# “智慧政务”中的文本挖掘应用

## 摘要

本文旨在于对某网络问政平台群众留言热点问题进行分析，通过群众留言记录以及相关部门的答复意见进行相关操作，实现对群众留言的一级标签的分类、热点问题挖掘以及制定答复意见评价方案三个方面的功能，以此分析出相关工作人员需要解决的问题，从而减轻相关部门的负担并且使得群众问题尽快获得解决。

**针对问题一：**本文建立了 Bert 模型改来解决群众留言的一级标签的分类问题，首先对群众留言信息进行冗余、缺失值消除处理；其次利用构建的 Bert 模型获得留言文本的特征向量，然后将提取的特征向量输入 softmax 回归模型中进行分类训练；结果基于 Bert 的群众留言分类模型的 F1 值达到 91.21%，随后发现对模型改进后的 ERINE 模型的 F1 值高达 94.14%，因此将利用 ERINE 模型解决分类问题。

**针对问题二：**本文定义了留言数量、互动情况、时效性三个评价指标，并建立了热度问题综合评价模型为：**热度指数=**$w_1$***留言数量+**$w_2$***互动指数+**$w_3$***时效性指数**。首先对留言信息进行清洗、分词等预处理，其次利用 TF-IDF 提取特征权重后进行 DBSCAN 密度聚类，然后求得热度指数公式进行热度排序，最后通过 TextRank 算法抽取每个热点问题的问题描述和地点人群，以此获得热度问题表。

**针对问题三：**本文定义了相关性、完整性、时效性、专业性四个评价指标，并建立了答复意见质量综合评价模型为：**答复意见质量评价=**$w_1$***相关性指数+**$w_2$***完整性指数+**$w_3$***时效性+**$w_4$***专业性指数**。首先对数据进行预处理，其次利用 TF-IDF 提取关键词后进行 Word2Vec 获取语义特征向量，并用余弦相似度求得相关性指数，然后文本熵求得完整性指数、时间间隔求得时效性指数、自定义词典匹配求得专业性指数，最后根据答复意见质量评价公式查看各答复意见的质量。

**关键词：**智慧政务；留言分类；热点问题；答复意见评价；Bert 模型；DBSCAN 聚类；TextRank 算法；TF-IDF 算法；Word2Vec 模型

# Abstract

The purpose of this paper is to analyze the hot issues of the public comments on a network political platform. Through the relevant operation of the records of the public comments and the replies of the relevant departments, the functions of the first-class label of the public comments, the mining of hot issues and the formulation of the evaluation plan of the replies are realized, so as to analyze the problems to be solved by the relevant staff, so as to reduce the Lighten the burden of relevant departments and solve the mass problems as soon as possible.

Aiming at the first problem: This paper establishes the Bert model to solve the problem of the first level label classification of the public message. Firstly, the redundant and missing value of the public message information is eliminated. Secondly, the feature vector of the message text is obtained by using the Bert model, and then the extracted feature vector is input into the softmax regression model for classification training. Results the classification of the public message is based on Bert The F1 value of the model is 91.21%. It is found that the F1 value of the improved erine model is as high as 94.14%. Therefore, the erine model will be used to solve the classification problem.

Aiming at the second problem: This paper defines three evaluation indexes: the number of messages, interaction and timeliness, and establishes a comprehensive evaluation model of the heat problem: the heat index = $w_1$* the number of messages + $w_2$* the interaction index +$w_3$* the timeliness index. Firstly, the message information is preprocessed by cleaning and word segmentation. Secondly, the feature weight is extracted by TF-IDF, and then the DBSCAN density clustering is carried out. Then the heat index formula is obtained to sort the heat. Finally, the problem description and location population of each hot issue are extracted by textrank algorithm to obtain the heat issue table.

In response to question 3: This paper defines four evaluation indexes: relevance, integrity, timeliness and professionalism, and establishes a comprehensive evaluation model for the quality of reply opinions as follows: quality evaluation of reply opinions = $w_1$* relevance index + $w_2$* integrity index +$w_3$* timeliness + $w_4$* professionalism index. First, preprocess the data, then extract the key words by TF-IDF, then get the semantic feature vector by word2vec, and get the correlation index by cosine similarity, then get the integrity index by text entropy, get the timeliness index by time interval, get the professional index by self defined dictionary matching, and finally check the quality of each reply according to the reply quality evaluation formula.



Keywords: smart government; message classification; hot issues; response evaluation; Bert model; DBSCAN clustering; textrank algorithm; TF-IDF algorithm; word2vec model

目录

# 1 问题重述

## 1.1 问题背景

近年来，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。而各类社情民意相关的文本数据量过大，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。其次，通过建立基于自然语言处理技术，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 问题重述

(1) 请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

(2) 请根据附件 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 Bert 模型的建立

已知问题要求根据附件 2 建立关于留言内容的一级标签分类模型，由于留言内容文本形式，因此本文将建立适用于文本分类的 Bert 模型对一级标签分类，并利用 F-score 评价分类结果。

建立 Bert 模型，首先 Bert 模型是基于 Transformers 转换器双向编码表征的模型，内部进行多次解码编码及即双向实现编码，最终获得特征向量并根据不同任务微调参数实现最终的目标。就如本文要实现句子的分类，先根据 Bert 模型生成一组特征向量，并通过一层全连接进行微调，然后损失函数是根据具体任

务自行设计，多分类就会运用 softmax 分类器来实现最终文本分类的目标。

### 2.1.1 Bert 模型简介

Bert 模型（Bidirectional Encoder Representations from Transformers）就是基于双向 Transformers 的 Encoder 编码器实现的，当然为了更好地增强语义地表达能力，Bert 模型新增了两个的预训练任务。Bert 模型其实就是对输入的文本利用双向 Transformers 的编码器训练出结合语义以及句子词与词之间关系的特征向量，然后利用前向反馈神经网络连接一个全连接层实现相应的目标。

### 2.1.2 Bert 模型架构

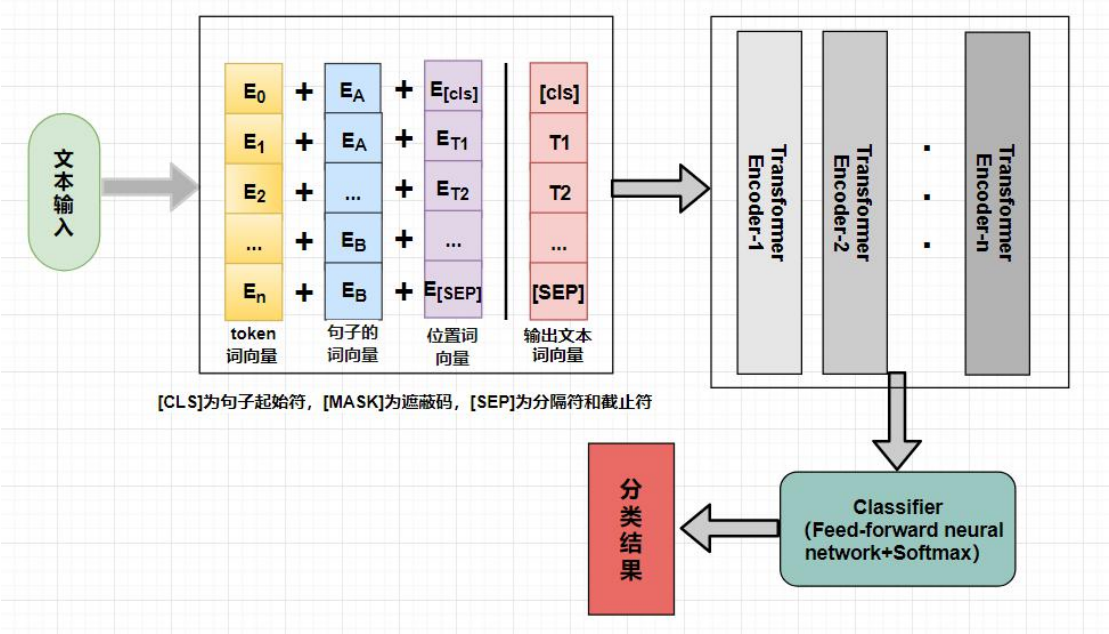


图 2-1 Bert 模型架构

如图 2-1 就是利用 Bert 模型解决分类模型的一个框架图，首先文本输入，其次对文本进行句子嵌入、位置嵌入、分割嵌入三个嵌入特征求和作为 Bert 模型的输入特征，然后 Bert 模型还进行 MLM、NSP 两个无监督学习的任务并利用 Transformer 的 Encoder 编码器进行特征向量化，最后通过前向反馈神经网络 Feed-forward neural network 连接一个全连接的 softmax 层，最终输出分类结果。其中模型中重点方法详解如下。

**Bert 模型的两个任务：**

（1）任务 1：遮蔽语言模型(Masked Language Model , MLM)

为了实现深度的双向表示，使得双向的作用让每个单词能够在多层上下文中间接的看到自己。其主要思想：随机屏蔽掉部分输入 token（替换为统一标记[MASK]），然后再去预测这些被屏蔽掉的 token，本文中会随机遮盖每个 sequence 中的 15%的 token。如图 2-2 所示。

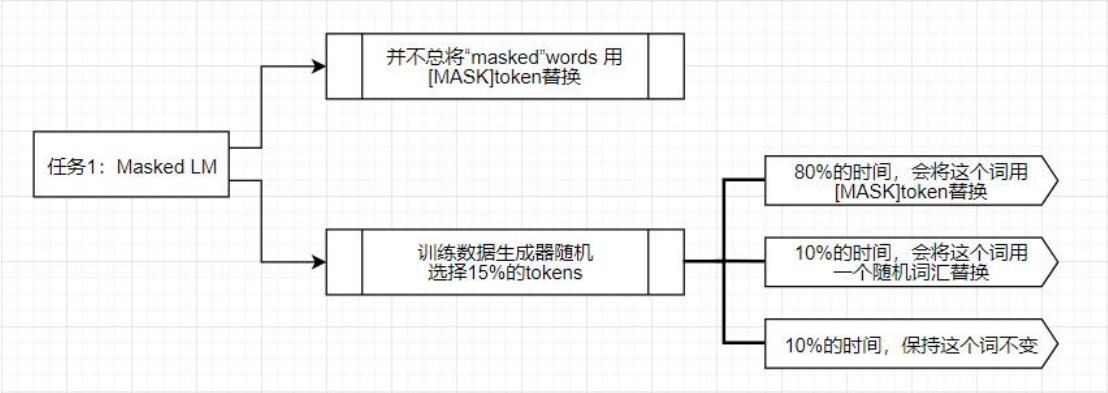


图 2-2 遮蔽语言模型

（2）任务 2：下一句子的预测(Next Sentence Prediction)

现在从句子的角度来考虑问题，为获取句子间的信息，预训练了一个二值化下一句预测任务，该任务可以从任何单语语料库中轻松生成，这点是语言模型不能直接捕捉到的。具体来说，选择句子 A 和 B 作为预训练样本：A 的下一句有 50%的可能是 B，另外 50%的可能是来自语料库的。如图 2-3 所示。

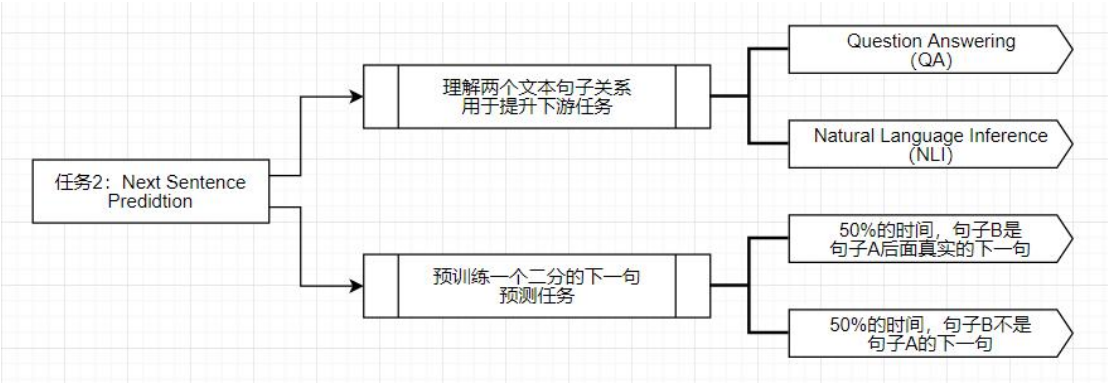


图 2-3 预测下一句子原理

**Bert 模型中 Transformer 的 Encoder 编码器：**

Encoder 的 Input 是一句话的字嵌入示，通过加上这句话里面每个字的位置信息，然后经过 Self-attention 层，帮助 Encoder 在编码某个字的时候通过查

看该字前后的字的信息，之后它的输出将会再通过一层 Add & Norm 层，Add 意为将 Self-attention 层的 Input 和 Output 进行相加，Norm 以为将相加过的 Output 进行归一化处理，这样可以使 Self-attention 层的 Output 存在固定的标准差和均值，标准差为 1，均值为 0，归一化后的向量列表将会再传入一层全连接的前馈神经网络，同样，Feed Forward 层也会有相对应的 Add & Norm 层处理，之后 Output 全新的归一化后的词向量列表。如图 2-4 所示。

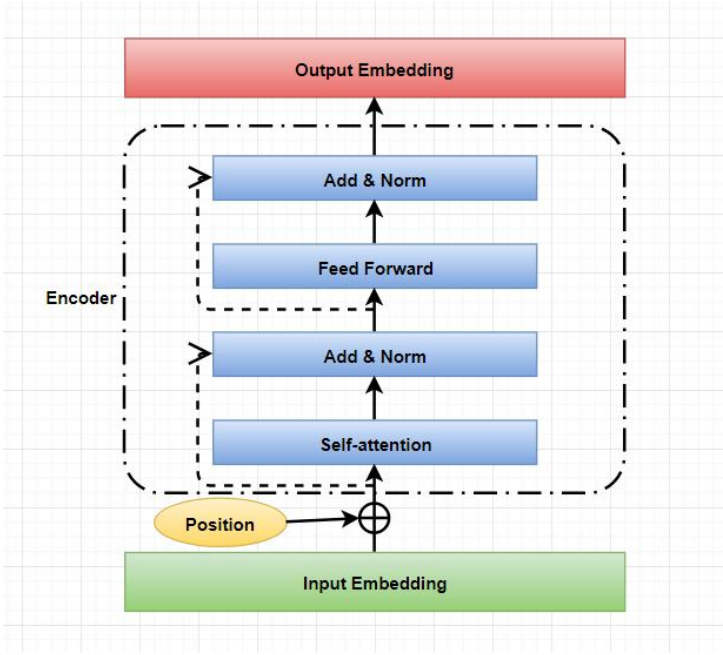


图 2-4 Transformer 的 Encoder 编码器

## 2.2 模型的求解

上面已经详述了 Bert 模型结构以及相关任务，现为实现分类目标，本文根据实际任务对 Bert 模型进行求解，求解流程如图 2-5 所示：

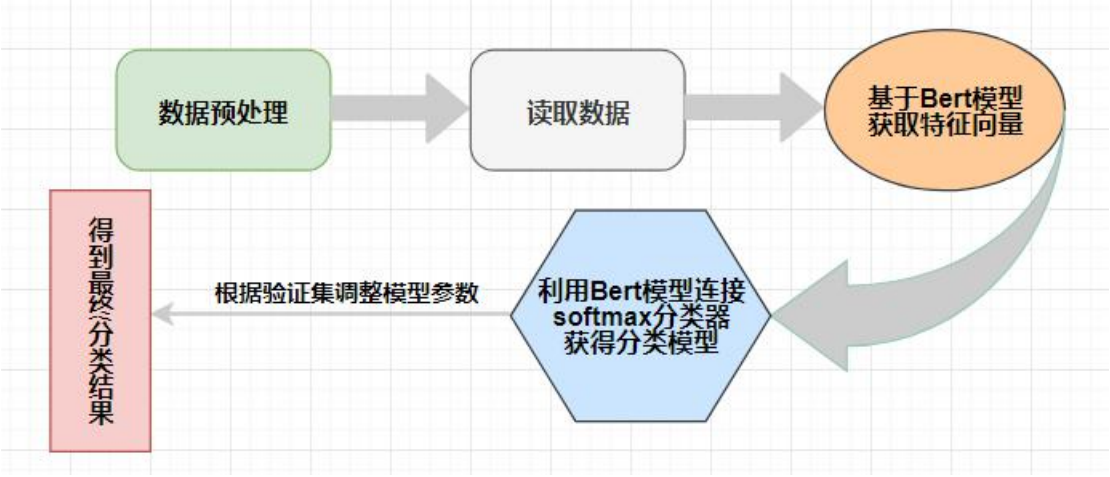


图 2-5 Bert 模型求解流程

### 2.2.1 数据预处理

首先，对附件一的数据进行分析，通过相关方法对数据进行留言信息进行冗余、缺失值消除处理。

其次，将附件 2 数据进行简单处理，首先读取留言详情及一级标签列，然后将文本标签转化为数字标签，其次将留言内容中存在的空格符以及转行等去除，最后将处理好的数据写进文档。

最后将数据集分为训练集、验证集、测试集并按 8：1：1 的比例分配。

处理后的部分数据示例如下：

请求查处 K7 县潇浦镇消江村三组村干部违法私卖土地..... .....消江村三组村民','2017 年 11 月 7 日 0
办理退休手续档案年龄与身份证年龄不符..... .....年龄与户口身份证年龄不符该怎么办？ 4
A9 市河镇头段非法排出大量污水，..... .....A9 市镇头镇金田社区（镇柏公路 1 公里处） 1
关于 A6 区县高塘岭实验小学小升初的问题','尊敬的孔玉成书记：..... .....谢谢。 3
A 市钱隆世家二期没达到交房标准，..... .....,并且存在房屋安全隐患和整体质量缺陷。 0
关于 D12 市 2017 新农保参保条件的咨询','..... .....年满 16 周岁的在校学生是否参保？谢谢！ 4
新计生奖励政策有待商榷！..... .....，应本着就高不就低原则，才相对合理、兼顾公平。 6
G8 县楚江街道办与嘉乐园一带焚烧有毒有害烟尘污染空气','..... .....二〇一七年十一月二十三日 1
强烈要求修建 K 市至通道的高速公路','尊敬的领导','..... .....。争取早日脱贫。 2
请求解决 K2 区楚江东路原味街项目部恶意托欠农民工工资的问题','..... .....，还我血汗钱！！！ 4
B 市潇楚实验学校只有 7 个班拥有普高学籍','..... .....。望学校加一改善以后不再出现这种情况。 3
请让 C 市的名校招生公开公平！','..... .....，让名校招生公开公平！ 3

### 2.2.2 Bert 模型的训练

首先由 2.2.1 数据简单预处理后作为 Bert 模型的输入文本，其次根据 2.1 Bert 模型对数据进行微调，然后训练得到特征向量，最后 softmax 对这些特征向量进行分类并训练模型。

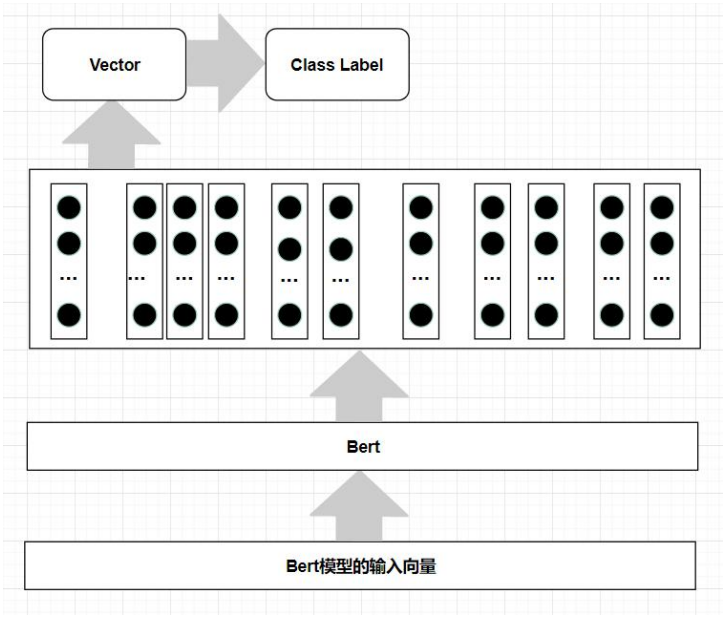


图 2-6 Bert 模型的特征向量化

■ **输入数据：**

输入训练数据集，并做嵌入向量化处理，然后假设处理后的训练集为 $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$，其中 $x_i$ 是每条留言内容文本且是以词为单位的 Bert 输入向量，$y_i$ 则是每条留言内容的所属类别标签，$i=1,2,\ldots,n$。

■ **输出分类：**

①使用 2.1 介绍的 Bert 模型对输入向量进过抽取语义特征，然后利用 Transformer 的 Encoder 编码器编码成特征向量 $V=(v_1,v_2,\ldots,v_n)$，其中 $v_i$ 表示每条留言内容的特征向量，$i=1,2,\ldots,n$。

②将上步骤获得特征向量 $V=(v_1,v_2,\ldots,v_n)$，如上图 2-6。通过前向反馈神经网络 Feed-forward neural network 连接 softmax 回归模型进行训练，最终获得文

本分类模型。

③获得分类结果，并通过验证数据集不断调整 Bert 模型参数。

**表 2-1 模型主要参数**

| 参数 | 含义 | 本文最终参数 |
|---|---|---|
| num_epochs | 模型迭代次数 | num_epochs=3 |
| batch_size | 每批训练集数据大小 | batch_size=16 |
| pad_size | 每句话处理成的长度(短填长切) | pad_size=300 |
| learning_rate | 学习率 | learning_rate=5e-5 |

利用 Bert 模型进行训练并测试，训练的迭代过程如下图，里面展示出训练数据和验证数据的损失值以及准确率，还有训练的时间，最终生成测试数据的损失值以及准确率，如图 2-7。

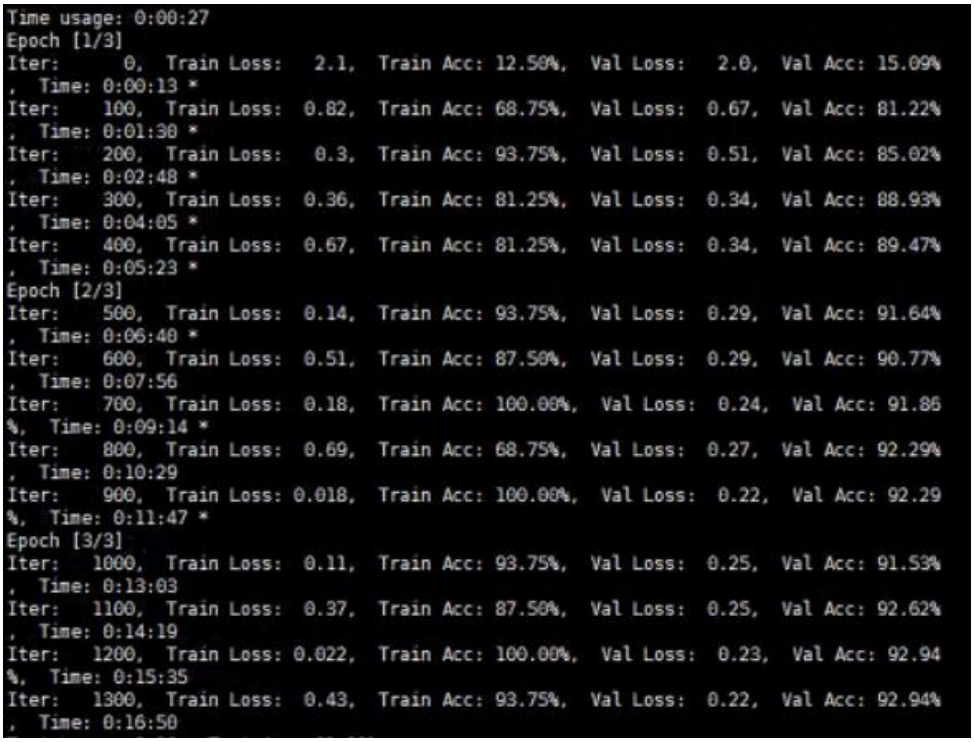


图 2-7 训练过程

Train Loss：训练过程中训练集的损失值；Train Acc：训练过程中训练集的准确率；

Val Loss：测试过程中验证集的损失值；Val Acc：测试过程中验证集的准确率；

Test Loss：最终测试集的损失值；Test Acc：最终测试集的准确率；

Epoch：一个 Epoch 是指将所有训练样本都训练一遍，该训练过程设置 Epoch=3。

Iter：是指使用 batch_size=16 个样本训练一次，每迭代 100 次就打印出相应结果，由于训练集有 7368 个样本，因此 Epoch 将迭代 400 多次并打印 5 次结果。

Time：训练过程所用的时间，最终用时 00:16:50

## 2.3 模型结果

### 2.3.1 评价指标

本文主要研究的问题是分类问题，而分类问题主要用于评价的指标有精准率 P 、召回率 R 以及 F1 值，而这些值的结果需要通过混淆矩阵来计算，其中混淆矩阵的介绍如表 2-2 所示：

表 2-2 混淆矩阵原理

| 混淆矩阵 | | 预测值 | |
|---|---|---|---|
| | | 真实预测 | 错误预测 |
| 真实值 | 真实为正例 | TP | FN |
| | 真实为反例 | FP | TN |

其中，TP（True Positive）为真阳性，即预测为正，实际也为正，

FP（False Positive）为假阳性，即预测为正，实际也为假，

FN（False Negative）为假阴性，即预测为假，实际也为正，

TP（True Negative）为真阴性，即预测为假，实际也为假。

■ 准确率 ACC

准确率 ACC 是指分类器所有预测正确样本占所有样本的比例，其计算公式为：

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{2-1}$$

■ 精准率 P

精准率 P 是指分类器真实预测且预测正确的样本占所有真实预测为正的样本的比例，其计算公式为：

$$P = \frac{TP}{TP + FP} \tag{2-2}$$

■ 召回率 R

召回率 R 是指分类器真实预测且预测正确的样本占所有真实为正的样本的比例，其计算公式为：

$$R = \frac{TP}{TP + FN} \tag{2-3}$$

■ **F-Score**

通常使用 F-Score 对分类方法进行评价：

$$\mathrm{F}_1 = \frac{1}{n}\sum_{i-1}^{n}\frac{2P_iR_1}{P_i + R_i} \tag{2-4}$$

其中 $P_i$ 为第 $i$ 类的查准率， $R_i$ 为第 $i$ 类的查全率。

$P_i$ 体现了模型对负样本的区分能力，$P_i$ 越高，模型对负样本的区分能力越强； $R_i$ 体现了模型对正样本的识别能力， $R_i$ 越高，模型对正样本的识别能力越强。F- score 是两者的综合，F- score 越高，说明模型越稳健。

### 2.3.2 评价结果

表 2-3 Bert 模型测试的混淆矩阵结果

| Class | 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 |
|---|---|---|---|---|---|---|---|
| 城乡建设 | 183 | 7 | 2 | 6 | 1 | 18 | 0 |
| 环境保护 | 2 | 92 | 0 | 0 | 0 | 0 | 0 |
| 交通运输 | 3 | 0 | 53 | 0 | 1 | 1 | 0 |
| 教育文体 | 0 | 0 | 0 | 146 | 0 | 3 | 1 |
| 劳动和社会保障 | 5 | 1 | 0 | 4 | 177 | 0 | 5 |
| 商贸旅游 | 1 | 0 | 1 | 2 | 2 | 95 | 6 |
| 卫生计生 | 0 | 0 | 0 | 0 | 6 | 3 | 94 |

Bert 模型训练出来对应 F-Score 结果，如图 2-8 所示：

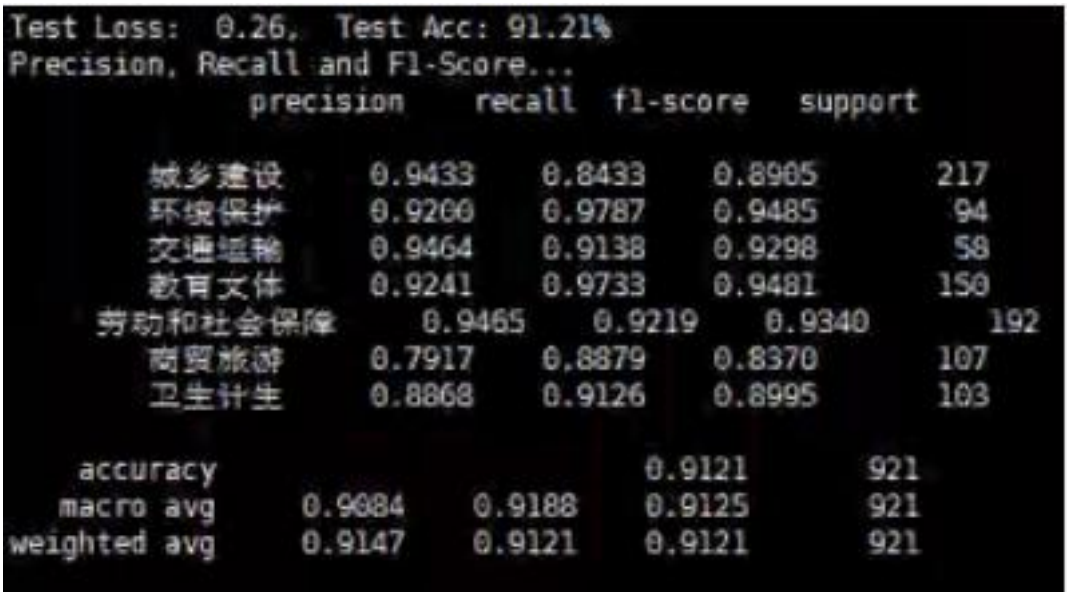


**图 2-8** Bert 模型训练的 F-Score

从图 2-8 中，可以看到 Bert 模型对群众留言的分类预测结果较好，对于测试集的预测准确率 ACC=91.21%，同样的我们可以发现每类的 F-score 值、精准率以及召回率都可达 80%以上。

## 2.4 模型的改进

### 2.4.1 ERNIE 模型

众所周知的，Bert 模型是由 Google 官方提出的，并运用于多方面且获得较好的结果，然而最近百度发布了增强知识的预训练模型 ERNIE，其在众多任务中都大幅度地超越了 Bert 模型，其跟 Bert 不同的是它有对训练数据的实体概念等先验语义知识进行建模，其中使得语义结构、词法结构、语法结构不被破坏，从而增强语义表达能力。

ERNIE 模型也是基于双向 Transformer 的 Encoder 的编码器实现的，不同就是还多了融合器，融合器即为知识编码器 K-Encoder，其不仅可以编码 token 和实体，也能融合异构特征，其实就是进行先验语义知识的建模。因此是 2.1 详述的 Bert 模型的改进版，简单地说根据建立的先验知识模型得知实体概念，然后在词

遮蔽任务时对整体实体语义的词连续遮蔽[MASK]。不同于 Bert 模型的融合层如图 2-9 所示。

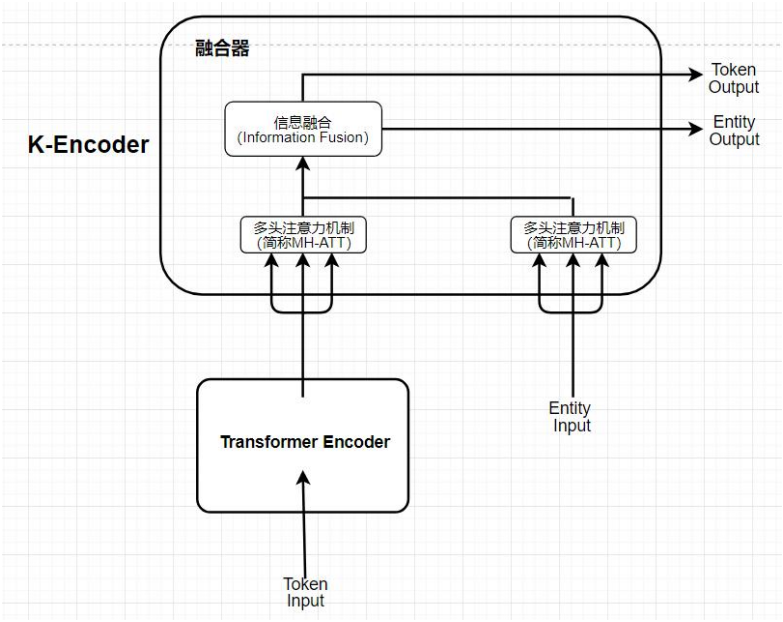


图 2-9 ERNIE 模型融合器

如下图 2-10，描述的即是利用 ERNIE 模型时对实体整体遮蔽，避免产生不必要的错误，同时增强语义表达。

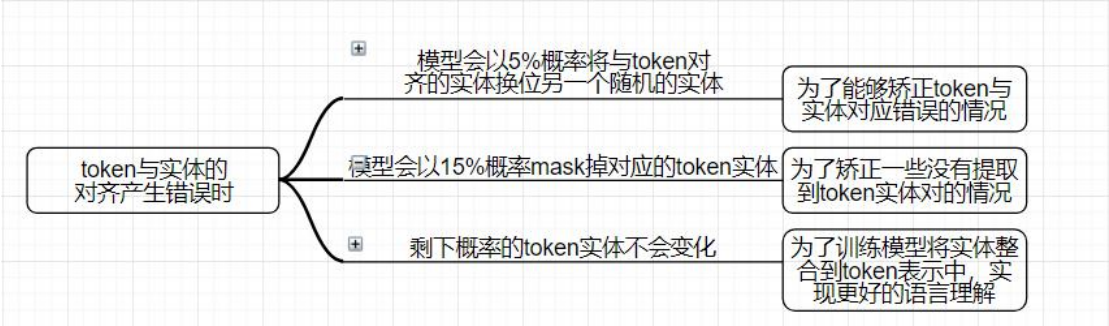


图 2-10 ERNIE 模型对实体整体遮蔽

## 2.4.2 改进后结果

利用 ERNIE 模型进行训练并测试，获得结果的评价测试仍然是利用 2.3.1 的评价指标——混淆矩阵，并获得每类的 F-score 值、精准率、召回率以及分类的准确率。

表 2-4 ERNIE 模型测试的混淆矩阵结果

| *Class* | 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 |
|---|---|---|---|---|---|---|---|
| 城乡建设 | 200 | 4 | 5 | 1 | 1 | 6 | 0 |
| 环境保护 | 3 | 91 | 0 | 0 | 0 | 0 | 0 |
| 交通运输 | 3 | 0 | 53 | 0 | 0 | 1 | 0 |
| 教育文体 | 4 | 0 | 0 | 146 | 0 | 0 | 0 |
| 劳动和社会保障 | 4 | 0 | 2 | 2 | 181 | 1 | 2 |
| 商贸旅游 | 4 | 0 | 0 | 4 | 0 | 97 | 2 |
| 卫生计生 | 0 | 0 | 0 | 0 | 2 | 3 | 98 |

ERNIE 模型训练出来对应 F-Score 结果，如图 2-12 所示：

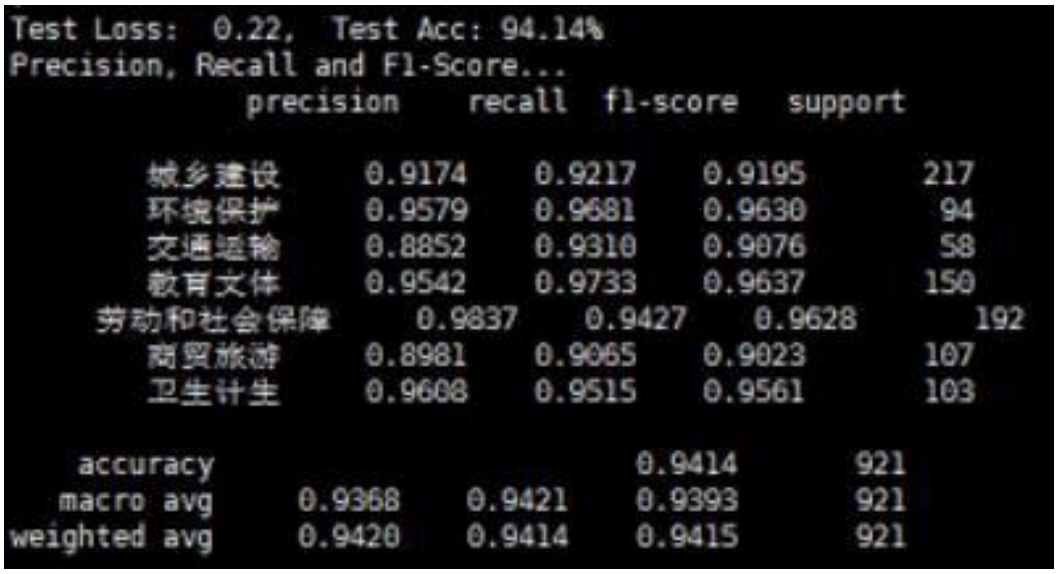


**图 2-12 ERNIE 模型训练的 F-Score**

从图 2-12 中，可以看到 ERNIE 模型对群众留言的分类预测结果比 2.3.2 Bert 模型训练的结果更胜一筹，对于测试集的预测准确率 ACC=94.14%，同样的我们

可以发现每类的 F-score 值、精准率以及召回率都可达 90%以上。然而，在该分类模型中，对于劳动和社会保障的分类效果格外突出，精准率率高达 98.37%。

### 2.4.3 结果对比

在上面两个模型的训练时为了能过进行对比，相应的参数设置时一样的，具体结果上面也详细叙述了对比结果表如下表 2-5。

**表 2-5 Bert 模型和 ERNIE 模型训练的 F-Score 对比**

| 模型 | Bert 模型 | | | | ERNIE 模型 | | | |
|---|---|---|---|---|---|---|---|---|
| 各指标 / Class | 精准率 P | 召回率 R | F-score | 准确率 ACC | 精准率 P | 召回率 R | F-score | 准确率 ACC |
| 城乡建设 | 0.9433 | 0.8433 | 0.8905 | | 0.9174 | 0.9217 | 0.9195 | |
| 环境保护 | 0.9200 | 0.9787 | 0.9485 | | 0.9579 | 0.9681 | 0.9630 | |
| 交通运输 | 0.9464 | 0.9138 | 0.9298 | | 0.8852 | 0.9310 | 0.9076 | |
| 教育文体 | 0.9241 | 0.9733 | 0.9481 | 91.21% | 0.9542 | 0.9733 | 0.9637 | 94.14% |
| 劳动和社会保障 | 0.9465 | 0.9219 | 0.9340 | | 0.9837 | 0.9427 | 0.9628 | |
| 商贸旅游 | 0.7917 | 0.8879 | 0.8370 | | 0.8981 | 0.9065 | 0.9023 | |
| 卫生计生 | 0.8868 | 0.9126 | 0.8995 | | 0.9608 | 0.9515 | 0.9561 | |

# 3 问题二：热点问题挖掘

## 3.1 问题分析

本文需要定义合理的热度评价指标，并利用热度评价指标判断问题是否为热点问题。故需要先定义影响问题热度的因素有哪一些。本文认为如果同一类问题留言数量越多那么越能证明该问题为热点问题，另外留言问题下的互动人数、问题反映的时间范围也能影响到问题的热度。

将数据读入后进行数据清洗，并对其进行结巴分词等处理以后采 TF-IDF 算法获取文本特征，再采用 DBSCAN 密度聚类算法对文本进行聚类，获得留言数量，并求得互动人数及时间间隔，最后根据自定义热度问题综合评价模型获得留言前五热度问题。

## 3.2 热度问题综合评价模型的建立

本文建立了的**热度问题综合评价模型**以此来判断一个问题是否为热点问题，热点问题是指在一段时间内群众集中反映的问题，故本文将热度评价指标定义为在较短时间内留言的数量越多、互动点赞人数越多则认为此类问题热度较高，因此热度评价指标包含以下三个指标：

（1）留言数量：每类问题得到的留言详情数量

（2）互动情况：每类留言详情下的点赞数和反对数的情况

（3）时效性：每类留言详情留言的时间间隔差

通过这三个指标对热度产生的不同的影响赋予权重值计算得到**热度指数**，利用热度指数的高低来评价问题热度的高低。

综上所述，本文给出的**热度问题综合评价模型**为：

$$P = w_1 * x_1 + w_2 * x_2 + w_3 * x_3 \tag{3-1}$$

其中 P 表示热度指数，$x_1$ 表示留言数量，$x_2$ 表示互动指数，$x_3$ 表示时效性指数，$w_1, w_2, w_3$ 分别表示各个指标所占权重。即公式 3-1 还可表示为 3-2：

$$\text{热度指数} = w_1 * \text{留言数量} + w_2 * \text{互动指数} + w_3 * \text{时效性指数} \tag{3-2}$$

## 3.3 模型的求解

模型求解的流程如图 3-1 所示。

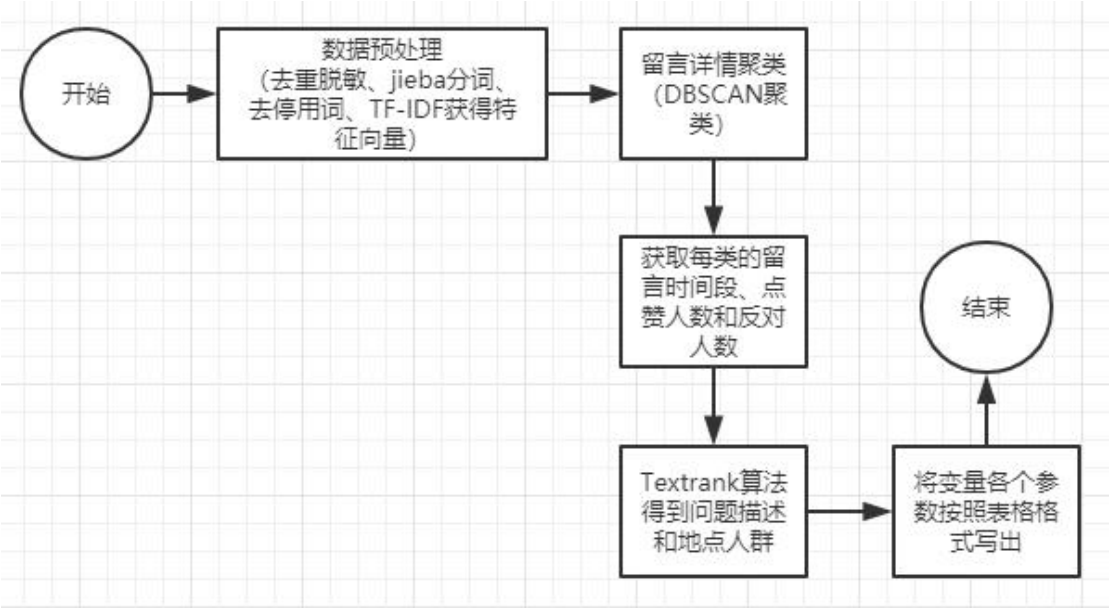


**图 3-1 模型求解流程**

### 3.3.1 数据预处理

为了实现留言详情问题的聚类，需要先预处理留言详情内容文本，再利用聚类算法对文本进行聚类。本文预处理包括：数据清洗、分词和特征提取。留言详情内容预处理框架图如图 3-2 所示：

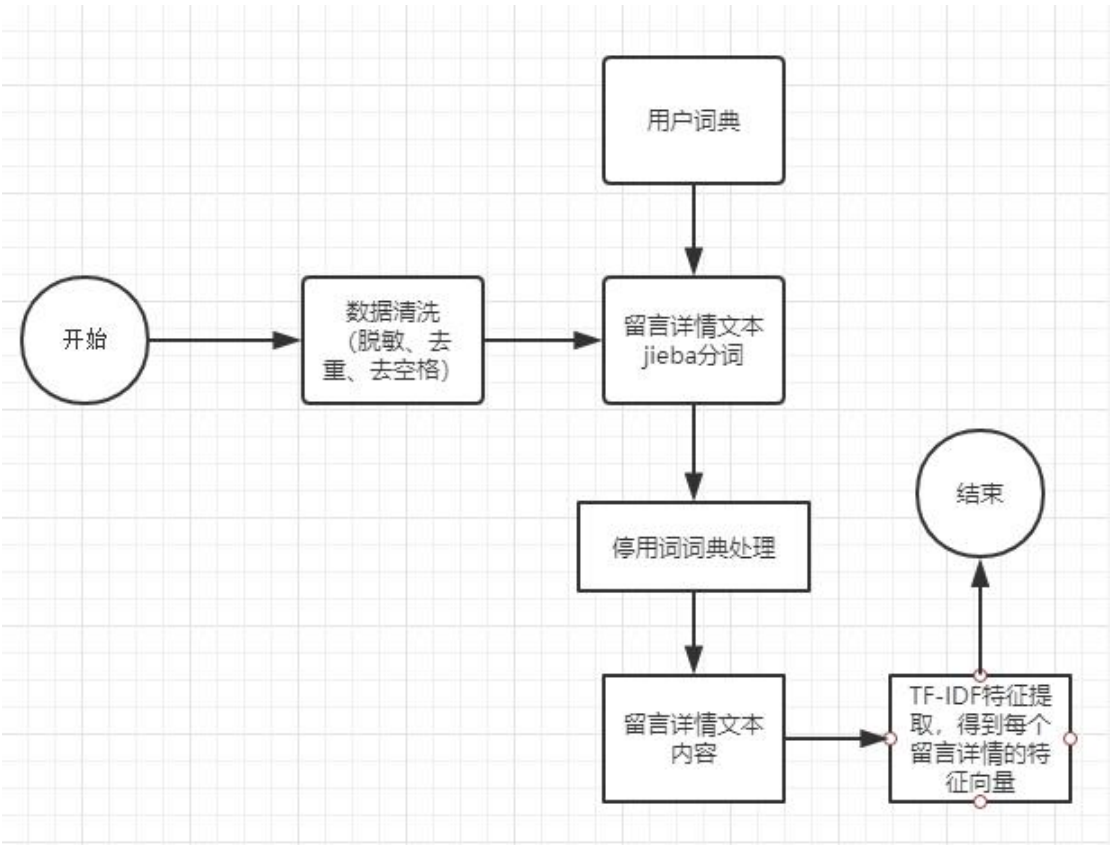


**图 3-2 留言详情内容预处理**

- **分词**

由于中文文本中词和词之间没有明显的界限，故从文本中提取词语时需要进行分词操作。本文采用 Python 开发的一个中文分词模块，即 jieba 分词，对数据中的留言详情文本进行中文分词。

- **去停用词**

在文本处理时，停用词是指功能比较普遍但相对没有实际含义的词语，除去停用词并不会对文本原本含义有改变。停用词通常是一些标点符号、单字和针对文本中的高频词语，比如中文文本中的“的、我、吗、你”等等，英文文本中的“this、a、am、any、can”等等。本文的停用词表有 1898 个停用词，并针对文本添加了“尊敬、领导、区长、市长、省长”等 8 个停用词。

- TD-IDF

将预料输入 DBSCAN 聚类模型进行训练，首先要先把文本符号语言表示成计算机能够理解的向量矩阵形式。针对本文的留言详情文本，我们采用词频-逆向文件频率模型（TF-IDF）对关键词进行匹配。

TF-IDF 算法简介：

TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术[12]。TFIDF 实际上是:TF ×IDF，其中 TF 表示词频，IDF 表示逆向文件频率，这两个参数共同决定着单词的重要性。若该文章中出现该单词的频率越高（TF），并且在其他文章中出现的频率低（IDF），则认为该单词能很好代表该文章，能作为文章分类的重要指标。

词的频率计算公式如下：

$$\text{词频（TF）} = \frac{\text{词在文段中出现的次数}}{\text{文章中总词数}} \tag{3-3}$$

逆向文件频率计算公式如下：

$$\text{逆向文件频率（IDF）} = \log\left(\frac{\text{语料库中文段总数}}{\text{包含该词的文段数}}\right) \tag{3-4}$$

计算得出 TF-IDF：

$$TF - IDF = TF * IDF \tag{3-5}$$

在忽略文本词语的重要性和词语在文本中出现的位置无关时，根据以上公式我们可以获取文本中每个词语的 TF-IDF 值，每个留言文本中词语的 TF-IDF 值构成一个词频向量矩阵，再对每个词频向量矩阵进行聚类。

- **数据标准化**

要将各个参数代入 式 3-4 中计算，先要给各个参数进行标准化。常见的标准化方法有三种：规范化方法、正规化方法和归一化方法。标准化数据能去除原始数据单位的限制，将数据转化为无量纲的数值，方便用以不同量级或者是单位的指标之间进行比较和加权。

本文将采用离差标准化（规范化）方法，对原始数据进行线性转化，使结果落在[0-1]区间内。离差标准化的转换函数为：

$$x^* = \frac{x - \min}{\max - \min} \quad (3\text{-}6) \qquad \text{或} \qquad x_* = \frac{\max - x}{\max - \min} \quad (3\text{-}7)$$

其中 max 指样本中最大的数值，min 为样本最小数值。

由于时间跨度越大热度指数应该越小，所以时间间隔差我们采用公式 3-7 进行标准化，留言数量和互动人数和热度指数成正相关，故用公式 3-6 进行表转化。

### 3.3.2 DBSCAN 聚类

通过对留言详情文本的清洗、分词和特质提取，采用聚类以此将众多留言归类。留言详情数目繁多，内容也是五花八门，会存在一个问题两种描述的留言文本，所以在对文本进行特征提取以后，使用基于密度的 DBSCAN 聚类算法。下面将对 DBSCAN 聚类算法模型介绍。

DBSCAN 聚类算法特点是不依赖于距离，而是依赖于密度来发现任意形状的簇，同时还可以过滤离群点，避免把离群点分布在某一个簇中。

DBSCAN 聚类过程：

```
（1）标记所有对象为 u;
（2）do
（3）    随机选择一个 u 对象 p;
（4）    标记 p 为 v;
（5）    If  p 为ε-领域至有 MinPts 各对象;
（6）    创建一个新簇 C，并把怕添加到 C;
（7）        令 N 为 P 的ε-领域中的对象的集合;
（8）        For N 中每个点 p1;
（9）            If p1 是 u;
（10）               标记 p1 为 v;
（11）               If p1 的ε-领域至少 MinPts 各点，把这些点添加到 N
（12）           If p1 还不是任何簇的成员，把 p1 添加到 C;
（13）        End for
（14）        输出 C;
（15）    Else 标记 p 为噪声;
（16）Until 没有标记为 u 的对象;
```

根据 DBSCAN 聚类过程可知，该算法先从某个核心点出发，然后不断向密度可达的区域扩张，从而得到一个包含核心点和边界点的最大化区域，区域中任意两点密度相连，即该相似的文本都会归为同一簇。在聚类时一般需

要输入三个参数。

（1）$\varepsilon$：半径

（2）MinPts：给定点在$\varepsilon$领域内成为核心对象的最小领域点数

（3）D：包含 n 个对象的集合

本文在聚类时，采用了余弦相似度来计算，故文本设置参数$\varepsilon$时，设置成 0.5，超过 0.5 的半径会使不同类的留言问题聚在一起。设置参数 Minpts 时，设置成 1，让单独的留言问题也可以自成一簇。

### 3.3.3 热度评价指标的求解

在上述已定义热度问题评价定义为在较短时间内留言的数量越多、互动点赞人数越多则认为此类问题热度较高，因此热度问题评价模型的指标含有留言数量、互动情况、留言间隔时段，下面将对这三个指标进行求解。

**（1）留言数量**，通过聚类的方式得到，把相同的的问题归为一类计算数量。上文已详述由 DBSCAN 聚类，然后计算聚类好的每一簇里的数量即为留言数量。

**（2）互动指数**，以同一类留言详情为单位，计算该留言问题分别获得点赞人数和反对人数并做标准化处理。根据赛题背景，定义热点问题是为了更好的针对性处理，如果该留言详情反对的人数多，则说明留言详情虚假或是夸大，故本文认为互动指数公式如下表示：

$$\text{互动指数} = \text{点赞数} - \text{反对数} \tag{3-8}$$

（3）**时效性指数**，以同一类留言详情为单位，将留言时间的间隔标准化。使其时间间隔越大标准化后的时效性指数越小。

$$\text{时效性指数} = \frac{\text{最早留言时间} - \text{该留言时间}}{\text{该类时间间隔}} \tag{3-9}$$

综上所述，本文给出的**热度问题综合评价模型**指标热度指数为：

$$\text{热度指数} = w_1 * \text{留言数量} + w_2 * \text{互动指数} + w_3 * \text{时效性指数} \quad (3\text{-}10)$$

其中 $w_1, w_2, w_3$ 分别表示留言数量、互动指数、时效性指数的权重，本根据实际情况给予合理的数值，留言数量即客观事实各群众的真实心声，故可作为主要影响因素，其次互动情况考虑到有些客观行为会影响其变动，最后时效性则为时间间隔大部分数据较短时间内，即差异性不大，因此综合考虑，$w_1, w_2, w_3$ 分别为 $w_1 = 0.9, w_2 = 0.08, w_3 = 0.02$ 。

### 3.3.4 Textrank 算法

通过上述热度指标模型公式即可获得热度排名，其次根据赛题要求需要绘制“热点问题表”，我们还需要自行获取问题描述，本文采用 Textrank 算法获得每类问题中的问题描述。下面将对 Textrank 算法模型进行介绍。

TextRank 是一种基于图的用于文本的排序算法，基本思想来自于 Google 的 PageRank 算法[11]。该算法对于词语可得到词语的排序，对于句子可得到句子的排序，所以 Textrank 算法不仅可以进行关键词提取，还可以进行自动文摘。该算法用于自动文摘时的思想是：PageRank 中的每个点代表每个句子，如果两个点之间的相似度大于设定的的阈值，则认为这两个点所代表的句子时相似的句子，这两个点之间会有一条有权边，边的权值表示句子的相似度。该算法的特点：算法可以脱离语料库的背景，对单篇文档进行分析就可以提取该文档的关键词或关键句。

Textrank 算法获取问题描述框架如下图 3-3 所示：

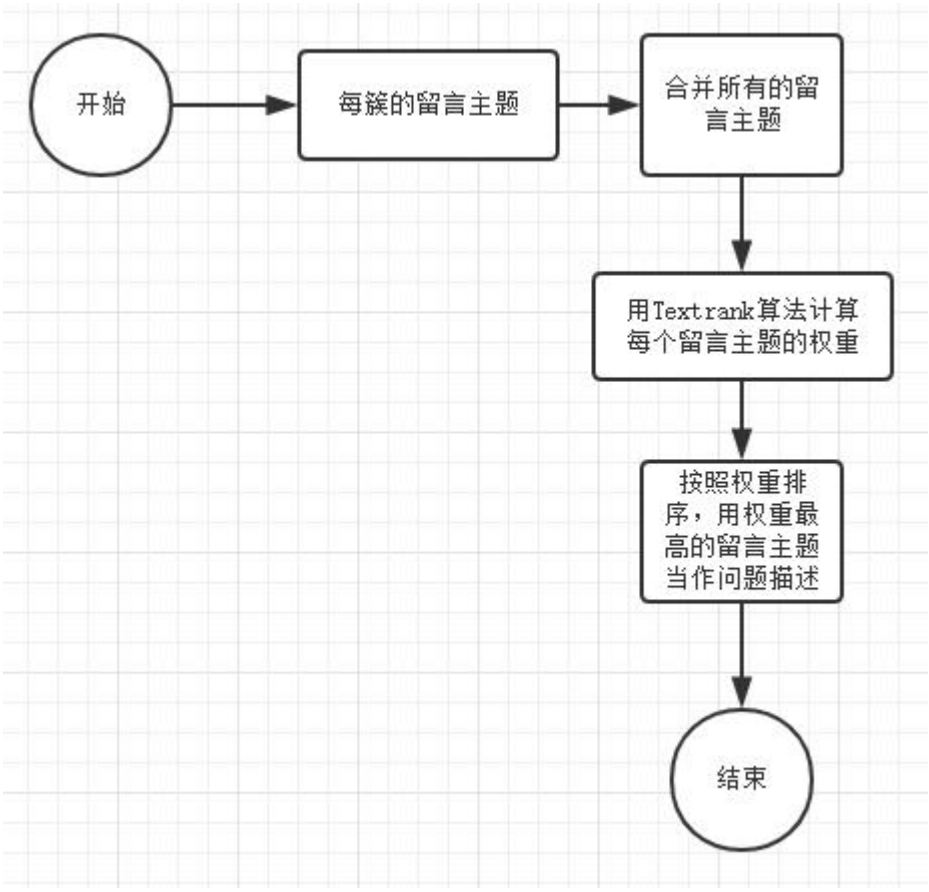


**图 3-3 获取问题描述框架**

Textrank 算法的过程步骤如下：

（1）预处理：分割文本中的句子，从而得到句子集合，然后对句子进行分词和去除出停用词等处理，得到关键词集；

（2）句子之间相似度的计算：

$$\text{句子间相似度} = \frac{\text{两个句子中都出现的词语数量}}{\log(\text{句子1中的总词数}) + \log(\text{句子2中的总词数})} \tag{3-11}$$

（3）句子权重的计算：

$$\text{句子1权重} = (1 - \text{阻尼系数}) + \text{阻尼系数} * \sum_{\text{和句子1向量的全部句子}} \frac{\text{句子1与句子2的相似度} * \text{句子2权重}}{\text{全部句子和句子2相连的句子边的权重和}} \tag{3-12}$$

由公式 3-5 多次迭代计算到收敛稳定后便可得到每个句子的权重得分。

（4）获得问题描述：根据得到的句子得分对句子进行排序，得分最高的排在最前，以此往下排序，抽取得分排序的前几个作为候选文摘句，本文将只抽取一个句子作为问题描述。

### 3.3.5 获取热度表相关列数值

根据题意可以知道热度表包括热度排名、问题 ID、热度指数、时间范围、问题描述，因此下面将陈述这些数值的求解方案。

**热度排名和问题 ID：**由求得的热度指数排序来作为热度排名和问题 ID

**时间范围：**反映着该类问题是否在短时间内被集中反映，根据聚类结果，以各类为单位，该类中问题的最早时间和最晚时间差即表示时间范围。

**问题描述：**通过 **Textrank 算法**获取，将各类文本为单位，根据 Textrank 算法计算该类文本中的留言主题，得到一个得分高的留言主题作为该类问题的描述。

## 3.4 模型结果

通过对文本留言详情进行 DBSCAN 聚类，如表 3-1 为部分聚类结果，其中 labels 为类别标签，若 labes 相同的文本则认为同一类别。

**表 3-1 DBSCAN 聚类部分结果**

| Index | 留言详情 | labels |
|---|---|---|
| 0 | A3区 一米阳… | 0 |
| 1 | 咨询 A6区 … | 1 |
| 2 | A7县 春华 … | 2 |
| 3 | A2区 黄兴路… | 3 |
| 4 | 市 A3区 中… | 4 |
| 5 | A3区 麓泉社… | 5 |
| 6 | A2区 富绿 … | 6 |
| 7 | 市 地铁 违… | 7 |
| 8 | 市 路 公交… | 8 |
| 9 | A3区 保利 … | 9 |
| 10 | A7县 特立 … | 10 |
| 11 | A3区 青青 … | 11 |
| 12 | 拆除 聚美龙… | 12 |
| 13 | 市利保 壹号… | 13 |

通过3.3模型求解获得的排名前五的热度问题表，具体结果如下表3-2所示。

**表3-2 热度问题表**

| 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 | 0.917226567799741 | 2019/07/07 至 2019/09/01 | A市伊景园滨河苑 | 投诉A市伊景园滨河苑捆绑销售车位 |
| 2 | 2 | 0.664824371750742 | 2019/11/02 至 2020/01/09 | A市A2区丽发新城小区 | 投诉丽发新城小区附近违建搅拌站噪音扰民 |
| 3 | 3 | 0.20264134210109 | 2019/01/06 至 2019/05/22 | A市辉煌国际城二期 | A1区辉煌国际城二期居民楼下商铺违法开饭店，维权近三月没有用 |
| 4 | 4 | 0.178368594545999 | 2019/01/06 至 2019/09/12 | A3区西湖街道茶场村五组 | A3区西湖街道茶场村五组何时启动拆迁 |
| 5 | 5 | 0.162541969293868 | 2019/01/07 至 2019/07/17 | A7县恒基凯旋门小区 | 反映A7县恒基凯旋门小区配套幼儿园公办或者普惠问题 |

热度问题表中每类问题相对应的热点问题明细表，部分展示如下表 3-3 所示。

**表 3-3 部分热点问题明细表**

| | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| 1 | 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 |
| 2 | 1 | 286304 | A909196 | 无视职工意愿、职工权益的 | 2019-08-23 10:23:23 | 广铁集团与A市政府及A市政工 | 0 | 0 |
| 3 | 1 | 188801 | A909180 | 投诉滨河苑针对广铁职工购 | 2019-08-01 00:00:00 | 尊敬的张市长，您好！我叫李 | 0 | 0 |
| 4 | 1 | 244243 | A909198 | 关于伊景园滨河苑捆绑销售 | 2019-08-24 18:23:12 | 广铁集团铁路职工定向商品房 | 0 | 0 |
| 5 | 1 | 199190 | A0009508 | 关于A市武广新城违法捆绑 | 2019-08-01 22:32:26 | 武广新城为铁广集团的定向商 | 0 | 0 |
| 6 | 1 | 268299 | A909193 | 惊！！A市伊景园滨河苑商 | 2019-08-21 15:32:33 | 伊景园滨河苑项目是广铁集团 | 0 | 0 |
| 7 | 1 | 255507 | A909195 | 违反自由买卖的A市伊景园 | 2019-08-20 12:34:21 | 广铁集团铁路职工定向商品房 | 0 | 0 |
| 8 | 1 | 207243 | A909175 | 伊景园滨河苑强行捆绑车位 | 2019-08-23 12:16:03 | 您好！A市武广新城片区的伊景 | 0 | 0 |
| 9 | 1 | 250514 | A0003200 | 广铁集团强制职工购房时捆 | 2019-08-20 10:39:17 | 广铁集团抢钱啦.还是武广新城 | 0 | 0 |
| 10 | 1 | 258037 | A909190 | 投诉伊景园滨河苑捆绑销售 | 2019-08-23 11:46:03 | 尊敬的领导:我是广铁集团铁路 | 0 | 0 |
| 11 | 1 | 195511 | A909237 | 车位捆绑违规销售 | 2019-08-16 14:20:26 | 对于伊景园滨河苑商品房，A市 | 0 | 0 |
| 12 | 1 | 280774 | A909199 | 反馈广铁集团铁路职工定向 | 2019-08-10 12:23:19 | 尊敬的领导，您好！我要反馈 | 0 | 0 |
| 13 | 1 | 276460 | A909170 | A市伊景园滨河苑捆绑销售 | 2019-08-24 17:23:11 | 尊敬的领导，您好，关于广铁 | 0 | 0 |
| 14 | 1 | 209571 | A909200 | 伊景园滨河苑项目绑定车位 | 2019-08-28 19:32:11 | 广铁集团铁路职工定向商品房 | 0 | 0 |
| 15 | 1 | 244528 | A909235 | 伊景园滨河苑开发商强买强 | 2019-08-21 19:05:34 | A市广铁集团伊景园滨河苑商品 | 0 | 2 |
| 16 | 1 | 209506 | A909179 | A市武广新城坑客户购房金 | 2019-08-02 16:36:23 | 您好！由A市广铁集团发起的定 | 0 | 0 |
| 17 | 1 | 195995 | A909199 | 关于广铁集团铁路职工定向 | 2019-08-10 18:15:16 | 尊敬的市政府领导，您好！我 | 0 | 0 |
| 18 | 1 | 224767 | A909176 | 伊景园滨河苑车位捆绑销售 | 2019-07-30 14:20:08 | 伊景园滨河苑车位捆绑销售！ | 0 | 0 |
| 19 | 1 | 271517 | A909238 | 开发商联合广铁集团捆绑车 | 2019-08-11 12:02:27 | 你好，本人购买伊景园滨河苑 | 0 | 0 |
| 20 | 1 | 190337 | A0009051 | 关于伊景园滨河苑捆绑销售 | 2019-08-23 12:22:00 | 投诉伊景园.滨河苑开发商捆绑 | 0 | 0 |
| 21 | 1 | 279941 | A909177 | 广铁集团职工商品房竟然捆 | 2019-08-28 09:30:20 | 领导好！A市广铁集团为职工提 | 0 | 0 |
| 22 | 1 | 218739 | A909184 | A市伊景园·滨河苑欺诈消 | 2019-08-24 00:00:00 | A市伊景园滨河苑强行捆绑车位 | 0 | 0 |
| 23 | 1 | 241373 | A0005378 | 强行要求捆绑车位，请有关 | 2019-08-14 09:15:22 | 这个楼盘名称是:伊景园.滨河 | 0 | 0 |
| 24 | 1 | 218709 | A0001066 | A市伊景园滨河苑捆绑销售 | 2019-08-01 22:42:21 | 伊景园滨河苑作为广铁集团定 | 0 | 1 |
| 25 | 1 | 196264 | A0009508 | 投诉A市伊景园滨河苑捆绑 | 2019-08-07 19:52:14 | A市伊景园·滨河苑现强制要求 | 0 | 0 |

# 4 问题三：答复意见的评价

## 4.1 答复意见质量综合评价模型建立

针对问政平台群众得留言相关部分做出答复意见，为了筛选出质量较好的答复意见，将以语言特征、文本信息特征、时间特征、专业术语特征等特征指标对答复意见与相关留言的质量进行评价。本文针对答复意见文本将给出以下几个指标：

**相关性**:答复意见和留言问题是否相关，是否在解决留言详情问题。

**完整性**:答复意见的文本内容是否足够完整。

**时效性**:答复的时间是否及时。

**文明性和专业性**:答复意见文本中语言是否专业和文明。

本文首先通过数据清洗处理，然后分析答复意见的各个特征制定答复意见质量评价的指标，最后根据各个指标的权重和求得每条答复意见的质量评价指数，以此来评价相关部门的答复意见的质量。

综上所述，最终获得**答复意见质量综合评价模型**为：

$$P = w_1 * x_1 + w_2 * x_2 + w_3 * x_3 + w_4 * x_4 \quad (4\text{-}1)$$

其中，$P$ 代表答复意见质量评价，$x_1$ 代表相关性指数，$x_2$ 代表完整性指数，$x_3$ 代表时效性指数，$x_4$ 代表专业性指数，$w_1, w_2, w_3, w_4$ 分别代表各指标所占权重，各指标具体权重值由 4.3.2-3 计算各指标的权重中求解获得。

同时公式 4-1 也可表示为公式 4-2：

$$\text{热度指数} = w_1 * \text{相关性指数} + w_2 * \text{完整性指数} + w_3 * \text{时效性指数} + w_4 * \text{专业性指数} \quad (4\text{-}2)$$

## 4.2 模型框架

根据相关部门的答复意见，结合以往前人的研究，如何挖掘答复意见中有效信息，并利用这有效信息对答复意见进行评价，正是本文要实现的主要目的，具体流程框架如下图 4-1 所示。

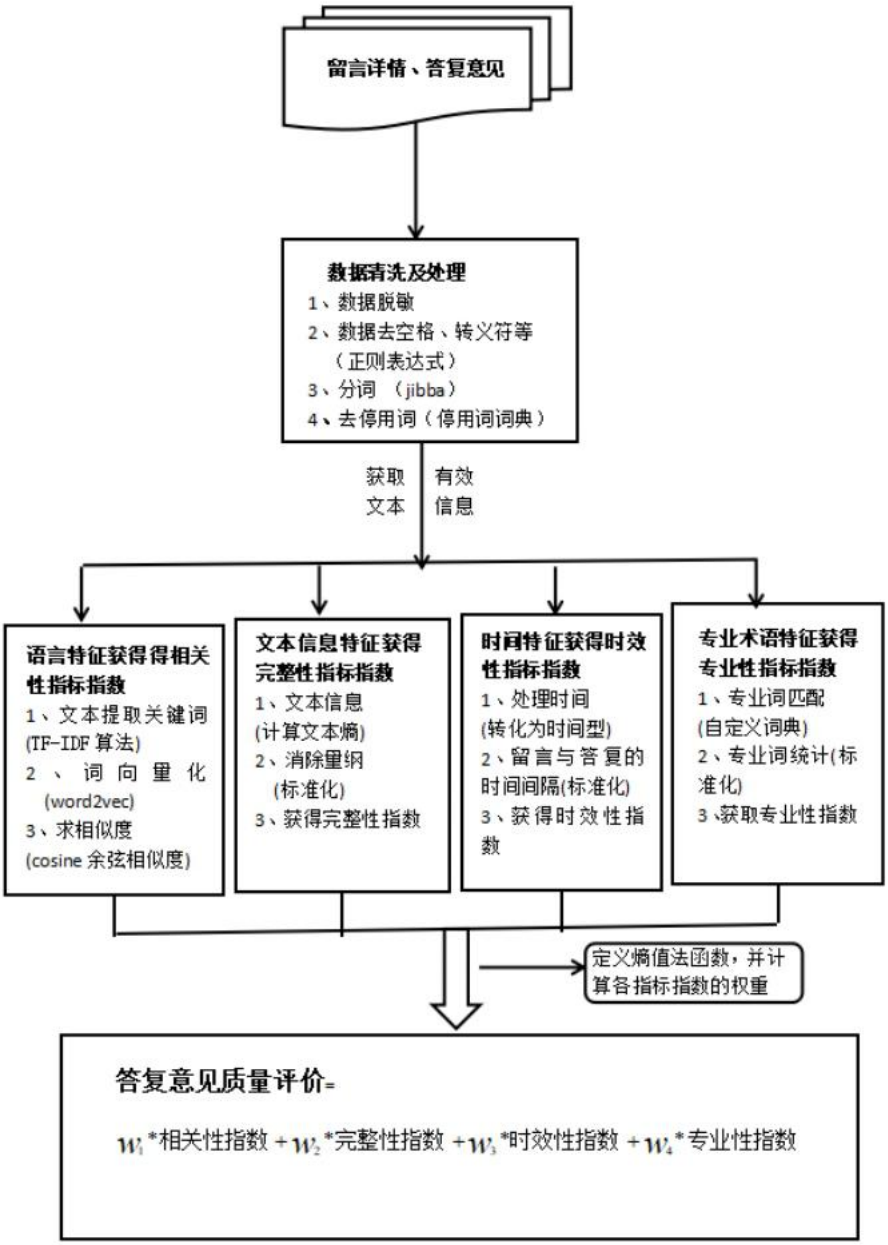


图 4-1 答复意见质量评价模型流程

## 4.3 模型求解

### 4.3.1 数据清洗与处理

- **数据脱敏**

由于数据来自真实社区群众的留言信息，难免会出现比较敏感的数字或词汇比如身份证等，所以问政平台会提前将个人身份信息进行加密，此数据中敏感信

息已经用*进行了加密，所以这些词对后续工作并没用什么用处，因此文章将这些词汇去除，即脱敏处理。

- **文本去空格**

由于留言及答复内容都以书信形式存在，因此存在大量空格或者制表符、转行符等无用的转义符，所以为了不影响下文特征的提取，先对数据利用正则化将空格符去除。

- **分词并去停用词**

此文章运用 jibba 分词对文本进行精准分词，其次 jibba 词库中可能有些网络用词或新词不存在，因此根据文本内容自定义了词典并加载到 jibba 词库中，为后续工作做准备。对文本分词后再根据停用词字典删除一些不必要的词汇，比如中文中的“我、的、了、地、吗”等，避免对文本造成负面影响。

## 4.3.2 答复意见质量综合评价模型求解

### 4.3.2-1 指标体系的制定及分析

为实现对答复意见质量的评价，本文将从答复意见的文本中获取相关的有效的信息，比如语言特征、内容评估特征、时间特征、专业术语特征等方面，并根据这些特征制定指标来作为评价答复意见质量的标准。如下图为答复意见评价质量指标体系图 4-2 所示。

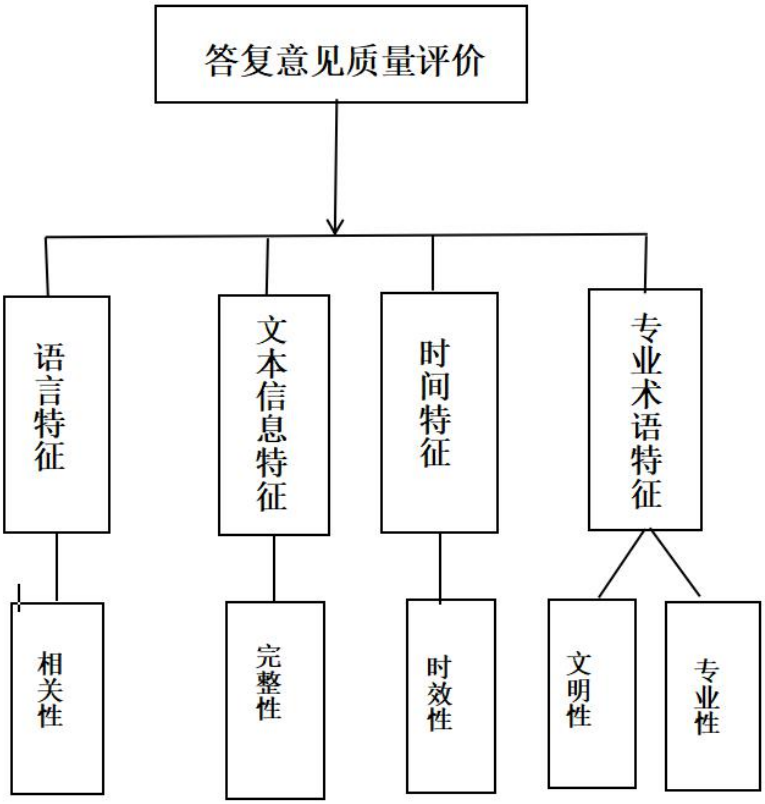


图 4-2 复意见评价质量指标体系

**1、语言特征**，本文定义语言特征为将含有语义的文本进行特征向量化，然后利用特征向量求得留言与答复意见的相似度。

**2、文本信息特征**，检索留言的问题答复意见中是否答复完整、篇幅是否过于简略，以此作为答复意见的准确性及完整性。

**3、时间特征**，通常认为留言的时间与答复的时间相差不大时，答复的意见往往比较有实用性及有一定的时效性。

**4、专业术语特征**，答复时引用相关章法或规则，而且用语较为礼貌性的，则认为词答复的文明性及专业性较高。

如上提及的六个指标即作为本文节评价答复意见质量的标准，那么六个指标分别由那些测量标准来衡量，详情如下表 4-1 所示。

表 4-1 各指标衡量标准

| 指标 | 衡量标准 | 具体详情 |
| --- | --- | --- |
| 相关性 | 1、符合问题的主题<br>2、正确分类<br>3、不含无关信息 | 答复意见是否符合留言详情所提问题的主题 |
| 准确性、完整性 | 1、完整并准确回答<br>2、提供事实、解释、论证 | 答复意见是否完整包含留言详情中所有问题 |
| 时效性 | 1、回答间隔时间合适<br>2、合理回答 | 答复意见的时间与留言详情的时间间隔 |
| 文明性 | 1、不含不良信息<br>2、态度谦和礼貌<br>3、不含脏话 | 答复意见中的语气礼貌程度 |
| 专业性 | 1、专业术语词汇<br>2、多处引用合理章法 | 答复意见是否由较为专业人士回答，引用的章法规则是否有利有据 |

### 4.3.2-2 相关特征提取及指标计算

### 1、语言特征提取并做相关性计算

首先对文本进行上述预处理操作，然后基于 TF-IDF 算法对文本进行关键词提取，再利用 Word2vec 算法实现词语的向量化，最后利用余弦相似度对文本进行相关性计算。具体流程如下图 4-3 所示。

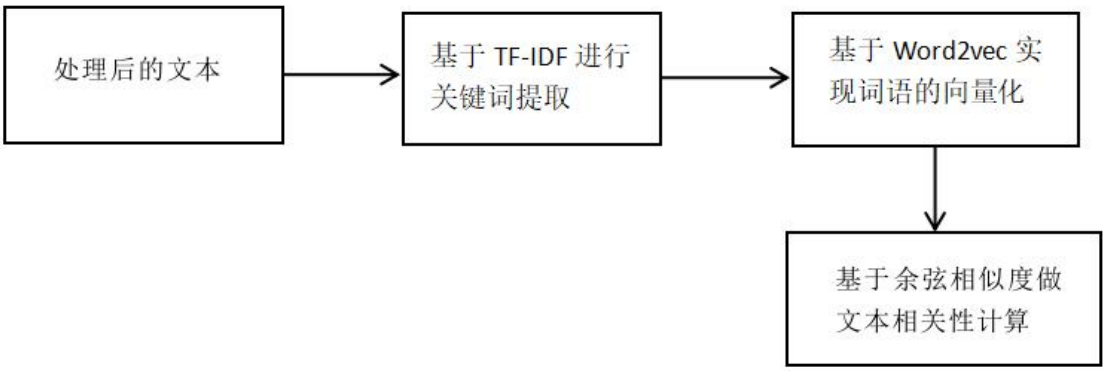


**图 4-3 计算相关性指数**

**①关键词的提取**

关键词的提取就是从文本里面把跟内容意义最相关的一些词语抽取出来。关键词的提取到目前为止，从算法角度看可以分类两类：

有监督学习算法，则是将关键词提取任务看作二分类问题：根据候选关键词的数据和提前设置好的特征，训练模型并使用模型预测新的候选词是否为关键词。

无监督学习算法，先抽取候选词，然后对各个候选词进行打分，然后输出 topK 个分值最高的候选词作为关键词。根据打分的策略不同，有不同的算法，例如 TF-IDF、TextRank、主题模型等算法；

本文节综合多方面考虑，最终选用 TF-IDF 作为本次抽取关键词的算法。TF-IDF 是 Term Frequency-Inverse Docunment Frequency 的缩写，即“词频-逆文本频率”，它由两部分组成。主要思想就是：如果某个词在一篇文档中出现的频率高，也即 TF 高；并且在语料库中其他文档中很少出现，即 DF 低，也即 IDF 高，则认为这个词具有很好的类别区分能力。

词频计算公式为：

$$tf_{ij} = \frac{n_{i,j}}{\sum_{k} n_{k,j}} \tag{4-3}$$

其中 $n_{i,j}$ 表示词 $i$ 在文档 $j$ 中出现的频次，$\sum_{k} n_{k,j}$ 表示文档 $j$ 的总词数

用直白的语言表示则为：

$$tf(\text{当前词}) = \frac{\text{当前词在文档中出现的次数}}{\text{文档总词数}} \tag{4-4}$$

逆文本频率计算公式：

$$idf_{i,j} = \log \frac{|D|}{|\{j : t_i \in d_j\}| + 1} \tag{4-5}$$

其中$|D|$表示文档中总文档数，$|\{j : t_i \in d_j\}|$表示文档集中文档$d_j$出现词$t_i$的文档个数，分母加一是为了避免文档集中没有出现词$t_i$，导致分母为零的情况。

用文字表示则为：

$$idf(\text{当前词}) = \log \frac{\text{总文档数}}{\text{包含当前词的文档个数}+1} \tag{4-6}$$

综合 TF 与 IDF 使用，权衡词频、逆文档频率两个方面俩恒力量词的重要程度，故得出 TF-IDF 算法的计算公式：

$$tf_{ij} \times idf_{ij} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|D|}{|\{j : t_i \in d_j\}| + 1} \tag{4-7}$$

通常，关键词的数量按照 TF-IDF 值降序排序，选出前几个值较大的作为关键词，也可以根据实际情况来确定数量。

**②文本向量化**

将文本转化为计算机认识的向量形式，此过程即叫文本向量化。文本向量化后即可度量两个文本的相似度。当然，文本向量化的方法也是多种的：

以词为单位的文本向量化：词集模型、词袋模型、n-gram、Word2Vec 算法；以句子作为单位的文本向量化：主题模型 LSA、NMF、PLSA 以及 LDA 等方法。

由于留言详情及答复意见的内容是存在一定的语境的，那么本文决定使用词

语语义可通过上下文信息来确定的分布式的词向量化--Word2Vec 算法。

Word2Vec 算法包含 CBOW（Continuous Bag-Of-WordsModel）和 Skip-gram 两种目标模型，Word2Vec 可以在百万数量级的词典和上亿的数据集上进行高效地训练；其次，该工具得到的训练结果即词向量（word embedding）可以很好地度量词与词之间的相似性。

Word2Vec 算法模型其实是简单化的神经网络，模型简图如图 4-4 所示。

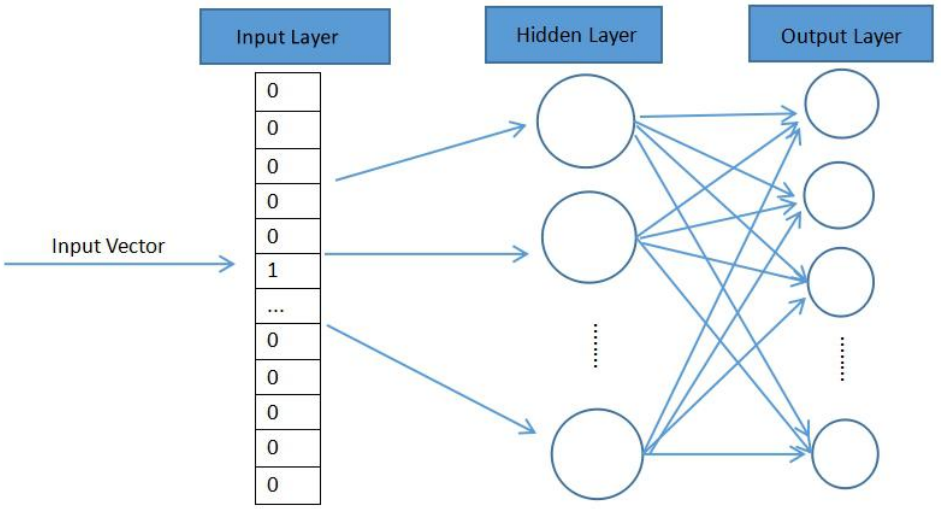


**图 4-4　Word2Vec 算法模型结构**

由上图 4-4 可以看到，很明显只是一个简单的神经网络结果，其中输入是 One-Hot Vector，Hidden Layer 是没有激活函数，也就是线性的单元。Output Layer 维度跟 Input Layer 的维度一样，用的是 Softmax 回归。

输入和输出的数据是由 Word2Vec 的模型决定的，CBOW 模型的训练输入是某一个特征词的上下文相关的词对应的词向量，而输出就是这特定的一个词的词向量。　Skip-Gram 模型和 CBOW 的思路是反着来的，即输入是特定的一个词的词向量，而输出是特定词对应的上下文词向量。CBOW 对小型数据库比较合适，而 Skip-Gram 在大型语料中表现更好。

因此根据留言或着答复的内容篇幅都属于小短文，本文将决定使用 Word2Vec 算法中 CBOW 模型进行词向量的训练。简易图如图 4-5 所示。

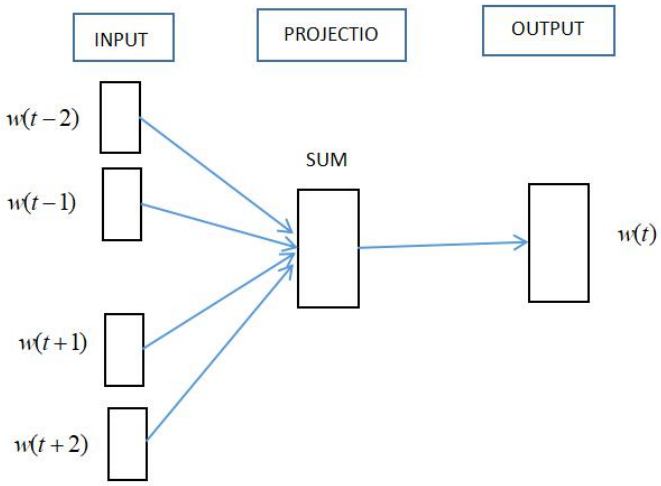


**图 4-5　CBOW 模型词向量训练**

模型假设原理，如图 4-6 所示：

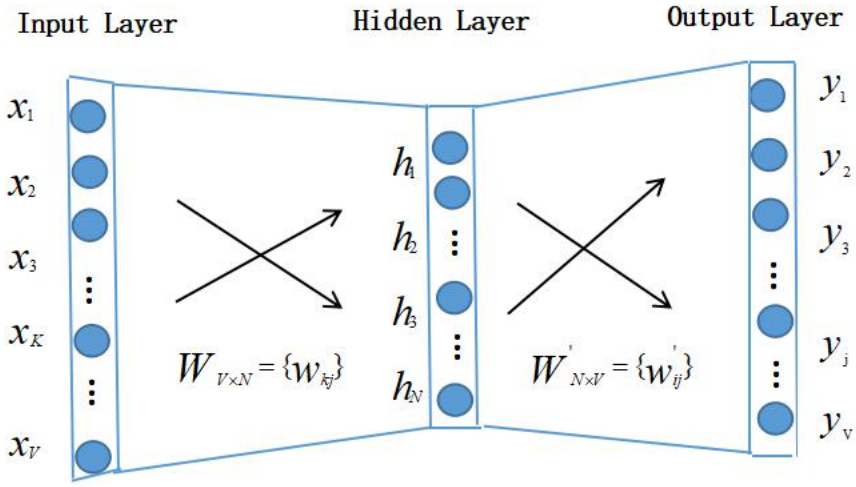


**图 4-6 模型假设**

- 文本词汇量的大小为 V，隐藏层的大小为 N ，相邻层的神经元是全连接的。
- 输入层是一个用 one-hot 方式编码的词向量其中只有一个为 1，其余均为 0。
- 输入层到隐藏层连接权重值用一个维的矩阵来表示，即

$$W=\begin{bmatrix} w_{11},w_{12},\cdots,w_{1n} \\ w_{21},w_{22},\cdots,w_{2n} \\ \cdots \\ w_{v1},w_{v2},\cdots,w_{vn} \end{bmatrix} \tag{4-8}$$

- 隐藏层到输出层连接权重用一个新的矩阵来表示，即

$$W'=\begin{bmatrix} w'_{11},w'_{12},\cdots,w'_{1n} \\ w'_{21},w'_{22},\cdots,w'_{2n} \\ \cdots \\ w'_{v1},w'_{v2},\cdots,w'_{vn} \end{bmatrix} \tag{4-9}$$

- 输出层使用 softmax 函数来计算词的后验分布

首先，词向量的训练包括两个步骤：一是对中文语料进行预处理；二是利用 gensim 库训练词向量。

- 预处理：格式转换，字体变换，jieba 分词。

- 训练词向量：Word2Vec 工具训练词向量，其中里面参数详情，如下表 4-1 所示。

**表 4-1 参数详情**

| 参数 | 参数含义 | 参数设置 |
|---|---|---|
| sentences | 要分析的语料 | 预处理好的语料 |
| size | 词向量维度 | size=192 |
| window | 词向量上下文最大距离，默认值 5 | window=5 |
| sg | Word2Vec 的模型选择，0 表示 CBBOW，1 表示 Skip-Gram | sg=0 |
| min-count | 词向量的最小词频，默认值 5 | min-count=5 |

| | | |
|---|---|---|
| iter | 随机梯度下降法中迭代最大次数，默认值 5 | Iter=5 |
| workers | 词向量使用的线程数 | Workers=9 |

最后，将训练好的词向量模型保存到相应文件，为后续做文本相似度做准备。

**③计算文本相似度并作为相关性指数**

有基于词向量（余弦相似度、曼哈顿距离、欧几里得距离、明式距离）；基于字符（编辑距离、simhash）；基于概率统计的（杰卡德相似系数）。那本文就使用基于词向量中的余弦相似度算法作为计算文本相似度。

余弦相似度，即为余弦距离，是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小度量。余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似。

以二维空间为例，如图上 a，b 是两个向量，要求它们的夹角 θ

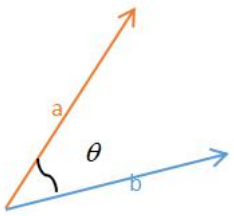


**图 4-7**

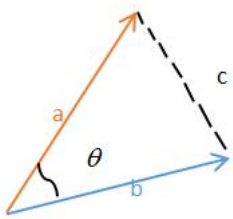


**图 4-8**

由余弦定理公式有：

$$\cos(\theta)=\frac{a^2+b^2+c^2}{2ab} \qquad (4\text{-}10)$$

假定 a 向量是[x1,x2]，b 向量是[y1,y2]，那么余弦定理公式则为：

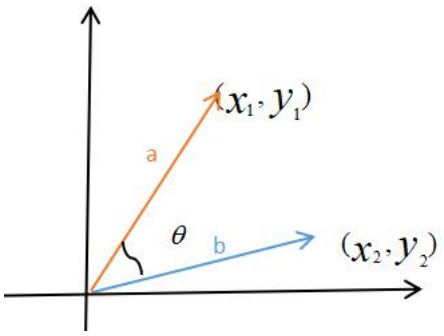


**图 4-9 余弦定理公式**

$$\cos(\theta) = \frac{x_1 x_2 + y_1 y_2}{\sqrt{x_1^2 + y_1^2} \times \sqrt{x_2^2 + y_2^2}} \qquad (4\text{-}11)$$

假定 a，b 不是二维的而是 n 维的，那么夹角的余弦公式为：

$$\cos(\theta) = \frac{\sum_i^n (A_i \times B_i)}{\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2}} = \frac{A \bullet B}{|A| \times |B|} \qquad (4\text{-}12)$$

因此，留言详情及答复意见的相关性则可由两个多维的词向量求夹角的余弦值来衡量，以此求得相关性指数。

## 2、文本信息特征提取并做完整性计算

通过预处理后文本通过计算文本熵提取特征信息，然后通过数据标准化消除数据量纲和数量级，最后获得文本的完整性指数。具体流程如下图 4-10 所示。

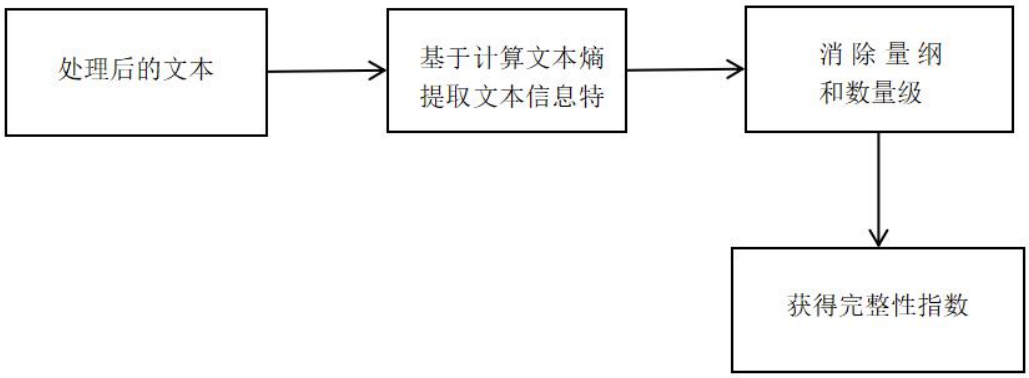

**图 4-10 完整性指数流程**

**①计算答复意见的文本熵**

描述内容所含信息量，内容越少那么其对应的文本熵就较小，所包含的内容就越少；相反，熵越大，则表示所含的信息量越大。本文要探究答复意见的完整性，如若信息量都低到一定程度，那也很难说明其的回答是完整，因此所求得的文本熵越大，则较大可能地认为，回答较为认真、不敷衍、甚至可认为较为完整。

文本熵先求得内容字符长度，然后根据信息熵计算公式获得。如公式 4-13：

$$H(X) = -\sum_{i=1}^{n} P_i \bullet \log_2 P_i \qquad (4-13)$$

**②消除不同量纲或数量级**

在上文已求得所有答复意见的文本熵，由于每个数据的文本熵差距较大，且每个数据的性质、量纲或者数量级都大不同，数据存在大差距，会削弱数值水平较低的数据，从而影响整体的结果，因此为了保证结果的可靠性，需要对求得的数据进行标准化处理。标准化方法有多种的，比如 min-max 标准化、log 函数转换、z-score 标准化、模糊量化法等。为后续各个不同单位或量级指标的指数能够进行加权，本文将使用 min-max 标准化对数据进行标准化映射到[0,1]，所以后文提及的标准化都为 min-max 标准化的方法。

min-max 标准化（Min--maxnormalization）也叫做离差标准化，就是对原始数据进行线性变换，使原始数据都映射到[0,1]区间，转换函数：

原始数据 $x_1, x_2, \cdots, x_n$，经过下公式 4-14：

$$y_i = \frac{x_i - \min\limits_{1 \le i \le n}\{x_j\}}{\max\limits_{1 \le j \le n}\{x_j\} - \min\limits_{1 \le j \le n}\{x_j\}} \qquad (4-14)$$

转换为 $y_1, y_2, \cdots y_n \in [0,1]$ 且无量纲的。

**③获得完整性指数**

将答复意见的文本熵都标准化后，其数据集的数据值大小都在[0,1]区间，因

此标准化后的数据则作为完整性指标的指数。

### 3、时间特征提取并做时效性计算

经过预处后的文本，提取出留言的时间和留言答复的时间，通过标准化数据消除量纲和单位的影响，最后获得时效性指数。具体流程如下图 4-11 所示。

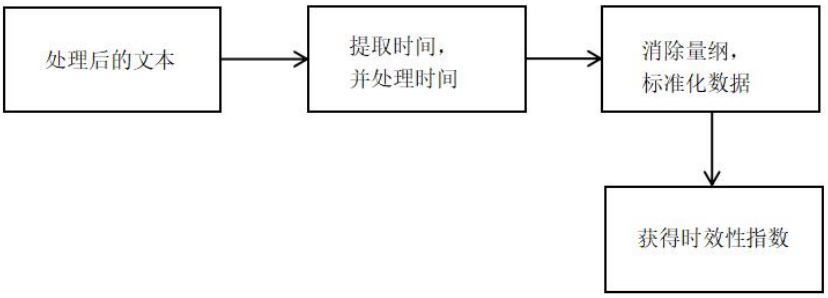


图 4-11 时效性计算

**①处理时间**

从文件中获取留言时间以及答复时间，但由于时间存在的形式是由留言者或答复者自行留下的，因此可能会有不同的类型，为了能同一处理，我们将时间数据同一转换为时间类型。比如非机器识别的时间类型 2019-04-08 08:37:20，转化为时间型 2019/04/08 08:37:20。

**②求时间间隔并消除量纲**

时间间隔即为留言时间与答复时间的间隔。首先时间间隔以天为单位，然后获得时间间隔根据上节描述的 min-max 标准化的方法进行量纲消除，将所有的时间间隔映射到[0,1]。

**③获得时效性指数。**

通常认为在某一时间的留言在较短时间内得到的答复，这条答复意见的实用性或者时效性较高，答复意见的建议较有利于提问者及时处理问题并得到解决。但目标的实现又与标准化后的数据恰恰相反，标准化的数据是原始数据越小对应标准化的数据仍然较小，那时效性想要达到时间间隔越小其标准化数据就越大，

那么本文将会在标准化后的数据再用 1 去减它，最总获得数据将是我们所想要的数据，因此上述时间间隔标准化并做一定处理后的数据即可作为答复意见质量评价时效性指标指数。

## 4、专业术语特征提取并做专业性、文明性计算

经过预处理后的文本，通过自定义专业词词典与之匹配，并统计后标准化化数据，消除量纲的影响，最后便可获得专业性指数。具体流程由下图 4-12 所示。

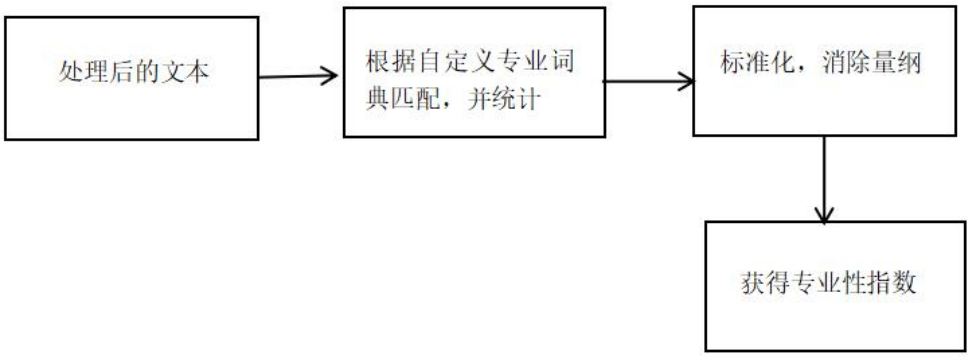


**图 4-12 专业性、文明性计算**

**①自定义专业词典并做匹配统计**

专业词典由人工自行筛选出来，一些谦和词语、一些敬辞、一些礼貌用语（比如尊敬的、您好、谢谢等），还有一些引用章法、规则（比如根据《机动车登记规定》等），一些具有条理性词汇（比如第一、第二、首先、其次等）。

自定义词典完成后，便进行词的匹配，查找答复意见中出现专业词典中的词的频数，聘书越高则认为专业水平较高、文明性较好。

**②标准化并获得专业性指数**

由上述的答复意见每条的相应专业词汇的频数进行标准化，标准化仍是使用前面提及的 min-max 标准化的方法。那么标准化后的数据值越大则认为专业性、文明性较高。因此标准化后映射到[0,1]区间的数据则作为答复意见质量评价的专业性指标的指数。

### 4.3.2-3 计算各指标的权重

本文研究的答复意见质量评价指标有文本相关性、答复完整性、时效性以及专业性等四个指标，由于不同指标的指数在数据上的表现程度不同，那么进行计算最终的答复意见评价

质量时所占的权重也会有所不同，因此本文将根据熵值法来对各个指标进行分配权重。

熵值法是用来计算指标的权重的一个经典算法，它是通过判断某个指标的离散程度，来判断此指标对综合指标的影响程度，若指标的熵值越大则离散程度则越大，那么说明该指标对综合评价的影响就越大。

熵值法的实现：

- 假设数据为 n*m 的矩阵 $A=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$ （4-15）

- 计算第 $i$ 项指标下第 $j$ 个记录所占的比重

$$P_{ji}=\frac{x_{ji}}{\sum_1^n x_{ji}},\quad (i=1,2,\cdots,m) \qquad (4\text{-}16)$$

- 求第 $i$ 项指标的熵值

$$e_i=-k*\sum_1^n p_{ji}*\log(p_{ji}),\quad k=\frac{1}{\ln(n)} \qquad (4\text{-}17)$$

- 计算第 $i$ 项指标的差异系数

$$d_i=1-e_j \qquad (4\text{-}18)$$

- 计算指标的权重

$$w_i = \frac{d_i}{\sum_1^m d_i} \quad (4\text{-}19)$$

根据上述的步骤，编辑熵值法函数，求得各指标权重值：

相关性　　0.533153
完整性　　0.410964
时效性　　0.009139
专业性　　0.046744

最终保留小数点后两位则为 0.53，完整性指数为 0.41，时效性指数为 0.01，专业性指数为 0.05。即可得出如下公式：

**答复意见质量综合评价=0.53*相关性指数+0.41*完整性指数+0.01*时效性+0.05*专业性指数** （4-18）

可以看出相关性指标占权重较大，其次为完整性指标，时效性即专业性占权较小，说明它们在综合评价影响不大，或者说是所有数据在着两指标的差异性不大。

## 4.4 模型结果

根据上述的解题思路以及解题流程，这节将根据给出的解题方法来现实题目要求：给出合理的答复意见质量综合评价模型，并应用。

下表将罗列一些留言详情以及答复意见，并将获得的各个指标指数一并呈现。如表 4-2 所示。

**表 4-2 答复意见各指标指数**

| 序号 | 留言主题 | 答复意见 | 相关性指数 | 完整性指数 | 时效性指数 | 专业性指数 | 质量评价指数 |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 80 | 质疑A市新城国际花都的装修价格以及[政府发文]【2018】53号文件 | 网友...您好！.....回复如下：一、根据《营业税改征增....》.....二、《商品房全装修销....》.....三、《商品房建筑面积预测报告书》..... | 0.8991 | 0.8869 | 0.9880 | 0.9188 | 0.8958 |
| 2420 | 请求L市合理设置城区的28路公交车终点站 | 尊敬的...：您好！.....现将相关情况回复如下：.....一、.....二、..... | 0.7187 | 0.6975 | 0.9677 | 0.8622 | 0.7189 |
| 1642 | 希望能给D12市檀山村电力线路扩容 | 网友：你好！你反映的问题，我们已转交电力部门。2019年3月28日 | 0.0763 | 0.2678 | 0.9977 | 0.6007 | 0.1879 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1425 | 希望 B9 市重视农村老人教育 | 网友：您好！2019 年 3 月 5 日 | 0.0040 | 0.0001 | 0.9090 | 0.0021 | 0.0113 |

如上可以清楚的看到，答复意见质量评价指数较高的，答复意见都有合理并认真地在回答群众留言地问题，并由举例各种规章制度增强说服力，那么质量评价质量较低地，我们也可以看到相应的回答都是较敷衍的，甚至没有任何内容，只答复了一个时间。

# 5 参考文献


[1]段丹丹，唐加山，温勇，袁克海.基于 BERT 的中文短文本分类算法的研究

[2]https://zhuanlan.zhihu.com/p/46887114?from_voters_page=true

[3]简书 https://www.jianshu.com/p/b8eab7a8ee2d

[4]CSDNhttps://blog.csdn.net/weixin_43199584/article/details/96477250

[5]知乎 https://zhuanlan.zhihu.com/p/46997268

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick.Mask R-CNN [EB/OL]. https://arxiv.org/abs/1703.06870,2017-03-20.

[7] Robert Tibshirani.Regression Shrinkage and Selection Via the Lasso[EB/OL]. https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x,1996

[8] van Griethuysen, J. J. M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R. G. H., Fillon-Robin, J. C., Pieper, S., Aerts, H. J. W. L. (2017). Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Research, 77(21), e104 - e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

[9] Waleed Abulla.Mask R-CNN for object detection and instance segmentation on Keras and Tensorflow[EB/OL]. https://github.com/matterport/Mask_RCNN,2017.

[10]360 百科 https://baike.so.com/doc/433640-459181.html

[11] 蒲梅,周枫,周晶晶,严馨,周兰江.基于加权 TextRank 的新闻关键事件主题句提 取[J].计算机工程,2017,43(08):219-224.

[12]简书 https://www.jianshu.com/p/471d9bfbd72f

[13]CSDNhttps://blog.csdn.net/weixin_41657760/article/details/9241092

5http://www.ruanyifeng.com/blog/2013/03/cosine_similarity.html

[14]CSDNhttps://www.cnblogs.com/huzixia/p/10403481.html

[15]知乎 https://zhuanlan.zhihu.com/p/46997268

# 附录 ：

**第一题：**

1、环境

python 3.7

pytorch 1.1

tqdm

sklearn

tensorboardX

2、数据说明

在相应路径的 Prompt 输入：python data partitioning.py

实现对附件 2.xlsx 数据的划分：按 8：1：1 的比例将数据划分为训练数据集、验证数据集、测试数据集，还有类别标签名单独一数据集。

数据集 数据量

训练集 7368

验证集 921

测试集 921

保存的相应位置：四个文件都存于 Bert-ERNIE 文件下的 THUCNews 目录下，

分别为 clss.txt（类文件）、train.txt(训练数据集）、dev.txt（验证集）、test.txt（测试集）

预训练语言模型：Bert 模型及 ERNIE 模型分别存在 Bert-ERNIE 文件下的 bert_pretain 目录下、ERNIE_pretrain 目下

3、训练并测试

#Bert

python run.py --model bert

# ERNIE

python run.py --model ERNIE

**第二题：**

1、数据说明

Hot Text 的目录下包含：

hot_text.py 代码文件，

newdic1.txt 自定义词典，

stopwords.txt 停用词词典，

原数据附件 3.xlsx

结果 Result 文件夹

2、运行代码

在相应路径的 Prompt 输入：python hot_text.py

将会展示聚类部分结果、

部分热度排名表、

前五热度问题表、

部分热度留言详情表等信息。

**第三题：**

1、数据说明

Reply_Quality 的目录下包含：

Reply_quality.py 代码文件，

tezhengdict.txt 自定义词典，

stopwords.txt 停用词词典，

原数据附件 4.xlsx

word2vec 文件夹

2、运行代码

在相应路径的 Prompt 输入：python Reply_quality.py

部分各指标指数、

各指标权重值、

部分数据及各指标指数。