“智慧政务”中的文本挖掘应用

摘要

为了解决近几年网络问政平台大力发展，留言文本数据量过大且需要人工划分和热点整理的问题，同时为了提高政府管理水平和施政效率对留言回复进行评价，本文首先使用 jieba 进行文本分析，再去掉文本停用词，最后使用 word2vec 进行处理。

针对留言分类问题使用 SVM，KNN 等构建分类模型，对比实验结果，基于 SVM 的分类模型 F1 值达到 0.931，可以很好的提高分类效率以及正确性，因此使用 SVM 进行模型构建。

针对热点问题使用词性标注及 TF-IDF 进行词语筛选，再用 DBSCAN 算法对留言内容聚类，将构建的热度指标应用于聚类结果，得到热点问题排名。

针对答复意见质量评估问题，本文使用基于 TF-IDF 的相关性模型求得相关度，然后对答复时间和留言时间作时间差运算进而求出答复及时性，在对相关度和及时性数据进行一定的标准化处理后，对相关度赋权 0.7，及时性赋权 0.3 从而计算出最终得分，得分越高答复质量越高。

实验结果得到分值较高的 F1-score，且正确率等方面也达到了较高的分值，最后的热点问题分析和答复意见质量评估也通过论证具有较强的泛化能力。

关键词：SVM，DBSCAN，TF-IDF,文本挖掘

Application of text mining in "intelligent government affairs"

Abstract

In order to solve the network asked ZhengPing stage developing in recent years, text messages and need large amount of data of artificial division and hotspot problem, at the same time in order to improve the government management level and efficiency of government to evaluate a message reply, this paper USES the jieba for text analysis, and then remove the text stop words, finally using word2vec processing, to leave a message using the SVM classification problems, such as KNN constructing classification model, comparing the experimental results, a classification model based on SVM F1 value reached 0.931, can well improve the classification efficiency and correctness, so using the SVM model building.In terms of hot issues, part of speech tagging and tf-idf are used for word selection. Then, DBSCAN algorithm is used to cluster the message content, and the heat index constructed is applied to the clustering results to obtain the ranking of hot issues.According to reply opinion quality evaluation problem, this article USES the correlation, the correlation model based on TF - IDF obtained and then the response time and message time lag operation, in turn, and reply you timely, for the standardization of relevance and timeliness of data must be processed, empowerment of relatedness 0.7, timeliness empowerment 0.3 to calculate the final score, score, the higher the quality of response.The experimental results obtained a higher f1-score, and the accuracy rate and other aspects also reached a higher score. The final analysis of hot issues and the quality assessment of response comments also demonstrated a strong generalization ability.

Keywords: SVM, DBSCAN, TF-IDF, Text Mining

目录

一、简介

1.1挖掘的意义

随着网络的普及我国网民的数量急剧增多，而政府为了更好的了解民意，服务民众，也开通了包括微信，微博，市长信箱，阳光热线在内的网络问政平台，民众可以足不出户提出对政府工作的建议，这种方式可以很好的了解基层群众的意见、增加民众对政府的认知、改善政府的工作方式，更好的汇聚民意。但是因为对于政府的建议没有统一的格式，不同的人有不同的方式进行建议，文本数量急剧攀升，而传统的依靠人工进行划分和热点整理的工作已经不能处理大量的文本，因此根据现有的技术，结合云计算，大数据，机器学习的方式进行留言分类处理，热度筛选处理已经是大趋势，可以有效提升政府的处理速度，同时可以推动政府更好的进行政策实施。根据热度筛选可以帮助政府筛选出某个具体的时间段民众最需要解决的问题，有助于提升政府形象，落实政府速度施政的政策。

为了更好的解决文本分类问题，构建自然语言处理的模型，我们借助机器学习的方法进行自然语言处理，目标是使机器可以根据对留言各个特征的理解，正确分类留言的类型，提高机器对语言的理解能力。机器学习对于文本分类，以及热度筛选的自然语言处理研究任务有积极的作用。因此以文本分类，文本挖掘，热度筛选为基础研究机器学习语言的技术有重要的研究跟应用价值。

1.2挖掘目标

针对问题一，我们要构建一个文本分类的挖掘模型，模型可以辅助人工进行文本分类，提高工作人员的分类效。具体到使用场景上，对于一条留言信息，模型根据留言主题、留言内容根据政务划分体系进行留言分类，输出一级留言标签。形式化说，我们的分类模型构建可以看成一个二元组<M，C>，二元组由留言信息 D 和分类 C 组成，根据输入的留言信息 D，输出 C 分类结果。

针对问题二，我们构建的是基于 DBSCAN 的留言聚类模型,模型可将大量用户留言归类处理,以便于解决同类问题,除此之外,再留言聚类之后为其定义合理的热度指标，对其热度值进行排名,便可挖掘出用户关心的热点问题,有利于政府更好的为人民服务。

针对问题三，我们需要构建一个评分系统对答复意见的质量进行评估，且该评分系统的评分需要对数据进行较为全面的分析,即可以从答复的相关性,完整性,可解释性，及时性等方面进行评估。除此之外，评分系统模型需要有较强的泛化能力。

1.3挖掘流程

针对问题一的构建文本分类模型，挖掘主要分为 2 部分，预处理部分、模型构建部分。其中预处理包括分词、去停用词、将词组向量化（word2vec），模型构建部分为核心步骤，为了得到准确的分类结果，将预处理得到的词向量使用支持向量机（SVM）、梯度提升（xgboost）等机器学习算法进行训练，最后根据训练的模型进行分类，输出分类效率。问题一挖掘流程图如图所示：

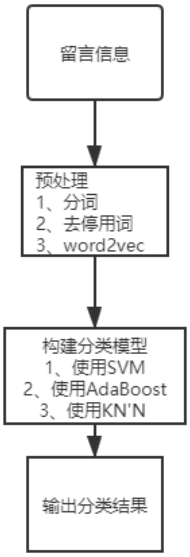

图 1 问题一挖掘流程

针对问题二，构建的 DBSCAN 留言聚类模型，仍是主要分为预处理部分和模型构建部分。预处理包括分词、去停用词、特征提取（TF-IDF）、降维等，模型构建部分将聚类结果按照定义的热度指标打分并排序，最终得到留言热点明细表，从中发现热点问题。

针对问题三，首先构建基于 TF-IDF 的文本相关性模型，对数据进行预处理后构建词库，之后计算两句子之间的相关性得出相关度，其次做时间差运算求出时间差作为及时性评判数据，最后对相关性和及时性进行标准化并且划分权重后计算评估得分。

二、 预处理

2.1 分词

分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。我们知道，在英文的行文中，单词之间是以空格作为自然分界符的，而中文只是字、句和段能通过明显的分界符来简单划界，唯独词没有一个形式上的分界符。因此中文文本词语之间没有明显的界限，且同一个词语可以有不同的表达方式跟理解，在处理中文文本的时候要进行分词，本文采用的是 Python 开发中的中文分词模块—jieba 分词，对留言内容以及留言主题中的每句话进行分词。

jieba 分词主要是基于统计词典，构造一个前缀词典；然后利用前缀词典对输入句子进行切分，得到所有的切分可能，根据切分位置，构造一个有向无环图；通过动态规划算法，计算得到最大概率路径，最终得到了切分的形式。其中最短路径匹配算法原理是，利用词典找到字符串中所有可能的词条，然后构造一个有向无环图，其中每个词条对应图中的一条有向边，并可利用统计的方法赋予对应的边长一个权值，然后找到从起点到最终点的最短路径，该路径上所包含的词条就是该句子的切分结果。

2.2 去停用词

在留言文本中有很多语气词，或者是没有实际含义的词，我们通常用停用词来表示这些词语，这些词语通常是一些语气词，单字，单字母以及高频的词语，比如中文中的“我，的，啊，吗，了，是”等，英文中常见的停用词包括“the、this、I、a、of”等，这些停用词不会对分类结果产生积极的影响，甚至可能干扰分类结果，因此在预处理阶段要删除，避免对文本分类模型造成不好的影响。本文使用的停用词表是，四川大学机器智能实验室停用词表。

2.3 word2vec

Word2vec，是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络，是用来训练完成重新建构语言的词文本的目标。在 word2vec 中词袋模型假设下，词的顺序是不重要的。训练完成之后，word2vec 模型可用来映射每个词到一个向量，可用来表示词对词之间的关系，该向量为神经网络的隐藏层。

为了构建文本处理模型，我们首先要将自然语言转变成计算机能够理解的形式。传统的独热编码（One-hot）形式，是把每个次表示成一个很长的向量，这个向量的维度是

词表的大小，其中绝大多数元素为 0，只有一个维度的值是 1，这个维度就代表了当前的词。这种方式有很大的不足，独热编码的维数根据词典的长度而定，存在着维度过多，过于稀疏，降维难的问题，计算很不方便。独热编码下的任意俩词都是孤立的，丢失了语言中语义关系。而 Word2vec 是 Google 在 2013 年提出的快速有效训练词向量的模型。作者的目标是要从海量的文档数据中学习高质量的词向量， 该词向量在语义和句法上都有很好地表现， 已经广泛应用于自然语言处理的各种任务中。Word2vec 包含了两种训练模型， 分别是 CBOW 和 Skip_gram 模型， 如图 2 所示。CBOW 模型利用上下文预测当前词， 而 Skip_gram 模型利用当前词预测其上下文。

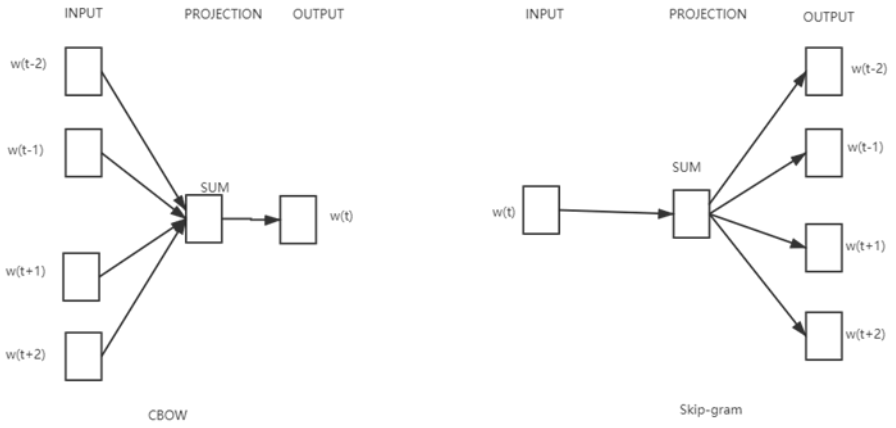

图 2 两种 word2vec 算法示意图

我们使用 Skip_gram 进行分类模型构建。即得到较好数据结果的前提是使下值最大：

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c} log_p(w_{t+j}|w_t)$$

其中，c 是窗口的大小，T 是训练文本的大小。基于的 Skip_gram 模型 如下式

$$\mathrm{p}(w_o|w_I)=\frac{\exp({v'_{w_o}}^T v_{w1})}{\sum_{w=1}^{W}\exp({v'_w}^T v_{w1})}$$

其中$v_w v'_w$是单词 w 的输入和输出向量，W 是词典的大小。

2.4 TF-IDF

为了标识一个文本，需要获取文本的特征。一般有三种常见的提取文本特征有方法：词频方法、文档频数法和 TF-IDF。在本文中，我们选用 TF-IDF 进行特征提取。在 TF-IDF 中，单词的重要性由两个因素共同决定，它与它在文档中出现的次数成正比，但它随着语料库中出现该词的频率越多而下降。

在某一文档中，词频（TF）是指词在文档中出现的次数。TF 的值往往偏向于词汇量较大的文件，即长文件。（如果用词频来决定一个词是否重要，那么长文本中的单词相同的频率往往会比短文本中的频率更高）。

词的频率的计算可由此公式算得：

$$词频(TF) = \frac{词在文档中出现的次数}{文档的总词数}$$

TF-IDF 的另一指标逆文档频率（IDF）是衡量单词总体重要性的指标，其值等于文档总数除以包含该单词的文档数量的商再取其商的对数。

词的逆文档频率的计算可由此公式算得：

$$逆文档频率(IDF) = \log\left(\frac{语料库的文档总数}{包含该词的文档数}\right)$$

为了避免某词可能从来都没出现在所有的文档中，而导致被除数为零，一般分母用（包含该词的文档数+1）代替。

最后可以计算的出每个词的 TF-IDF 值为：

$$TF - IDF = TF * IDF$$

某词在某文档中是高词频，而在整个文档集中，该词又是低文档频数，那么该词可以得到一个较高权重的 TF-IDF 值。TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。因此，TF-IDF 有助于降低常见的词语特征。

通过 TF-IDF 特征提取之后，在所有的留言内容中，每一条都有一个向量标识。向量上的每一个值都是一个词的 TF-IDF 值。其向量获得方式为首先统计出所有的词，把每个词当成向量的每一个维度，如果该文档中有某词，就在某词的维度上计算它的 TF-IDF 值；如果不存在某词，那么某词的维度上的值就为 0。用这种方式对所有的留言进行特征提取，提取的结果是一个稀疏矩阵。

三、 模型构建

针对问题一，文章使用了 KNN、SVM、AdaBoost 三种算法来进行分类,，通过对比能够看出不同算法在留言信息数据集上的分类效果。

针对问题二，本文使用词性标注及 TF-IDF 进行词语筛选，再用 DBSCAN 算法对留言内容聚类，将构建的热度指标应用于聚类结果，得到热点问题排名。

针对问题三，我们使用 TF-IDF 算法以及时间差运算对相关性和及时性做出得分，经过标准化后划分权重最后得出评分结果，从而达到对答复意见的质量评估。

3.1 留言信息分类模型

3.1.1 分类算法

KNN 算法是根据样本周围的一个或者多个邻近样本来确定自身所属的类别的，核心思想就是如果一个样本点在特征空间中的 K 个最相邻的样本大多数属于某一个类别，那么该样本点就是属于这个类别。KNN 算法适用于类域交叉或者重叠较多的待分类的样本集。

支持向量机 (SVM) 通过将二维线性不可分的样本空间映射到多维空间中，使其在特征空间中变得可分。支持向量机在小样本、线性不可分、高维模式识别上表现出优势，SVM 算法流程图如图 3 所示。

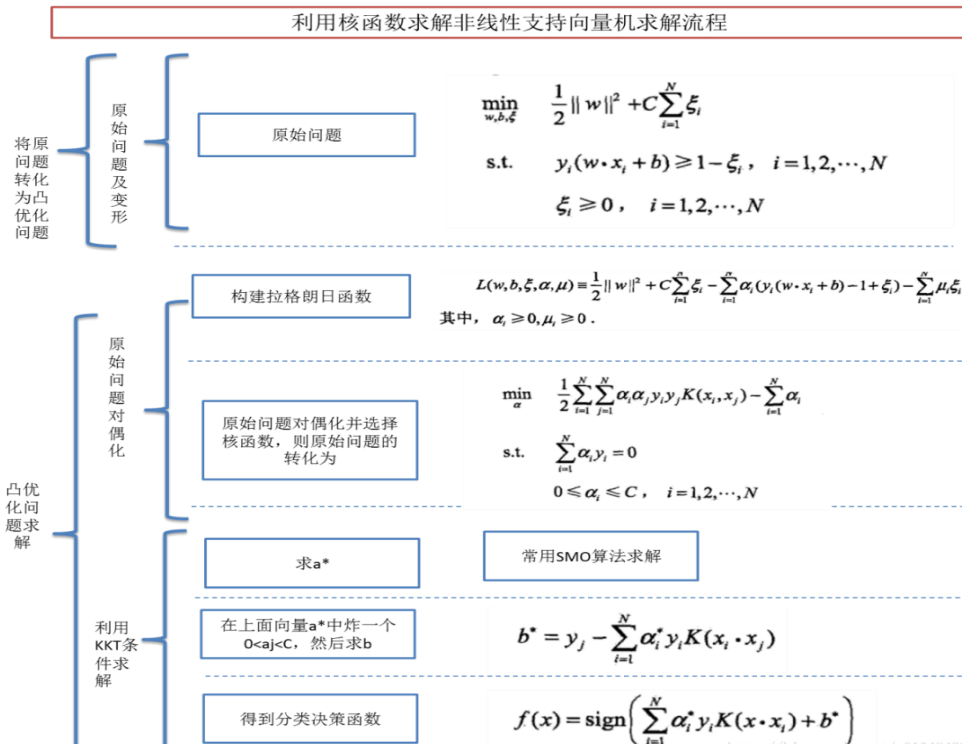

图 3：SVM 算法流程

AdaBoostM1 可多次迭代训练弱分类器，把错分类样本数量权值进行调整，提高分类模型的泛化性能。每一次训练好的弱分类器将会参与下一次迭代训练。根据上一次迭代结果增大误分类为多数类的样本点在训练集中所占权值，同时把正确分类样本点的权值减少，并进入下一次迭代，可以有效的提高分类器的分类性能。下一次迭代产生的分类器更加关注上一个分类器分类错误的样本，增加了样本分类的正确率，最后根据每次迭代产生的分类器进行投票决定分类结果。每次迭代产生的分类器根据分类错误率来计算最后组成强分类器时所占比重。分类错误率越低，权重越高。因此使用 AdaBoostM1 可以有效降低分类错误率，提高少数类别的分类准确率，使模型的泛化性能更好。

随机森林作为弱分类器，对于一个输入样本会产生多个分类结果，最终分类结果由随机森林中每个树投票产生。随机森林因为它是随机选取特征子集，因此减少了维度过多带来的影响，使模型训练的效果更好，增加鲁棒性。AdaBoostM1 框架如图 4 所示。

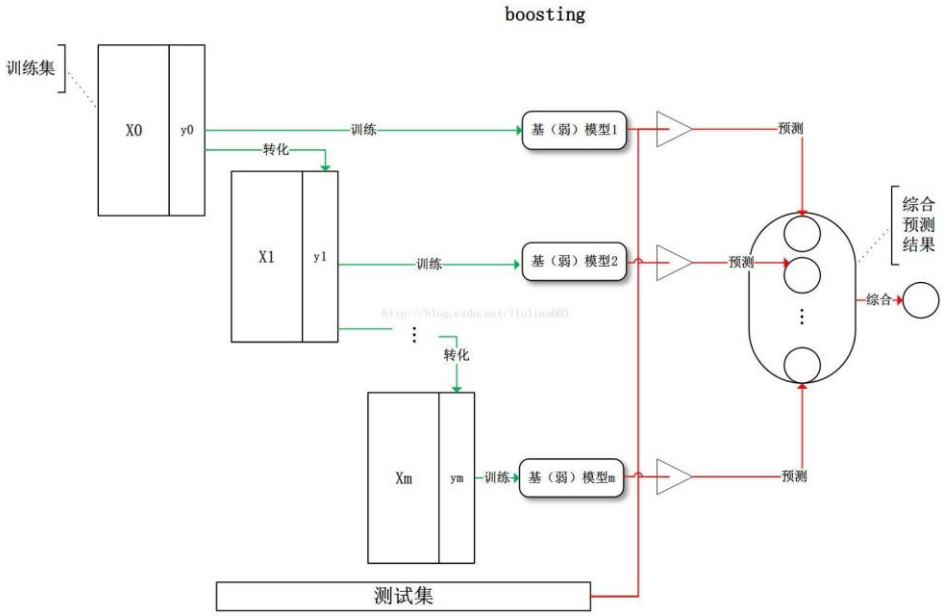

图 4　AdaBoostM1 框架流程图

3.1.2 模型性能评价准则

不同的分类模型有不同的实验结果，本文通过计算不同分类模型的精度、查准率、查全率和 F1 值，来判断分类模型在该数据集上的分类效果。主要依据 F1 的值进行模型好坏的判断。

本文使用计算不同分类模型的混淆矩阵，精度、查准率、查全率和 F1 值，来判断 分类模型在该数据集上的分类效果。混淆矩阵也叫做误差矩阵，用来比较分类结果跟实　际结果，是可以可视化描绘出分类器性能的指标。

漏报率、准确率计算公式：

漏报率=把某类攻击样本判断为正常样本的数量（FP）/该类攻击样本总数

准确率=所有预测类别正确的样本数量(TP+TN)/所有类别总的样本数量

ROC 曲线横轴是伪阳性率 FPR，纵轴真阳性率 TPR。

$$\mathrm{TPR} = \frac{TP}{(TP+FN)} \qquad (2)$$

真阳性率或真正性率表示模型把正常样本预测为正常样本的数量与所有预测为正　常样本数量的比值。

$$\mathrm{FRP} = \frac{FP}{(FP+TN)} \qquad (3)$$

伪阳性率或假正性率 FPR 表示模型把正常样本预测为攻击类型的数量与所有预测　为攻击类型的样本的数量的比值。

ROC 曲线通常用来表示模型分类器的效果，在最佳状态下，ROC 应该在左上角，这 表示在较低假阳率的情况下有高真阳性率。

F1 的公式如下如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{4}$$

3.2 热点问题分析提取模型

3.2.1 基于 DBSCAN 聚类算法的留言内容聚类

通常情况下，绝大多数基于分层或划分的聚类算法都是通过一般的距离方法来计算两个对象之间的距离，并由此形成类簇，这类方法优点就是方法简单易于运用，但是它们只能发现有规则的球形聚类，而无法发掘其他形状的聚类，这也是这类方法的一个局限性。其次就是，基于划分的聚类算法在面对数据集中的噪声数据处理效果较差，而基于密度的聚类算法在对噪声点应对效果较好以及对噪声数据不敏感。

目前，很多新提出的聚类方法都结合了密度聚类方法的思想以吸纳其部分优点，密度聚类算法通常是衡量数据集中的相邻对象构成簇的密度，当数据密度超过了事先给定的阈值，则可以判定这些数据是在一个类簇之中，当然一般情况下都会给类簇设定一个最少包含对象数据的阈值，只有当类簇满足类簇密度以及最少包含对象数这两个阈值时，其才能被确定为一个类簇。正是因为如上特点，基于密度的聚类方法较其他密度聚类算法具有对噪声数据不敏感和可以发现任意形状的类簇的优势。在密度聚类方法中较为著名要属 DBSCAN 算法以及 OPTICS 算法。

但是密度聚类算法也有不少缺点，首先，当数据分布比较稀疏离散时，其聚类效果会比较差。其次，当数据量比较大时内存等相关硬件消耗过大，最后则是聚类最少包含对象数(Minpts)以及扫描半径(Eps)，这两个输入参数选择是否恰当关系到聚类的最终质量。

DBSCAN 聚类过程如下图所示：

```
(1) 标记所有对象为unvisited;
(2) do
(3)     随机选择一个unvisited对象p;
(4)     标记p为visited;
(5)     if p的ε-邻域至少有MinPts个对象
(6)         创建一个新簇C，并把p添加到C;
(7)         令N为p的ε-邻域中的对象的集合;
(8)         for N中每个点p1
(9)             if p1是unvisited
(10)                标记p1为visited;
(11)                if p1的ε-邻域至少有MinPts个点，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unvisited的对象;
```

图 5 DBSCAN 聚类过程

可见，聚类通过获取密度核心由往较高的地方延展将相近的内容合并为同一个簇。在聚类过程中需要输入三参数：

1）D：一个包含 n 个对象的数据集

2） ε：半径参数

3）MinPts：邻域密度阈值

在计算过程中，使用“余弦相似度”来计算距离，所以本文在设置 ε 参数时，一般都设置为 0.3-0.5 之间，因为超过 0.5 的半径值会使不属于同类留言聚在一起，小于 0.3 则无法识别相同事件的留言。设置 MinPts 参数时，可以人为的假定一个阈值。

3.3 答复质量评估模型

3.3.1 基于 TF-IDF 的相关性分析

正如该词所示，TF-IDF 代表词频-逆文档频率，用于计算在文档语料库使用查询中哪些词可能更受偏好。TF-IDF 计算每个单词的权值出现在文档中的百分比。具有 TF-IDF 得分更高的单词，意味着与它们出现的文档有很紧密的关系，也表示如果一个句子出现了一些词组，另一个与之对应的句子若含有这些词组越多，则它们相关性也就越高。在问题三中，我们提取出“留言详情”和“答复意见”后，计算出“留言详情”和“答复意见”的 TF-IDF 值并且结合余弦定理，我们就可以计算出“留言详情”和“答复意见”之间的相似度了。

首先，把“留言详情”进行分词处理之后，将所有的句子通过 TF-IDF 化为向量空间模型，做以下符号定义：文章 D 中出现所有词语的集合标记为 W = (w1,w2 …wM)。通过

TFIDF 算法,可以得到包含句子 d 中每个词语 TF-IDF 值的向量,记做 t = (t1,t2 …tM),其中 t1 表示 w1 在 d 中的 TF-IDF 值。

接下来可以将 S1 ：“留言详情”，S2 ：“答复意见”两者表示为 TFIDF 值的向量：

S1=(t11,t12…t1n)

S2=(t21,t22…t2n)

之后利用余弦定理计算相关性：

$$\cos\theta = \frac{\sum_{i=1}^{n} t_{1i} * t_{2i}}{\sqrt{\sum_{i=1}^{n}(t_{i1})^2} * \sqrt{\sum_{i=1}^{n}(t_{2i})^2}}$$

当余弦值越接近 0 时越无关，当余弦值越接近 1 时越相关。

3.3.2 及时性

当一个留言出现时，答复者回复留言的及时性往往是对留言者极其重要的，而及时性也体现了答复者的工作质量从而影响答复质量。

对于问题 3 所给的数据，我们将留言时间和答复时间提取出来，将答复时间减去留言时间，我们可以得到答复时间和留言时间的时间差，时间差以天为单位。

3.3.3 完整性

对于一个留言，答复意见的完整性是对答复质量评估的一个有效因素，完整的答复模式应该包含一些固定的模板如：敬语，说明情况，解决方式，给予承诺等。对于问题三中所给的数据，在计算出相关性和及时性及得出最后得分的时候后，我们可以利用完整性进行模型的检验，得分高的答复意见完整性是否良好，得分低的答复意见完整性又怎么样，完整性可以作为一个检验项。

3.3.4 评分模型

在得到的相关性和及时性数据中，相关性数据在 0—1 区间上上的数据且数据较为集中所以我们队相关性数据进行归一化处理:

x’=(x-min)/(max-min)

而由于及时性数据区间跨度大，受极值影响大，所以对时间差进行标准化处理:

设均值为 a，标准差为 b，则：

y’=（y-a）/b

对所得数据进行规范处理后，我们对相关性和及时性进行对比予以权重，由于相关性是通过留言详情和答复意见联系得来，所以相关性是评估答复意见质量的根本，由此我们决定赋予相关性 0.7 的权重，赋予及时性 0.3 的权重。因为及时性的值是答复时间和留言时间的差，及时性的值越小，答复质量越高，得分越高，所以最后得出评价模型公式：

score=0.7x-0.3y

四、验证与评估结构分析

本节中我们将对上文中我们建立的模型进行实验验证，通过与传统方法以及其他深度学习方法进行比较分析，从而说明我们模型的合理性与有效性，同时介绍与分模型的最优参数调整过程与各类参数对模型性能的具体影响。

本文训练集与验证集划分的方法是 10 折交叉验证法，把数据集中的数据均等分为 10 份，选取一份做验证集，其余九份做训练，依次迭代 10 次。最后使用 10 个模型的平均实验结果作为整个模型的结果，使用测试集进行测试，可以很好的提升实验准确率。

表 1：不同模型的实验结果

分类器	TPR	FPR	F1	ROC
AdaboostM1	0.902	0.024	0.910	0.930
KNN	0.622	0.101	0.639	0.721
SVM	0.931	0.013	0.931	0.987

通过表 1 可以看到，SVM 的 F1 值最好，因此选用 SVM 构建模型。

SVM 模型得到的混淆矩阵如下图 5 所示：

```
=== Confusion Matrix ===

    a    b    c    d    e    f    g   <-- classified as
 1919   49   29   10    2    0    0 |    a = 城乡建设
   73  850   11    4    0    0    0 |    b = 环境保护
   54   12  530   13    4    0    0 |    c = 交通运输
   22    7   17 1495   45    3    0 |    d = 教育文体
    1    1    6   37 1882   16   26 |    e = 劳动和社会保障
    0    4    3   11   39 1122   36 |    f = 商贸旅游
    0    2    0    1   32   63  779 |    g = 卫生计生
```

图 5：SVM 模型下的混淆矩阵

针对问题二构建的聚类模型得到的聚类结果如下图所示，其中白色表示离群点，即认为该留言构不成热点问题。

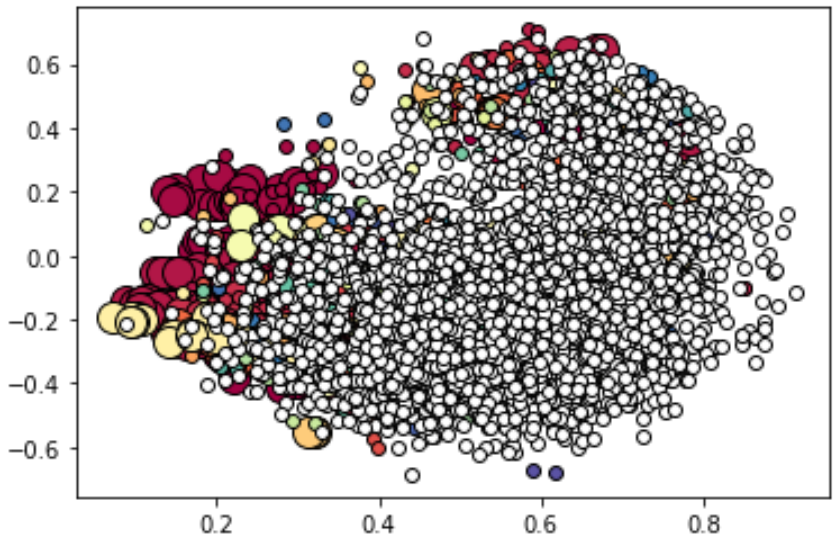

图 6 聚类结果

通过综合考虑多方面因素得到了如下所示的热度指标：

$$hot_{degree} = \frac{cluster\_num}{sum(cluster\_num)} * \omega_1 + \frac{agree_{num}}{sum(agree_{num})} * \omega_2 + \frac{duration}{3} * \omega_3$$

其中：统计出的每类中留言的数目，记为 cluster_num，其权重为 ω1；每条留言均有赞同和反对选项，此处我们用赞同数减去反对数得到净赞同数，记为 agree_num，其权重为 ω2；从每类留言中分别找出最早留言日期和最晚留言日期，二者相减得到此类问题的持续时间，记为 duration，其权重为 ω3，当持续时间在 30 天以内，duration=1；当持续时间在 30-180 天范围内，duration=2；当持续时间在 180 天以上，duration=3。

在综合考虑三者对热度的影响后，取ω_1=0.4，ω_2=0.4，ω_3=0.2.

根据上述定义的热度指标计算出热度指数，根据热度指数对留言类别进行排名，再从每一类选出有代表性的问题作为该类的热点问题。综合来看，选出的前五名热点问题如下表所示：

表 2：留言热点表

热度排名	问题 ID	热度指数	时间范围	地点\人群	问题描述
1	217032	0.243375	2019/1/11 至 2019/7/8	A 市 58 车贷案件嫌疑人	严惩 A 市 58 车贷特大集资诈骗案保护伞

2	360112	0.234098	2018/5/17 至 2019/9/8	A 市经济学院	A 市经济学院强制学生实习
3	280966	0.233741	2019/1/10 至 2020/1/6	A 市王陵国家考古公园	请加快 A 市王陵国家考古公园建设
4	229554	0.229644	2019/2/10 至 2020/1/6	西地省聚利人普惠投资有限公司	聚利人普惠公司涉嫌诈骗巨额资金
5	202909	0.22885	2019/1/1 至 2020/1/3	A 市万润滨江天著企业	万润滨江牟取精装暴利

针对问题三，我们对最后的得分结果排名前 100 的数据中任意取一个分析：

1）抽取留言编号为“25528”，得分为 0.62，排名为 22 的数据进行分析，其留言时间为“2016/4/13 8:47:18”，答复时间为“2016/4/20 9:28:38”,时间差仅为 7 天，而全部数据平均的时间差将近 20 天，所以及时性得分优越。

留言内容为“政府近期是否有规划 A 市四环线，经安沙镇到 A9 市的快速通道？如果有什么时候动工建设？毛塘铺恒广国际物流中心才刚刚开园不久，就已经严重超负荷，长货车经常性的停靠在 107 干道，严重导致了 107 国道的车辆流通，是否有二期扩建规划？107 国道本身车道并不宽，国际物流园增加了大量的货车车流，是否规划新的道路干线贯通 A7 县以北，缓解车流压力。”

答复意见为“网友“UU0082018”您好！您的留言已收悉，现回复如下： 1、政府近期是否有规划 A 市四环线，经安沙镇到 A9 市的快速通道？如果有什么时候动工建设？ 根据最新的 A 市总体规划，在北三环北侧规划了一条东西向快速路，即北横线。往东可至 A9 市，往西可达 A6 区、A8 县。北横线的建设由 A 市交通运输局负责，该道路正处在初步设计阶段，开工建设具体时间请咨询 A 市交通运输局。 2、毛塘铺恒广国际物流中心才刚刚开园不久，就已经严重超负荷，长货车经常性的停靠在 107 干道，严重导致了 107 国道的车辆流 通，是否有二期扩建规划？ 目前物流园的车流，仅只能通过 107 国道上北三环进行疏散，确实对 107 国道交通造成一定影响。现万家丽北路与北三环互通项目已经开建，预计今年年底建成通车，将缓解 107 国道现有交通压力。 3、107 国道本身车道并不宽，国际物流园增加

了大量的货车车流，是否规划新的道路干线贯通 A7 县以北，缓解车流压力。 根据《安沙镇总体规划修改(2015-2020),在京港澳高速与东八路之间,还规划了一条贯穿镇域南北的交通性干道,往北可以与北横线相接。感谢您对我们工作的理解和支持！ 2016 年 4 月 19 日”

对比两者内容不难发现答复的质量非常好，这也肯定了我们的模型评估方案的优良。

再从得分结果中对排名最后 100 的数据任意抽取一组进行分析：

2）抽取留言编号为“119092”，得分为-0.58，排名为倒数 52 名的数据进行分析，其留言时间为“2019/6/17 11:09:30”，答复时间为“2019/9/24 9:52:07”，时间差为 98 天，远高于平均水平，所以及时性得分低。

留言内容为“永迥路白牛路交汇处名为向阳宾馆长期以来油烟污水严重污染环境，则面与宾馆后面油烟污水直接外排地面和室外无任何保护处理设施,严重影响周边居民的正常生活环境与身体健康。而且向阳宾馆还经营餐饮早、中、晚餐吃的人还不少，本人书读得少不知道在这样的环境做出来的饭菜能不能吃,在这样的环境卫生许可证和营业执照是怎样办出来的或者有没有经营许可证。希望相关部门与政府出面管一管冶一冶,这样的环境都没人管，据说此老板有背景，也不知道是没人管还是不敢管还是有保护伞罩着。”

答复意见为““UU0081838”您好！帖文中反映的问题已转交有关单位调查核实。感谢你的留言！2019 年 6 月 24 日”

在这次数据中留言内容非常详细,但是答复意见并不具备和留言内容相关的话语且可以看出答复质量较差，所以总得分低。

对于得分高的数据中很容易发现答复及时且认真,而得分较低的数据答复很慢且答复质量不怎么高。总而言之，此评分模型效果良好且泛化能力强。

五、未来改进

对于答复质量的评估模型，若加上如下方案将会式模型评估能力更强。

1）提取留言内容中的相关关键词，如建议等，将留言分为咨询类、建议类、投诉类以及求助类。具体每类回复模式如下：

2）咨询类留言回复模式：尊敬问候；依据政策或法规条文回答;说明此前与留言者通过电话取得联系，已询问对方是否满意；留下部门联系方式，以便留言者后续提问。

3）建言类留言回复模式：尊敬问候；对建言所针对的问题做出回应；分析建言的现实操作性；对建言表示感谢。

4）投诉类留言回复模式：尊敬问候；表明调查主体部门；分点说明调查结果是否与描述相符；对于符合实际情况的内容，可以迅速解决的，立马解决后回复；需要长期解决的，提出解决方案和解决进度，并做出承诺；电话联系，告知结果且询问是否满意；留下部门联系方式，以便留言者后续关注。

5）求助类留言回复模式：尊敬问候；表明调查主体部门；可以直接解决的，迅速解决；需要个人参与解决的，则提供建议；电话联系，告知结果且询问是否满意;留下部门联系方式，以便留言者后续求助。

针对上述留言回复标注模式，将留言回复内容进行切分，对应格式相匹配，以判断留言回复的完整性。

参考文献

[1] https://github.com/fxsjy/jieba

[2] www.tensorflow.org/

[3] https://github.com/Samurais/insuranceqa-corpus-zh

[4] https://baike.baidu.com/item/泛化能力

[5] An.Y, Chen.X.L, Han.C.P, Li.Z.M, Liu.L, Yang.R, Extracting New Word with Mutual Information and Logistic Regression*, Data Analysis and Knowledge Discovery, 2019, No.8,105-113.

[6] Hu.Y, Liu.L, Zhong.M.S, Chinese Text Segmentation Based on Quantified Conceptual Relations Estracted from Chinese Dictionary, Computer Engineering and Applications, 2008,Vol.21,No.44,25-29.

[7] Li.L, Zhao.M.G, "Forecast Analysis of Industrial Production Index Based on ARIMA Model,EconomicResearchGuide,2020,No.3,30-36.

[8] Shen.J, Yang.Y.P, "Dynamic Collaborative Filtering Recommender Model Based on Rolling Time Windowsand its Algorithm", Computer Science, 2013, Vol.40, No.2, 206209.

[9] An.Y, Chen.X.L, Han.C.P, Li.Z.M, Liu.L, Yang.R, Extracting New Word with Mutual Information and Logistic Regression*, Data Analysis and Knowledge Discovery, 2019, No.8,105-113.