问政平台群众留言的数据挖掘分析

摘要

近年来，随着网络问政平台的发展，网络问政平台已经逐渐成为政府了解民意、汇聚民智、凝聚民气的重要渠道。因此，运用自然语言处理技术和数据挖掘技术对群众留言信息进行分析具有重大的意义。

对于问题一，首先对文本进行预处理，包括去重、去空值、去 X 序列；然后利用 jieba 分词的精确模式对文本进行分词，对分词结果去停用词及向量化处理；再然后拆分训练集和测试集，运用 LSTM 模型训练数据进行文本多分类。最后运用 F-Score 模型对分类结果进行评价，得到整体分类结果的精确度为 84.26%。

对于问题二，首先对群众留言根据留言主题进行 K-means 聚类；然后根据群众留言建立热度评价指标，分别为被提及次数，点赞数，反对数；再根据聚类结果统计指标数据，运用熵权法计算每个指标的权值；最后运用 TOPSIS 模型对每一类问题进行打分并排序，求得排名前五的热点问题。

对于问题三，首先根据群众留言的问题答复建立回答质量评价指标，分别为回复的相关性和及时性，并运用余弦相似度计算相似度得分，利用 Excel 计算及时性得分；然后根据每个回复的指标得分，运用熵权法求得每个指标的权值；再然后运用线型加权法求得每个回复的质量得分；最后通过计算所有回复质量得分的平均值来评价总体答复意见的质量。

关键词：中文分词 LSTM TOPSIS 模型 K-means 聚类 熵权法 热度评价模型

Abstract

In recent years, with the development of the online political inquiry platform, the online political inquiry platform has gradually become an important channel for the government to understand public opinion, gather people's wisdom, and gather people's popularity. Therefore, the use of network natural language processing and data mining technology is of great significance to the study of mass message information.

For question one, the first step is to preprocess the text, including deduplication, null value removal, and X sequence removal. Then use the precise mode of jieba word segmentation to segment the text, and to stop words and vectorize the segmentation results. Then split the training set and test set, and use the LSTM model training data for classification. Finally, the F-Score model is used to evaluate the classification results.We get the accuracy of the overall classification results is 84.26%.

For question two, firstly, K-means cluster the mass messages according to the subject of the message, and then establish a heat evaluation index based on the mass messages, which are the number of times mentioned, the number of likes, and the number of objections. Then according to the statistical index data of the clustering result, the weight value of each index is calculated using the entropy weight method. Finally, the TOPSIS model is used to rank the problems, and the top five hot issues are obtained.

For question three, the first step is to establish an evaluation index for the quality of responses based on the answers to questions left by the masses, which are relevance and timeliness of responses. The cosine similarity is used to calculate the similarity score, and Excel calculates the timeliness score. According to the index score of each reply, the weight value of each index is obtained using the entropy weight method, and finally the quality score of each reply is obtained using the linear weighting method. Evaluate the quality of the overall response opinion by calculating the average of all response quality scores.

Keywords: Chinese word segmentation LSTM TOPSIS K-means clustering entropy weight method thermal evaluation model

目录

一. 前言

1.1 挖掘意义

随着越来越多的问政平台的发展，网络问政平台逐渐成为了政府了解民意、汇聚民智、凝聚民气的重要渠道。但是随着各类社情民意相关的文本数量不断攀升，给留言划分和热点整理工作带来了极大挑战。

构建群众留言挖掘模型，对群众留言进行分类并挖掘热点问题，能让政府及时了解最受民众关注和最亟需解决的问题，同时也能对给群众留言的答复意见进行质量把控。因此，根据问政平台的群众留言文本进行挖掘具有一定的理论意义和实用价值。

1.2 挖掘目标

本次建模的目标是根据收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见，运用 Python、MATLAB、Excel 软件，采用 LSTM 模型和 K-means 聚类方法建立模型，达到以下三个目标：

1) 利用文本分词和 LSTM 模型，根据留言内容及内容分类标签体系的一级标签对群众留言进行文本多分类；
2) 根据群众留言，定义热度评价指标，通过 K-means 聚类结果，运用熵权法计算热点评价指标的权值，最后运用 TOPSIS 模型对热点问题进行排序，得出排名前五的热点问题；
3) 首先根据相关部门对留言的答复意见，定义回答质量评价指标，其次运用余弦相似度计算相似度得分，运用 Excel 计算及时性得分，并通过熵权法求得每个指标的权值，最后运用线型加权法计算回复的质量得分。

1.3 挖掘流程

文本的挖掘流程如图 1-1 所示。

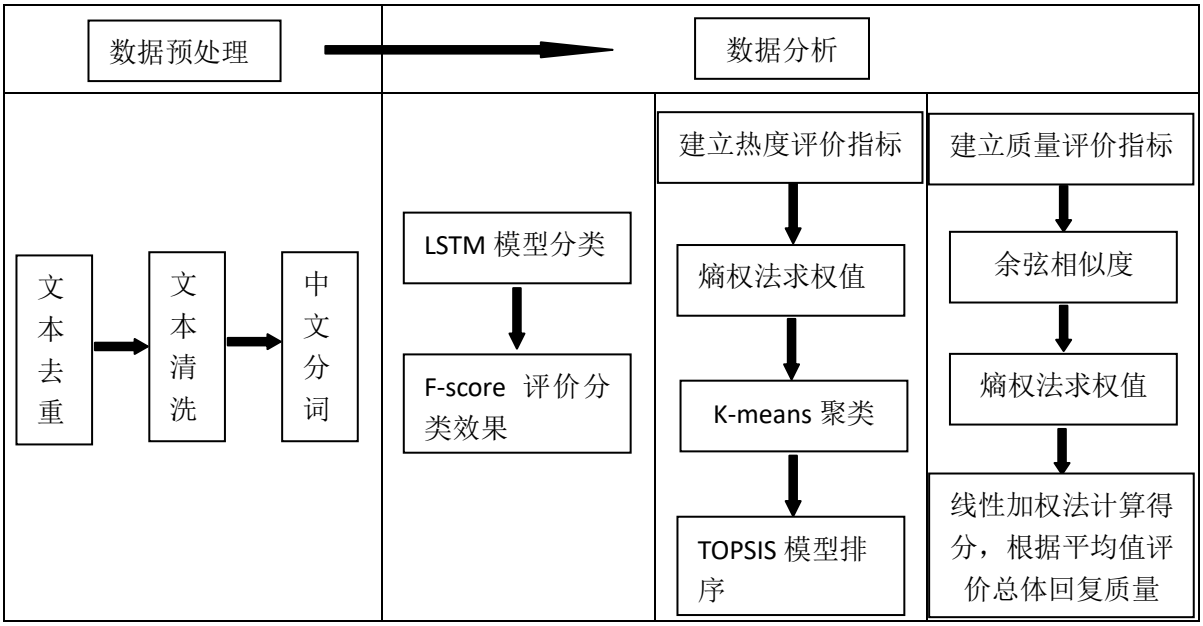

图 1-1 总体流程图

二. 基于 LSTM 的文本分类

问题一的流程图如图 2-1 所示。

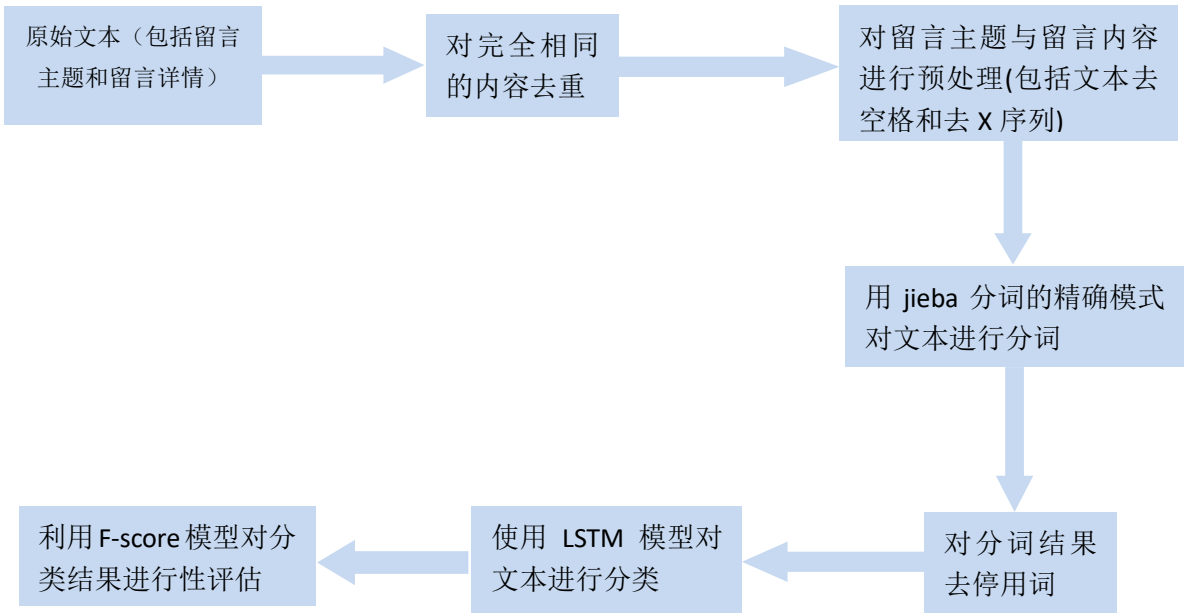

图 2-1 问题一处理流程

2.1 数据预处理

2.1.2 文本去重

在对文本分类之前，要对文本进行去重处理，因为本文使用的数据来源于网络问政平台的群众留言，群众都是主动反映问题，所以一般不会存在大幅度的文本重复问题。而即便大部分文本重复也不能说明该文本没有价值。所以这里，我们去重是指去除完全相同的文本。避免一个居民重复反映同样的问题造成的影响。

2.1.3 文本清洗

(1)去除空值

在文本中有许多空格，若不去除会对文本分类产生影响，所以在分词之前需先对文本去除空格。我们使用 python 的 re 模块获取文本去除空值后的内容。

(2)去除 X 序列

在文本中有一些数字，这些数字对文本分类没有任何帮助，且会对分类造成影响，比如以下一段截自文本中的内容：

> 3 月 6 日拨打市长热线 **12345** 反映问题，**12345** 转接给环保部门，环保部门工作人员进行了记录。10 日拨打 K 市 K2 区环境监测站 **8413602**（下称环境监测站）询问处理情况，工作人员不了解我反映的问题，再次告知，同时了解到环保部门工作任务繁忙，处理这类问题需要一周时间才有答复（中国公民真可怜）。

在这段文本中出现的“12345”，“8413602”这些号码在分类时没有任何作用，所以我们也先用 python 的 re 模块去除文本中这种类似的数字串，称为 X 序列。

2.2 文本分词

在对文本预处理完后，接下来就是对文本进行分词，我们使用 Python 的 jieba 分词的精确模式对文本进行大致的分词。但因为 jieba 词库中一些分词与现实生

活中人们的习惯的分词存在差别。比如下列一部分对留言主题的分词结果：

0,"['西湖', '建筑', '集团', '占', '道', '施工', '安全隐患']"

22,"['市中坡', '山', '公园', '内溜狗', '有损', '景区', '环境', '应', '严禁']"

29,"['农村信用', '合作', '联社', '208', '户', '合伙', '建房', '工程', '招投标', '问题']"

78,"['申请', '市公', '租房', '问题', '咨询']"

79,"['城市居民', '保障', '房', '相关', '政策']"

211,"['校园', '暴力事件', '屡屡', '发生']"

266,"['县七甲坪', '镇应', '发展', '赶尸', '傩', '文化', '哭', '嫁', '民俗风情']"

在这些分词中，划线的词语的分词都与现实生活中的分词不太符合，所以要对分词的词库进行添加，使这些词语能够符合现实的词语表达，比如：‘占道’，‘公园内’，‘公租房’，‘保障房’，‘校园暴力’。还有一些词语是比较特殊的名字，比如传统习俗的名字‘哭嫁’。如果单单分词成‘哭’，‘嫁’，则与它真实想要表达的内容存在较大差别。所以要将这些词语重新分词加入词库。

2.3 停用词过滤

在分词之后，会发现分词结果中存在大量无意义的词，比如：‘了’，‘的’，‘地’等，还有一些符号，数字之类的词语，不仅对文本分类没有作用，还会增加之后的工作量，所以要将这些停用词过滤。我们首先是使用传统的停用词库进行大致的去停用词，并生成词云，查看分词情况，结果如图 2-2 所示：

图 2-3　更新停用词库后的词云

由图 2-3 可知，在重新去除停用词后的词云，能够更加明显的突出民众想要反映的问题。

2.4 文本分类

由于民众给出的留言中存在普遍留言较长的现象，所以为了能够更好地理解群众留言的意思，进行正确的分类，应该结合留言的上下文进行分析，而 LSTM 模型就能做到这一点。因为相对于其他神经网络而言，LSTM 模型的优势在于能够长远结合上下文进行分析。所以本文选择采用 LSTM 模型来进行文本分类。

2.4.1 LSTM 模型原理[1]

LSTM 的关键结构就是细胞状态，LSTM 使用“细胞状态”来刻画神经元的记忆中不太容易衰减的部分，并围绕细胞状态构造长期记忆，其原理如图 2-4 所示。

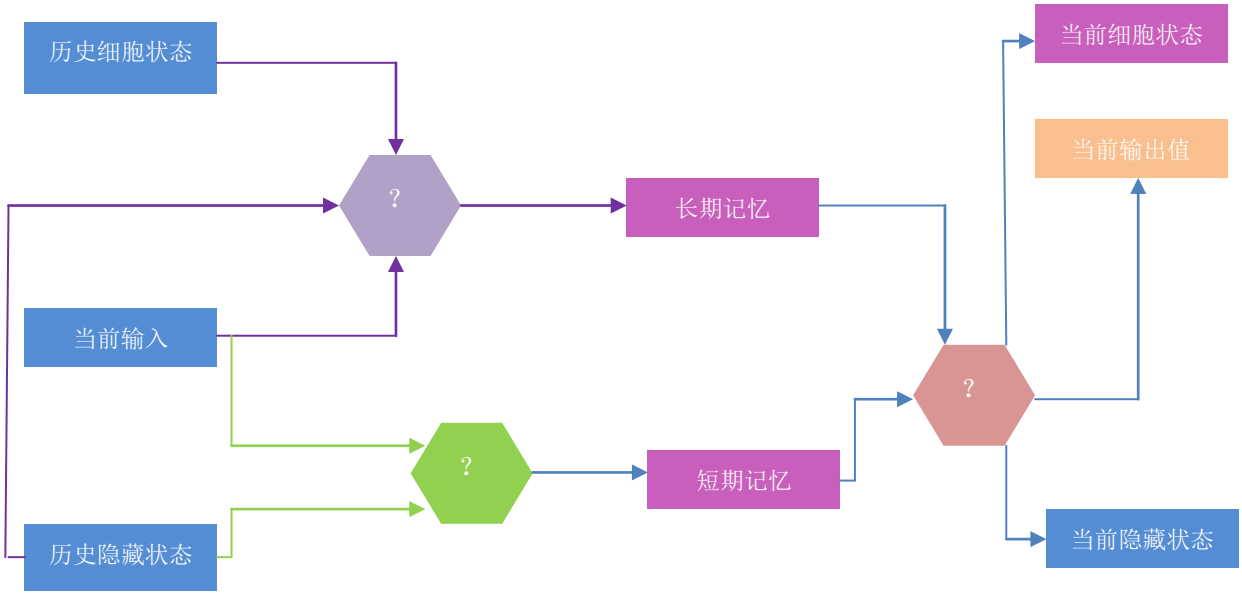

图 2-4 LSTM 模型原理

LSTM 以细胞状态向量的形式，来存储神经元对过往序列的宏观理解和记忆，然后把上一个时间步的细胞状态、当前时间步的输入、上一时间步的隐藏状态综合起来，构造出神经元的历史长期记忆。紫色六角形所代表的是构造长期记忆的模块结构；绿色六角形所代表的是用于构造短期记忆的模块结构；棕色六角形所代表的是用于综合长期记忆和短期记忆、计算当前时间步输出等的模块结构。

在 LSTM 的一个细胞中，主要包含三个不同的门结构：遗忘门，输入门和输出门。这三个门用来控制 LSTM 的信息保留和传递，最终反映到细胞状态，如图 2-5 所示：

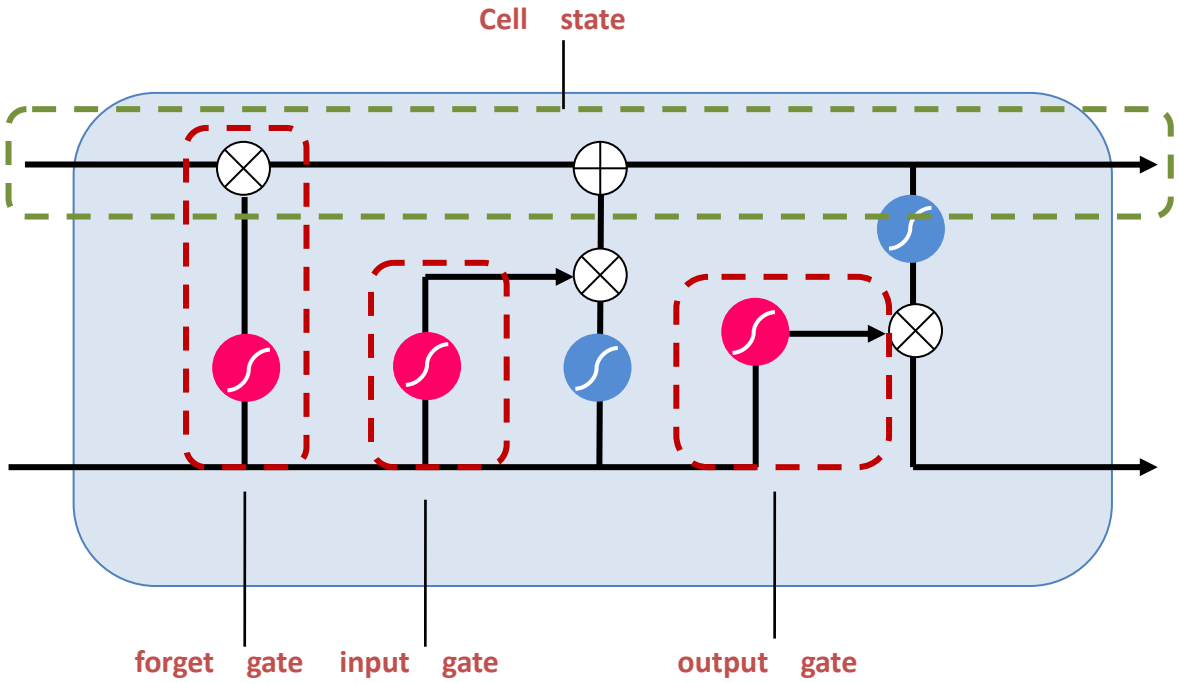

图 2-5 LSTM 门的结构

(1)遗忘门

决定细胞状态中丢弃什么信息。把和拼接起来，传给一个 sigmoid 函数，该函数输出 0 到 1 之间的值，这个值乘到细胞状态上去。sigmoid 函数的输出值直接决定了状态信息保留多少。比如当我们要预测下一个词是什么时，细胞状态可能包含当前主语的性别，因此正确的代词可以被选择出来。当我们看到新的主语，我们希望忘记旧的主语。

(2)输入门

在遗忘门中细胞状态已经被忘记了一部分，接下来就是考虑应该把哪些信息新加到细胞状态中。这里又包含 2 层：一个 tanh 层用来产生更新值的候选项，tanh 的输出在[-1,1]上，说明细胞状态在某些维度上需要加强，在某些维度上需要减

弱；还有一个 sigmoid 层（输入门层），它的输出值要乘到 tanh 层的输出上，起到一个缩放的作用，极端情况下 sigmoid 输出 0 说明相应维度上的细胞状态不需要更新。在那个预测下一个词的例子中，我们希望增加新的主语的性别到细胞状态中，来替代旧的需要忘记的主语。

(3)输出门

输出值跟细胞状态有关，把输给一个 tanh 函数得到输出值的候选项。候选项中的如果细胞状态告诉我们当前代词是第三人称，那我们就可以预测下一词可能是一个第三人称的动词。

2.4.2 LSTM 算法流程[2]

LSTM 流程图如图 2-6 所示：

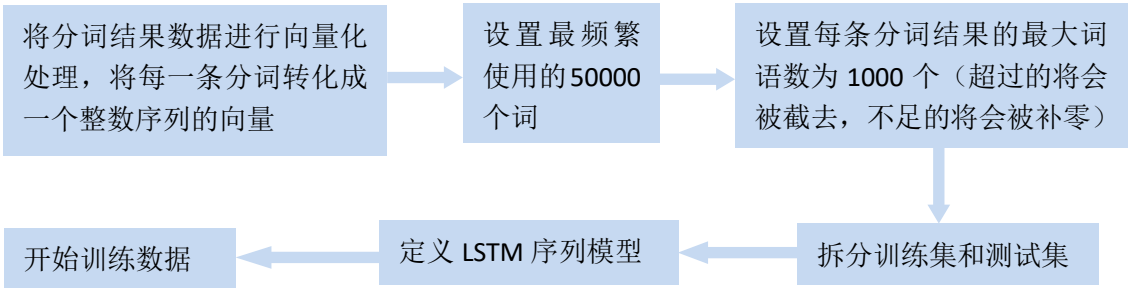

图 2-6 运用 LSTM 模型进行文本多分类

2.5 模型评价

在对群众留言进行分类后，需要评估分类结果的准确程度，因为所给数据中有给出群众留言的正确分类，所以本文可以通过所给的数据和自己的分类结果，使用 F-Score 模型对本文的分类结果进行评估。

2.5.1 F-Score 模型的建立

(1) TP、TN、FP、FN 解释说明

TP：True Positive, 被判定为正样本，事实上也是正样本。

FP：False Positive,被判定为正样本，但事实上是负样本。

TN：True Negative, 被判定为负样本，事实上也是负样本。

FN：False Negative,被判定为负样本，但事实上是正样本。

表 2-1 TP、TN、FP、FN 解释

真实情况	预测结果	
	正例	反例
正例	TP（真正例）	FN（假反例）
反例	FP（假正例）	TN（真反例）

(2) 各类参数指标的计算

Accuracy：表示预测结果的精确度，等于预测正确的样本数除以总样本数。

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

precision: 准确率，又称为查准率，表示预测结果中，预测为正样本的样本中，正确预测为正样本的概率。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

recall: 召回率，又称为查全率，表示在原始样本的正样本中，最后被正确预测为正样本的概率。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1 指标（F1-score）:F1-score 表示的是 precision 和 recall 的调和平均评估指标。

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

2.6 分类结果及 F-Score 求解分析

本文首先对各个类别的数据量进行统计，结果如图 2-7 所示。由图 2-7 可知，

城乡建设问题、劳动和社会保障问题都是群众反映较多的问题，有将近两千条留言；交通运输问题群众反映最少，只有大概六百条留言。各类目分布如图 2-7 所示。

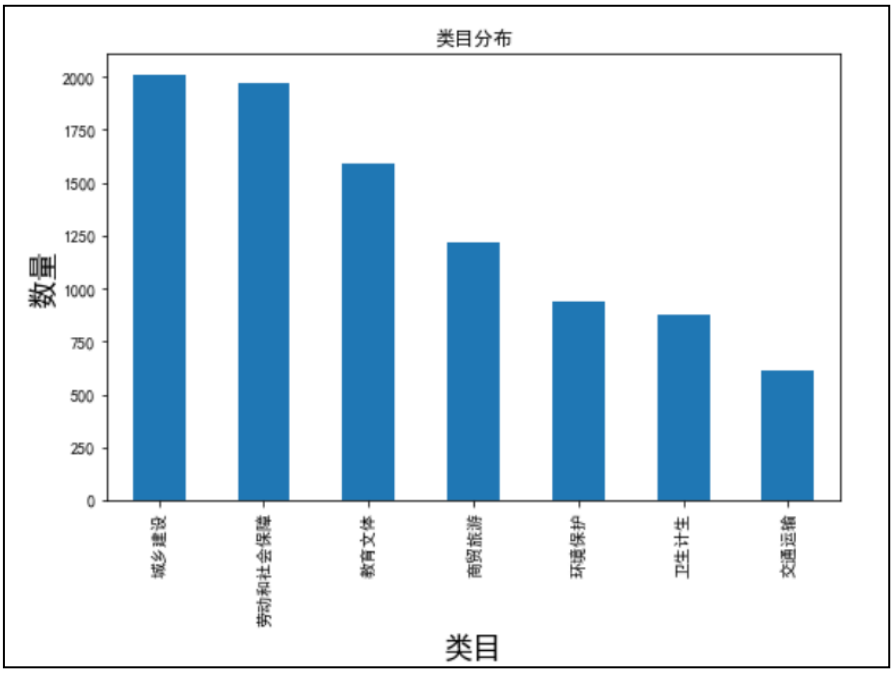

图 2-7　类目分布

运用 LSTM 模型训练数据后，为了检查训练效果，画出损失函数趋势图和准确率趋势图，分别如图 2-8，2-9 所示。

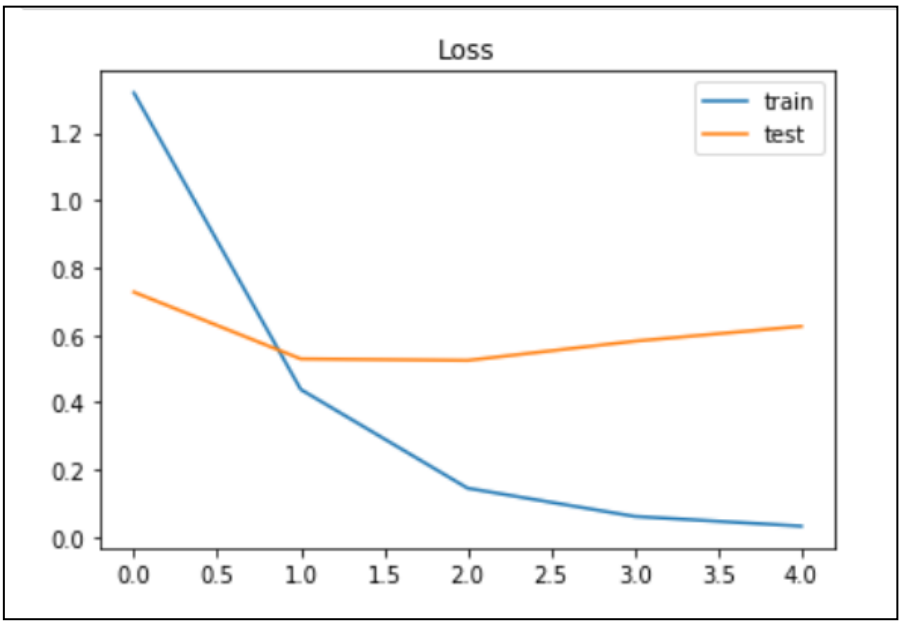

图 2-8　损失函数趋势

由图 2-8 可知,随着训练周期的增加,模型在训练集中损失越来越小,这是典型的过拟合现象,而在测试集中,损失随着训练周期的增加由一开始的从大逐步变小,再逐步变大。

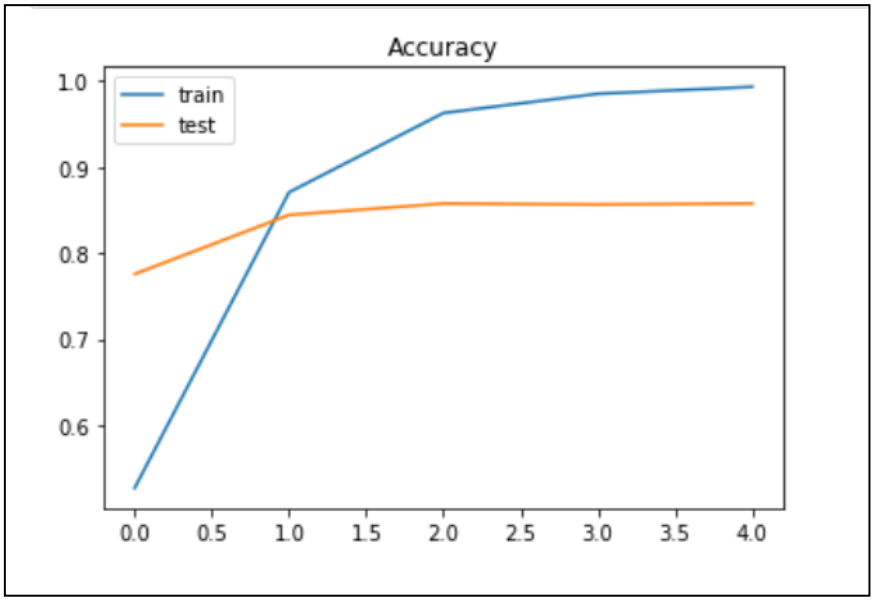

图 2-9 准确率趋势

由图 2-9 可知,随着训练周期的增加,模型在训练集中准确率越来越高,这是典型的过拟合现象,而在测试集中,准确率随着训练周期的增加由一开始的从小逐步变大，再逐步变小。

对分类后的结果运用 F-score 模型进行评价，得到结果如表 2-2 所示：

表 2-2 评价结果

Accuracy：0.8427				
	Precision	Recall	F1-score	support
城乡建设	0.82	0.78	0.80	200
环境保护	0.92	0.82	0.87	102
交通运输	0.80	0.70	0.74	73
教育文体	0.91	0.91	0.91	170
劳动和社会保障	0.84	0.94	0.89	172
商贸旅游	0.71	0.82	0.76	114
卫生计生	0.93	0.84	0.88	90
Accuracy(精确率)			0.84	921
Macro avg(宏平均)	0.85	0.83	0.84	921
weighted avg(加权平均)	0.85	0.84	0.84	921

由表 2-2 可以看出整个分类结果的精确度大概为 84.27%，而从不同分类的 F1 分数上看,“教育文体”类的 F1 分数最大为 0.91，“交通运输”类的 F1 分数最差只有 74%，原因可能是因为“交通运输”分类的训练数据最少,使得模型学习的不够充分,导致预测失误较多。

三．基于 K-means 聚类的热点问题挖掘

问题二的流程图如图 3-1 所示。

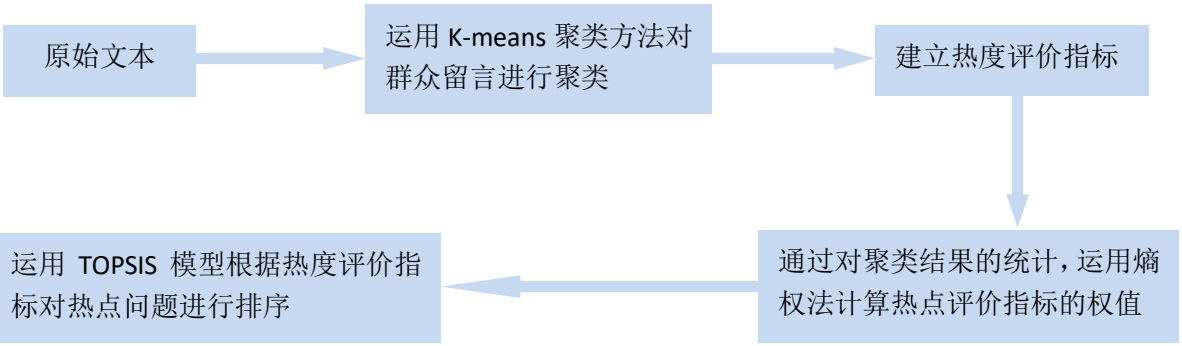

图 3-1 问题二流程图

3.1 对群众留言进行 K-means 聚类

本文首先运用 K-means 聚类方法对群众留言进行聚类，将类似问题分为一类问题，后续再对每一类问题进行热度评估。

3.1.1 K-means 聚类原理[3]

假设有一个包含 n 个 d 维数据点的数据集$X = \{x_1, x_2, \dots, x_i, \dots, x_n\}$，其中$x_i \in R^d$，K-means 聚类将数据集 X 组织为 K 个划分$C = \{c_k, i = 1,2, \dots, K\}$。每个划分代表一个类$c_k$，每个类$c_k$有一个类别中心$\mu_k$。选取欧氏距离作为相似性和距离判断准则，计算该类中每个点到聚类中心μ_k的距离平方和。

$$J(c_k) = \sum_{x_i \epsilon c_k} |x_i - \mu_k|^2$$

聚类目标是使各类总的距离平方和$J(C) = \sum_{k=1}^{K} J(c_k)$最小。$J(C)$的表达式如下：

$$J(C)=\sum_{k=1}^{K}J(c_k)=\sum_{k=1}^{K}\sum_{x_i\in c_i}|x_i-\mu_k|^2=\sum_{k=1}^{K}\sum_{i=1}^{n}d_{ki}|x_i-\mu_k|^2$$

其中，$d_{ki}=\begin{cases}1, & 若x_i\in c_i\\ 0, & 若x_i\notin c_i\end{cases}$。根据最小二乘法和拉格朗日原理，聚类中心$\mu_k$应该取为类别$c_k$内各数据点的平均值。

K-means 聚类的算法步骤如下：

1. 从 X 中随机取 K 个元素，作为 K 个簇各自的中心；
2. 分别计算剩下的元素到 K 个簇中心的相异度，将这些元素分别划归到相异度最低的簇；
3. 根据聚类结果，重新计算 K 个簇各自的中心，计算方法是取簇中所有元素各自维度的算术平均数；
4. 将 X 中全部元素按照新的中心重新聚类；
5. 重复第 4 步，直到聚类结果不再变化；
6. 将结果输出。

K-means 聚类流程图如下：

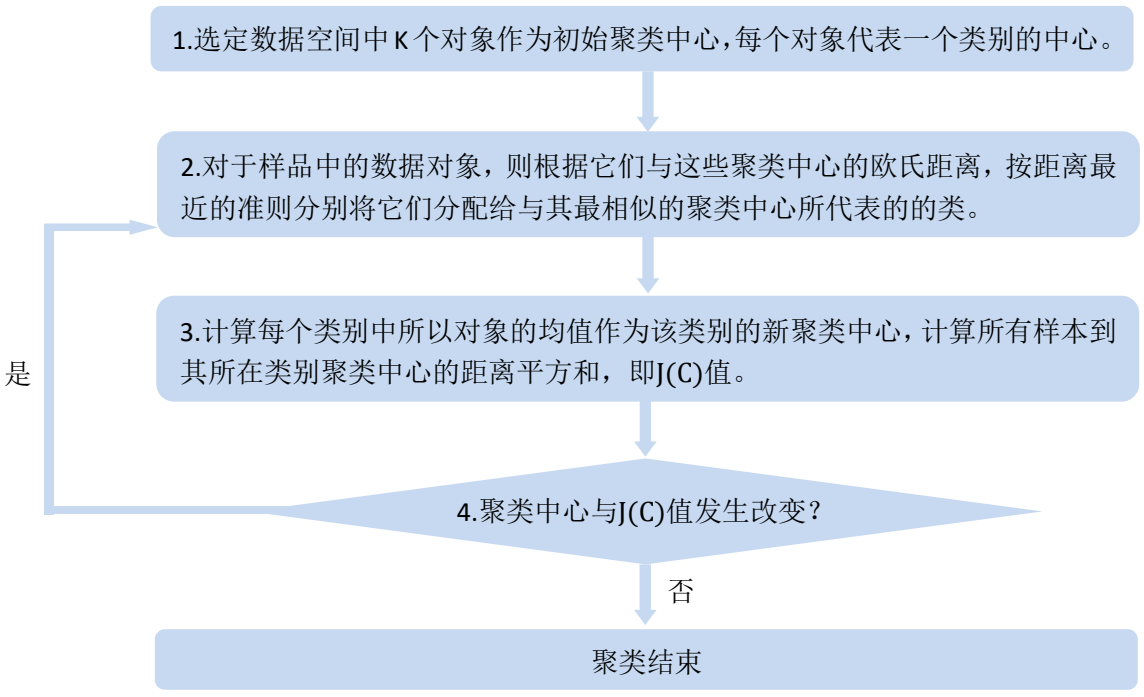

图 3-2 利用 K-means 算法进行文本聚类流程

3.1.2 K-means 聚类确定 K 值

使用 K-means 聚类方法，K 值的选取至关重要。若 K 值太大，会出现将一类问题归为几类的情况；若 K 值太小，会出现将几类问题归为一类的状况，所以需要对 K 值不断进行调试，得到聚类效果最好、最适合的 K 值。

调试流程如下：

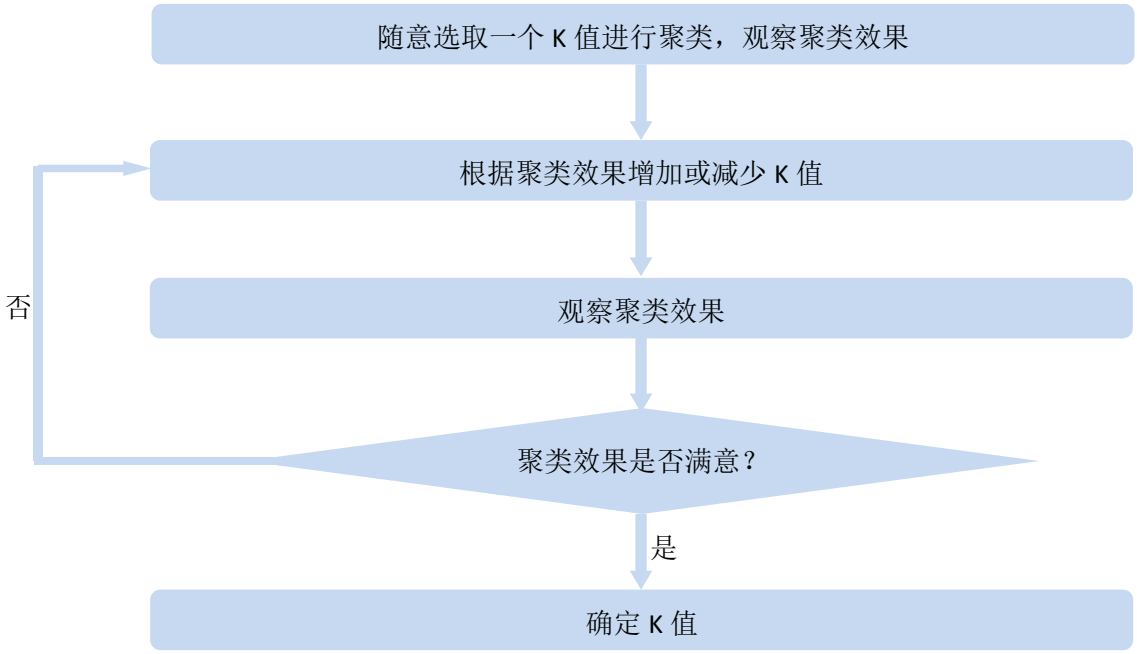

图 3-3 调试流程

本文通过调试，最终确定 K=285。

3.2 建立热度评价指标

通过对数据的分析，本文认为可以从所给数据中提出的体现热度的指标有三个，分别是：问题被提及次数，点赞数，反对数。

问题被提及次数能够直观的看出群众对问题的反映强烈度与讨论热度。问题被提及次数越多，说明问题对群众的影响越大，被解决的迫切程度越大。

点赞数可以看出群众们对此问题的认同度，是否也受到同样问题的困扰，点赞数越多，说明此问题的讨论热度越高。

反对数可以看出，有些问题只是个别人的希望，只能给个别人带来益处，并不能给广大群众带来好处。反对数越高，说明此问题群众认同感越低。

所以本文以这三个指标对热点问题进行排序。问题被提及次数即为每一类问题中问题的条数；点赞数即为每一类问题中每个问题的点赞数之和；反对数即为每一类问题中每个问题的反对数之和。

3.3 运用熵权法计算热度评价指标的权值

3.3.1 熵权法的基本原理

熵权法就是利用信息熵，计算出各个指标的权重，为多指标综合评价提供依据。按照信息论基本原理的解释，信息是系统有序程度的一个度量，熵是系统无序程度的一个度量。如果指标的信息熵越大，该指标提供的信息量越大，在综合评价中所起作用理当越大，权重就应该越高。

熵权法计算权重的步骤：

1.对 m 个待评项目，n 个评价指标，形成原始数据矩阵

$$R=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & & r_{2n} \\ \vdots & & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{pmatrix},$$

其中r_{ij}为第 j 个指标下第 i 个项目的评价值。

1. 求各指标值权重

（1） 计算第 j 个指标下的第 i 个项目的指标值的比重p_{ij}；

$$P_{ij}=r_{ij}/\sum_{i=1}^{m} r_{ij}$$

（2） 计算第 j 个指标的熵值e_j；

$$e_j=-k\sum_{i=1}^{m} p_{ij}\cdot \ln p_{ij} \quad \text{其中，} k=1/\ln m$$

（3） 计算第 j 个指标的权值w_j；

$$w_j = (1 - e_j) / \sum_{j=1}^{n} (1 - e_j)$$

3.3.2 熵权法计算结果[5]

通过对聚类结果进行统计，去除聚类效果特别差的分类，然后通过统计每一类问题的提及次数，点赞数和反对数总和，得到原始数据矩阵。

最终计算出权值为：

$w_{提及次数} = 0.5271$

$w_{点赞数} = 0.3850$

$w_{反对数} = 0.0879$

3.4 运用 TOPSIS 模型对热点问题排序[8]

在筛选掉聚类效果不好的 152 个聚类问题之后，对剩下的问题 113 个聚类问题运用 TOPSIS 模型对热点问题进行排序，得到排名前五的热点问题。

3.4.1 TOPSIS 模型基本原理[4]

TOPSIS 法是一种常用的组内综合评价方法，能充分利用原始数据的信息，其结果能精确地反映各评价方案之间的差距。基本过程为基于归一化后的原始数据矩阵，采用余弦法找出有限方案中的最优方案和最劣方案，然后分别计算各评价对象与最优方案和最劣方案间的距离，获得各评价对象与最优方案的相对接近程度，以此作为评价优劣的依据。

TOPSIS 模型算法步骤：

Input：原始数据集$X = \{x_1, x_2, \dots, x_n\}$

各指标权重$w = \{w_1, w_2, \dots, w_m\}$

Process：

1.对原始数据集中的指标属性同向化X'；

2.构造向量归一化后的标准化矩阵$Z=\{z_1,z_2,\ldots,z_n\}$；

3.**for Z 的每一列 Z_i do**

4.最劣方案Z^-的第 i 维度 $\leftarrow Z_i$ 元素最小值

5 .最优方案Z^+的第 i 维度 $\leftarrow Z_i$ 元素最大值

6 .**end for**

7 .**for $z_i \in Z$ do**

8 .z_i与最优方案的接近程度$D_i^+ \leftarrow D_i^+=\sqrt{\sum_{j=1}^{m} w_j(Z_j^+-z_{ij})^2}$

9 . z_i与最劣方案的接近程度$D_i^- \leftarrow D_i^-=\sqrt{\sum_{j=1}^{m} w_j(Z_j^--z_{ij})^2}$

10. z_i与最优方案的贴近程度$C_i \leftarrow C_i=\frac{D_i^-}{D_i^++D_i^-}$

11.**end for**

12.根据C_i大小进行排序

Output 各数据样本 TOPSIS 评价结果

3.4.2 TOPSIS 模型求解

通过TOPSIS模型对每一类问题进行综合打分得到排名前五的热点问题结果如表 3-1 所示。

表 3-1 排名前五的热点问题

排名	问题	被提及次数	点赞数	反对数
1	A5 五矿万境一系列问题	9	2109	0
2	A 市丽发新城小区附近搅拌机站扰民	40	261	0
3	A 市伊景园捆绑销售车位	47	22	1
4	关于地铁线路问题	24	36	2
5	A7 县金科时代车位及停车问题	7	18	1

具体热点问题数据见附件。

四．建立答复质量评估模型

问题三的流程图如图 4-1 所示。

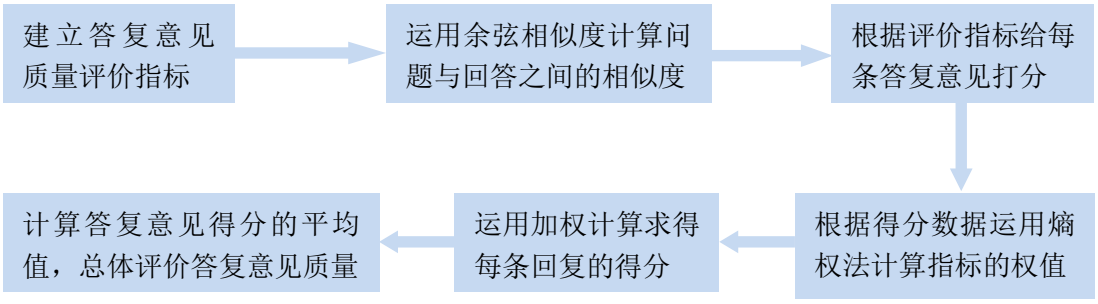

图 4-1　问题三处理流程图

4.1 建立答复意见质量评价指标[7]

本文通过对留言回复的分析，认为对于一个答复意见来说，首先最重要的是答复与问题的相关性，其次是答复的及时性，所以本文根据这两个指标来建立答复意见质量评价指标。

对于及时性，通过了解国家信访条例，本文定义了及时性打分标准为：

$$F_{及时性}=\begin{cases}1 & 0\le t_{天}<7\\ 0.9 & 7\le t_{天}\le 15\\ 0.8 & 15\le t_{天}<30\\ 0.6 & 30\le t_{天}<90\\ 0.4 & 90\le t_{天}<180\\ 0.2 & 180\le t_{天}<365\\ 0 & 365\le t_{天}\end{cases}$$

其中，$F_{及时性}$为答复意见的及时性得分，$t_{天}$为答复时间与提问时间之间的间隔天数。

4.2 余弦相似度计算相关性得分[6]

对于相关性，本文运用余弦相似法计算问题与答复意见之间的相似度。

余弦相似度，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的度量。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，这就是“余弦相似性”。

向量 a 和向量 b（n 维）的夹角的余弦计算如下：

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(X_i \times Y_i)}{\sqrt{\sum_{i=1}^{n}(X_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} = \frac{a \cdot b}{||a|| \times ||b||}$$

相关性得分的计算步骤如下：

1. 运用 jieba 分词对问题及回复意见进行分词；
2. 找出问题与答复意见中的关键词；
3. 分别从问题与答复意见中取出若干个关键词，合并成一个集合，计算问题与答复意见对于这个集合中的词的词频；
4. 运用 TF-IDF 算法生成问题与答复意见各自的词频向量；
5. 计算两个向量的余弦相似度，值越大就表示越相似。

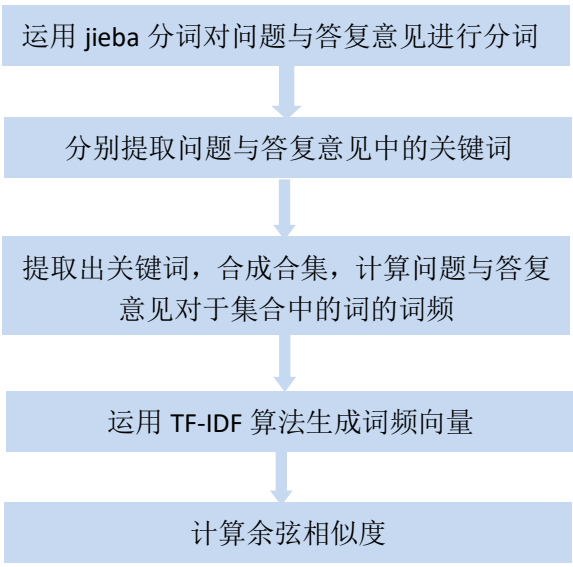

图 15：计算文本相似度

4.3 运用熵权法计算评价指标的权值

通过最终计算出的每条回复得分数据，采用熵权法计算每个指标的权值，结果如下：

$$W_{及时性} = 0.1854$$

$$W_{相关性} = 0.8146$$

4.4 线性加权法计算答复意见质量得分

通过 Excel 统计计算得出所有答复意见质量得分的平均值为 0.51，其中最高得分为 0.93，最低得分为 0.03。

表 4-1 答复意见得分

	平均得分	最高得分	最低得分
	0.51	0.93	0.03
相关性		0.94	0.04
及时性		0.9	0

通过最终得分可以看出总体答复意见处于中等水平，答复质量有高有低，最高得分与最低得分相差 0.9，说明对于群众留言的答复质量参差不齐，还需要改进。

五．总结与展望

5.1 总结

为了减轻人工对群众留言分类及热点问题发现的工作量，本文基于对留言文本的挖掘，构建模型，让政府工作人员能够更轻松，更高效率的处理群众问政留言。根据不同的问题，我们分别建立的相应的模型，通过实验验证了本模型的可行性，基本实现本赛题设立的目标。

5.2 展望

本文的热度评价模型只考虑了一个问题的被提及的次数，点赞数及反对数。并没有考虑到一个热点问题的时间问题，导致筛选出的热点问题跨度较大。命名实体识别是信息提取、问答系统、句法分析、机器翻译、面向 Semantic Web 的元数据标注等应用领域的重要基础工具，在自然语言处理技术走向实用化的过程

中占有重要地位。命名实体识别的任务就是识别出待处理文本中三大类(实体类、时间类和数字类)、七小类(人名、机构名、地名、时间、日期、货币和百分比)命名实体。所以可运用命名实体识别筛选出一定时间内被多次反映的问题，来完善热度评价模型。

六．参考文献

[1] 《LSTM——起源、思想、结构与“门”》https://zhuanlan.zhihu.com/p/115026734
[2]《基于 LSTM 的中文文本多分类实战》
https://blog.csdn.net/weixin_42608414/article/details/89856566?depth_1-utm_source=distribute.pc_relevant.none-task&utm_source=distribute.pc_relevant.none-task
[3]《网络招聘信息的数据挖掘与综合分析》
https://max.book118.com/html/2019/0503/8035104117002021.shtm
[4]《数学建模方法——带权重的 TOPSIS 法》
https://blog.csdn.net/limiyudianzi/article/details/103410150
[5]《熵权法求权重的 Matlab 实现》
https://blog.csdn.net/andy20081251/article/details/23192465
[6]《使用余弦相似度算法计算文本相似度》
https://www.cnblogs.com/airnew/p/9563703.html
[7]袁健，刘瑜．基于混合式的社区问答答案质量评价模型[J]．计算机应用研究，2017, 034(006):1708-1712.
[8]魏德志，陈福集，林丽娜．基于 MFIHC 聚类和 TOPSIS 的微博热点发现方法[J].计算机应用研究, 2018, v.35;No.318(04):60-63+87.