基于自然语言处理的“智慧政务”文本挖掘模型

摘　要

在大数据时代，随着微信、微博、阳光热线等网络问政平台逐渐成为政府了解民意、解决民生问题等的重要渠道，各类民意相关问题的文本数据量也随之激增，依靠政府工作人员对各类留言问题分类以及整理的传统方式面临着前所未有的挑战，智能处理政务需求日益激增，另一方面也由于深度学习在自然语言处理（Natural Languague Processing）领域日新月异的发展，建立基于自然语言处理的智慧政务系统已经是社会治理的创新趋势。

对于本题给出的“智慧政务中文本挖掘应用”，我们对问题重新剖析，进一步分析处理并重新定义为：自然语言处理文本分类问题，使用Pytorch框架建立了TextCNN模型 ，在海量数据中可以快速定位并对数据进行分类、给出相关热点问题，并建立一套完整的评价方案。

在数据预处理阶段，我们对源数据首先进行了筛选，清洗了数据中的非法字符，使用jieba中文分词,抽取主干部分进行分词,建立适用于解决该问题的词典，使用TF-IDF计算字频特征提取

针对第一问，对于留言内容的分类，考虑到实际数据情况，按照 6:2:2 的比例将附件 2 的数据分割为三个数据集，即训练集、验证集、测试集，借助分类算法在已知的训练集基础上学习到分类规则，初步建立分类模型用模型进行分类：利用验证集和测试集对前一步的 TextCNN 模型进行验证测试，通过多次对实验模型参数的合理调整，模型的准确率从 86.14%提升到 89.58%，性能显著提高,验证了该文本挖掘模型的准确性。

针对第二问，在对热点问题的挖掘处理时，我们对热度评价的指标进行了合理的定义：将每簇留言问题数量/时间跨度以及点赞数和反对数三个衡量标准综合考虑进去。借助 Python 爬虫技术获取到附件 3 中留言问题所在城市的所有街道社区、公共交通设施等信息，使用自定义词典对留言内容进行 jieba 分词处理，使用 TF-IDF 进行词频特征提取，然后利用相似度分析、DBSCAN 聚类分析，得出热点问题，并对其进行降序排序，借助TextRank4zh库根据聚类分析的不同类别提取出每个热点问题的关键句，通过Hanlp命名实体识别得到地点/人群信息，整理得出热点留言明细表。

针对第三问，建立答复意见的合理评价机制过程中，我们考虑从答复意见质量的相关性、完整性、及时性、可解释性四个角度进行评价，将答复意见和留言内容进行相似度比较、分析留言内容是否符合一套完整的回复规范、挖掘出留言时间和答复时间的时间间隔、对留言内容进行语义分析，考察答复意见是否讲法律、重事实、有机构，并从四个角度分别量化评价标准，从而建立一套完整的答复意见评价机制。

关键词：jieba 分词；TF-IDF；TextCNN；DBSCAN 聚类；TextRank

Intelligent Government Text Mining Model Based on Natural Language Processing

Abstract

In the big data era, as the network political platform such as WeChat, Weibo, Sunshine Hotline gradually becomes an important channel for the government to understand public opinions, solve people's livelihood problems, and so on, the amount of text data of various kinds of public opinions related issues has also increased dramatically. Relying on the traditional way of classifying and sorting out various types of message problems by government staff, it is facing unprecedented challenges. Intelligent processing of government needs dayOn the other hand, due to the rapid development of in-depth learning in the field of natural language processing (Natural Language Processing), the establishment of an intelligent government system based on natural language processing has become an innovative trend in social governance.

For the "Intelligent Government Text Mining Application" given in this topic, we reanalyze the problem, further analyze, process and redefine it as: Natural language processing text classification problem, using Pytorch framework to build a TextCNN model, which can quickly locate and classify large amounts of data, give relevant hot issues, and set up a complete evaluation program.

In the data preprocessing stage, we first filtered the source data, cleaned the illegal characters in the data, used the Jieba Chinese word segmentation, extracted the main cadre sub-division for word segmentation, built a dictionary suitable for solving the problem, and calculated the word frequency feature extraction using TF-IDF.

For the first question, regarding the classification of message content, considering the actual data situation, the data of Appendix 2 is divided into three datasets according to the 6:2:2 ratio, namely training set, validation set and test set. With the help of the classification algorithm, the classification rules are learned from the known training set, and a preliminary classification model is established to classify the text of the previous step: using validation set and test set to classify the text of the previous stepCNN model validation test, through several reasonable adjustments to the experimental model parameters, the

accuracy of the model is improved from 86.14% to 89.58%, the performance is significantly improved, and the accuracy of the text mining model is verified.

For the second question, when dealing with the mining of hot issues, we have made a reasonable definition of the index of heat evaluation: take into account the number/time span of each cluster of message questions, as well as the number of points and inverse logarithm.With the help of Python crawler technology, we get all the street communities, public transportation facilities and other information of the city where the message problem is located in Appendix 3. We use custom dictionary to Jieba the message content, TF-IDF to extract the word frequency feature, then use similarity analysis, DBSCAN cluster analysis to get hot issues, and sort them in descending order. With the help of TextRank4zh library rootKey sentences for each hot topic are extracted from different categories of cluster analysis, location/population information is obtained through Hanlp named entity identification, and a detailed list of hot spot messages is sorted out.

For the third question, in the process of establishing a reasonable evaluation mechanism for the response to comments, we consider four aspects to evaluate the quality of the response comments: correlation, integrity, timeliness and explanability. We compare the similarity between the response comments and the message content, analyze whether the message content meets a complete set of response specifications, dig out the message time and the time interval between the response time,This paper makes a semantic analysis of the message content, examines whether the reply opinions are legal, factual and institutional, and quantifies the evaluation criteria from four perspectives, so as to establish a complete evaluation mechanism for the reply opinions.

Keywords: jieba participle; TF-IDF; TextCNN; DBSCAN clustering; TextRank

目录

一、问题分析

随着网络问政平台的逐渐推广以及政务留言数据量的激增，政府部门通常采用的人工分类数据方式以及热点政务问题整理方法的不足之处逐渐显现：分类准确率较低、工作效率不高、耗时费力等，更多情况下，我们希望可以对网络问政平台的政务意见留言快速定位、分类并且找到热点问题及时回复，同时对政府部门的工作进行合理的评价，因此我们希望基于自然语言处理的智慧政务文本挖掘可以在这一方面提供帮助。

对于此题需要根据所给出的标签体系，建立分类模型训练数据、验证模型、测试数据获得关于留言内容的一级标签分类模型并评价其准确率，运用数据挖掘方法分词、词频提取、相似度比较、聚类分析得出热点问题，最后从答复的相关性、及时性、完整性、可解释性对政府答复意见进行合理评价。

根据附件 2 给出的 60%留言数据分析并提取出分词，用于训练数据学习分类规则，建立关于留言内容的一级分类模型，再用 20%的留言数据验证建立的分类模型准确性，最后利用余下的 20%的留言数据来测试模型，保证分类模型具有较高的准确率。

根据附件 3 的内容，利用 Hanlp 命名实体识别和爬虫算法得到地点、人群、机构等信息，分词后进行相似度比较，聚类分析得到热点问题，并使用留言问题数量/时间跨度和综合点赞数和反对数得到热度指数，最后利用 TextRank4zh 库提取热点问题关键句。

对附件 4 给出的答复意见和留言详情进行相似度分析，利用时间差函数挖掘出答复相隔时间，并且从相关性、完整性、及时性、可解释性等角度建立一套关于答复意见质量的完整评价机制。

二、模型假设

假设 1：所有留言内容均完整且真实可靠，数据有挖掘价值

假设 2：假设数据清洗后的留言内容主题均包含在所给分类标签中

假设 3：是否为热点留言问题与其在附件中出现的先后顺序不相关

三、数据准备

3.1 数据描述

通过观察分析所给附件数据，可以发现数据量达到万条以上，且附件 2、附件 3、附件 4 的留言内容和回复意见均为文本格式，分词处理后才可以对其进行分析，同时附件 2 存在同一个人短期内重复留言情况，如果不做数据清洗工作，对后续聚类结果的质量会带来很大的影响，地点/人群信息的提取也尤为重要，分词阶段不能合适提取出地名公共交通设施等信息将会导致 TF-IDF 词频统计时出现误差，必然得不到真正的热点留

言问题，同时为准确计算出时间跨度，也需要标准化日期格式，综上所述本文首先需要对数据进行预处理。

3.2 数据预处理

高质量的数据集是数据挖掘模型匹配和优化的基础，对于整个挖掘过程来说也是至关重要的，在本模型中，我们主要采用将自然语言处理的问题转换为机器学习的方式进行学习，在数据预处理阶段完成对数据集的处理和数据清洗，减少重复处理数据的问题出现，建立适用于解决该问题的词典,抽取主干部分进行使用jieba中文分词，标准化日期格式，去除同一用户短期内相似度较高的留言。

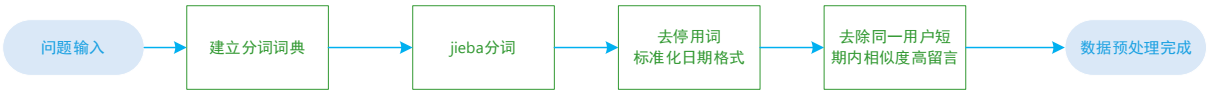

图3-1数据预处理过程

3.2.1 jieba 分词

中文文本与英文文本的明显区别是中文词语之间没有明显的界限，从长句中提取词语需要采用分词处理，本文采用Python的jieba分词，对留言问题详情和答复意见中的每一句话进行中文分词。

jieba分词基于前缀词典实现高效的词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG），采用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，对句子从右往左反向计算最大概率，因为中文句子的重心经常落在后面（右边），因为通常情况下形容词太多，后面的才是主干。因此，从右往左计算，正确率要高于从左往右计算，这里类似于逆向最大匹配），P(NodeN)=1.0，P(NodeN-1)=P(NodeN)*Max(P(倒数第一个词))…依次类推，最后得到最大概率路径，得到最大概率的切分组合。某留言详情分词结果如下图所示：

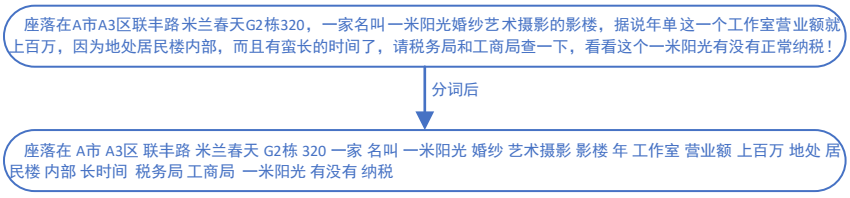

图3-2某留言详情分词结果展示

3.2.2 去停用词

在文本挖掘处理中，停用词通常是指对表征文本几乎没有作用的词语，比如英文中的“a,the,of,and,or”等,仅对语法结构起到支撑作用，中文的“那么，然后，的，但是”等，对表示文本特征毫无作用，因而对于停用词在数据预处理阶段就必须删除，降低对文本的负面影响，不仅能够提高有效词的密度，还能大大降低文本的维度。

3.2.3 TF-IDF 词频统计

TF-IDF（词频-逆文档频率）算法是一种统计方法，用以评估一个词对于一个文件集或一个语料库中的其中一份文件的重要程度。词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降[1]。其中词频 TF（Term Frequency）词频：一个词在文章中出现的次数，即二者之间的相关性，计算时用该词比上文章的总次数。IDF（Inverse Document Frequency）逆文档频率：一个词语的普遍次数，其主要思想就是，如果包含某个词 Word 的留言越少，则这个词的区分度就越大，也就是 IDF 越大。对于如何获取一条留言内容的关键词，我们可以计算这条留言内容出现的所有名词的 TF-IDF，TF-IDF 越大，则说明这个名词对留言内容的区分度就越高，取 TF-IDF 值较大的几个词，就可以当做留言内容的关键词。

$$TF(\text{词频}) = \frac{\text{某个词在留言中出现的次数}}{\text{留言中出现次数最多的词语次数}}$$

$$IDF(\text{逆文档频率}) = \log(\frac{\text{语料库中的留言总数}}{\text{包含某个词的留言数}+1})$$

$$TF-IDF = TF(\text{词频}) * IDF(\text{逆文档频率})$$

3.2.4 标准化日期格式

在对热点留言问题挖掘时，需要使用聚类后同簇留言时间跨度，建立完整的评价机制时，我们从及时性的角度对答复意见进行评价，同样需要使用到时间跨度，在此我们选择时间差函数标准化时间格式为：%Y/%m/%d %H:%M:%S，以降低对后续模型效果的影响。

四、问题求解

4.1 问题求解流程

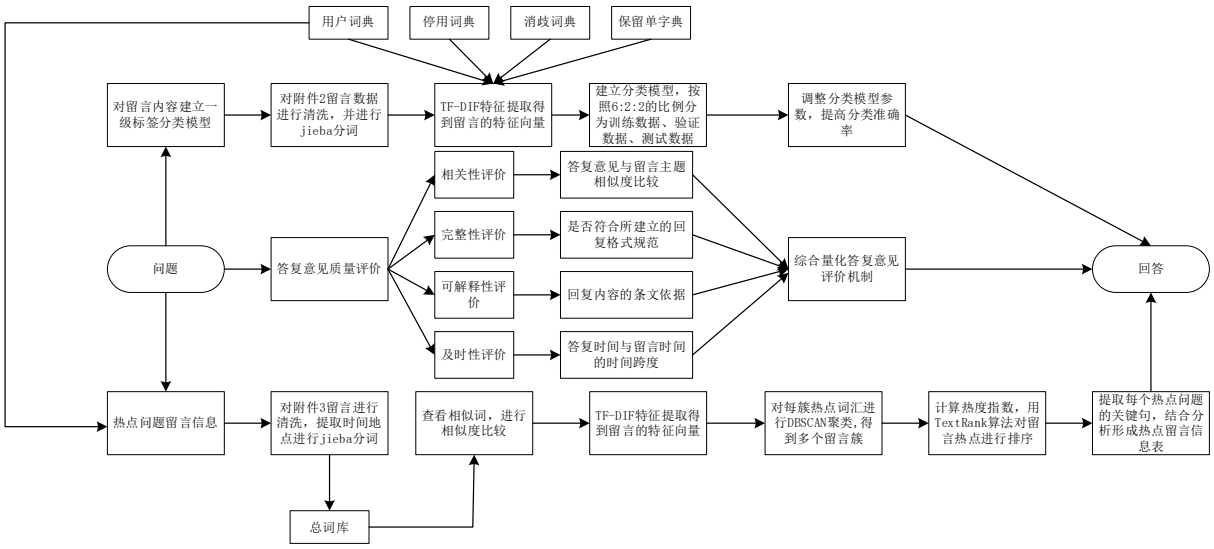

图 4-1 问题求解流程图

由问题求解流程图，可知在对问题的求解过程中存在以下主要任务：

（1） 在数据预处理的基础上，利用 TextCNN 模型对文本建模，将清洗后的数据合理划分，分别为训练集、验证集、测试集，多次调整参数来提高模型准确率。

（2） 使用聚类方法得到不同的簇，初步完成对热点问题的划分，借助 TextRank 算法提取出每簇热点问题的关键句,综合考虑每簇留言数量/时间跨度以及点赞数和反对数，量化热点问题的评价指标，并形成热点问题明细表。

（3） 在问题 1 和 2 的方法和标准下，从相关性、完整性、及时性、可解释性四个角度通过相似度比较、构造答复意见规范标准、计算时间跨度、提取留言回复的条文依据等方法全面多角度来量化答复意见评价方案。

4.2 针对特定地点、人群的提取处理

4.2.1 HanLP 命名实体识别

命名实体识别本质上是一个模式识别任务, 即给定一个句子, 识别句子中实体的边界和实体的类型，是自然语言处理任务中一项重要且基础性的工作，识别有特定意义的实体并将其区分为人名，机构名，日期，地名，时间等类别的。Han Language Processing（HanLP）作为常用的命名实体识别方法，在本题中对特定地点和人群的提取中有着至关重要的作用，调用 HanLP 方法时需要注意：

(1) 命名实体识别模块的输入是单词列表，输出是命名实体的边界和类别，例如：

```
recognizer = hanlp.load(hanlp.pretrained.ner.CONLL03_NER_BERT_BASE_UNCASED_EN)
recognizer(["President", "Obama", "is", "speaking", "at", "the", "White", "House"])
[('Obama', 'PER', 1, 2), ('White House', 'LOC', 6, 8)]
```

(2) 中文命名实体识别是字符级模型，需要用 list 将字符串转换为字符列表。输出格式为(entity, type, begin, end)

(3) 使用基于 BERT[^bert]的最准确的模型 MSRA_NER_BERT_BASE_ZH，来浏览该模型的评测指标，继而准确识别特定地名、人群。

4.2.2 特定地点的定位发现

在对数据的进一步处理过程中我们发现：特定地点和人群的划分经常会影响分词的准确率，进而影响最终分类结果，因此我们对留言内容中出现的地点、人群建立专用词典。通过对比留言内容的出现地点，调用百度地图多次验证，确认所有政务留言内容属于湖南省长沙市附近的政务留言，同时为了进一步验证我们的发现，我们在该市找到诸多与附件中完全吻合的地点，这一点也与该文本挖掘模型的假设 1 相对应。留言内容中出现地点所在城市的地图如下：

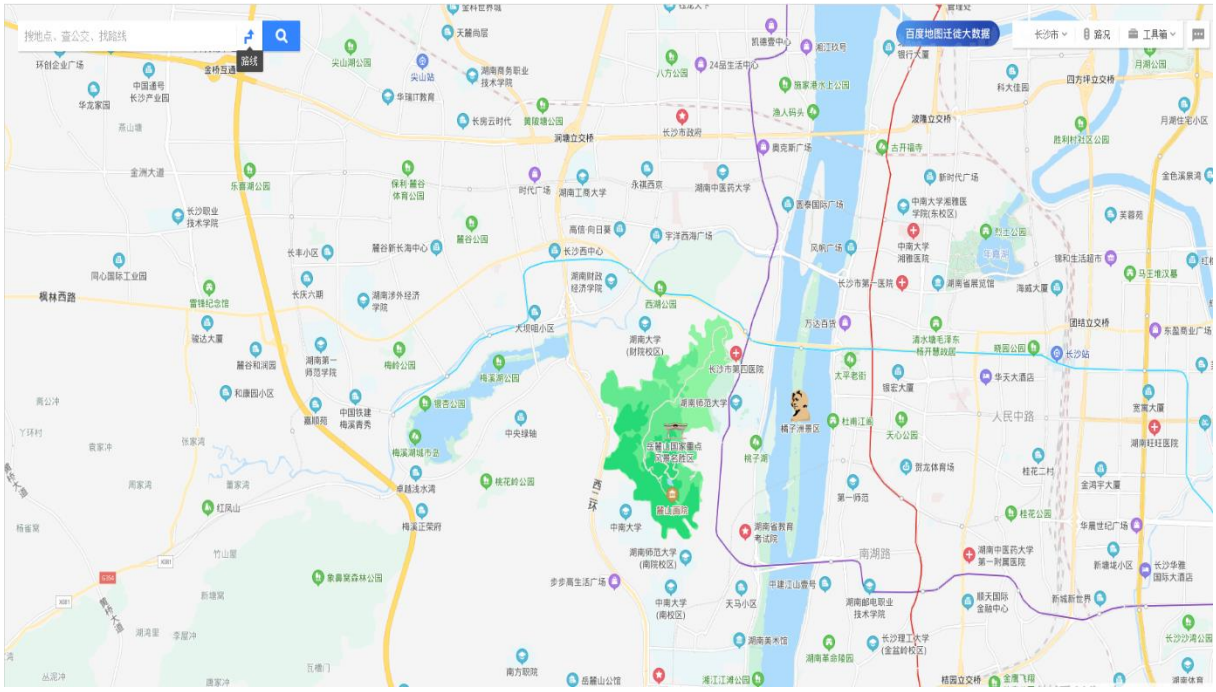

图 4-2 留言内容中出现地点所在城市的地图

4.2.3 爬虫技术获取 A 市地点信息

在掌握具体地点信息后，我们从安居客、图吧两个网站上爬取了 A 市附近所有楼盘、小区、街道、社区名称以及公交车站、地铁站、道路等交通设施名，进一步完善用户自定义的 jieba 分词词典，爬取过程如图 4-3 所示：

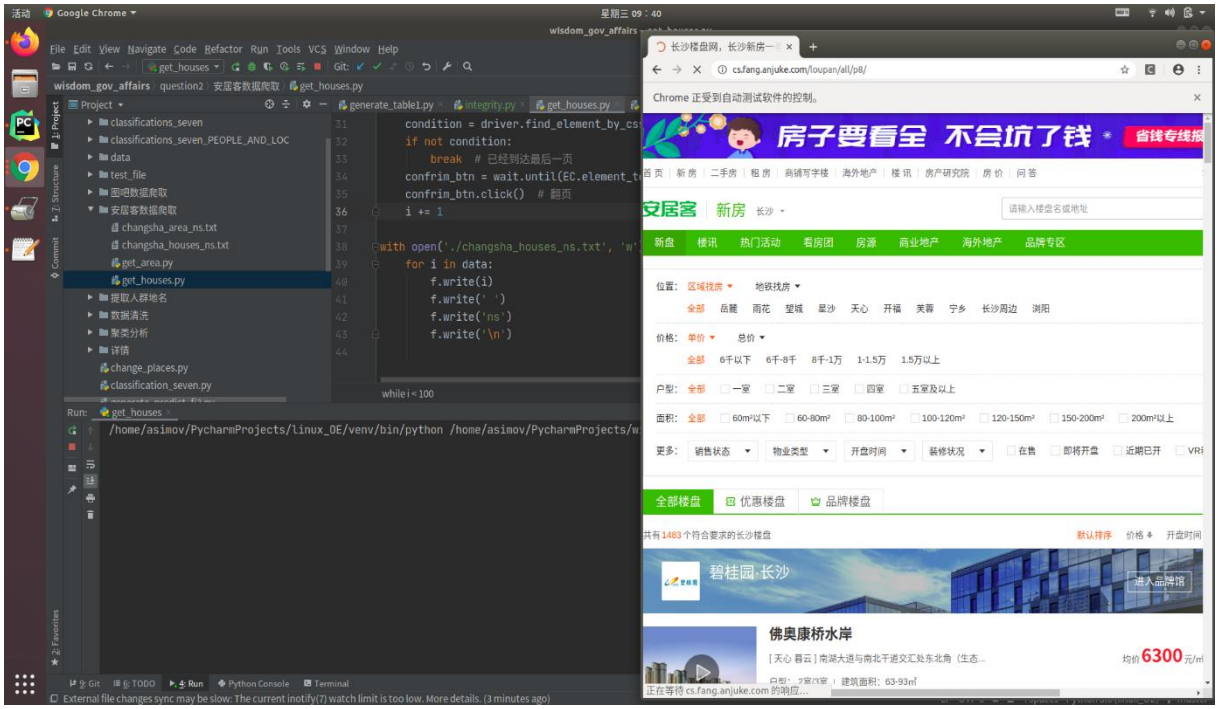

图 4-3 爬虫技术获取 A 市地点信息

我们对比使用 A 市地点信息的自定义词典前后的留言内容分词效果，分词准确率得到极大程度提高，例如对【A7 县深业睿城开发商违法出售人防车位，请政府查处】的前后分词对比：

（1）未使用 A 市地点信息的自定义词典的分词结果：【A7 县深业睿城开发商违法出售人防车位，请政府查处】➔【A7 县 深业 睿城 开发商 违法 出售 人防 车位 请政府 查处】

（2）使用 A 市地点信息的自定义词典的分词结果：【A7 县深业睿城开发商违法出售人防车位，请政府查处】➔【A7 县 深业睿城 开发商 违法 出售 人防 车位 请 政府 查处】

4.3 基于自然语言处理的“智慧政务”文本挖掘过程详解

4.3.1 基于 TextCNN 的留言内容一级标签分类

(1) 模型原理及介绍

卷积神经网络（CNN）最早被应用于计算机视觉领域，是一类包含卷积计算且具有深度结构的前馈神经网络[2]，2014 年 Yoon Kim 在《Convolutional Neural Networks for Sentence Classification》中对 CNN 输入层做出调整，从而提出了 TextCNN，使得 CNN 网络能够应用到文本分类任务中。TextCNN 是一种采用卷积神经网络提取文本 n-gram 特征，最大池化，全连接然后进行分类的一种新型模型，其架构如下图所示。

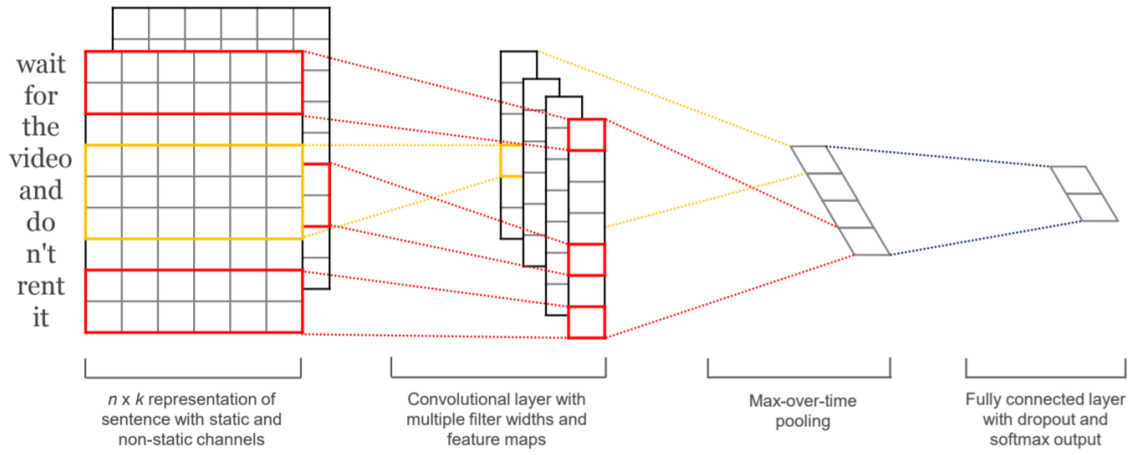

图 4-4 TextCNN 模型架构示图

如上图，TextCNN 模型由嵌入层、卷积层、最大池化层和全连接层组成。

① 嵌入层

嵌入层是一个 $n \times k$ 的矩阵，其中 n 为句子中单词数，k 为每个词对应的词向量维度。每个词向量可以是预先在其他语料库中训练好的，也可以作为未知的参数由网络训练得到。这两种方法各有优势，预先训练的词嵌入可以利用其他语料库得到更多的先验知识，而由当前网络训练的词向量能够更好地抓住与当前任务相关联的特征。

② 卷积层

由于句子中相邻的单词关联性总是很高的，因此可以使用一维卷积，即文本卷积与图像卷积的不同之处在于只在文本序列的一个方向（垂直）做卷积，卷积核的宽度固定为词向量的维度。

③ 最大池化层

池化层工作与卷积层大同小异，不同于卷积层将卷积核中数据相加；池化层的池化核将池化核中数据求平均或者只保留最大值。Yoon Kim 采用了最大池化[3]。max-pooling 只会输出最大值，对输入中的补零做过滤。

④ 全连接层

最后一层全连接的 softmax 层，输出每个类别的概率。为了防止过拟合，可以采用正则化的方法，在倒数第二层加入 dropout 方法，防止隐藏层过拟合，并用 L2 范数约束权重向量[4]。

（2） 模型的建立与训练

① 模型搭建与运行环境

TextCNN 模型搭建所使用的软硬件配置如下表所示。

表 4-1 建模软硬件配置

软/硬件配置	版本/型号
操作系统	Windows 10 X86
CPU	Intel(R) Core(TM) i7-7700HQ
内存	8.00GB
显卡	Geforce GTX 1060
Python	3
Pytorch	1.2

② 基于 Pytorch 框架的 TextCNN 模型搭建

Pytorch 框架是 Torch 的 Python 版本，不仅具有反向自动求导技术，能够快速且容易地更改神经网络而且易于准确定位错误，有利于深入理解内部实现。我们采用成熟的 Pytorch 框架，搭建 TextCNN 模型实现文本分类，如下图所示。

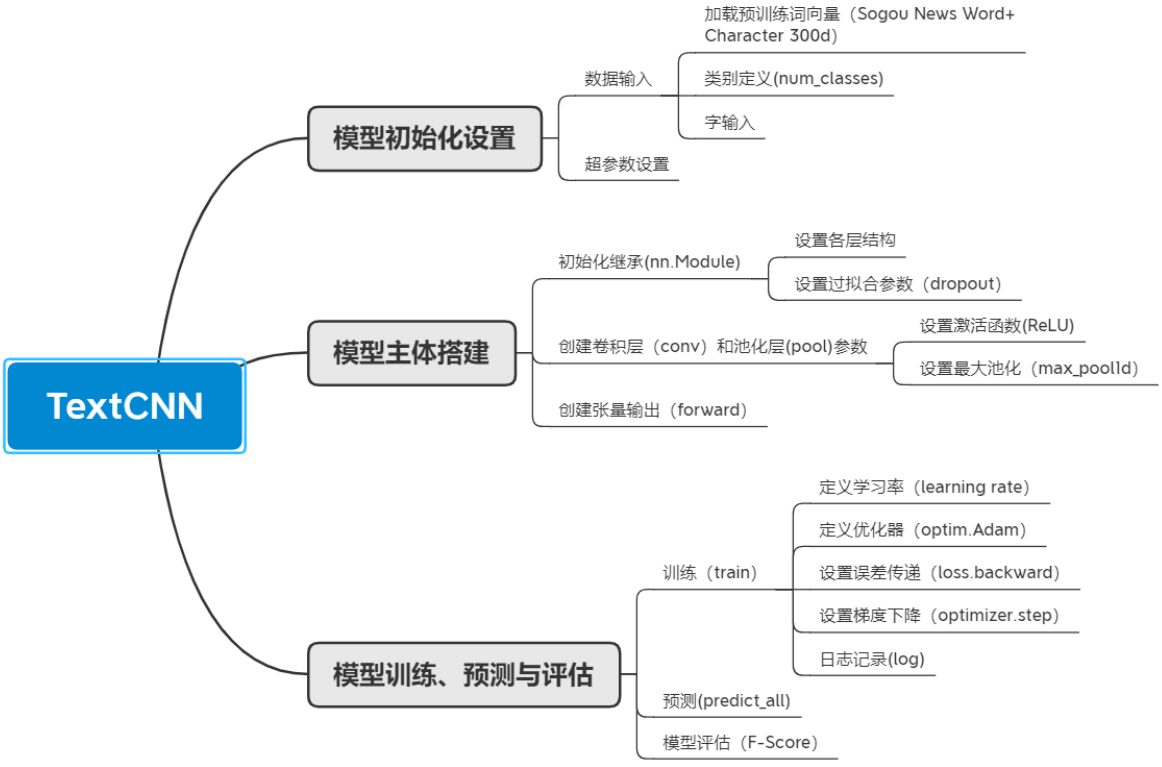

图 4-5 TextCNN 网络搭建详情示意图

③ TextCNN 模型参数设定

通过对模型的训练，参考模型训练效率、训练效果得到调参结果如下表所示。

表 4-2 TextCNN 参数设定

参数	参数值
Dropout	0.5
Num_epoch	31
Batch_size	128
Pad_size	460
Learning_rate	1e-4
Filter_size	(2,3,4)
Num_filters	512

④ Pad_size 参数的确定

对附件 2 中留言详情内容进行无效数据清洗，得到下图文本长度统计结果。我们采用字输入，遂选择 75%留言长度作为模型 Pad_size 的值。

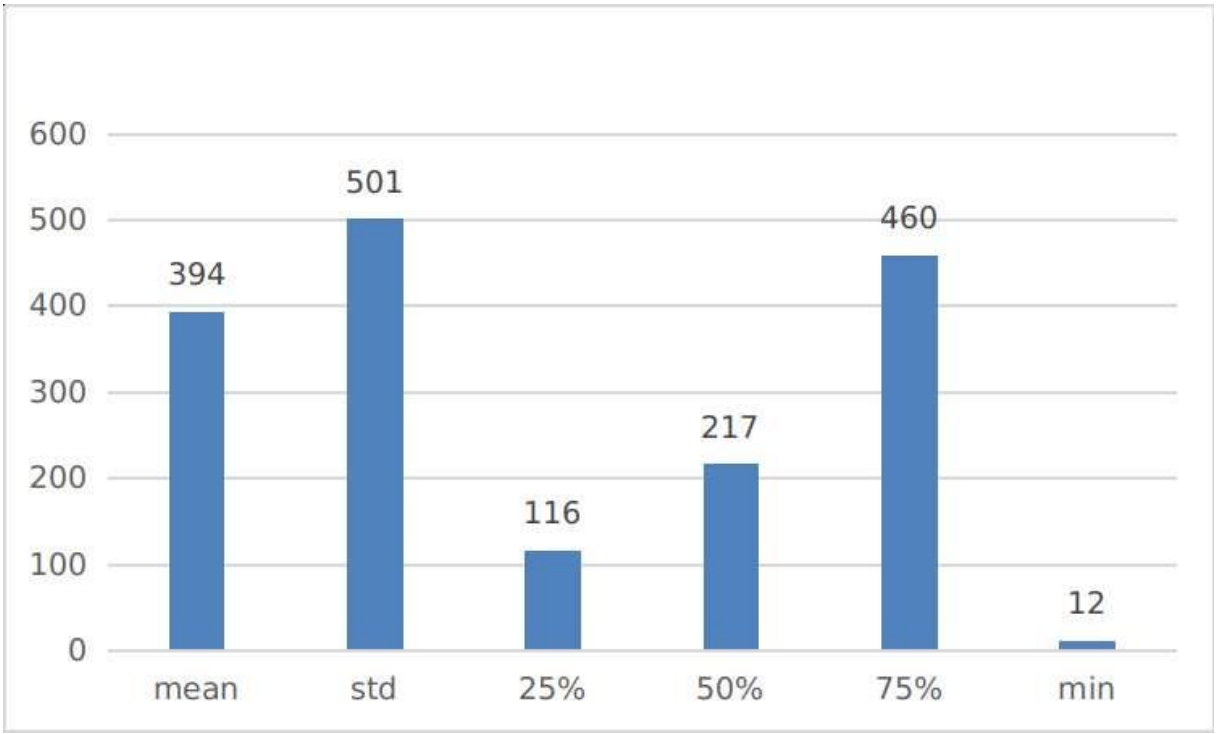

图 4-6 留言详情文本长度统计示图

(3) 模型训练结果评价

对于评价本文中模型优劣的指标，我们通过计算模型的分类准确率(Accuracy)、精确率(Precision)、召回率(Recall)得到题目中叙述的精确度和召回率的调和平均值(F1-Score)，还进一步生成了模型的混淆矩阵(Confusion Matrix)，以上指标详细定义如下：

首先，定义四类基础指标 TP、TN、FP、TN。

$$TP = True\ Positive$$

$$FN = False\ Negative$$

$$FP = False\ Positive$$

$$TN = True\ Negative$$

（1） 准确率(Accuracy)：分类模型中所有判断正确的结果占总观测值的比重

$$acc = \frac{TP + TN}{TP + TN + FP + FN}$$

（2） 精确率(Precision)：模型预测为 Positive 的所有结果中，模型预测正确的比重

$$P = \frac{TP}{TP + FP}$$

（3） 召回率(Recall)：真实值为 Positive 的所有结果中，模型预测正确的比重

$$R = \frac{TP}{TP + FN}$$

（4） F1-Score：精确度和召回率的调和平均值

$$F1 - Score = \frac{2PR}{P + R}$$

模型结果：

经过模型参数调优与训练，最终得到本文搭建的 TextCNN 模型在使用 Sougou News Word+Character 300d 作为预训练词向量、C 题正式数据作为数据集时，模型准确率提高明显，模型误差梯度下降较快，如下图所示模型准确率可到达：89.58%，各分类精确率、召回率、F1-Score 值分别如下表所示。证明，基于 Pytorch 框架建立的 textCNN 文本分类模型具有优异性能，可满足实际分类需求。

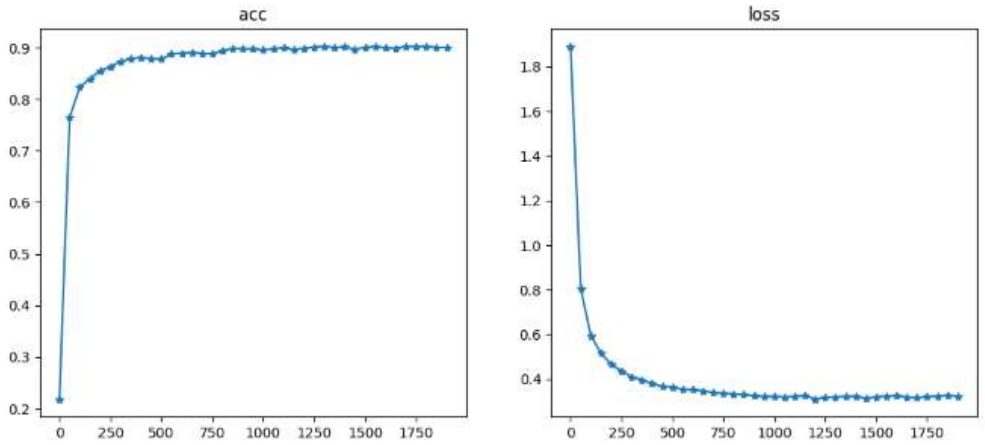

图 4-7 模型 ACC 与 LOSS 变化示图

表 4-3 各分类字母表示表

城乡建设	A
环境保护	B
交通运输	C
教育文体	D
劳动和社会保障	E
商贸旅游	F
卫生计生	G

表 4-4 各类别评价参数表

	Precision	Recall	F1-Score	support
A	0.8322	0.9005	0.8650	402
B	0.9274	0.8830	0.9046	188
C	0.9029	0.7623	0.8267	122
D	0.9519	0.9340	0.9429	318
E	0.9400	0.9543	0.9471	394
F	0.8333	0.8436	0.8384	243
G	0.9345	0.8920	0.9128	176

表 4-5 混淆矩阵

	A	B	C	D	E	F	G
A	362	9	7	5	9	9	1
B	20	166	0	1	0	1	0
C	17	4	93	0	0	11	0
D	6	2	0	297	5	6	2
E	5	0	0	4	376	2	7
F	23	1	3	4	6	205	1
G	2	0	0	1	4	12	157

4.3.2 基于 DBSCAN 聚类的热点留言问题挖掘

（1） 聚类分析

聚类分析是一种无监督的学习方式，用无标签的数据来训练分类模型，仅利用某种距离计算将多个数据对象划分成集合的过程，每个集合即为一个簇，簇内的相似度尽可能高，簇间的相似度尽可能低（相异度尽可能高）。

聚类分析处理的数据通常都是没有标号的数据，但是仍然需要对其分簇，也由于聚类分析可以挖掘出事先未知的群组，在数据分析经常被使用。

（2） 基于数据挖掘的角度的四种聚类分析方法

① **划分聚类**：给定一个 n 个对象的集合，划分方法构建数据的 k 个分区，其中每个分区表示一个簇。大部分划分方法是基于距离的，使用均值或者中心点等代表簇中心，给定要构建的 k 个分区数，首先创建一个初始划分，然后使用一种迭代的重定位技术将划分对象重新定位，逐渐逼近局部最优解。

② **层次聚类**：层次聚类可以分为凝聚和分裂的方法；凝聚也称自底向上法，开始便将每个对象单独为一个簇，然后逐次合并相近的对象，直到所有组被合并为一个族或者达到迭代停止条件为止。分裂与凝聚方法相反，是自顶向下的方法，开始将所有对象划分为同一个簇，多次迭代分成若干个更小的簇，直到每个对象都满足某个特定条件终止或者在一个单独的簇中。层次聚类是一种层次分解，不能纠正错误的合并或划分，但可以集成其他的技术，常见的凝聚层次聚类算法有 AGNES,分裂层次聚类算法有 DIANA。

③ **基于密度的聚类**：其主要思想是只要“邻域“中的密度超过某个阀值，就继续增长给定的簇，将高密度点连成一片，形成各个簇。也就是说，对给定簇中的每个数据点，在给定半径的邻域中必须包含最少数目的点。这样的主要好处就是过滤噪声，剔除离群点，常见的基于密度聚类算法有 DBSCAN 算法。

④ **基于网格的聚类**：它把对象空间量化为有限个单元，形成一个网格结构，所有的聚类操作都在这个网格结构中进行，这样使得处理的时间独立于数据对象的个数，而仅依赖于量化空间中每一维的单元数。基于网格的聚类分析使用一种多分辨率网格数据结构，能快速处理数据。

（3） 基于密度的 DBSCAN 热点留言问题聚类

在使用 TF-IDF 对留言内容文本特征提取后，由于 DBSCAN 聚类分析可以过滤离群点、产生任意形状的簇[5]，降低聚类的偏差，DBSCAN 聚类分析为此题较为合适的选择。解题过程如下：

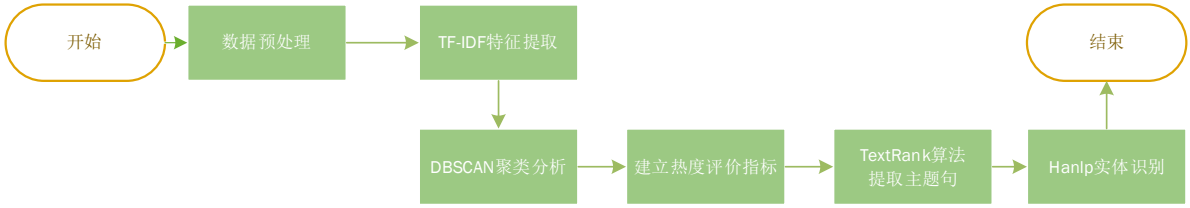

图 4-8 问题 2 解题过程

在聚类过程中，需要通过获取密度核心，由密度核心延展，把留言内容相近的留言问题归为以同一个簇，在经过数据预处理阶段以及借助命名实体识别算法、爬虫获取特定地点、人群信息后，对所得数据进行 DBSCAN 聚类分析。具体参数说明如下：

（1） eps：DBSCAN 算法参数，即我们的 ϵϵ-邻域的距离阈值，和样本距离超过 ϵϵ 的样本点不在 ϵϵ-邻域内。默认值是 0.5.一般需要通过在多组值里面选择一个合适的阈值。eps 过大，则更多的点会落在核心对象的 ϵϵ-邻域，此时我们的类别数可能会减少，本来不应该是一类的样本也会被划为一类。反之则类别数可能会增大，本来是一类的样本却被划分开。

（2） min_samples：DBSCAN 算法参数，即样本点要成为核心对象所需要的 ϵϵ-邻域的样本数阈值。默认值是 5. 一般需要通过在多组值里面选择一个合适的阈值。通常和 eps 一起调参。在 eps 一定的情况下，min_samples 过大，则核心对象会过少，此时簇内部分本来是一类的样本可能会被标为噪音点，类别数也会变多。反之 min_samples 过小的话，则会产生大量的核心对象，可能会导致类别数过少。

（3） metric：最近邻距离度量参数。可以使用的距离度量较多，一般来说 DBSCAN 使用默认的欧式距离（即 p=2 的闵可夫斯基距离）就可以满足我们的需求。

聚类过程图解如下：

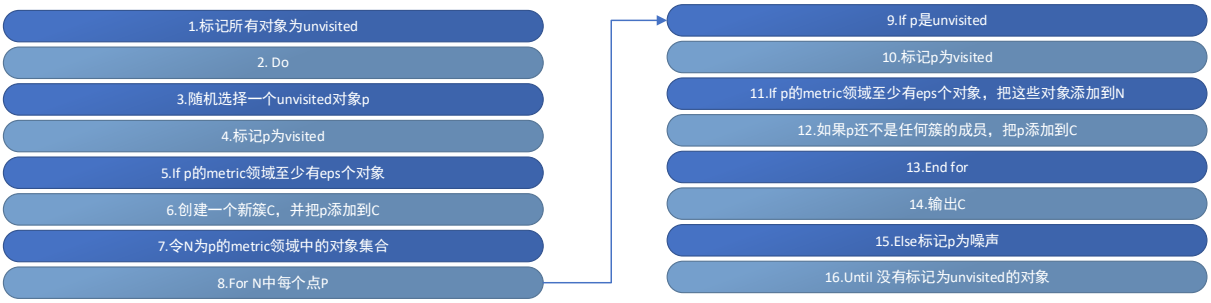

图 4-9 聚类过程图解

不同参数的聚类结果如图 4-10 所示，经过多次试验调整，在参数 **DBSCAN**(eps=0.9, min_samples=4)得到的聚类分析效果最好，聚类分析效果如图 4-11 所示：

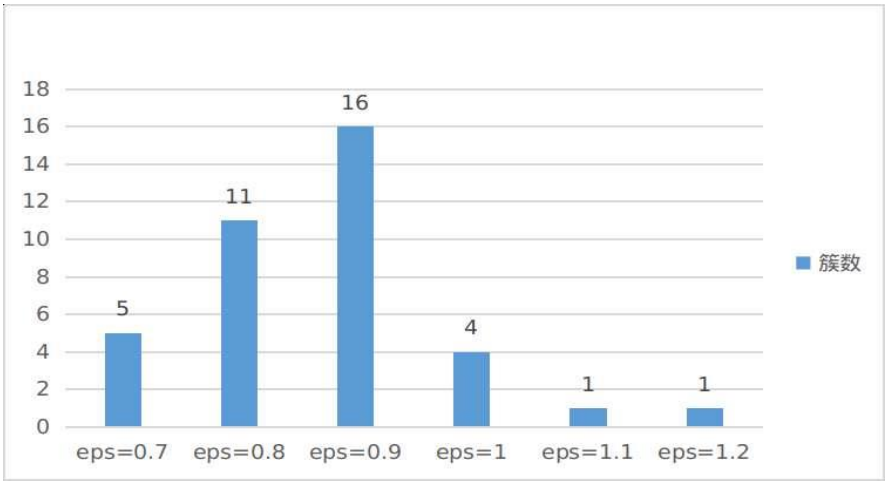

图 4-10 不同参数下的聚类簇数

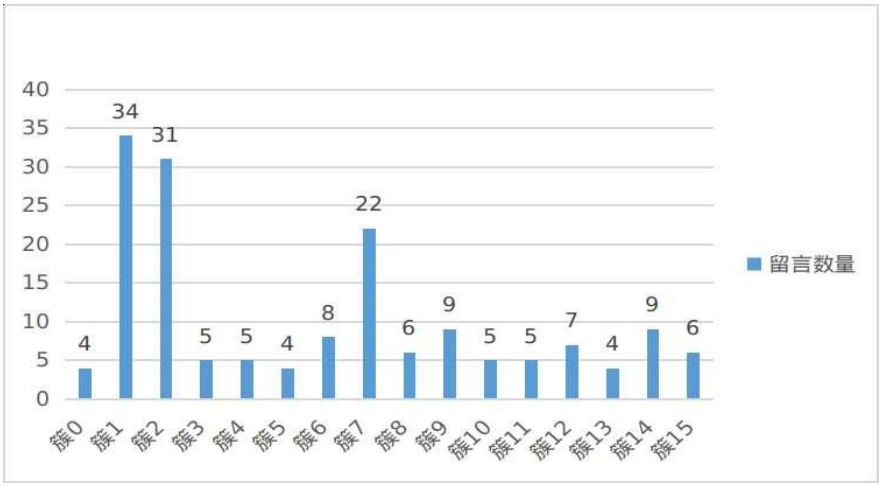

图 4-11 指定参数下每簇留言数量

（4） 合理的热度评价指标

对热点留言问题评价机制进行合理定义时，我们首先清洗同一用户 30 天内的相似度高于 75%的留言内容，将两个个体的特征向量化，然后通过余弦公式计算两者之间的相似性即可，热度评价指标如下图所示：

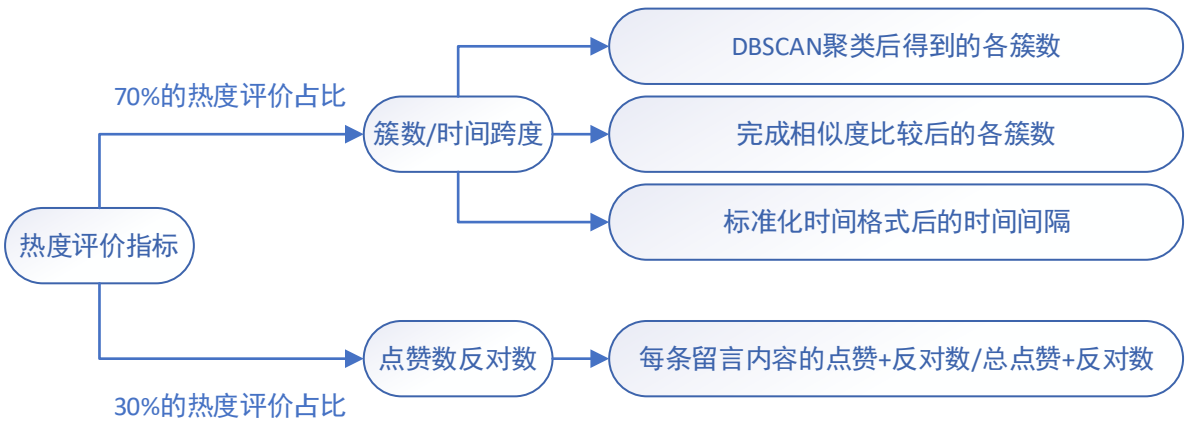

图 4-12 热度评价指标

（5） 基于 TextRank 算法的关键句提取

利用 TextRank 算法对聚类后每簇留言主题的关键句提取过程如图 4-13 所示：

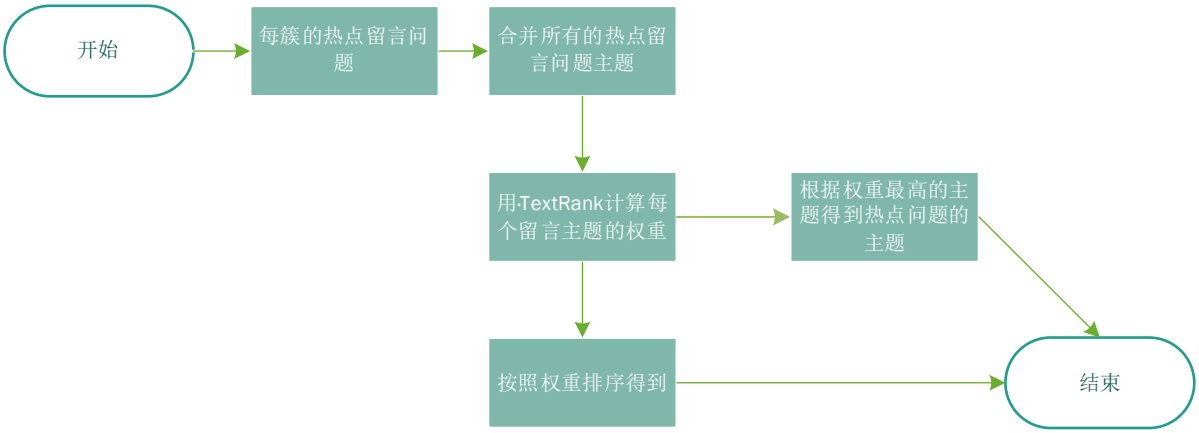

图 4-13 对聚类后每簇留言主题的关键句提取过程

TextRank 算法属于文本排序算法，由谷歌的网页重要性排序算法 PageRank 算法改进，它能够从一个给定的文本中提取出该文本的关键词、关键词组，并使用抽取式的自动文摘方法提取出该文本的关键句[6]，在本问题中我们采取 textrank4zh 模块来提取聚类后属于同簇的留言内容关键句，给出 TextRank 的计算公式：

$$S(v_i)=(1-d)+d\sum_{(j,i)\in\varepsilon}\frac{w_{ji}}{\sum v_k\in out(v_j)w_{jk}}S(v_j)$$

即 TextRank 中一个留言关键句 i 的权重取决于与在 i 前面的各个点 j 组成的(j,i)这条边的权重，以及 j 这个点到其他其他边的权重之和[7]，通过以上数据，构建出候选关键句图，把同簇内的 2 个关键句组成的边，和以及其权值记录下来，套用 TextRank 的公式，迭代传播各节点的权值，直至收敛。对结果中的 Rank 值进行倒序排序，得到我们需要的关键句。

TextRank 得到的前 5 条热点留言问题下表所示（完整热点留言表见附件热点问题表，此处仅展示主要热点留言问题主要内容）：

表 4-6 热点留言问题主要内容

热度排名	热度指数	问题描述
1	0.5856	丽发新城小区附近的搅拌站噪音严重扰民
2	0.4997	A4 区洪山公园规划了十几年了，到底什么时候开工建设？
3	0.4748	A7 县星沙四区凉塘路旧城改造要拖到何年何月才能动工
4	0.4644	投诉 A 市伊景园滨河苑捆绑车位销售
5	0.4520	咨询 A 市人才购房补贴通知问题

从词云图可得，出现频率较高的地名有 A 市、A7 县等，进一步验证了我们热度评价指标的合理性，得出的热点问题表也具有较高准确性。

图 4-14 留言主题词云图

（6） 基于 HanLP 的地点/人群的提取

使用 HanLP 命名实体识别，从 textrank4zh 模块提取的关键句中识别地点、职业、机构名，提取出热点问题的地点/人群,如下表所示：

表 4-7 热点留言问题地点/人群

热度排名	问题 ID	时间范围	地点/人群
1	12	2019/02/13 至 2019/03/28	南山十里天池玉佩路
2	1	2019/02/28 至 2020/01/26	丽发新城小区
3	10	2019/01/15 至 2019/07/25	洪山公园
4	2	2019/02/28 至 2019/09/01	伊景园滨河苑
5	9	2019/02/14 至 2019/09/09	星沙凉塘路

4.3.3 综合考量的答复意见评价机制的实现

（1） 评价机制设计

建立答复意见的合理评价机制过程中，我们考虑从答复意见质量的相关性、完整性、及时性、可解释性四个角度进行评价机制设计。

将答复意见和留言内容进行相似度比较、分析留言内容是否符合一套完整的回复规范、挖掘出留言时间和答复时间的时间间隔、对留言内容进行语义分析，考察答复意见是否讲法律、重事实、有机构，并从四个角度分别量化评价标准，从而建立一套完整的答复意见评价机制。

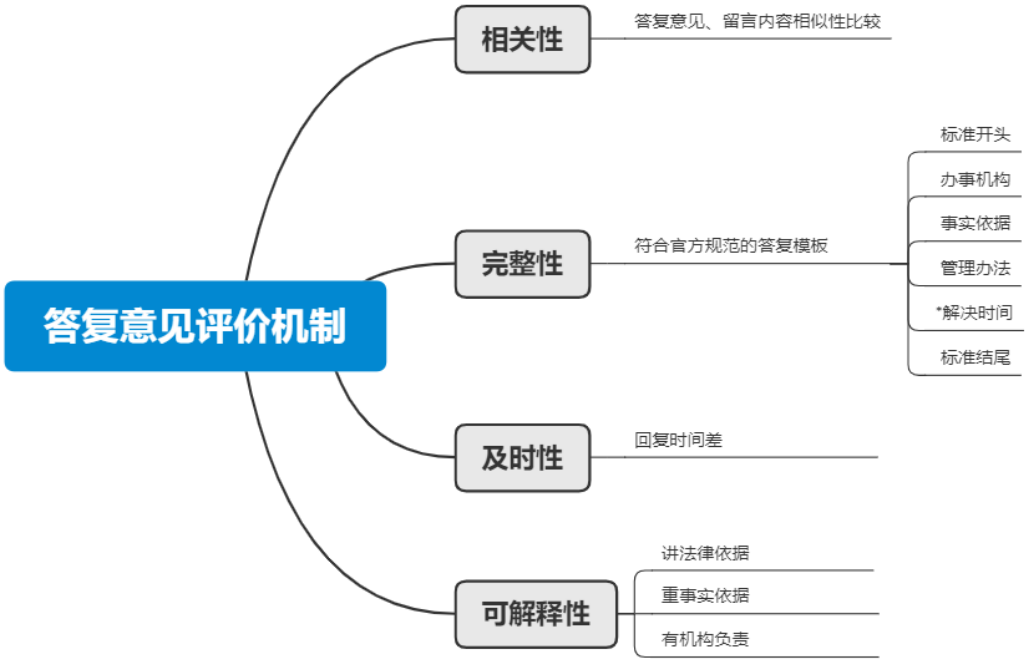

图 4-15 答复意见评价机制示意图

（2） 评价机制技术实现

我们使用正则表达式对答复意见进行初步清洗，实现标准开头、结尾地提取与分隔。

① 相关性实现：

对初步清洗的答复意见与用户留言详情，使用 TF-IDF 算法中的词频统计方法，进行相似度比较，建立相似度分级标准，实现两者相似度结果评级，评级结果如下图所示。

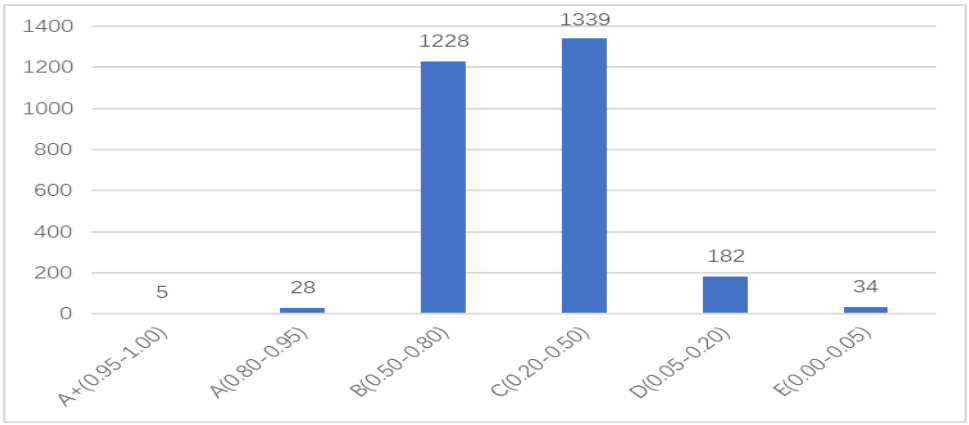

图 4-16 回复意见与用户留言相似度分级示意图

② 完整性实现

通过对答复意见的深度分析，我们使用正则表达式提取了答复意见的标准开头、结尾，并对开头、结尾实现分词处理，逐一验证了答复意见开头、结尾的标准性；通过 Hanlp 对答复意见中的机构名称进行命名实体识别、提取；利用正则表达式分别提取答复意见

中的事实依据、办法依据，从而建立了一套符合完整性的答复结构：标准开头、办事机构、事实依据、管理办法、标准结尾。

③ 及时性实现

根据用户留言时间与答复意见回复时间差，建立意见评价机制及时性回复分级标准。

④ 可解释性实现

对答复意见进行语义识别，依据事实、法律法规、办事条例、办事机构与用户留言内容的内在联系，评价答复意见是否具有可解释性。

（3） 评价机制算法实现

① 各角度评级系统设计如下表所示，每一角度设置 A+——E 六个等级，分别代表优异、优、良、一般、合格、不合格，满足下述相应算法要求，每一角度会被赋予相应等级。

表 4-8 各角度算法设计

项目	相关性	完整性	可解释性	及时性
标准	相似度	开头、结尾、办事机构 法律、事实、解决时间	讲法律、事实 有机构	时间差
E	0.00-0.05	无任何内容 或含开头、结尾中的一项	无任何内容	未回复
D	0.05-0.20	有开头、结尾	有机构管理	>30
C	0.20-0.50	有开头、结尾 含法律、事实、机构中的一项	有事实依据	7-30
B	0.50-0.80	有开头、结尾 含法律、事实机构中的两项	有法律支撑	4-7
A	0.80-0.95	有开头、结尾 含办事机构且依法、事实管理	有法律 有事实依据	<4
A+	0.95-1.00	满足 A 中条件且明确给出解决时间	三者具备	当天

② 整体评级算法设计如下图所示。依据回复意见的受众群体主体性，意见回复的侧重程度表现为可解释性>相关性>及时性>完整性，首先将各角度等级均匀量化映射在 0-1 区间，设置角度重要性权重 0.3 0.3 0.2 0.2，计算四个角度加权和。对加权和按照是否出现 A+与 E 等级，进行级间调优，得出最终回复意见等级。

图 4-17 整体评级算法流程示意图

依据该评级算法对留言答复进行评级，证明该评级算法实际有效，合理刻画了答复意见的质量，证明基于答复相关性、完整性、可解释性、及时性角度的答复评价方案切实可行。部分数据展示如下图，详细评级结果见附件评价方案.xls。

留言编号	相关性	完整性	可解释性	及时性	权值	评级
2549	B	C	C	C	0.4	C
17461	B	C	A+	A	0.8	A
17562	B	C	C	B	0.4	C
17570	C	C	A+	B	0.6	B
17576	C	D	C	A	0.4	C
17580	B	D	B	A	0.6	B
17589	B	C	C	C	0.4	C
17609	B	C	C	A	0.4	C
17743	D	C	C	C	0.2	D
17745	C	C	D	B	0.4	C
17789	C	C	D	B	0.4	C

图 4-18 评级后部分数据展示

五、总结及评估

5.1 基于自然语言处理的文本挖掘模型总结

本文主要从基于自然语言处理的文本挖掘完成对政务留言问题的分析，实现了对留言内容的分类、热点留言问题的挖掘并给出一套完整的答复意见评价机制。主要实现了以下任务：

第一：利使用Pytorch框架建立了TextCNN模型 ，并对所建立的模型进行合理的评价并多次调整参数达到最优输出结果。

第二：利用 jieba 分词、TF-IDF 算法、DBSCAN 对所给留言内容进行数据预处理、特征提取、以及聚类分析，利用爬虫算法实现对地点名称、公共交通设施名称等的爬取。

第三：使用 TextRank 算法对每簇热点留言问题的关键句提取，形成热点问题表。在最后从相关性、完整性、及时性、可解释性四个角度建立完整的回复意见评价机制。

5.2 模型的提升与改进

由本文上述内容可知，我们建立的文本挖掘模型可以出色完成所给任务，但是在下一步过程中，我们仍然会继续改进所建立的模型：更换TextCNN模型的预训练词向量，更好地提高模型分类的准确率，拓宽答复意见评级机制的参考角度，如考虑问题的解决时间是否明确、答复意见是否积极正面等。

六、参考文献

[1]张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.

[2]刘春磊,武佳琪,檀亚宁.基于 TextCNN 的用户评论情感极性判别[J].电子世界,2019(03):48+50.

[3] Kim Y . Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014.

[4]林荣华. 基于卷积神经网络的句子分类算法[D].浙江大学,2015.

[5]李群,袁津生.基于 DBSCAN 的最优密度文本聚类算法[J].计算机工程与设计,2012,33(04):1409-1413.

[6]Mihalcea R, Tarau P. TextRank: Bringing order into texts[C]. Association for Computational Linguistics, 2004.

[7]陈志泊,李钰曼,许福,冯国明,师栋瑜,崔晓晖.基于 TextRank 和簇过滤的林业文本关键信息抽取研究[J/OL].农业机械学报:1-11[2020-05-08].