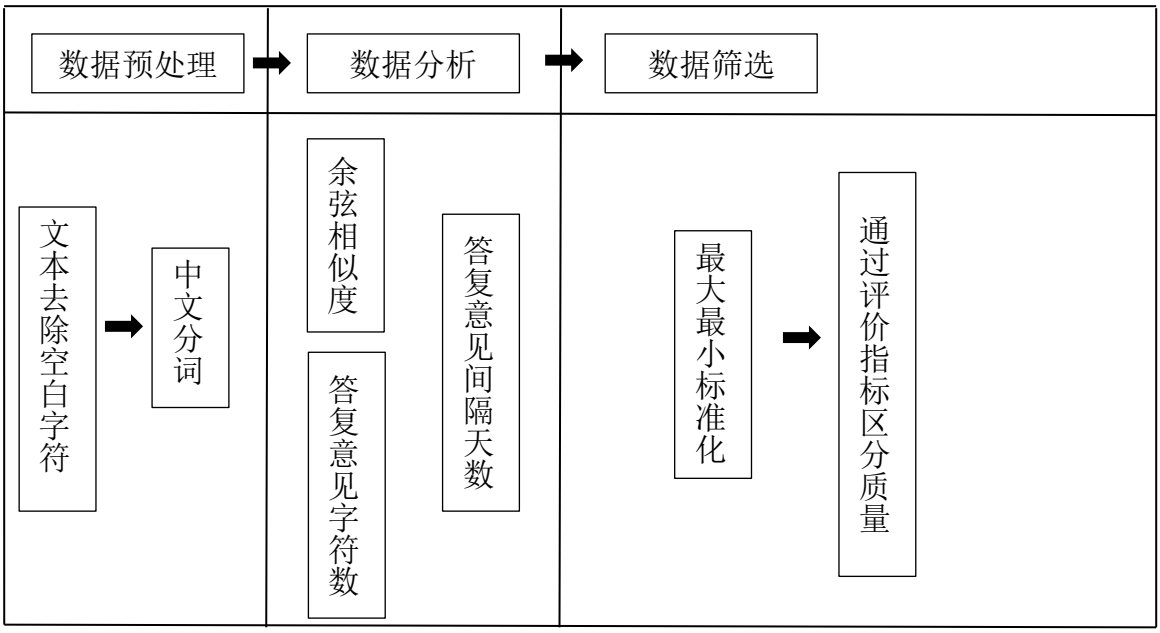
（1）相关性：留言信息和答复意见的相关性决定了答复意见的质量，答复意见需要对留言信息进行针对性的回答，因此计算留言信息和答复意见的文本相似度。

（2）完整性：对答复意见的文本长度进行评价，答复意见较短，内容可能会不够详细，信息量不足，不能完全解答留言意见，所以答复文本程度在一定程度上体现答复意见的质量。

（3）时效性：对答复意见的留言时间进行评价，越早被处理的信息，一定程度上可以体现工作者对留言的重视程度，因此答复意见时间越短，质量可能越高。

（4）组合特征，构建基于多指标的答复意见质量评价模型。

4.2 总体流程图



4.3 数据预处理

读取答复意见的所有信息，提取留言意见和答复意见发现有一些多余的空白格影响模型评价，因此对空白字符进行清除。统一时间的格式，对时间进行读取。

4.4 模型建立

4.4.1 模型简介

答复意见评价模型根据相关性、完整性、时效性三项指标建立。分别对应留言意见和答复意见的文本相似度、以及答复意见的文本长度、间隔时间，评估公式为：

$$Score = W1 * score1 + W2 * score2 + W3 * score3 + b \quad （公式4.4.1-1）$$

其中 score1 为留言意见和答复意见的文本相似度，score2 为答复文本长度，score3 为时间间隔长度。W1 为 score1 权重，W2 为 score2 权重，W3 为 score3 权重。b 为截距。

4.4.2 相关性指标

答复意见相关性指标是通过计算留言信息与答复意见的文本相似度求相关性。答复意见需要对留言信息进行针对性的回答，因此计算留言信息和答复意见的文本相似度。在求文本相似度过程中，使用 **TF-IDF 算法，**指标使用 Jaccard 相似系数，求留言和答复意见的相关性。

Jaccard 相似系数（Jaccard similarity coefficient）用于比较有限样本集之间的相似性与差异性。Jaccard 系数值越大，样本相似度越高。与 Jaccard 系数相关的指标是 Jaccard 距离，用于描述集合之间的不相似度，Jaccard 聚类越大，样本相似度就越低[8]。该问题我们使用 JS 距离计算留言和答复建议的相关性，即 sa 和 sb 的相似度（sa：留言意见，sb：答复意见）：

$$score1 = len（sa \ \& \ sb）/ \ len（sa \ | \ sb） \quad （公式4.4.2-1）$$

4.4.3 完整性指标

完整性评价即对答复意见的文本长度进行评价。答复意见较短，内容可能会不够详细，信息量不足，不能完全解答留言意见，所以答复文本程度在一定程度上体现答复意见的质量。留言越长，越详细，则得分越高。

score2=文本长度

4.4.3 时效性指标

时效性指标即评价留言时间和答复时间的间隔。间隔时间越短，则得分越高。对答复意见的留言时间进行评价，越早被处理的信息，一定程度上可以体现工作者对留言的重视程度，因此答复意见时间越短，质量可能越高。间隔天数：获取留言与答复的时间间隔，将所有留言的时间格式设置一致。

score3=答复时间-留言时间

4.4.4 标准化

分别使用最小最大标准化，z-score 标准化，消除量纲影响。最小最大标准化也称为离散标准化，是对原始数据的线性变换，将数据值映射到[0,1]之间。离差标准化保留了原来数据中存在的关系，是消除量纲和数据取值范围影响的最简单方法。公式为：

$$X* = (x - x.min) / (x.max - x.min) \quad \text{（公式4.4.4-1）}$$

零-均值规范化也称标准差标准化，经过处理的数据的均值为 0，标准差为 1。转化公式为：

$$X* = (x - \mu) / \delta，x = (a - a.mean()) / a.std() \quad \text{（公式4.4.4-2）}$$

其中，μ 为原始数据的均值，δ 为原始数据的标准差，这是当前用得最多的数据标准化方式。

4.4.5 分类器分类

针对附件 4 的数据，人工设置标签，用 0-1 分别表示低质量与高质量的留言信息。针对问题 3，本项目采用机器学习常用算法 SVM、逻辑回归模型、随机梯度下降和决策树构建了答复意见质量评价分类器。不同的分类器的分类效果不同，可以比较其结果挑选最适合的一种分类方法。挑选评分最高的一种分类器计算其权重和截距，建立评分公式：

$$Score = W1 * score1 + W2 * score2 + W3 * score3 + b \quad \text{（公式4.4.5-1）}$$

4.5 模型求解

4.5.1 分类器算法评分

表 4.5.1-1

分类器	SVM	逻辑回归模型	随机梯度下降	决策树
最小最大标准化-算法评分	0.83	0.70	0.73	0.7
0-1 标准化-算法评分	0.87	0.87	0.80	0.73

由结果可知，SVM 分类器是几种模型中算法评分是最高的，因此，选用 SVM 分类器对应的权重和截距建立评分公式。

4.5.2 SVM 分类器所得权重和截距

表 4.4.2-1

W1	W2	W3	b
0.16	0.67	-0.12	0.42

由表格 4.4.2-1 可知，W1=0.16，W2=0.67，W3=-0.12，b=0.42.其中，W1、W2 为正数，W3 为负数，表示了在答复意见评估公式中 score1（文本相似度），score2（答复意见完整性）是正相关关系，score1，score2 得分越高，总得分越高。Score3（时效性）与总得分呈负相关，即 score3 越高，总得分越低。又因为 W1、W2、W3 中，W2 的绝对值最大，表示答复意见完整性是三项指标中对总得分影响最大。

4.5.2 建立评估公式

$$Score = 0.16 * score1 + 0.67 * score2 - 0.12 * score3 + 0.42 \quad （公式4.5.2）$$

即为针对答复意见的质量评价模型。其中，score1 代表相关性指标，score2 代表完整性指标，score3 代表时效性指标。

综上，实现质量评价模型的建立与求解，通过三个评价指标，实现针对答复意见的质量评价，对于相关部门工作质量的提高具有一定的价值。

5 模型分析与改进

针对问题 1，对群众留言进行中文文本多分类训练，本项目采用的是基于 BERT 模型的分类模型，首先，使用 BERT 模型训练短文本数据，再通过 Softmax 分类器对训练得到的词向量进行分类处理，最终对模型的分类效果进行评价时，得到测试的 F1-Score 为 0.959，即模型的性能较高。

但就分类模型本身来说，从 BERT 层向 Softmax 层传输词向量的结构较为简单直接，若要改进模型，可尝试提取从 BERT 层获得的词向量的上下文相关特征进行深度学习，再对提取出的信息分配权重，突出重点信息，最终输出到 Softmax 层，实现中文文本多分类。

针对问题 2，对相关人群或地点的识别的准确率还不是很高，可以将基于 jieba 的 TextRank 算法换为其他实体识别有关的算法。

针对问题 3，对答复意见的质量评价模型过程中，我们手动为部分数据设立标签，使用不同的标准化方式和不同的分类器挑选得分最高的一项求其权重和截距，但其具有一定的主观性，因此在使用分类器分类过程中会存在误差。同时，在对全部数据进行评分过程中，因为还存在其他的影响因素，所以也会存在误差。但是使用该评分公式能快速大量的对不同的数据进行评分，筛选出不同质量的答复意见，减轻工作量，减少人为评价质量时间。改进方法可从质量评价指标方面设立更多的有效指标，类似可解释性指标、文本情感指标等，可以减少误差，更有效的区分不同质量的留言评价。

6 参考文献

[1]於张闲,胡孔法.基于 BERT-Att-biLSTM 模型的医学信息分类研究*.计算机时代[J].2020 年.第 3 期:1-4.

[2] Jacob Devlin,Ming-Wei Chang,Kenton Lee,Kristina Toutanova.BERT:Pre-training of Deep Bidirectional Transformers for Language Understanding[J].2018.

[3]wuliytTaotao.混淆矩阵、准确率、精确率/查准率、召回率/查全率、F1 值、ROC 曲线的 AUC 值[EB].https://www.cnblogs.com/wuliytTaotao/p/9285227.html，2020.4.9.

[4] 爱知菜.对 Python 中文分词模块结巴分词算法过程的理解和分析 [EB].https://blog.csdn.net/rav009/article/details/12196623,2020.4.15

[5] lyn5284767.sklearn 基础（一）文本特征提取函数 CountVectorizer()和 TfidfVectorizer()[EB].https://blog.csdn.net/lyn5284767/article/details/85316931,2020.4.17
[6] owolf.DBSCAN 聚类算法[EB].https://www.jianshu.com/p/d2eddc733c4d,2020.4.23.
[7] 机器学习与数据挖掘.使用 TextRank 算法为文本生成关键字和摘要[EB].https://www.letianbiji.com/machine-learning/text-rank.html,2020.4.27
[8] 华校专,王正林.Python 大战机器学习：数据科学家的第一个小目标.北京:电子工业出版社,2017.