“智慧政务”中的文本挖掘应用

摘要

本文旨在基于原始文本数据和人工标记的文本分类标签，采用文本挖掘的分析方式构建对应模型，使得计算机能够解决文本分类、热点问题挖掘、答复意见的评价体系构建三方面的任务。目前针对上述文本任务主要采取人工标记的方式进行，耗时耗力且准确率低，所以通过计算机完成上述任务对“智慧政务”体系的完善具有重要意义

针对问题一，基于深度学习 **ERNIE 模型**，实现了一级分类标签模型。将数据预处理后输入模型进行训练，保存验证集损失值最小的模型作为最优模型。将测试集输入最优模型得到对应的一级分类标签，采用 F-score 指标对结果进行评价，计算出 **F-score 的值约为 0.91，查准率约为 0.91，查全率约为 0.92。**

针对问题二，本文建立了热度评价指标模型：**热度指标=(w1*留言数量+w2*平均互动人数)/间隔天数**。将预处理后的数据经过 **Birch 聚类**算法聚类成多类问题，再计算每类问题的留言数量、平均互动人数、间隔天数等特征并进行标准化处理，最后对特征加权求和计算出每类问题的热度评价指标，热度评价指标最大的五个问题即为所需热点问题。

针对问题三，本文从答复的相关性、完整性、可解释性等方面建立了答复质量评价模型：**评价指标=(w1*语义相关度+w2*法律条款数量+w3*文本长度)/时间间隔**。对预处理后的每条问题和答复计算其语义相关度、法律条款数量、文本长度、时间间隔并进行标准化处理，再进行加权求和计算每条答复的评价指标，并结合分布直方图设置阈值将答复分为**低质量答复和高质量答复**。

关键词：智慧政务；ERNIE；TF-IDF；Birch 聚类；

Abstract

This article aims to build a corresponding model based on the original text data and manually-labeled text classification labels, using text mining analysis methods, so that the computer can solve the tasks of text classification, hot spot problem mining, and evaluation system construction of reply opinions. At present, the above text tasks are mainly carried out by manual labeling, which is time-consuming, labor-intensive, and has low accuracy. Therefore, completing the above tasks through a computer is of great significance to the improvement of the "smart government" system.

For problem one, based on the deep learning ERNIE model, a first-level classification label model is implemented. The data is preprocessed and input into the model for training, and the model with the smallest loss value in the validation set is saved as the optimal model. The test set is input into the optimal model to obtain the corresponding first-level classification label, and the results are evaluated using the F-score index. The value of the F-score is calculated to be about 0.91, the precision rate is about 0.91, and the recall rate is about 0.92.

To solve the second problem, this paper establishes a model for evaluating the heat index: heat index = (w1 * number of messages + w2 * average number of people interacting) / interval days. The pre-processed data is clustered into multiple types of questions through the Birch clustering algorithm, and then the characteristics of the number of messages, the average number of interactions, and the number of days between each type of questions are calculated and standardized. Finally, each type is weighted and summed to calculate each type The hot evaluation index of the problem, the five hottest evaluation indexes are the hot issues needed.

For question three, this article establishes a response quality evaluation model from the aspects of relevance, completeness, and interpretability of the response: evaluation index = (w1 * semantic relevance + w2 * number of legal clauses + w3 * text length) / interval . Calculate the semantic relevance, number of legal clauses, text length, time interval of each question and answer after preprocessing and standardize it, then perform weighted sum to calculate the evaluation index of each answer, and

set the threshold value in conjunction with the distribution histogram The responses are divided into low-quality responses and high-quality responses.

Key word: Smart government affairs；ERNIE; TF-IDF; Birch clustering;

目录

1 问题重述

1.1问题背景

近年来，许多网络问政平台逐步成为政府了解民意的重要渠道，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。

1.2要解决的问题

1) 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。
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，定义合理的热度评价指标，并给出评价结果，按表 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 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”。
3) 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 问题一：群众留言分类

2.1问题分析

问题一旨在根据题目所提供的文本数据及人工标记的一级分类标签，建立文本分类模型并训练模型，最终得到效果较好的文本分类模型。

2.2数据预处理

由于本文使用的模型可以直接对字符级数据进行训练，因此在数据预处理阶段仅需将无用的标点符号使用正则表达式去除，表达式如下：

$$\text{rule} = u\text{"[\^a-zA-Z0-9\u4E00-\u9FA5]"} \quad (1)$$

2.3模型建立

■ ERNIE 简介

ERNIE 通过建模海量数据中的词、实体及实体关系，学习真实世界的语义知识。相较于 BERT 学习原始语言信号，ERNIE 直接对先验语义知识单元进行建模，增强了模型语义表示能力，以 Transformer 为网络基本组件，以 Masked Bi-Language Model 和 Next Sentence Prediction 为训练目标，通过预训练得到通用语义表示，再结合简单的输出层，应用到下游的 NLP 任务，在多个任务上取得了 SOTA 的结果。其可用于文本分类、序列标注、阅读理解等任务。

相比于 BERT 基于字的随机 mask，ERNIE 改进了 masking 策略，使其能够基于短语和命名实体进行随机 mask，保留了更多的知识和语义。

根据附件 2 中的留言主题和一级分类标签，本文利用 ERNIE 模型对留言进行文本分类。

■ ERNIE 程序架构

ERNIE 是一个持续的语言理解预训练框架，其中可以逐步构建和通过多任务学习来学习预训练任务。在此框架中，可以随时逐步引入不同的自定义任务。例如，利用包括命名实体预测，话语关系识别，句子顺序预测在内的任务，以使模型能够学习语言表示。

最终建立的程序架构如下

图 1　ERNIE 模型的程序架构

2.4模型求解

1. 实验条件

实验的硬件环境为服务器：Ubuntu 16.04、GPU：Nvidia GTX 1080Ti；

实验的软件环境为：python 3.7、Tensorflow 1.3.0、Keras 2.3.1。

2. 数据选择

本次实验将原始数据打乱后按 7:2:1 的比例划分为训练集、验证集、测试集。其中训练集和验证集输入模型中进行训练，测试集用于最优模型的结果评估。

3. 算法流程

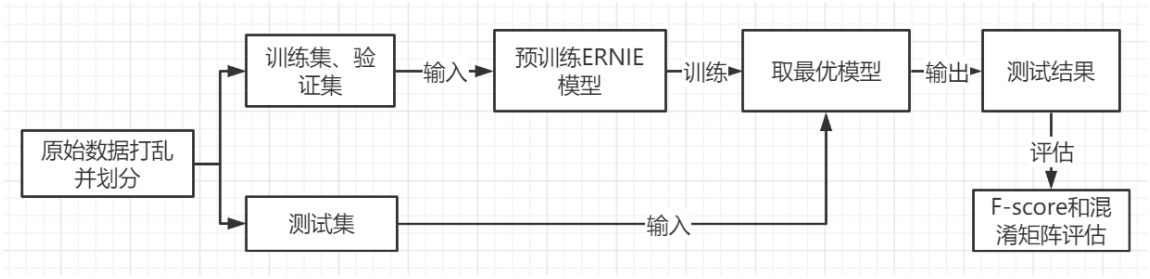

图 2 问题一的算法流程图

4. 模型的训练

模型的主程序是 run.py 文件，在命令行中输入 python run.py --model ERNIE 即可开始训练模型。训练前可通过 models 文件夹下的 ERNIE.py 文件修改模型参数，本文训练模型的主要参数及解释如下表所示：

表 1 模型主要参数

参数	值
NUM_EPOCHS（训练周期数）	10
BATCH_SIZE（批次大小）	64
PAD_SIZE（句子截取长度）	32
LEARNING_RATE（学习率）	5e-5
HIDDEN_SIZE（隐藏层神经元个数）	768

部分训练过程如下表：

表 2　　部分训练过程

Epoch [7/20]
Iter:1220，Train Loss: 0.12, Train Acc: 96.88%, Val Loss: 0.48, Val Acc: 89.58%, Time: 0:08:55
Iter:1230，Train Loss: 0.01, Train Acc: 100.00%, Val Loss: 0.48, Val Acc: 89.20%, Time: 0:08:59
Iter:1240, Train Loss: 0.014, Train Acc: 100.00%, Val Loss: 0.49, Val Acc: 89.58%, Time: 0:09:04
Iter: 1250, Train Loss: 0.0029, Train Acc: 100.00%, Val Loss: 0.5,Val Acc: 89.09%, Time: 0:09:08
Iter:1260,Train Loss: 0.0033, Train Acc: 100.00%,Val Loss: 0.51, Val Acc: 89.36%, Time: 0:09:12
No optimization for a long time, auto-stopping...

5. 模型的测试

从训练过程中可以看出，训练集的准确率已经完全拟合，验证集准确率约 89%，损失函数的值约 0.5，接着使用测试集对保存的最优模型进行测试，得到的准确率约为 91%，损失函数的值约为 0.3。

2.5模型评价

■ **F-score**

文本多分类问题采用 F-score 指标进行评价，它是将查准率和查全率进行加权调和平均得到，常用于评价分类模型的好坏，公式如下：

$$F-score=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_I+R_i} \qquad (2)$$

其中，P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。F-score 的取值范围是[0,1]，取值越接近 1 表明模型的分类效果越好。在本实验中，计算出 **F-score 的值为 0.91**。

■ **混淆矩阵**

表 3　　ERNIE 文本分类模型的混淆矩阵

预测类别 / 真实类别	**类别 1**	**类别 2**	**类别 3**	**类别 4**	**类别 5**	**类别 6**	**类别 7**
类别 1	64	0	0	0	2	0	0
类别 2	1	193	5	1	3	3	5
类别 3	0	4	79	0	1	1	0

类别 4	0	2	2	90	10	5	2
类别 5	6	3	0	5	170	1	6
类别 6	0	3	1	2	2	145	2
类别 7	0	0	0	1	3	0	95

从表中可以看出，分类错误主要集中在类别 2、4、5、6 上，其余类别的准确率较好。结合 F-score 指标和混淆矩阵，说明模型能对绝大多数文本进行正确的分类，模型的总体效果较好。

3 问题二：热点问题挖掘

3.1问题分析

问题二需要将语义相近的文本内容聚合成热点问题，同时给出热点问题的评价方案，最终整理成指定的文件格式输出。

3.2解题思路

首先对数据进行预处理后采用**TF-IDF 算法**构建文本权重矩阵，再采用**Birch 层次聚类算法**将文本进行聚类；

其次需要构建热度指数模型计算各类问题的热度指数，再根据热度指数取出排名前五的问题作为热点问题。

最后，题目中要求输出的热点问题表中，**地点/人群**和**问题描述**两个参数同样需要构建模型进行获取。**地点/人群**的获取采用 **TextRank 算法**。将同一类别的文本合并后，根据 TextRank 算法计算每个词语的重要程度并根据词性过滤出最重要的三个，如地点获取仅选取词性为 ns(地点名词)、nz(专有名词)两类，人群选择词性为 n(名词)的词语；**问题描述**则是根据 **PageRank 算法**计算同一类别的留言主题中得分最高的一条作为问题描述。数据进行预处理后，使用词向量的平均值构建各句子间的相似度矩阵，最后采取 PageRank 算法构建图结构并计算出各个主题的得分。

解题流程如下：

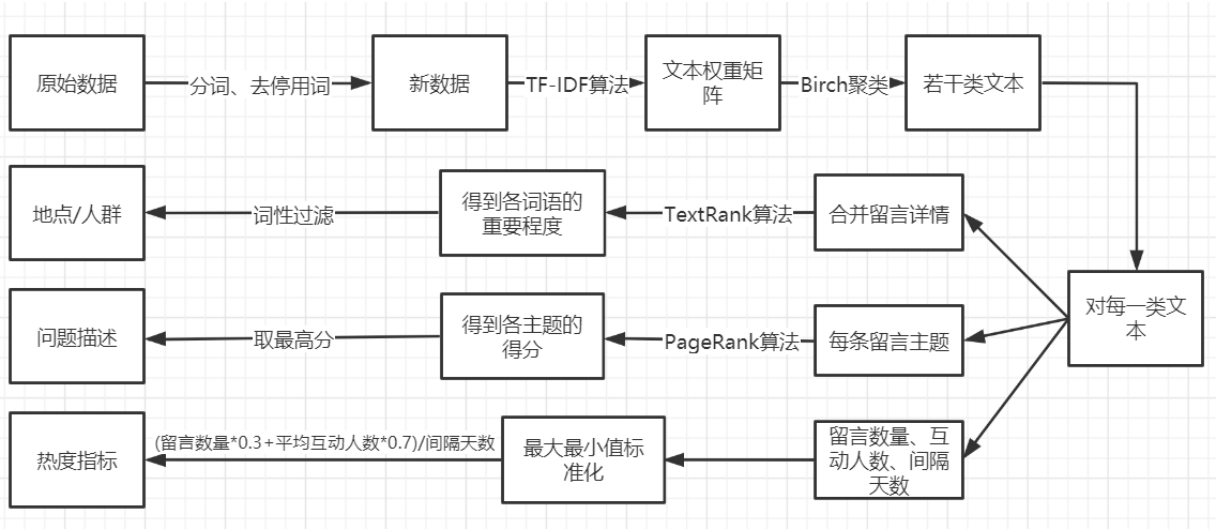

图 3 问题二的算法流程图

3.3热度指标模型

热度指数用来判断问题是否为热点问题，本文认为，热点问题应符合**留言数量多、互动人数多、间隔天数短**三个特点。互动人数越多，越能说明问题的受关注度高，因此需要赋予其更高的权重，最终确定 w1 为 0.3，w2 为 0.7；间隔天数短更能符合“热点”这一要求。互动人数采用参与点赞和反对的总人数代替，再除去留言数量得到平均互动人数，计算公式如下：

热度指标=(w1*留言数量+w2*平均互动人数)/间隔天数 （3）

构建的程序架构如下：

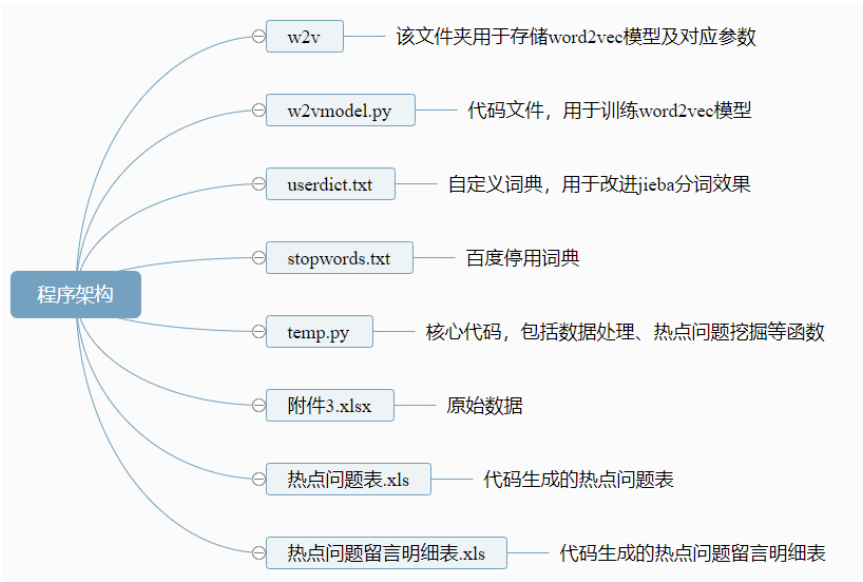

图 4 热度指数模型的程序架构

直接运行 temp.py 文件即可针对附件 3 数据生成热点问题表及留言明细表。

3.4算法介绍

■ TF-IDF 算法

TF-IDF（term frequency－inverse document frequency，词频-逆向文件频率）是一种用于信息检索（information retrieval）与文本挖掘（text mining）的常用加权技术。

TF-IDF 是一种统计方法，用以评估字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。字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。其主要思想为，假设某个词在一篇文档中出现的频率高（即词频较高），并且在其他文档中出现较少（即逆文档频率高），则认为这个词条具有很好的类别区分能力。

（1） TF 是词频（Term Frequency）

词频（TF）：表示词条(关键字）在文档中出现的频率。

计算公式如下：

$$\mathrm{t}f_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{4}$$

$$\mathrm{T}F_W = \frac{\textbf{在某一类中词条出现的次数}}{\textbf{该类中所有的词条数目}} \tag{5}$$

其中，n_{ij} 是该词在文件中出现的次数，分母则是文件中所有词汇出现的次数总和。

（2）IDF 是逆向文件频率（Inverse Document Frequency）

逆向文件频率 (IDF)：某一特定词语的 IDF，可以由总的文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。如果包含词条 t 的文档越少，IDF 越大，则说明词条具有很好的类别区分能力。

公式如下：

$$\mathrm{id}f_i = \log \frac{|D|}{|\{j:t_i \in d_j\}|} \tag{6}$$

其中，$|D|$是语料库中的文件总数，$|\{j:t_i \in d_j\}|$表示包含词语 t_i 的文件数目。如果该词语不在语料库中，就会导致分母为零，因此一般情况下会将分母加 1。

$$\mathrm{IDF} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含词条w的文档数+1}}\right) \quad (7)$$

（2） TF-IDF 实际上是：TF*IDF

某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。即公式为：

$$\mathrm{TF} - \mathrm{IDF} = \mathrm{TF} * \mathrm{IDF} \quad (8)$$

■ Birch 聚类算法

BIRCH（Balanced Iterative Reducing and Clustering Using Hierarchies）全称是：利用层次方法的平衡迭代规约和聚类。BIRCH 是一种聚类算法，它最大的特点是能利用有限的内存资源完成对大数据集的高质量的聚类，同时通过单遍扫描数据集能最小化 I/O 代价。

BIRCH 算法利用了一个树结构来帮助我们快速的聚类，这个数结构类似于平衡 B+树，一般将它称之为聚类特征树(Clustering Feature Tree，简称 CF Tree)。这颗树的每一个节点是由若干个聚类特征(Clustering Feature，简称 CF)组成。适用于数据量大，特征列少的数据集。

■ PageRank 算法

PageRank 算法是 Google 用于研究网页重要性评价的一种方法，是 Google 用来衡量一个网站的好坏的唯一标准。其核心思想是利用网页链接到其它页面和其它页面链接到本页面的情况来计算页面的分值，并对其进行分值排序。在使用 PageRank 算法对网页重要性进行排序时，一般会假设两个前提：

1. **数量假设：**假设存在一个网页，如果有大批的链接与它相连，则表示这个页面非常重要；本文中的节点，假设有大量的其它节点与它相连，则表示该节点非常受到信任。
2. **质量假设：**在互联网中，各网页的内容质量参差不齐，如果一个高质量的页面与该页面相连，则传递给该页面的权重应该很高；同样，如果有一个信任度很高的节点信任本节点，则传递给本节点的权重就会很高。

基于这两个假设，算法基于下式对页面等级进行评估时，对于每一个页面的权重赋予相同的值，然后通过不断的递归迭代计算，更新每一个页面的得分，直到页面得分稳定。

$$PR(u) = \sum_{v \in B_u} \frac{PR(v)}{L(v)} \quad (9)$$

其中 L(v)表示网页 v 的出链数量，PR(v)表示网页 v 的 PageRank 值，B_u表示网页 u 的入链集合。从该公式不难看出，每个页面的 PageRank 值是由其所有入链网页的值 PageRank 累加得到。

PageRank 算法规定一个页面只能对其他页面进行一次投票，如下图所示。所以图中页面 B 只能给页面 A 投半票，即页面 B 只能将它链接权重值的一半赋予页面 A。同理页面 D 只能将其链链接权重值的三分之一赋予页面 A。

则节点 D 的 PR 值计算公式如式所示:

$$PR(D) = \frac{PR(A)}{3} + \frac{PR(B)}{2} \quad (10)$$

- Textrank 算法

TextRank 自动摘要算法源于 Google 公司提出的 PageRank 算法，主要思想是把文档划分为由若干文本单元(词语或者句子)构成的节点，文本单元间的相似度构成节点间的边形成图模型，利用 PageRank 算法对图模型进行迭代计算，直到节点的累加权重收敛，然后根据累加权重值对所有节点进行排序，输出关键句，算法流程如下所示。

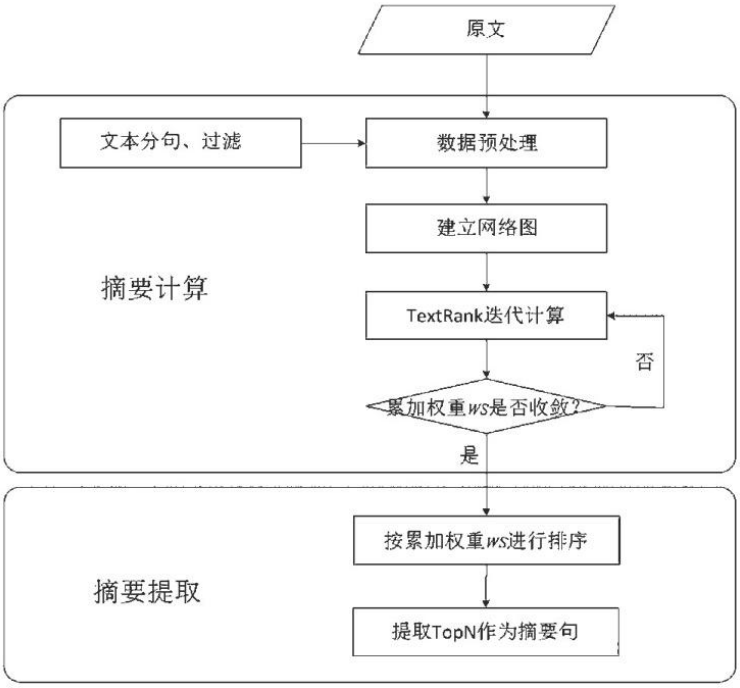

图 5 TextRank 摘要算法流程

3.5结果评估

在本实验中，设置类簇数量 K=100 进行聚类分析，并计算热度指数所得排名前五的热点问题如下表：

表 4　　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	38	262.9	2019/07/02 至 2019/09/01	A 市滨河车位	无视职工意愿职工权益 A 市伊景园滨河苑车位捆绑销售
2	100	78.8	2019/11/13 至 2020/01/15	新城 A 市搅拌站	丽发新城小区旁建搅拌厂扰民
3	73	51.7	2019/01/08 至 2019/07/08	A 市南昌车贷	A 市 58 车贷恶性退出案件发布案情进展通报
4	55	35.9	2019/02/07 至 2019/09/06	A 市奥园高铁	渝长厦高铁长赣高铁征地路线 A6 区周边小区影响
5	19	22.3	2019/02/21 至 2020/01/07	A 市西地省平台	A 市余易贷 P2P 诈骗平台受害者民生艰难

从上表中可以看出，热度指数最高的问题是车位捆绑销售的事件，其次是搅拌厂扰民的问题，以这两个问题为例，查看其对应的留言是否均说明同一个问题，第一个热点问题对应的部分结果如下：

表 5　　部分热点问题详情

问题 ID	留言编号	留言用户	留言主题
38	188801	A909180	投诉滨河苑针对广铁职工购房的霸王规定
38	190337	A00090519	关于伊景园滨河苑捆绑销售车位的维权投诉
38	191001	A909171	A 市伊景园滨河苑协商要求购房同时必须购买车位
38	192739	A909188	请政府救救广铁集团的职工吧
38	195511	A909237	车位捆绑违规销售
38	195995	A909199	关于广铁集团铁路职工定向商品房伊景园滨河苑项目的问题
38	196264	A00095080	投诉 A 市伊景园滨河苑捆绑车位销售
38	199190	A00095080	关于 A 市武广新城违法捆绑销售车位的投诉
38	200085	A000104234	A 市市政建设开发有限公司对广铁职工住宅项目的操作合法吗？
38	204960	A909192	家里本来就困难,还要捆绑买卖车位
38	205277	A909234	伊景园滨河苑捆绑车位销售合法吗？！
38	205982	A909168	坚决反对伊景园滨河苑强制捆绑销售车位
38	207243	A909175	伊景园滨河苑强行捆绑车位销售给业主
38	209506	A909179	A 市武广新城坑客户购房金额并且捆绑销售车位
38	209571	A909200	伊景园滨河苑项目绑定车位出售是否合法合规
38	212323	A00020702	广铁集团要求员工购房时必须同时购买车位

从留言主题中可以很清晰地看出，内容均和车位捆绑销售有关。

第二个热点问题对应的部分结果如下：

表 6　　部分热点问题详情

问题 ID	留言编号	留言用户	留言主题
100	188809	A909139	A 市万家丽南路丽发新城居民区附近搅拌站扰民
100	189950	A909204	投诉 A2 区丽发新城附近建搅拌站噪音扰民
100	190108	A909240	丽发新城小区旁边建搅拌站
100	190523	A00072847	A 市丽发新城违建搅拌站，彻夜施工扰民污染环境
100	190802	A00072636	A 市丽发小区建搅拌站，噪音污染严重
100	203393	A00053065	A 市丽发新城小区侧面建设混泥土搅拌站，粉尘和噪音污染严重
100	213464	A909233	投诉丽发新城小区附近违建搅拌站噪音扰民
100	213930	A909218	A2 区丽发新城附近违规乱建混凝土搅拌站谁来监管？
100	216824	A909214	搅拌站大量加工砂石料噪音污水影响丽发新城小区环境
100	217700	A909239	丽发新城小区旁的搅拌站严重影响生活
100	225217	A909223	A2 区丽发新城附近修建搅拌厂严重影响睡眠
100	231136	A909204	投诉 A2 区丽发新城附近建搅拌站噪音扰民
100	233158	A909242	丽发新城小区旁建搅拌厂严重扰民！
100	235362	A909215	暮云街道丽发新城小区附近水泥搅拌站非法经营何时休
100	238212	A909203	丽发新城小区附近建搅拌站合理吗？
100	243692	A909201	丽发新城小区附近的搅拌站噪音严重扰民

同样可以从留言主题中看出，内容均与搅拌站扰民有关。

通过上述两个例子可以看出，所挖掘出的问题确实是群众大量反映的热点问题，本实验在文本聚类、热点问题挖掘上的表现较好。

4 问题三：答复意见的质量评价

4.1问题分析

问题三旨在根据现有的留言信息和答复信息，并结合答复时间等信息，建立模型评估答复意见的质量。

4.2模型建立

首先需要对答复意见的质量进行定义，本文将答复质量分为**高质量答复和低质量答复**，模型如下：

评价指标 =(w1*语义相关度+w2*法律条款数量+w3*文本长度)/时间间隔（11）

该模型主要从以下四个方面衡量文本质量：

1、**语义相关度**。指留言详情和答复意见的语义是否相近或相似，主要通过 TF-IDF 算法将文本向量化后，计算余弦相似度所得，语言相关度越高说明答复

的相关性越好；

2、**答复意见的文本长度**。在大多数情况下，文本情况越长，则所包含的信息量越大，同时也代表答复的完整性更好；

3、**法律条款数量。**在答复中提到相关法律条款数量越多，说明答复的可解释性越好。

4、**处理留言的时间间隔**。对于留言中的问题，留言者越早收到答复越能更好的处理问题，代表答复方的效率较高，此处采用“天”为单位。

在上述四个特征中，语义相关度的权重最高，法律条款数量其次，文本长度最低。因此，权重定义为：w1=0.5，w2=0.3，w3=0.2.

解题流程图如下：

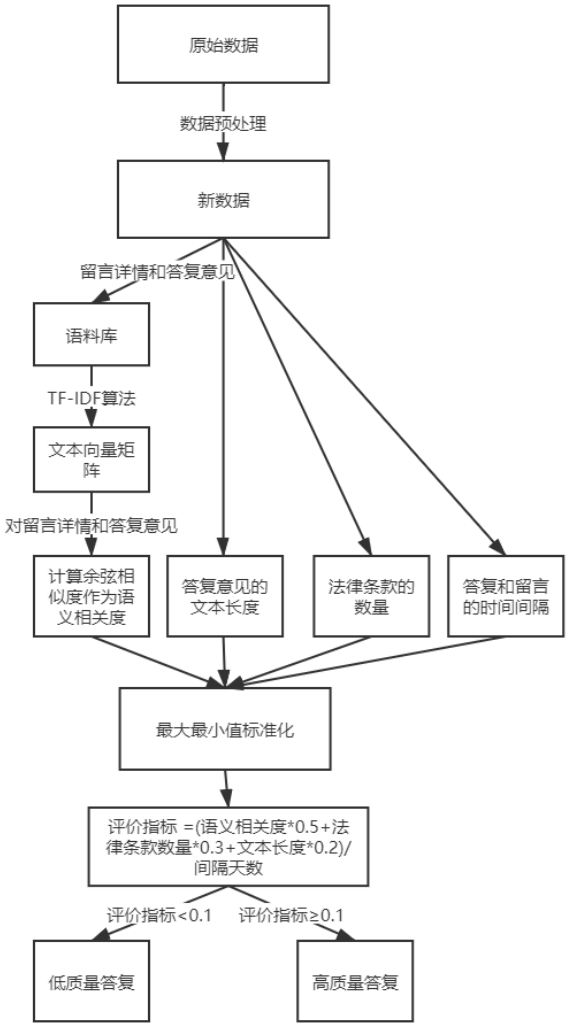

图 6 问题三的解题流程图

制作的程序架构如下图：

图 7 答复质量评价模型的程序架构图

直接运行 temp.py 文件即可根据质量评价模型从附件 4 中抽取十条低质量答复生成 random_result.txt 文件。

其中，阈值的确定是通过观察评价指标的分布直方图，并不断缩小观察窗口，最终确定 0.1 为阈值较为合适，评价指标低于 0.1 被视为低质量答复，共得到 77 条低质量答复。阈值越大，被判断为低质量的答复越多，对答复的要求越高。

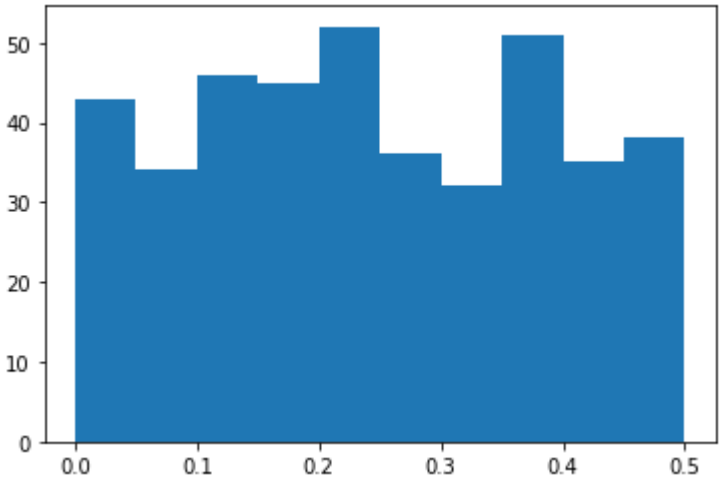

图 8 评价指标[0,0.5]区间的数量分布图

4.3模型评价

对程序随机抽取 10 条低质量答复进行分析：

表 7 随机抽取的十条低质量答复

序号	低质量答复
1	“叮了咚”网友： 您好，请您详细咨询公安局户政科。（0000-00000000）
2	“UU008780” 您好！您反映的问题我局已收悉，已派执法人员到大庆

	坪乡现场调查取证，如存在占用农田行为，我局将依法立案查处。
3	网友：您好！留言已收悉
4	您的留言已收悉。关于您反映的问题，已转市文旅广电局调查处理。
5	“UU0081637”您好！您反映的情况已转相关部门核实、办复。感谢您的留言。
6	“UU0081222” 您好！根据相关政策规定，参保人员的养老金是从社会保障行政部门审批退休的次月起开始发放。 如有疑问，欢迎致电0731-0000-00000000。 2017 年 12 月 7 日
7	网友：您好！留言已收悉
8	2018 年 12 月 12 日
9	您好，留言收悉。您的问题请致电西地省律师协会办公室0731-0000-00000000 咨询。谢谢。 2018 年 3 月 2 日
10	网友：您好，留言收悉，现回复如下：此事正在报送审批中，感谢您对M2 县教育的关心和支持，祝您生活愉快！

结合生活实际可以看出，除第 2、6 条答复相对完整一些外，其余答复基本存在未解答问题、解答不全面的情况，均可认定为低质量答复，模型结果较好。

5 模型的改进与局限性

5.1模型的改进

在本文中，对于第一问中文本分类的正确率达到了 91%，但仍然还有提升空间。比如由于数据量较小使得模型的泛化能力不足，可以使用 EDA 模块对数据进行同义词替换、随机插入、随机交换等方式新增数据，使模型更加健壮；第二问中热点问题的挖掘，对于热点问题有多种不同的定义方式，应当针对不同地使用场景使用不同的权重改进模型；第三问中对于答复意见的质量评+价，直接构建模型的效果并非最佳，可以先通过人工对数据进行标注，再将其放入模型中进行训练，可达到更好的效果。另外二、三问中计算语义相似度主要使用 TF-IDF+余弦相似度计算，可使用更复杂的深度神经网络代替。

5.2模型的局限性

在本文中，第一问的文本分类模型受限于数据量的大小，模型在泛化能力上有所不足，模型准确率有待提高；第二问中的热点问题挖掘模型，局限性主要体现在对于热点问题的定义以及，不同的场景下效果会有所不同；第三问中低质量答复的筛选主要取决于阈值的选取，准确率和召回率受其影响较大。

参考文献

[1]戴炳荣,姜胜明,李顿伟,李超.基于改进 PageRank 算法的跨链公证人机制评价模型[J/OL].计算机工程:1-9[2020-03-31].

[2]石元兵,周俊,魏忠.一种基于 TextRank 的中文自动摘要方法[J].通信技术,2019,52(09):2233-2239.

[3]杨秀璋,夏换,于小民,武帅,赵紫如,窦悦琪.基于特征词典构建和 BIRCH 算法的中文百科文本聚类研究[J].计算机时代,2019(11):23-27+31.