基于 TF-IDF 算法的 LSTM 模型解决文本挖掘问题

摘 要

互联网为我们敞开了言论自由的大门，微博、微信、市长邮箱逐渐成为政府和百姓交流沟通的有效工具。随着大数据，云计算的普遍发展，这种网络留言交流的方式普遍流行，需要处理的文本也越来越多。文本数量过多时，人工进行留言分类显得浪费时间，效率低下。因此基于这种情况，我们探索一种基于自然语言处理技术的方式来进行留言分类，热点问题挖掘，更高效地解决民生问题。

针对第一问，我们首先对数据进行预处理，用 Word2Vec 工具计算出每个词的词向量，然后再代入 LSTM 模型中，将数据集分成训练集和测试集，通过 LSTM 模型对训练集的训练学习，从而对测试集中的每个留言主题进行预测，得到其对应的一级标签，并进行对比验证，得到的准确率为 83.35%，未出现严重的过拟合现象，从而完成关于留言内容的一级标签分类模型的建立。

针对第二问，用第一问中建立好的模型给附件 3 贴上对应的一级标签，完成归类；之后对每条留言进行关键词提取，运用 Python 计算每个关键词的 TF-IDF 权重值；并用 Python 将同一条留言里对应的所有关键词赋予同一个 ID 号，计算每一个 ID（即每一条留言）的 TF-IDF 总权重；最终利用每条留言的 TF-IDF 总权重、反对数及点赞数加权求和作为热度评价指标，得到排名前 5 的热点问题以及对应的具体留言信息结果。

针对第三问，要给附件 4 中相关部门对留言的答复意见质量设定一套评价方案，我们从相关性、完整性、可解释性以上三个角度进行评定：

①相关性角度：我们利用基于加权 Word2Vec 的句向量去计算留言和答复的文本余弦相似度，并以此给出相似评分值；

②完整性角度：根据我们对完整性的定义，仍利用 TF-IDF 算法提取关键词，并定义完整率，使之成为完整性评价方案角度的指标；

③可解释性角度：我们通过先筛选出相似性评分和完整率均较高的答复内容以缩小范围，再从中查找并计数有多少对书名号“《》”和其他表示引经据典的符号来最终计算可解释性评分；

最后从上述三个角度分析出来的结果，加权求和得到最终的答复意见质量评价方案。

关键词 Word2Vec LSTM 模型 TF-IDF

Abstract

The Internet has opened the door to freedom of speech for us, and Weibo, WeChat, and the mayor's mailbox have gradually become effective tools for communication between the government and the people. With the general development of big data and cloud computing, this online message exchange method is generally popular, and more and more texts need to be processed. When there are too many texts, manual message classification is a waste of time and inefficient. Therefore, based on this situation, we explore a method based on natural language processing technology for message classification, hot spot problem mining, and more efficient solutions to people's livelihood problems.

For the first question, we first preprocess the data, use the Word2Vec tool to calculate the word vector of each word, and then substitute it into the LSTM model, divide the data set into a training set and a test set, and train the training set through the LSTM model. Learn, so as to predict each message topic in the test set, get its corresponding first-level label, and compare and verify it. The accuracy rate is 83.35%, and there is no serious over-fitting phenomenon, thus completing the message content Establishment of a first-level label classification model.

For the second question, use the model created in the first question to attach the corresponding first-level label to Annex 3 to complete the classification; after that, extract the keywords for each message and use Python to calculate the TF-IDF of each keyword Weight value.Finally use the total TF-IDF weight of each message, The weighted sum of the number of objections and the number of likes is used as a heat evaluation index, and the top 5 hot issues and corresponding specific message information results are obtained.

Regarding the third question, a set of evaluation schemes should be set for the quality of the response comments of the relevant departments in Annex 4 from three perspectives: relevance, completeness, and interpretability:① Relevance angle: We use sentence vectors based on weighted Word2Vec to calculate the cosine similarity of the text of the message and the reply, and give a similarity score value based on this; ②Integrity perspective: According to our definition of integrity, the TF-IDF algorithm is still used to extract keywords and define the integrity rate, making it an indicator of the integrity evaluation program perspective;③Interpretability angle: we first narrowed down the scope of the response content with a high similarity score and completeness rate, and then looked up and counted how many "《》"and other symbols representing the citations to finally calculate Explanatory score;Finally, the results analyzed from the above three angles, weighted summation to get the final response opinion quality evaluation program.

目录

1 问题重述

1.1 问题背景

近年来，随着微信、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，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。

附件给出了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。请利用自然语言处理和文本挖掘的方法解决下面的问题。

1.2 需解决的问题

本文将题述的三个问题提炼出以下主要求解的信息，并建立相应的数学模型或合适的算法进行分析与求解。

问题一：群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。

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。

问题二：热点问题挖掘

根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，并给出评价结果，按格式给出排名前 5 的热点问题，并整理出“热点问题表.xls”，如表 1；

表 1：热点问题表

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	…	2019/08/18 至 2019/09/04	A 市 A5 区魅力之城小区	小区临街餐饮店油烟噪音扰民
2	2	…	2017/06/08 至 2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
…	…	…	…	…	…

并且按表 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”。

表 2：热点问题留言明细表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	360106	A235367	A 市魅力之城小区底层商铺营业到凌晨，各种噪音好痛苦	2019/08/26 01:50:38	2019 年 5 月起，小区楼下商铺越发嚣张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音严重影响了小区居民休息……	0	0
…	…	…	…	…	…	…	…
1	360109	A0080252	魅力之城小区底层门店深夜经营，各种噪音扰民	2019/09/04 21:00:18	您好：我是魅力之城小区的业主，小区临街的一楼是商铺，尤其是餐馆夜宵摊等，每到凌晨都还在营业，每到晚上睡觉耳边都充斥着吆喝……	0	0
2	360110	A110021	A 市经济学院寒假过年期间组织学生去工厂工作	2019/11/22 14:42:14	西地省 A 市经济学院寒假过年期间组织学生去工厂工作，过年本该是家人团聚的时光，很多家长一年回来一次，也就过年和自己孩子见一次面，可是这样搞……	0	0
2	360111	A1204455	A 市经济学院组织学生外出打工合理吗？	2019/11/5 10:31:38	学校组织我们学生在外边打工，在东莞做流水线工作，还要倒白夜班。本来都在学校好好上课，十月底突然说组织到外省打工……	1	0
…	…	…	…	…	…	…	…

问题三：答复意见的评价

针对附件 4：相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

2 问题分析

2.1 问题一：群众留言分类

实际上，这是一个文本分类(多分类)问题，而常用的文本分类，大体上分为基于传统机器学习的文本分类模型，基于深度学习的文本分类模型；我们可以采用最经典的 Word2Vec 工具，该工具在 NLP 领域具有非常重要的意义，我们用 Word2Vec 工具先计算每个词的词向量，然后再代入 LSTM 模型中，将数据集分成训练集和测试集，通过对训练集的训练学习，从而对测试集中的每个留言主题进行预测，并进行对比验证。

在问题一可能存在的一些难以处理的点，比如说：文本语义带来的词语交叉、多分类问题带来的难度(转化为多个二分类)、数据不平衡带来的影响(数据增强)、长文本的无意义表达太多(是否转为短文本、关键句)，这些都将是我们主要解决的问题。

在一级标签分类完成后，可以对该分类方法进行评价，这里我们采用 F-Score 对分类方法进行评价：

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

其中，P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

2.2 问题二：热点问题挖掘

将问题二拆解为以下几个子任务：

子任务 1:问题识别

即如何从众多留言中识别出相似的留言。这里可以用第一问中计算出来的词向量进行相似度计算，从而找到相似的留言。

子任务 2：问题归类

把特定地点或人群的数据归并，即把相似的留言归为同一问题（即给予相同的问题 ID），结果对应表 2。

子任务 3：热度评价

热度评价指标的定义和计算方法，对指标排名之后得出对应表 1，即可得到筛选出来的排名前 5 的热点问题。

问题 2 中我们主要解决以下难点：一是地点、人群的识别(表达多样化)；二是相似的计算复杂(特征多、两两之间计算相似计算量大)。

2.3 问题三：答复意见的评价

要解决的问题主要是针对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。

而相关性可以从答复意见的内容是否与问题相关来分析；完整性可以从是否满足某种规范来考虑；可解释性可以从答复意见中内容的相关解释，是否引经据典，若不能解决该问题，是否有说明为什么不发解决该问题，若可解决，则可以观察在提供方法时，是否让人可以接受。

3 假设与符号

3.1 模型假设

1.假设所有留言都在附件 1 中的一级标签范围内，无其他额外的一级标签。
2.假设附件 2 中所给的原有标签都是正确的。

3.2 符号说明

表 3：符号说明表

符号	意义
w_i	第 i 个索引词
A_i	表示向量 A 的第 i 维度分量

$n_{i,j}$	第i个关键词在某一文档中出现的次数
|D|	表示语料库中的文件总数

注：未列出符号及重复的符号以出现处为准

4 模型建立与求解

4.1 基于问题一的求解

4.1.1 模型的建立

我们尝试了不同模型进行文本多分类处理，分别基于机器学习和深度学习的方法进行处理。不同模型得出不同结果进行比较，全方面地对问题进行分析。并且通过对比出各个模型优缺点及准确率，最后找到如下模型为最佳模型。

- 基于 Word2Vec 的 LSTM 文本分类模型

（1）词向量文本分类方法的背景

在 NLP 任务中，由于当下互联网发展过快，人们想从大量文本中挖掘有用的信息，而计算机无法直接识别人类语言，自然语言处理技术应运而生。其中统计语言模型是很重要的一个环节，它是 NLP 中很重要的一个环节，是实现各个任务的基础。

简而言之，统计语言模型根据一个句子来计算其概率模型的，通过语料库，由贝叶斯公式链式生成句子中各个词的语言参数：

$$p(W)=p(w_1^T)=p(w_1,w_2\ldots,w_T) \tag{2}$$

$$p(w_1^T)=p(w_1)p(w_2\mid w_1)p(w_3\mid w_1^2)\ldots p(w_T\mid w_1^{T-1}) \tag{3}$$

Word2Vec 也称 Word Embedding，中文的叫法是“词向量”或“词嵌入”，是一种计算非常高效的，可以从原始语料中学习字词空间向量的预测模型。NLP 中所有任务都是拿词语开刀，词语组成句子，句子组成文章，把握词语的特性就能把握整个文本的特性。NLP 中的词语是一个抽象化的数学表达形式，或者说嵌入到数学空间中，这种方法就称为词嵌入，也就是 Word2Vec 模型的基础。

传统方法处理词向量通常是用一段长向量来表示句子，向量长度即为词典的大小，用 0,1 的位置表示该词在词典中的位置。这样的方法表示简洁，清晰易懂，但是易受到维数灾难的影响，也不能表示词与词之间的联系。在 1968 年 Hinton 提出了 Distributed REpresentation 的方法，可以解决上述模型的缺点，即用一个普通向量来表示句子，其向量长度通常为 50 到 100 个维度；当然一个词怎么表示成这么样的一个向量需要通过训练得到，Word2Vec 是最常见的一种训练方法。每个词在不同的语料库和不同的训练方法下，得到的词向量是不一样的。

而 Word2Vec 模型又分为以下两种模型，分别是 Skip-gram 和 CBOW 模型。

①若是用一个词语作为输入，来预测其周围的上下文，那这个模型叫做 Skip-gam 模型。

Skip-gram 模型的原理图如下所示：

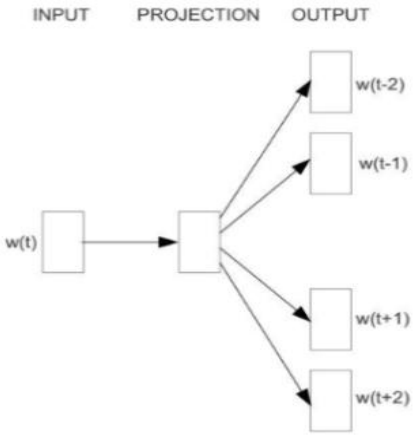

图 1：Skip-gram 模型原理图

Skip-gram 模型分为输入层，映射层和输出层，$w(t)$ 为输入，$w(t-1),w(t-2),w(t-3)\ldots$为在 $w(t)$ 为前提情况下的预测的上下文，其根据贝叶斯概率模型得出相关结果。

②若是拿一个词语的上下文作为输入，来预测这个词语本身，则是 CBOW 模型。

CBOW 模型的原理图如下所示：

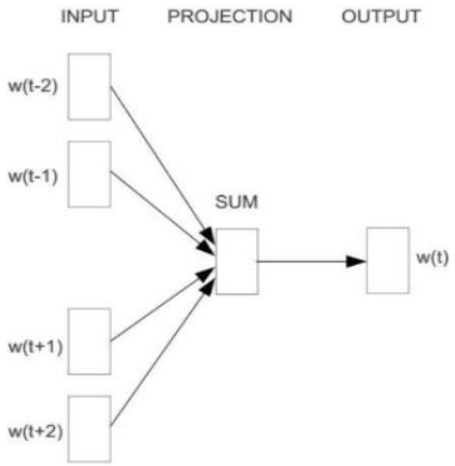

图 2：CBOW 模型原理图

CBOW 模型分为输入层，映射层和输出层，$w(t)$ 为输入，$w(t-1),w(t-2),w(t-3)\ldots$ 为在 $w(t)$ 为前提情况下的预测的上下文，由贝叶斯概率模型得出相关结果。不难发现，CBOW 和 Skip-gram 模型本质上是一样的。

（2）LSTM 模型

（2.1）一提及 LSTM 模型，我们就不得不谈一下 RNN（递归神经网络）。

①概述

Recurrentneural network，循环神经网络，在普通多层 BP 神经网络基础上，增加了隐藏层各单元间的横向联系，通过一个权重矩阵，可以将上一个时间序列的神经单元的值传递至当前的神经单元，从而使神经网络具备了记忆功能，对于

处理有上下文联系的 NLP、或者时间序列的机器学习问题，有很好的应用性。

②优缺点

优点：模型具备记忆性。

缺点：不能记忆太前或者太后的内容，因为存在梯度爆炸或者梯度消失。

③模型

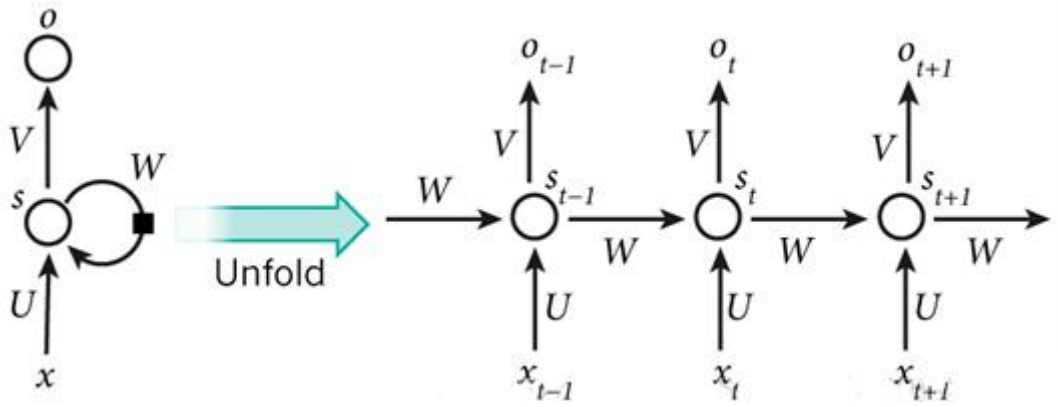

图 3：RNN 模型

下式 4 是输出层的计算公式，输出层是一个全连接层，也就是它的每个节点都和隐藏层的每个节点相连。V 是输出层的权重矩阵，g 是激活函数。式 5 是隐藏层的计算公式，它是循环层。U 是输入 x 的权重矩阵，W 是上一次的值作为这一次的输入的权重矩阵，f 是激活函数。

我们可以用下面的公示来表示循环神经网络的计算方法：

$$o_t = g(Vs_t) \tag{4}$$

$$s_t = f(Ux_t + Ws_{t-1}) \tag{5}$$

（2.2）LSTM 的原理

①概述

Longshort term memory，循环神经网络的变形结构，在普通 RNN 基础上，在隐藏层各神经单元中增加记忆单元，从而使时间序列上的记忆信息可控，每次在隐藏层各单元间传递时通过几个可控门（遗忘门、输入门、候选门、输出门），可以控制之前信息和当前信息的记忆和遗忘程度，从而使 RNN 网络具备了长期记忆功能，对于 RNN 的实际应用，有巨大作用。

②优缺点

优点：比 RNN 具备长期记忆功能，可控记忆能力。

缺点：网络结构上比较复杂，门多，对效率又影响。于是，在此基础上，更加简化实用的结构形式 GRU(gate recurrent unit)提高了效率，它把输入门和遗忘门进行了合并，把 S_t 和 C_t，即记忆单元和输出单元进行了合并。

③模型

LSTM 也设计两个门控制记忆单元状态 c 的信息量：一个是遗忘门（forget gate）。所谓的“遗忘”，也就是“记忆的残缺”。它决定了上一时刻的单元状态有多少“记忆”可以保留到当前时刻；另一个是输入门（input gate），它决定了当前时刻的输入有多少保存到单元状态。

我们说过，LSTM 是由三个门来实现的。实际上，为了表述方便，很多文献还添加了一个门，叫候选门（Candidate gate），它控制着以多大比例融合“历史”

信息和“当下”刺激。

最后，LSTM 还设计了一个输出门（output gate），来控制单元状态有多少信息输出。下面对这 4 个门分别进行简单的原理介绍。

本问题中 LSTM 模型的遗忘门：

$$f_t = \sigma(W_f^T \times s_{t-1} + U_f^T \times x_t + b_f) \tag{6}$$

本问题中 LSTM 模型的输入门：

$$i_t = \sigma(W_i^T \times s_{t-1} + U_i^T \times x_t + b_i) \tag{7}$$

本问题中 LSTM 模型的候选门：

$$C_t' = \tanh(W_c^T \times s_{t-1} + U_c^T \times x_t + b_c) \tag{8}$$

于是，记忆单元的模型函数就是：

$$C_t = f_t \times C_{t-1} + i_t \times C_t' \tag{9}$$

本问题中 LSTM 模型的输出门是：

$$O_t = \sigma(W_o^T \times s_{t-1} + U_o^T \times x_t + b_o) \tag{10}$$

最终的时间序列上的输出量是：

$$s_t = O_t \times \tanh(C_t) \tag{11}$$

④代价函数

原理与 RNN 一样。

类似这种：

$$\min J(\theta) = \sum_{t=1}^{T} loss(\hat{y}^{(t)}, y^{(t)}) \tag{12}$$

⑤求解模型参数的方法

BPTT（backpropagation through time）算法是针对循环层的训练算法，它的基本原理和 BP 算法是一样的，也包含同样的三个步骤：

（i）首先确定参数的初始化值，然后前向计算每个神经元的输出值；不过它的输出值比 RNN 和 NN 要多，因为有几个门，对于 LSTM 而言，依据前面介绍的流程，按部就班地分别计算出 f_t，i_t，c_t，o_t 和 s_t。

（ii）反向计算每个神经元的误差项值，它是误差函数 E 对神经元 j 的加权输入的偏导数；与传统 RNN 类似，LSTM 误差项的反向传播包括两个层面：一个是空间上层面的，将误差项向网络的上一层传播。另一个是时间层面上的，沿时间反向传播，即从当前 t 时刻开始，计算每个时刻的误差。

（iii）计算每个权重（即参数）的梯度。

最后再用随机梯度下降算法更新权重。

4.1.2 模型的建立与计算过程

（1）数据处理

由于不同的文本间的关联信息各不相同，因此我们选择使用附件 2 中留言文本计算一套新的 Word2Vec 词向量以方便后续建模。先对所给的附件 2 数据进行处

理。根据问题 1 中的背景，我们可知标签分类以留言内容作为依据，先对留言数据进行整理，步骤如下：

①对数据集进行预处理

首先利用 jieba 分词算法对留言内容进行分词处理，利用基于前缀词典对文本进行词图扫描，建立如下图的所有可能生成词情况所构成的有向无环图（DAG），并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。

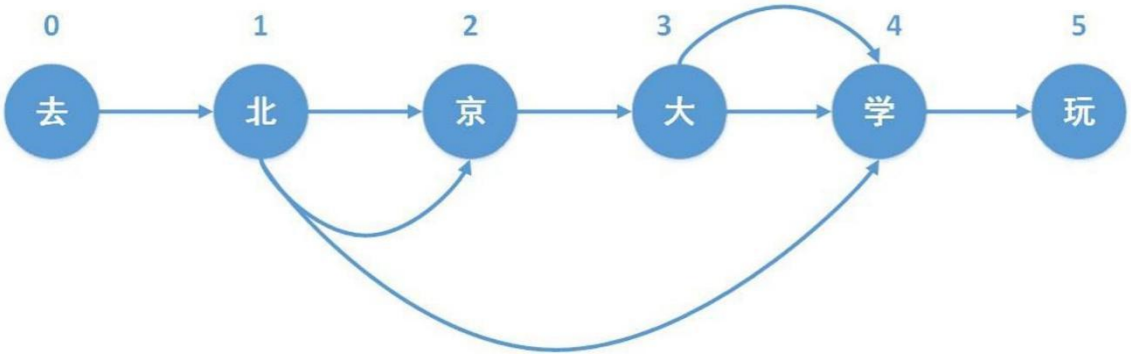

图 4：可能生成词情况有向无环图

接着提取特征词、建立词索引。将不利于训练的样本剔除，清洗数据，有助于准确度的提升。通过数据预处理，先建立 Word2Vec 训练需要的语料库，对其进行编码；根据特征词挖掘出留言内容的关键信息。

②二次采样，降低高频词概率

为了提高文本分类的准确性和效率，一般先剔除决策意义不大的词语，二次抽样即为对文本数据中一般会出现一些高频词，在背景窗口中，一个词和较低频词同时出现比和较高频词同时出现对训练词嵌入更有益，因此，词嵌入模型时可以对词进行二次采样。

通常进行二次采样如下所示：记数据集中每个索引词 w_i 有一定的概率被丢弃，这个概率可以表示为：

$$P(w_i)=\max(1-\sqrt{\frac{t}{f(w_i)}},0) \tag{13}$$

其中 $f(w_i)$ 为数据集中词 w_i 与总词数之比，t 为一常数，可以令 $t=10^{-4}$。按照上述公式可知，当词的出现频率越高时，被丢弃的概率就越大，完成了二次采样的目的。

经过二次采样后，数据冗余的情况下减少，可以进行训练的有效词数量增加。

（2）Word2Vec 的求解

我们通过 Word2Vec 模型可以先对各个问题进行相关性分析，得到关键词对应的词图。Word2Vec 模型分为以下两种模型，分别是 Skip-gram 和 CBOW 模型。我们采取的是 Skip-gram 模型，即使用一个关键词作为输入，来预测其周围的上背景词出现的概率，根据 Skip-gram 模型的原理图，可知 Skip-gram 模型分为输入层，映射层和输出层，$w(t)$ 为输入，$w(t-1),w(t-2),w(t-3)\ldots$ 为在 $w(t)$ 为前提情况下的预测的上下文，其根据贝叶斯概率模型得出相关结果。在本文中，我们使用 Mxnet 来实现 Skip-gram 的 Word2Vec 模型：

①提取中心词和背景词

采用非固定长度的向量进行背景词的提取，按要求来看，每个句子至少有两个词才可以提取出一对“中心词-背景词”。再将中心词排除在背景词之外，以“使安全隐患非常大”为例，可得到如下结果：

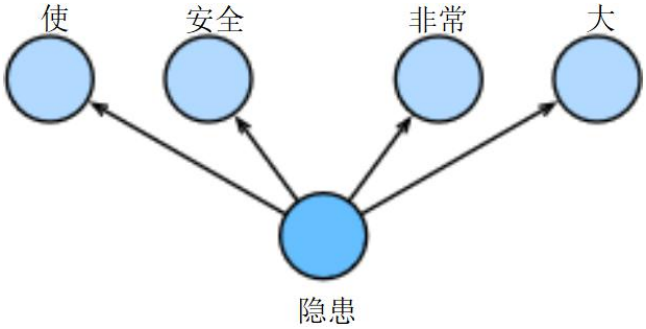

图 5：Skip-gram 模型关心给定中心词生成背景词的条件概率

②负采样

负采样的一种形式就是带权采样。通常情况下，训练一个神经网络意味着使用一个训练样本就要稍微调整一下所有的神经网络权重，也就是说，每个训练样本都会改变神经网络中的权重。负采样的意义在于使得每个训练样本仅改变一部分神经网络的权重，减轻训练负担提高效率。

我们随机选择较少数目的“负”样本改变权重，并且我们仍然为我们的“正”样本更新权重。“负”样本的选择取决于词出现的概率，对于出现频率高的词，我们更倾向于选择其为负样本。

③构建模型训练数据

Skip-gram 模型的参数每个词所对应的中心词向量和背景词向量。数据集的格式如下图所示：

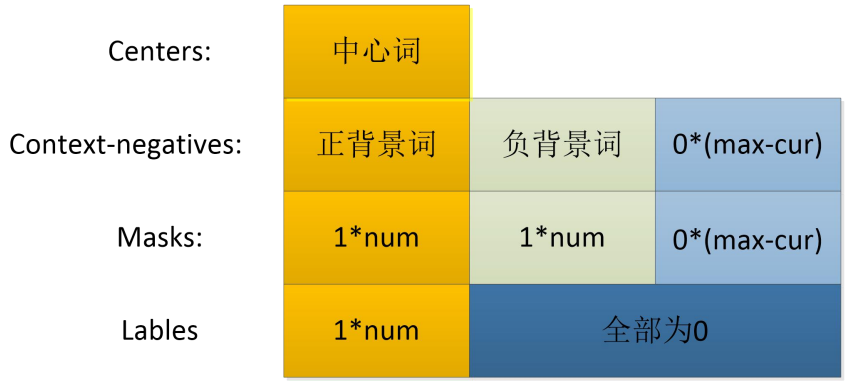

图 6：Word2Vec 训练原理图

训练中我们通过最大化似然函数来学习模型参数，即最大似然估计。这等价于最小化以下损失函数：

$$-\sum_{t=1}^{T}\sum_{-m\le j\le m, j\neq 0}\log P(w^{(t+j)}|w^{(t)}) \qquad (14)$$

按照上述目标，我们选择采用随机梯度下降来进行训练，根据实际样本，设置嵌入层大小，中心词嵌入层，背景词和负样本嵌入层。对每一层数据进行训练，以其中 200 个字词为例，将结果可视化可以得到：

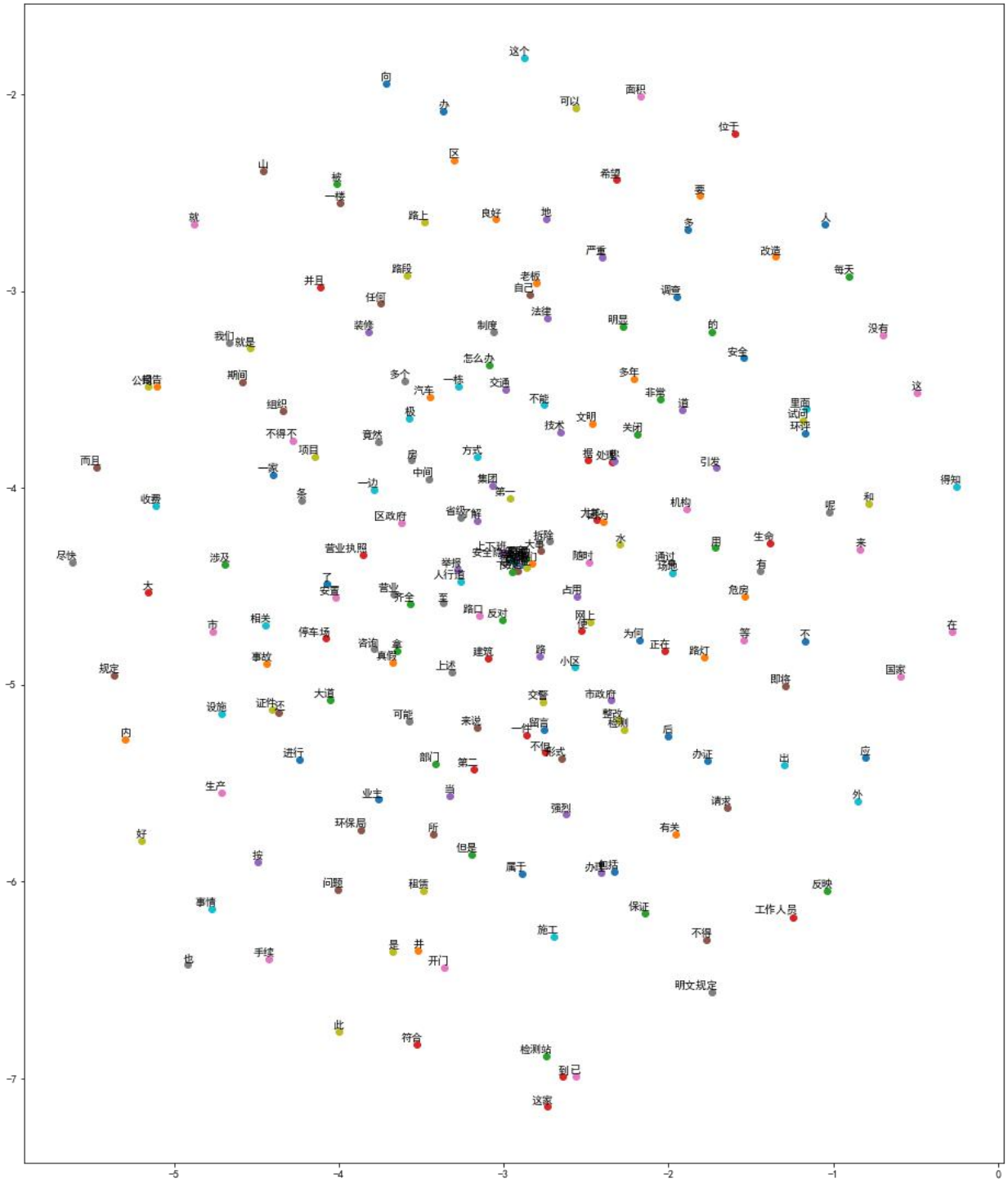

图 7：Word2Vec 词向量可视化结果示意图

将每个字词对应的词向量拆分成训练集和测试集代入 LSTM 模型中，作为模型的第一层嵌入层(Embedding)。

（3）代入 LSTM 模型进行求解

将 Word2Vec 的处理结果作为输入代入 LSTM 模型中，我们得到最终的算法流程图，绘制如下：

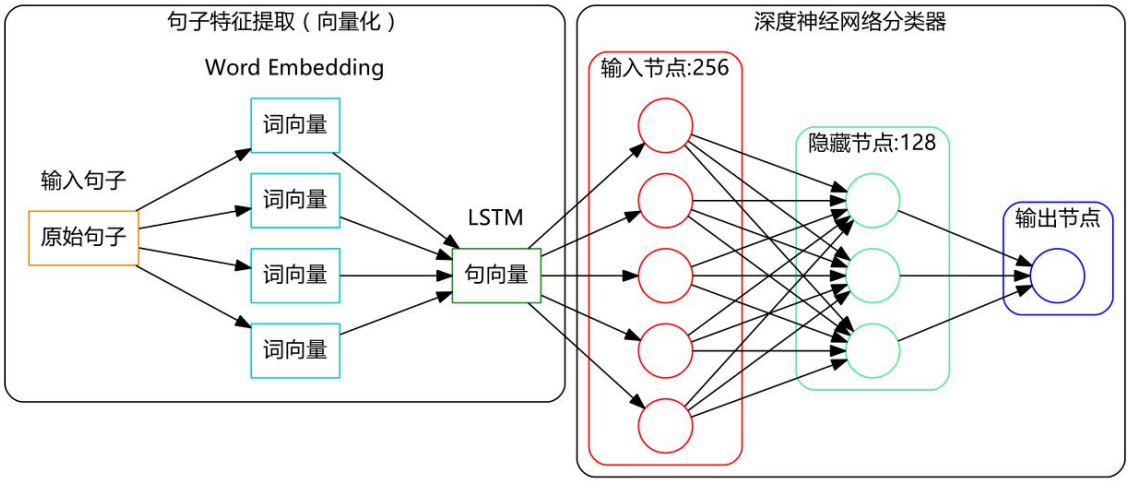

图 8：基于 word2vec 的 LSTM 文本分类模型示意图

具体求解步骤如下：

首先我们将数据集分成不同的十份，进行交叉训练：

	classification	comments
3331	交通运输	G8县维新镇申通、圆通、中通快递每件收费3元
7970	商贸旅游	对食品相关产品工业产品生产许可证范围问题的咨询
1365	城乡建设	L3县危房改造补贴款要怎么查询
5703	劳动和社会保障	G市城区做液化气生意人在工作场所和工作中抽烟
7182	商贸旅游	A7县松雅有大量外地人员从事传销活动，请严厉打击
8947	卫生计生	办出生医学证明无着落，小孩上户难如上青天
8984	卫生计生	A市妇幼产一科主治医生言语极其恶劣
6589	劳动和社会保障	要求解决原L市电子研究所退休人员养老待遇问题
7593	商贸旅游	投诉K8县公交车乱收费
3029	交通运输	高速公路C4市收费站出入口的ETC总是坏

图 9：随机选取的样本图

统计出整个数据集中各个分类所占的比重，并根据分类我们得出 15141 个不同的词语，以下是统计分类后的类目和对应的数量分布：

	classification	counts
0	城乡建设	2009
1	劳动和社会保障	1969
2	教育文体	1589
3	商贸旅游	1215
4	环境保护	938
5	卫生计生	877
6	交通运输	613

图 10：统计分类（表格形式）

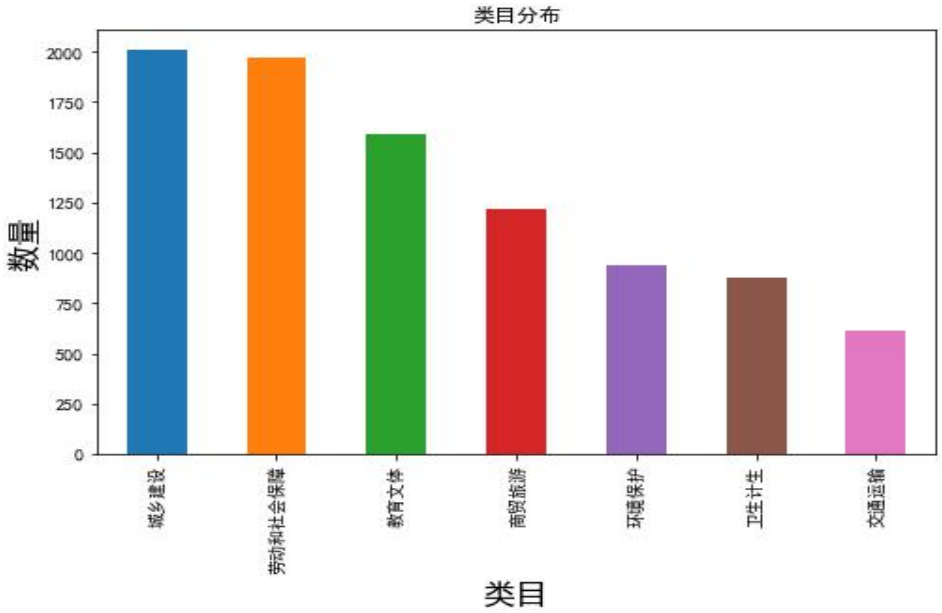

图 11：统计分类柱状图

我们设置 5 个 epochs，batch_size 设置为 25，得到以下的训练结果如下图所示：

```
- 81s 11ms/step - loss: 1.3580 - accuracy: 0.5029 - val_loss: 0.6874 - val_accuracy: 0.8034
- 79s 11ms/step - loss: 0.4059 - accuracy: 0.8849 - val_loss: 0.5174 - val_accuracy: 0.8299
- 79s 11ms/step - loss: 0.1412 - accuracy: 0.9610 - val_loss: 0.5363 - val_accuracy: 0.8372
- 80s 11ms/step - loss: 0.0617 - accuracy: 0.9828 - val_loss: 0.5615 - val_accuracy: 0.8335
- 81s 11ms/step - loss: 0.0331 - accuracy: 0.9929 - val_loss: 0.6679 - val_accuracy: 0.8191
```

图 12：训练过程结果图

从图中结果可以看出，训练的损失值大约在 0.7-0.8 之间，测试集的准确率为 **83.35%**，未出现严重的过拟合现象。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。

下图分别表示损失值和准确率的关系图：

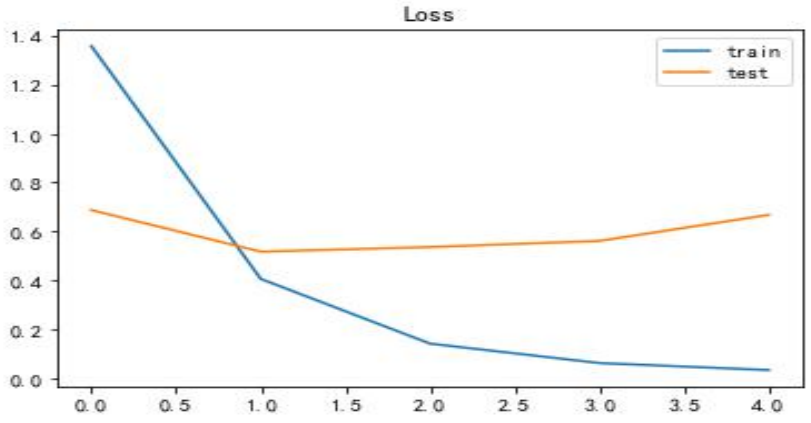

图 13：损失值随网络层数的变换

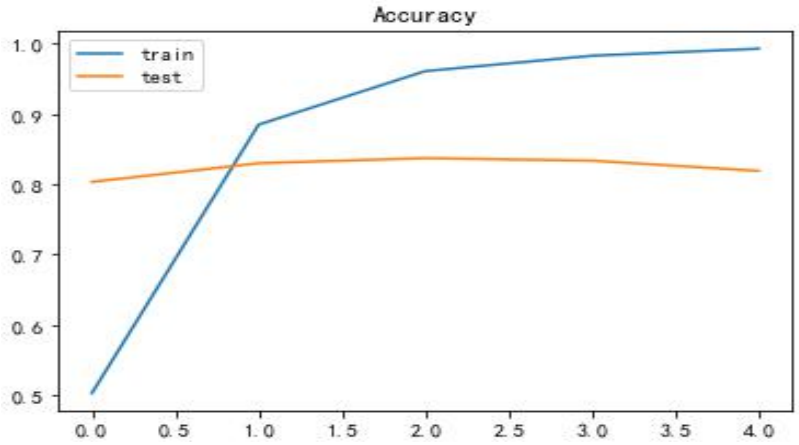

图 14：准确率随网络层数的变换

我们将测试集得到的预测结果用下图来表示，对角线上的元素代表的是预测正确的样本例子，其余为预测错误的情况：

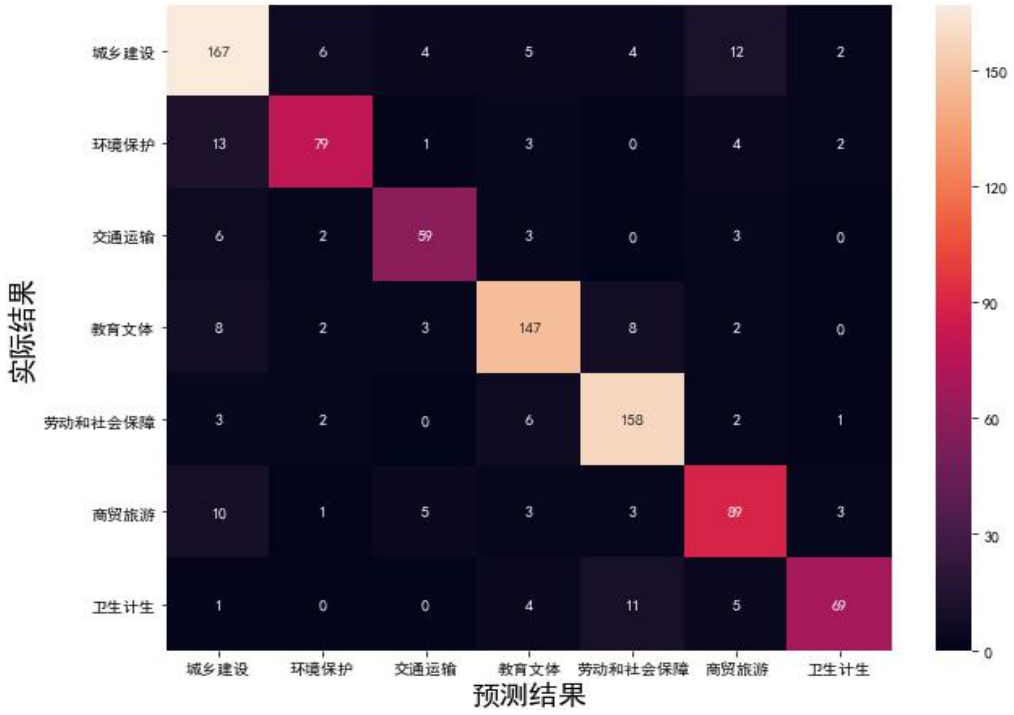

图 15：预测结果直观图

通过表格的数据来展示训练结果：

accuracy 0.8338762214983714

	precision	recall	f1-score	support
城乡建设	0.80	0.83	0.82	200
环境保护	0.86	0.77	0.81	102
交通运输	0.82	0.81	0.81	73
教育文体	0.86	0.86	0.86	170
劳动和社会保障	0.86	0.92	0.89	172
商贸旅游	0.76	0.78	0.77	114
卫生计生	0.90	0.77	0.83	90
micro avg	0.83	0.83	0.83	921
macro avg	0.84	0.82	0.83	921
weighted avg	0.84	0.83	0.83	921

图 16：F1 分类分数表

4.1.3 模型的评价与改进

第一问的模型是对于附件 2 中的留言主题转化为 Word2Vec 词向量进行后续处理，本模型的优点在于用 LSTM 模型预测出来的对应的一级标签准确率较高，为 83.35%，未出现严重的过拟合现象，说明模型建立地较完善。

当没有引入注意力 Attention 模型时，若输入句子比较短的时候问题不大，但是如果输入句子比较长，此时所有语义完全通过一个中间语义向量来表示，单词自身的信息已经消失，可想而知会丢失很多细节信息，所以可以在改进模型时，与注意力 Attention 模型相结合。

4.2 基于问题二的求解

4.2.1 基于 TF-IDF 算法的原理

TF-IDF 算法是一种用于信息检索与数据挖掘的常用加权技术。TF 的意思是词频(Term-frequency),IDF 的意思是逆向文件频率（inverse Document frequency）。

TF-IDF 是传统的统计算法，用于评估一个词在一个文档集中对于某一个文档的重要程度。它与这个词在当前文档中的词频成正比，与文档集中的其他词频成反比。

首先说一下 TF（词频）的计算方法，TF 指的是当前文档的词频，

$$\mathrm{TF}_{\mathrm{i,j}}=\frac{n_{i,j}}{\sum_k n_{k,j}} \tag{15}$$

在这个公式中，分子表示的是该词在某一文档中出现的次数，分母表示在该文档中所有关键词出现的次数之和，即：

$$\text{词频}（TF）=\frac{\text{某个词在文章中的出现次数}}{\text{文章的总词数}} \tag{16}$$

然后来说下 IDF(逆向词频)的计算方法，IDF 指的是某个词汇普遍性的度量，某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到，计算公式如下：

$$\mathrm{IDF}_{\mathrm{i}}=\log\frac{|D|}{|\{j:t_i\in d_j\}+1|} \tag{17}$$

即： 逆文档频率$（IDF）=\log（\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}+1}）$。

其中，$|\mathrm{D}|$表示语料库中的文件总数，$|\mathrm{j}：\mathrm{t_i}\in\mathrm{d_j}|$表示包含词语的文件数目（即 $n_{i,j}\neq 0$ 的文件数目）如果该词语不在语料库中， 就会导致分母为零，因此一般情况下使用$1+|\mathrm{j}：\mathrm{t_i}\in\mathrm{d_j}|$。如果一个词越常见，那么分母就越大，逆向文件频率就越小越接近 0。

4.2.2 基于 TF-IDF 算法的求解过程

（1）数据预处理

根据题目要求，需要将附件 3 中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，于是我们运用第一问建立好的模型，用题目所给的已经分好标签的附件 2 作为训练集，对附件 3 进行预测，给出对应的一级标签，方便下一步操作，结果如下：（只给出部分数据，完整处理结果将由附件上传）

表 4：已贴标签的附件 3（部分数据）

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	一级标签
188006	A000102948	A3区一米阳光婚纱艺术摄影是否合法纳税了？	2019/2/28 11:25:05	为地处居民	0	0	教育文体
188007	A00074795	咨询A6区道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	都未曾更换	0	1	城乡建设
188031	A00040066	反映A7县春华镇金鼎村水泥路、自来水到户的问题	2019/7/19 18:19:54	日天还没黑就	0	1	城乡建设
188039	A00081379	A2区黄兴路步行街大古道巷住户卫生间粪便外排	2019/8/19 11:48:23	清扫。没有解	0	1	城乡建设
188059	A00028571	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	2019/11/22 16:54:42	诉业主，态	0	0	环境保护
188073	A909164	A3区麓谷社区单方面改变麓谷明珠小区6栋架空层使用性质	2019/3/11 11:40:42	府调规、改	0	0	交通运输
188074	A909092	A2区富绿新村房产的性质是什么？	2019/1/31 20:17:32	业主了，然	0	0	城乡建设
188119	A00035029	对A市地铁违规用工问题的质疑	2019/5/27 16:04:44	还扣钱，扣	0	0	城乡建设
188170	A88011323	A市6路公交车随意变道通行	2019/12/23 8:50:24	机并未按地	0	0	城乡建设
188249	A00084085	A3区保利麓谷林语桐梓坡路与麓松路交汇处地铁凌晨2点施工扰民	2019/9/17 4:25:00	居也是苦不	0	0	城乡建设
188251	A00013092	A7县特立路与东四路口晚高峰太堵，建议调整信号灯配时	2019/10/19 11:02:40	少两到三个	0	0	教育文体
188260	A00053484	A3区青青家园小区乐果果零食炒货公共通道摆放空调扰民	2019/5/31 17:06:13	炒货公共通道	0	0	城乡建设
188396	A00047580	关于拆除聚美龙楚在西地省商学院宿舍旁安装变压器的请求	2019/4/15 16:23:09	中小学校校园	2	1	城乡建设
188399	A00097934	A市利保壹号公馆项目夜间噪声扰民	2019/7/3 6:23:25	在施工中，	0	0	环境保护
188409	A0003274	A市地铁3号线星沙大道站地铁出入口设置极不合理！	2019/6/19 10:14:39	、星沙四区、	0	4	城乡建设
188414	A00096844	A4区北辰小区非法住改商问题何时能解决？	2019/8/1 7:20:31	、不做太多	0	0	城乡建设
188416	A00029753	请给K3县乡村医生发卫生室执业许可证	2019/06/20 20:38:47	件下来啊。	0	0	卫生计生
188451	A00013004	A7县春华镇石塘铺村有党员家开麻将馆	2019/4/11 17:54:25	了。但是石	0	2	劳动和社会保障
188455	A00035902	咨询异地办理出国签证的问题	2019/5/16 15:20:43	的户籍所在	0	0	劳动和社会保障
188467	A00050188	投诉A市温斯顿英语培训学校拖延退费！	2019/3/28 19:57:19	推辞的态度	0	1	教育文体

（2）定义合理的热度评价指标

第二问求解的基本思想是先用 TF-IDF 算法，对每一条留言提取关键词，并计算每个词对应的 TF-IDF 权重值，为了将每一条留言里提取出来的关键词对应到同一个问题 ID 号里，我们用 Python 进行相应的编程，并将计算结果和各自的 ID 号写入表格中，结果如下图：（由于数据太大，以下截图均只截取部分数据）

表 5：计算 TF-IDF 值结果图（部分数据）

ID	关键词	权重
1	一米阳光	1.88679
1	A3	1.707824
1	艺术摄影	1.707824
1	婚纱	1.426951
1	纳税	1.178815
1	合法	1.036884
1	是否	0.749203
2	A6	1.086797
2	门牌	1.017512
2	公示	0.792627
2	城乡	0.675725
2	命名	0.630081
2	成果	0.610275
2	咨询	0.607101
2	初步	0.574497
2	道路	0.546713
2	规划	0.541989
2	问题	0.360319
3	水泥路	1.564298
3	A7	1.494346

计算出每个关键词的 TF-IDF 权重值后，将同一个 ID 号里的关键词权重值进行相加，作为对应的留言的 TF-IDF 总权重，这里我们同样运用 Python 进行求和编程，结果如下：

表 6：TF-IDF 权重值求和结果图（部分数据）

ID	TF-IDF权重值求和
1	9.6942904
2	7.443635322
3	9.634284779
4	10.55423937
5	8.28031644
6	8.510151663
7	8.31665804
8	7.274932455
9	8.815263399
10	10.09278042
11	9.221192595
12	9.346689339
13	8.551183078
14	9.323541267
15	9.01807033
16	8.586036456
17	9.124879017
18	11.35822522
19	7.199269052
20	8.517536276

根据附件 3 中所给的因素，我们不仅需要考虑 TF-IDF 权重值的求和结果，还需要考虑用户对留言的反对数和点赞数，需要结合这三种因素制定出第二问的热度评价指标，给不同的因素分配不同的权重，我们认为反对数和点赞数的权重一致，均赋予 10%的权重，最后计算出整体权重值，再用整体权重值进行排序，结果如下表：

表 7：最终热度问题排序图（部分数据）

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	TF-IDF权重值求和	一级标签	关注数	整体权重
208636	A00077171	A市A5区汇金路五矿万境K9县存在一系列问题	2019/8/19 11:34:04	人，请问有	0	2097	9.305727738	城乡建设	2097	219.0057277
223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	配套入学，	5	1762	7.25917437	教育文体	1767	183.9591744
220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	消息总是先	0	821	9.381846084	教育文体	821	91.48184608
217032	A00056543	严惩A市58车贷特大集资诈骗案保护伞	2019/2/25 9:58:37	东、苏纳弟	0	790	10.25630044	商贸旅游	790	89.25630044
194343	A000106161	承办A市58车贷案警官应跟进关注留言	2019/3/1 22:12:30	并没有跟进	0	733	9.262204216	城乡建设	733	82.56220422
263672	A00041448	A4区绿地海外滩小区距长赣高铁最近只有30米不到，合理吗？	2019/9/5 13:06:55	我如下问题	0	669	8.120300039	城乡建设	669	75.02030004
193091	A00097965	A市富绿物业丽发新城强行断业主家水	2019/6/19 23:28:27	地摊上买的	0	242	9.537283222	城乡建设	242	33.73728322
262052	A00072424	关于A6区月亮岛路沿线架设110kv高压线杆的投诉	2019/3/26 14:33:47	令《建设项	0	78	9.367622459	环境保护	78	17.16762246
284571	A00074795	建议西地省尽快外迁京港澳高速城区段至远郊	2019/1/10 15:01:26	浏高速出口	0	80	8.499681288	交通运输	80	16.49968129
226723	A00040222	A市三一大道全线快速化改造何时启动？	2019/9/15 15:31:19	，打通机场	0	66	8.631066225	城乡建设	66	15.23106623
272089	A00061602	关于A6区月亮岛路110kv高压线的建议	2019/4/9 17:10:01	体操学校，	2	55	9.210310931	环境保护	57	14.91031093
203187	A00024716	咨询A9市高铁站选址的问题	2019/8/1 13:48:57	东进的步伐	53	10	8.240118841	交通运输	63	14.54011884
281898	A00096623	A市长房云时代多栋房子现裂缝，质量堪忧	2019/2/25 15:17:38	处理，陷入	5	55	8.348466983	城乡建设	60	14.34846698
200667	A00079480	请问A市为什么要把和包支付作为任务而不让市场正当竞争？	2019/1/16 17:01:25	作者也不理	0	78	6.16681241	劳动和社会保障	78	13.96681241
244178	A00057874	希望A市地铁四号线北延线“同心路站”设在雷峰大道上	2019/1/30 23:59:12	站设在雷峰	0	38	9.60764252	城乡建设	38	13.40764252
202909	A00052058	A市无良万润滨江天著牟取精装暴利	2019/1/3 19:37:36	邵市频道《	0	28	10.56569706	城乡建设	28	13.36569706
279062	A00027836	建议加大A7县东六线榔梨段拆迁力度	2019/1/17 19:25:45	地（正钢机	1	42	8.960998268	城乡建设	43	13.26099827
227371	A00011929	对西地省聚利网的联名控诉书	2019/4/11 21:16:21	地进行沟通	0	16	11.65993636	劳动和社会保障	16	13.25993636
239595	A00057814	建议A市经开区收回东六路恒天九五工厂地块，打造商业综合体	2019/11/8 15:48:07	这里潜力	0	44	8.787037474	城乡建设	44	13.18703747
280425	A00032687	A市经开区泉塘小塘路路灯太暗，建议提质改造	2019/11/8 16:38:44	路灯太暗，	0	29	10.17362391	城乡建设	29	13.07362391
236295	A00053343	A7县广圣大酒店4楼瑞生堂足道涉黄	2019/1/18 15:35:48	88，888，	7	8	11.55729697	城乡建设	15	13.05729697

从中得出排名前 5 的热点问题和每个热点问题对应的具体留言，并按题目中的要求整理出对应的两个表格（分别作为附件上传）。

以下展示排名前 5 的热点问题：

表 8：热点问题表

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1	263.3003507	2019/1/8 至2019/10/25	西地省A市聚利人普惠投资有限公司	58车贷恶性退出立案近半年没有发过一次案情通报
2	2	219.0057277	2019/1/15 至2019/11/11	A市A5区汇金路五矿万境K9县24栋	A市A5区汇金路五矿万境K9县存在一系列问题
3	3	183.9591744	2019/3/12 至2019/12/10	A市	配套入学的问题
4	4	75.02030004	2019/1/30 至2019/9/6	A4区绿地海外滩小区	渝长厦高铁的长赣高铁征地路线对周边小区影响巨大
5	5	33.73728322	2019/1/15 至2019/6/19	A2区富绿新村	建筑楼房物业、环境问题

以及相应热点问题对应的留言信息如下：

表 9：热点问题对应的具体留言表（部分数据）

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	220711	A00031682	请书记关注A市A4区58车贷案	2019/2/21 18:45	消息总是失	821	0
1	217032	A00056543	严惩A市58车贷特大集资诈骗案保护伞	2019/2/25 9:58	卡、苏纳弟弟	790	0
1	194343	A000106161	承办A市58车贷案警官应跟进关注留言	2019/3/1 22:12	并没有跟进	733	0
1	268251	A000106090	西地省58车贷立案近半年毫无进展，单位回复让人心寒	2019/2/2 15:03	管和资产，请	25	0
1	240554	A00029163	A市58车贷老板跑路美国，经侦拖延办案	2019/2/10 20:58	是涉嫌保护伞	6	0
1	272413	A000106062	西地省A市58车贷恶性退出，A4区立案已近半年毫无进展	2019/1/14 20:23	但我们出借	2	0
1	207791	A000106183	西地省聚利人普惠投资有限公司涉嫌诈骗巨额资金	2019/3/9 10:23	提现于8月16	15	0
1	214238	A00061787	请问A4区公安派出所对58车贷一案办案的进度如何了	2019/1/20 22:28	想问一下：请	2	1
1	274796	A00047120	西地省九九富达集团涉嫌诈骗	2019/10/25 8:25	梅国齐福明林	1	0
1	226265	A000106448	恳请A市经侦公正办理58车贷案件，还我们受害人一个公道	2019/5/28 15:08	但这样的经侦	3	0
1	234320	A000106592	不要让A市因为58车贷案件而臭名远扬	2019/7/8 17:16	看还是一如	0	0
1	264119	A00084445	58车贷立案五个月过去，A4区公安分局未公布过任何案情	2019/1/19 9:47	员，未查封资	0	0
1	253735	A00025441	西地省利聚人普惠投资有限公司涉嫌诈骗巨额资金	2019/4/14 22:57	于8月16日处	1	0
1	254532	A000106062	A市58车贷恶性退出立案近半年没有发过一次案情通报	2019/1/14 22:08	在58官网上	3	0
1	223787	A00034861	西地省58车贷案件创造全国典型诈骗案，立案至今无公告	2019/1/11 21:12	产4.任由犯罪	0	0
1	275171	A00034861	西地省展星投资有限公司涉嫌诈骗	2019/1/8 21:34	，用尽各种	0	0
1	263353	A000106808	西地省九九富达实业发展集团涉嫌诈骗	2019/9/4 21:34	到王梅国齐	1	0
1	229554	A00025441	西地省聚利人普惠投资有限公司涉嫌诈骗巨额资金	2019/4/14 22:56	于8月16日处	0	0
1	272858	A00061787	A市58车贷恶性退出案件为什么不发布案情进展通报？	2019/1/16 23:21	立案近半年	0	0
1	218132	A000106090	再次请求过问A市58车贷案件进展情况	2019/1/29 19:15	办案警官毛	0	0
2	208636	A00077171	A市A5区汇金路五矿万境K9县存在一系列问题	2019/8/19 11:34	，请问有人	2097	0

4.2.3 求解过程评价与改进

本小题选择运用 TF-IDF 算法来提取关键词主要是由于 TF-IDF 实现简单，相对容易理解。但是，TF-IDF 算法提取关键词也有缺点，它需要选取质量较高且和所处理文本相符的语料库进行训练，而且不能反应词的位置信息，在对关键词进行提取的时候，词的位置信息，例如文本的标题、文本的首句和尾句等含有较重要的信息，应该赋予较高的权重。

改进方面可以选择在计算 TF-IDF 权重值之前，先运用文本聚类模型，初步根据已有的一级标签为聚类中心，再进行聚类，根据聚类结果疏密程度，分别对前 5 个较密的聚类中心开展 TF-IDF 权重值计算，重新得到排名前 5 的热点问题以及对应的热点问题的具体留言信息。

4.3 基于问题三的求解

4.3.1 相关性评价

（1）评价方案原理：

第三问从三个角度上建立评价方案，第一个角度是相关性。

相关性我们选取用文本词向量之间的余弦相似度来衡量。

在计算余弦相似度之前需要先计算基于加权 Word2Vec 的句向量，具体计算公式如下：

$$sen\_vec = \frac{\sum_{i=1}^{m} vec_i * e^{weight(i)}}{m} \quad (18)$$

其中，sen_vec 表示句向量，m 表示每个样本中的词的个数，vec_i 表示每个词的词向量，weight(i)表示每个词的权重，权重可以根据 TF-IDF，信息增益等方法求得。

而余弦相似度的原理如下：

余弦相似性通过测量两个向量的夹角的余弦值来度量它们之间的相似性。0 度

角的余弦值是 1，而其他任何角度的余弦值都不大于 1；并且其最小值是-1。从而两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。两个向量有相同的指向时，余弦相似度的值为 1；两个向量夹角为 90° 时，余弦相似度的值为 0；两个向量指向完全相反的方向时，余弦相似度的值为-1。这结果是与向量的长度无关的，仅与向量的指向方向相关。余弦相似度通常用于正空间，因此给出的值为 0 到 1 之间。

注意这上下界对任何维度的向量空间中都适用，而且余弦相似性最常用于高维正空间。例如在信息检索中，每个词项被赋予不同的维度，而一个文档由一个向量表示，其各个维度上的值对应于该词项在文档中出现的频率。余弦相似度因此可以给出两篇文档在其主题方面的相似度。

两个向量间的余弦值可以通过使用欧几里得点积公式求出：

$$a \cdot b = |a| \cdot |b| \cos\theta \tag{19}$$

给定两个属性向量和 ，其余相似性由点积和向量长度给出，如下所示：

$$similarity = \cos(\theta) = \frac{A \cdot B}{|A| \cdot |B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \tag{20}$$

这里的 A_i 和 B_i 分别代表向量 A 和 B 的各分量。

而从上式中可以看出，给出的相似性范围是从-1 到 1：

- -1 意味着两个向量指向的方向正好截然相反
- 1 表示它们的指向是完全相同的
- 0 通常表示它们之间是独立的

而在这之间的值则表示中间的相似性或相异性。对于文本匹配，属性向量 A 和 B 通常是文档中的词频向量。余弦相似性，可以被看作是在比较过程中把文件长度正规化的方法。

在信息检索的情况下，由于一个词的频率（TF-IDF 权）不能为负数，所以我们定义留言主题与对应的回复文本的余弦相似性范围为从 0 到 1。

（2）求解过程：

我们首先利用第一问求取词向量的方法计算出所有字词的 word2vec 词向量，我们简单地采用用所有字词词向量的平均值分别表示评论和回复句子的词向量。通过计算各句子向量间的余弦相似度即可得到最终的结果。以其中两个评论及恢复为例，如下：

表 10：相似度评分表（部分数据）

留言编号	留言用户	留言详情	答复意见	相似度评分
2549	A00045581	2019 年 4 月以来，位于 A 市 A2 区桂花坪街道的 A2 区公安分局宿舍区（景蓉华苑)出现了一番乱象，该小区的物业公司美顺物业扬言要退	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2 区景蓉花苑物业管理有问题”的调查	0.978559

		出小区，因为小区水电改造造成物业公司的高昂水电费收取不了(原水电在小区买……	核实情况向该网友答复如下：您好,首先感谢您对我们工作的信任和支持……	
139551	UU0082040	领导您好我们永丰镇绍塘村拆迁户的小孩，到底能不能在临近的湾田小学入学。我们房子拆了这么多年，租住乡里乡亲的房子也不少年头，宅基地迟迟没有交付也就算了，好不容易熬到小孩要上学了……	网友：您好，留言收悉，现回复如下：M2 县城城区义务教育阶段公办学校 2019 年秋季招生范围划分表学校初中 M2 县永丰中学城中路以东城区部分；定源群工站……。	0.797367

如上表留言编号为 2549 的留言及答复可以看出，有关留言提及的“A2 区景蓉花苑物业管理有问题”、“小区停车收费问题”以及“物业公司去留问题”这三个问题答复意见都有做一一有针对性的回复，因此该答复与留言的相似度评分较高。而例如留言编号为 139551 的留言中反应的是拆迁户孩子的就学问题，不仅留言中提及的宅基地问题没有任何的回应，并且留言中已经指出“县教育部门在新生入学划分学校上的表达这么模糊”，而回复仅仅又把表达模糊的范围重复了一遍，既没有地区针对性，也没有提出实际的对应解决问题办法，属于答复与留言相关性不强的回复，因此相似度评分不高。综上，该模型可以帮助我们快速找到回复准确以及不准确的回复。

4.3.2 完整性评价

（1）评价方案原理：

完整性评价，采用 TF-IDF 进行评价。TF-IDF 算法可以用来计算文件中某个词的权重。通过将文件中所有关键词的权重相加，再加上文件中关键词的密度，可以用来表示文件的权重。

完整性可以理解为答复意见是否覆盖了留言的每个内容，如果答复的内容包含了留言的内容，我们就称该答复是一个完整的答复，反之则称为不完整的答复。

针对于每条留言和答复意见，我们采用 TF-IDF 算法进行关键词提取，每条留言和答复意见分别提取主要内容进行比对，计算出二者之间重叠的关键词。为了方便起见，我们定义完整率计算如下：

$$F = \frac{\text{相互匹配的关键词个数}}{\text{答复意见总的关键词个数}} \tag{21}$$

（2）求解过程：

容易看出，完整性的取值范围在 0-1 之间，且与答复留言的完整性成正相关。留言内容和答复意见的提取结果如下图所示：

留言内容		答复内容	
1 物业公司	weight:0.407261	1 业主大会	weight:0.363416
1 小区	weight:0.394934	1 业委会	weight:0.343120
1 投票	weight:0.345175	1 业主	weight:0.206068
1 A2	weight:0.209733	1 胡华衡	weight:0.156271
1 业委会	weight:0.184201	1 A2	weight:0.156271
1 高昂	weight:0.153463	1 区景蓉	weight:0.156271
1 收费	weight:0.130817	1 花苑	weight:0.142548
1 业主大会	weight:0.121936	1 反映	weight:0.141525
1 公安干警	weight:0.112299	1 感谢您	weight:0.134158
1 2019	weight:0.104866	1 物业管理	weight:0.122044
1 景蓉华苑	weight:0.104866	1 物业公司	weight:0.121380
1 美顺	weight:0.104866	1 留言	weight:0.120264
1 0.64	weight:0.104866	1 桂花	weight:0.117395
1 以交	weight:0.104866	1 意见	weight:0.115257
1 20	weight:0.104866	1 来信	weight:0.112780
2 几下	weight:0.323442	2 施工	weight:0.392590
2 一段	weight:0.257834	2 坪塘	weight:0.252740
2 方便	weight:0.257175	2 排水	weight:0.218428
2 2018	weight:0.243975	2 换填	weight:0.217359
2 挖机	weight:0.243975	2 土方	weight:0.180213
2 这路	weight:0.243975	2 管线	weight:0.172841
2 来台	weight:0.219698	2 道路	weight:0.164014
2 围栏	weight:0.215683	2 大道	weight:0.136311
2 很大	weight:0.208162	2 项目	weight:0.133168
2 店面	weight:0.191854	2 A00023583	weight:0.108680
2 修好	weight:0.190117	2 区潇楚	weight:0.108680
2 太长	weight:0.183580	2 洋湖	weight:0.108680
2 动工	weight:0.173521	2 潇楚	weight:0.108680
2 不怎么	weight:0.171334	2 因该	weight:0.108680

图 17：留言与答复结果图

通过 TF-IDF 算法得到的结果，我们计算答复留言的完整率，即可得到对于答复留言完整性的一套评价。例如针对前十条留言内容和答复意见，我们可以计算出其完整率分别为：

留言	完整率
留言1	0.43
留言2	0.36
留言3	0.53
留言4	0.33
留言5	0.16
留言6	0.33
留言7	0.47
留言8	0.36
留言9	0.69
留言10	0.6

图 18：部分留言的完整率

类似地，我们可以计算附件三中所有留言的完整率，分析出有关完整性的一个量化评价。

4.3.3 可解释性评价

可解释性即为答复的内容能够合理全面的解释留言内容，是对于留言内容和答复内容深层次的语义理解。我们采用自定义可解释性来建立评价方案。

可以通过从留言内容和答复情况，建立关键词与答复内容的映射关系。先根据 TF-IDF 算法提取出留言内容中的关键词，通过映射关系，查找到答复情况中的

对应的关键词的信息，先用相似性是否达到一定的指标来判断下一步是否继续进行比较，再用答复是否完整，若完整率高则比较是否具有可解释性。具体我们可以通过先筛选出相似性评分和完整率均较高的答复，然后通过查找计数有几对“《》”等能表示用到了法律条文或者是其他引经据典的符号，并适当的为可解释性添分。以下提供部分数据进行分析：

表 11：可解释性评分表（部分数据）

留言编号	留言用户	留言详情	答复意见	可解释性评分
12189	UU008948	尊敬的易书记：农村环境卫生确实存在很大问题，政府很重视，现在已经开始第二个三年整治行动。但是，我们管理的办法还是人治思路，拿起绩效考核的大棒，一级一级压任务……	网友：您好！留言已收悉	0.0670
4043	UU008187	维护文明交通劝导员合法权益给付经济补偿A市政府自2007年开发设立文明交通劝导员公益性岗位至今十多年来，违反《劳动合同法实施条例》第12条，出台的《A市文明交通劝导员管理办法》损害有关人员合法权益。国家和省财政厅相关文件规定公益性岗位补贴和社保补贴……	……根据《A市公益性岗位管理办法》（长办发[2012]44号）文件第二章第六条规定……《A市政府投资开发的公益性岗位申报表》……《A市企业（单位）招用就业困难人员就业社会保险补贴实施办法》（长劳社发[2009]95号）文件第四章第七条规定……A2区人社局于2012年1月至2014年12月（共36个月)对持有《就业失业登记证》……。	0.97597

通过上述选取的部分数据比较来看，留言编号为12189的用户所获取的答复意见就过于简略，并无可解释性；而留言编号为4043的用户所获取的答复意见就较为详细，并且用了三对“《》”表示引用了三篇法律条文中的内容，并且是基于相似性评分和完整率均较高的前提下，所达到的可解释性高。

4.3.4 最终方案

将上述三个角度生成的各自的评价指标得分后，先用 SPSS 进行无量纲归一化处理，然后再分别进行加权求和，权重均设定为$\frac{1}{3}$，得到的最终分数为该答复意见的质量评分。

4.3.5 该评价方案的优缺点

（1）相关性：

该评价方案，在相关性的角度上选择采用余弦相似度而不是欧氏距离的优异性在于：

余弦相似度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。相比距离度量，余弦相似度更加注重两个向量在方向上的差异，而非距离或长度上。与欧几里德距离类似，基于余弦相似度的计算方法也是把用户的喜好作为 n-维坐标系中的一个点，通过连接这个点与坐标系的原点构成一条直线（向量），两个用户之间的相似度值就是两条直线（向量）间夹角的余弦值。因为连接代表用户评分的点与原点的直线都会相交于原点，夹角越小代表两个用户越相似，夹角越大代表两个用户的相似度越小。同时在三角系数中，角的余弦值是在[-1, 1]之间的，0 度角的余弦值是 1，180 角的余弦值是-1。

借助三维坐标系来看下欧氏距离和余弦相似度的区别：

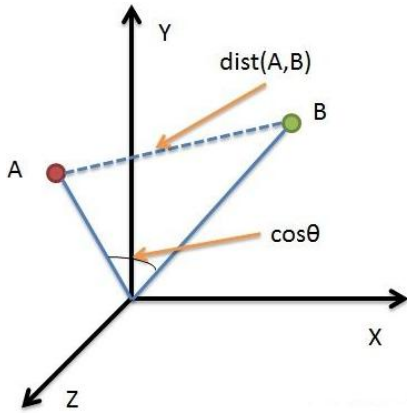

图 19：欧氏距离和余弦相似度示意图

从图上可以看出距离度量衡量的是空间各点间的绝对距离，跟各个点所在的位置坐标（即个体特征维度的数值）直接相关；而余弦相似度衡量的是空间向量的夹角，更加的是体现在方向上的差异，而不是位置。如果保持 A 点的位置不变，B 点朝原方向远离坐标轴原点，那么这个时候余弦相似度 cos θ 是保持不变的，因为夹角不变，而 A、B 两点的距离显然在发生改变，这就是欧氏距离和余弦相似度的不同之处。

根据欧氏距离和余弦相似度各自的计算方式和衡量特征，分别适用于不同的数据分析模型：欧氏距离能够体现个体数值特征的绝对差异，所以更多的用于需要从维度的数值大小中体现差异的分析，如使用用户行为指标分析用户价值的相似度或差异；而余弦相似度更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分用户兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题，因为余弦相似度对绝对数值不敏感。

所以我们采取用余弦相似度来衡量留言主题与对应回复的相关性。

但若要更精准的判断其之间的相关性，应该对余弦相似度做出相应的调整。

在余弦相似度的介绍中说到：余弦相似度更多的是从方向上区分差异，而对绝对的数值不敏感。因此没法衡量每个维数值的差异，会导致这样一个情况：比如用户对内容评分，5 分制，X 和 Y 两个用户对两个内容的评分分别为(1,2)和(4,5)，使用余弦相似度得出的结果是 0.98，两者极为相似，但从评分上看 X 似乎不喜欢这个内容，而 Y 比较喜欢，余弦相似度对数值的不敏感导致了结果的误差，需要修正这种不合理性，就出现了调整余弦相似度，即所有维度上的数值都减去一个均值，比如 X 和 Y 的评分均值都是 3，那么调整后为(-2,-1)和(1,2)，再用余弦相似度计算，得到-0.8，相似度为负值并且差异不小，但显然更加符合现实。

（2）完整性：

优点：完整性的评价方案符合了我们对于完整性的定义，即完整率表征答复内容关键词与留言内容关键词重合的情况。这样的量化过程直观易懂，具有代表性。TF-IDF 提取出每个留言内容和答复内容的关键词，概括了这两部分的主题，抓住了主要矛盾去解决实际问题。

缺点：由于 TF-IDF 提取的是关键词，没有完充分体现词与词之间的联系，有时在比对的时候容易把近义词，官方用语和口头用语分开来判断，造成了完整率总体偏小。另一方面，TF-IDF 提取的是整个内容的关键词部分，其全部的含义并不完全能够关键词表示，在计算完整率之前语句信息有一定的丢失。

（3）可解释性：

优点：用 TF-IDF 可以概括留言的关键内容，建立在关键词上的映射关系，可以针对留言反映的每个问题进行相应的解释，具有全面性。可解释性定义为能够从答复内容的集合里找到对应关键词占所有关键词的比重，抓住关键词解释的主要矛盾，强调答复对于关键词的一一对应的关系，并且直接从寻找书名号等有关引经据典的符号上入手能较快速和直接地反映可解释性。

缺点：模型只针对答复对于留言的问题覆盖是否具有可解释性，没有考虑答复本身提出的解决方法是否具有可实现性。

5 参考文献

[1] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. InProceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies

[2]文峤. 基于文本特征提取方法的文本分类研究[J]. 电脑知识与技术，2018，14（18）：188-189，192.

[3] 张雪松，贾彩燕. 一种基于频繁词集表示的新文本聚类方法 [J] . 计算机研究与发展，2018，55（1）：102-112.

[4] MIKOLOV T，SUTSKEVE R I，CHEN K，et al. Distributed representations of words and phrases and their compositionality [J] . Advances in Neural Information Processing Systems，2013，10（26）：3111-3119.

[5] 周顺先，蒋励，林霜巧，等.基于 Word2vector 的文本特征化表示方法 [J] . 重庆邮电大学学报：自然科学版，2018，30（2）：272-279.

[6]贺益侗. 基于 doc2vec 和 TF-IDF 的相似文本识别[J].电子制作，2018(18)：37-39.

［7］黄承慧，印鉴，侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法［J］. 计算机学报，2011，34（5）：856-864.
［8］高明霞 ，李经纬.基于 word2vec 词模型的中文短文本分类方法［J］. 山东大学学报：工学版，2018，11（2）：159-163.
［9］汪静，罗浪，王德强. 基于 Word2vec 的中文短文本分类问题研究［J］.计算机系统应用，2018，27（5）：209-215.
［10］https://blog.csdn.net/yu444/article/details/86764352
［11］http://www.ruanyifeng.com/blog/2013/03/cosine_similarity.html
［12］徐维林，朱宗，高丽，等.基于主题模型的网络微博舆情分析［J］.软件导刊，2016，15（5）：153-154.