基于 NLP 的“智慧政务”文本挖掘系统

摘要

随着近年来人工智能的快速发展，我们已经逐步进入了“智能化”时代，生活中方方面面都在“智能化”，网络问政平台也逐渐成为公众反映问题的主要渠道。为解决传统人工处理的方法中效率低、差错率高的问题，提高政府工作效率，“智慧政务”的概念应运而生。本套系统就是为了提高“智慧政务”的水平、拓展“智慧政务”的功能而设计实现的。

针对问题一：本系统以附件 1、2 中的数据作为基本参考，经过预处理、中文分词、结构化表示、权重策略、分类方法以及模型评价六个步骤，并结合测试结果进行了四种分类器的调参优化和功能分析，最终形成了一个基于逻辑回归算法的中文文本分类系统。

针对问题二：本系统在子系统 1 的基础之上，增加了词性相关的权重矩阵，将文本结构化后首先进行了初步的主题分类，而后在每种主题之下应用 DBSCAN 聚类算法找出特定问题。最后，在传统热度模型的基础上增设了时间维度，给出了热度评价指标的计算公式，构建了一个热点问题挖掘系统。

针对问题三：本系统通过分析附件 4 中的数据，先定义了打分评判的标准，主要从相关性，完整性，可解释性，时效性四个方面进行综合打分评定。接着进行了数据的预处理，剔除了与研究相关性不大的数据项。针对四个方面的判定，分别构建了基于 BERT 的相关性分析模型，基于 TextRank 的关键词提取模型，基于 word2vec 的情感分析模型等。最终得到了一个完整的打分系统。

最后我们分析了模型的优缺点，并做了总结反思以及适度的推广。

关键词：NLP,逻辑回归算法,DBSCAN 聚类算法,BERT,word2vec

Text mining system of "intelligent government affairs" based on NLP

Abstract： With the rapid development of artificial intelligence in recent years, we have gradually entered the era of "intelligent". All aspects of life are "intelligent", and online political platform has gradually become the main channel for the public to report problems. In order to solve the problem of low efficiency and high error rate in the traditional manual method and improve the working efficiency of the government, the concept of "intelligent government affairs" came into being. This system is designed to improve the level of "intelligent government" and expand the function of "intelligent government".

In view of the problem a: data in this system to attachment 1, 2, as the basic reference, after pretreatment, Chinese word segmentation, structured representation, weight strategy, classification method and model of evaluation of six steps, and combined with the test result has carried on the four kinds of the analysis of the adjustable parameter optimization and function of classifier, eventually forming a Chinese text classification system based on logistic regression algorithm.

For problem 2: based on subsystem 1, this system adds the weight matrix related to part of speech. After the text is structured, it first conducts the preliminary topic classification, and then applies DBSCAN clustering algorithm to find out specific problems under each topic. Finally, a time dimension is added on the basis of the traditional heat model, the calculation formula of heat evaluation index is given, and a hot issue mining system is constructed.

Aiming at question 3: by analyzing the data in appendix 4, this system first defines the criteria for rating and evaluation, and conducts comprehensive rating from four aspects: relevance, integrity, interpretability and timeliness. Then the data is preprocessed and the data items which are not relevant to the research are removed. According to the judgment of four aspects, the correlation analysis model based on BERT, the keyword extraction model based on TextRank, and the emotion analysis model based on word2vec were constructed respectively. The result was a complete scoring system.

Finally, we analyzed the advantages and disadvantages of the model, and made a summary of reflection and moderate promotion.
Keywords:NLP,Logistic Regression,DBSCAN,BERT,word2vec

目录

1 问题引入

近年来，随着网络技术的发展，微博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。由于传统的政务系统主要依赖相关工作人员人工处理市民的留言信息以及反应的问题，效率较低，在面对大量网上留言时难免力有未逮，所以利用电脑来处理这些数据成为了相关部门的迫切需求。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。

想要利用人工智能来处理留言，我们需要解决很多问题。一方面，中华文化博大精深，汉语语言的复杂表达形式，使得很多句子即使拥有类似的用词也会因为句子结构不同而有不同的含义。另一方面，留言反映的问题不仅和语句内容本身有关，还与很多环境因素有紧密的联系，包括时政、时间，这需要在模型中设计合理的权重规则。

为解决上述问题，提高本系统的准确性和易用性，我们可以参考人工处理留言的步骤。作为处理人员，首先要读懂留言内容，即构建基于 NLP 的中文文本分类系统。分类完毕后，我们通过将文本进行聚类提取，根据定义的热度指标对问题进行排名，达到挖掘热点问题的目的。最后，群众反映的问题都会得到答复意见的反馈，我们结合了实际中的网络问政平台对相关回复的要求，针对答复意见质量，设计了一套较为完备的多指标综合评分系统。

2 问题解决

2.1 问题一：群众留言分类

中文文本分类主要分为六个步骤：预处理、中文分词、结构化表示、权重策略、分类方法以及结果评价。流程如图 1 所示。

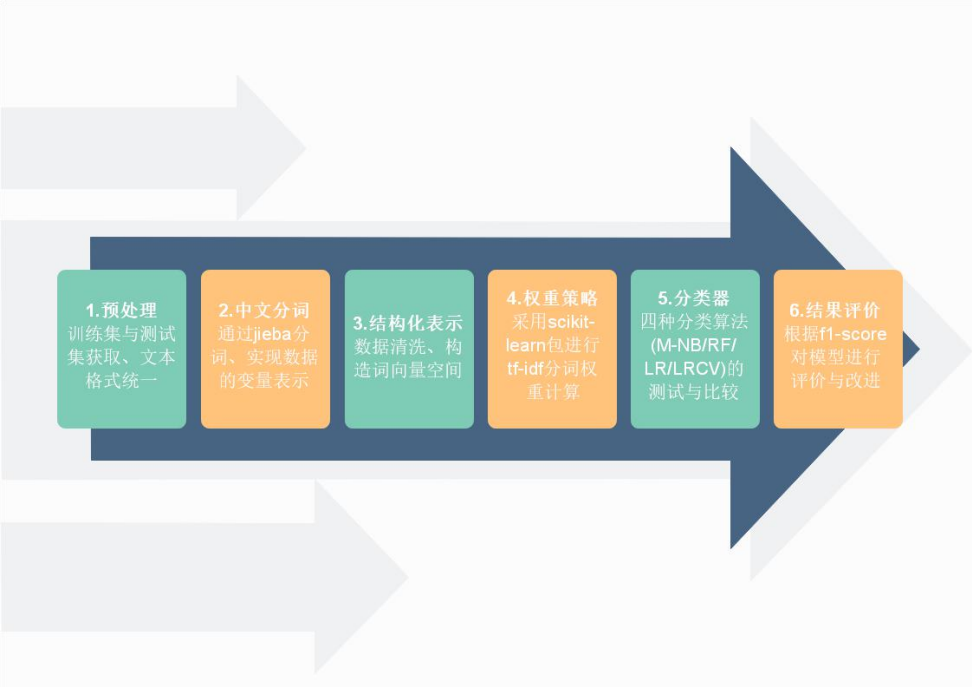

图 1 群众留言分类流程图

2.1.1 预处理

首先，我们要从给出的附件 2 中提取待分类的文本，包括“留言主题”和“留言详情”，并按照表中给出的“一级标签”分别存在不同文件夹下，共计 9210 条。为保证模型学习的准确度和检验模型的可靠性，取训练集和测试集为 8：2（即 7368：1842）。

2.1.2 中文分词

预处理后的语料库是连续的句子，现在需要对这些文本进行分词，只有这样，才能在基于单词的基础上，对文本进行结构化表示。相较于英文，中文分词有其特有的难点，最终完全解决中文分词的算法是基于概率图模型的条件随机场（CRF）。

在实现中，我们采用 python 第三方库 jieba 来进行中文分词，得到分词后的训练集和测试集语料库[1]。接下来需要把这两个数据集表示为变量，从而为之后程序调用提供服务，这里我们选择 Scikit-Learn 库中的 Bunch。Bunch 是一种存储键值对的数据结构，类似于 python 中的字典。格式示例见表 1。在 Bunch 对象里创建了 4 个 list 成员：

(1) target_name：整个数据集的一级分类标签集合。

(2) filenames：文件名集合。

(3) label：每个文本的标签。

(4) contents：分词后的文本内容。

表 1 Bunch 存储结构示例

target_name	filenames	label	contents
交通运输	test_corpus_seg/交通运输/2960.txt	交通运输	...几万名 网约车 违法 司机 在 跑车...
	test_corpus_seg/交通运输/2961.txt	交通运输	...A 市 公交 的 变化 有目共睹...
	test_corpus_seg/交通运输/2962.txt	交通运输	...实现 东塘 路口 的 全互通...
卫生计生	test_corpus_seg/卫生计生/8469.txt	卫生计生	...将 民营 医院 定位 为高新技术 企业...
	test_corpus_seg/卫生计生/8475.txt	卫生计生	...为什么 女方 乡 计生干部 不 同意 ？...
商贸旅游	test_corpus_seg/商贸旅游/7302.txt	商贸旅游	...挖掘 乡镇 旅游 项目整合 C 市 全域 4A 景区...
	test_corpus_seg/商贸旅游/7320.txt	商贸旅游	...滴水洞 景区 既然 设有 门票 售票处...

2.1.3 结构化表示

在这一部分，我们需要生成词向量，并将训练集和测试集统一到同一词向量空间下。

首先进行数据清洗，即过滤掉文本中“垃圾词汇”，以压缩词向量空间的规模。这些词通常意义模糊，无助于语义分析和特征提取，如某些编码符号，“的”、“了”等词频较高却无实际意义的虚词。我们构造一个停用词表（stop words）来存放这些词，见图 2。一般来说，停用词表越全面，处理后的文本信息冗余度越小，数据集受到的干扰就越小，越有利于提高模型的准确度。

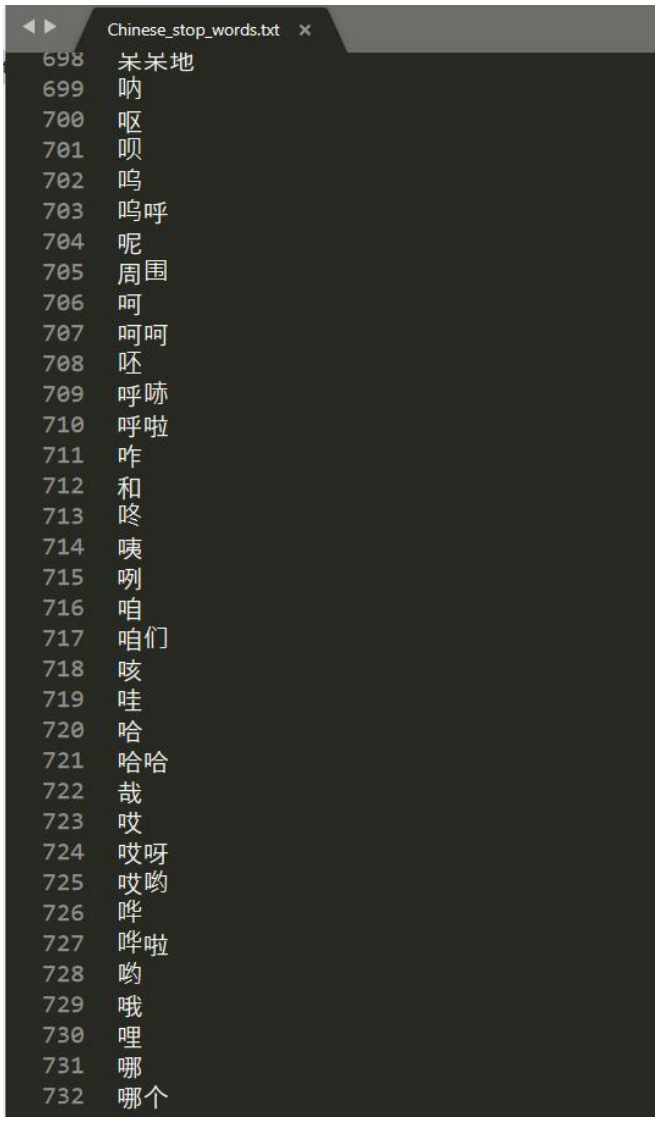

图 2 停用词表示例

其次，我们引入向量空间模型（VSM：Vector Space Model），把对文本内容的处理简化为向量空间中的向量运算，从而可以通过计算向量之间的相似性来度量文本间的相似性[2]。文本用 $D(document)$ 表示，特征项（即出现在文本 D 中且具有代表性的基本语言单位，具体方法 TF-IDF 将在 2.1.4 中详细讲解）用 t_i 表示。这样，文本可表示为 $D(t_1,t_2,...,t_n)$。

在建立向量空间模型时，一般会给每个特征项一个权值 w_i（本题中取 TF-IDF 值）来刻画该特征项对该文本的重要性。因此，特征项加权后的文本向量化表示为： $D(t_1,w_1,...,t_n,w_n)$，简写为 $D(w_1,...w_n)$。

于是，每个文本都能被表示成维数为$|n|$的向量。实际情况下特征项非常多，每个向量非常稀疏。那么，如何比较查询$Q=Q(w_{1q},w_{2q},...,w_{nq})$与某一文本$D_j=D_j(w_1,w_2,...,w_n)$之间的相似度呢？最常用的相似性度量方式是余弦距离（cosin measure）。

$$Sim(D,Q)=\frac{D\bullet Q}{\|D\|\times\|Q\|}=\frac{\sum_{i=1}^{n}D_i\times Q_i}{\sqrt{\sum_{i=1}^{n}(D_i)^2}\times\sqrt{\sum_{i=1}^{n}(Q_i)^2}} \quad (1)$$

其中，D_i、Q_i分别代表D和Q的各分量。分母称作规范化因子（normalization factor），使文本得分不受其长度的影响。由于词频（TF-IDF 权）不能为负数，余弦相似度$Sim(D,Q)$的范围为$[0,1]$。

2.1.4 权重策略

我们已经阐述过文本向量化的原理。接下来，我们引入 TF-IDF 方法进一步提取文本的关键词并计算权重。

TF-IDF 的原理是，根据字词在文本中出现的次数和在整个语料中出现的文本频率来计算一个字词在整个语料中的重要程度。其优点是在滤掉一些常见却无关紧要的词的同时保留影响更大的重要字词。具体讲解如下。

(1) 计算词频

词频指的是一定时间内特征词语在该文件中出现的次数。对它进行了归一化处理，以防止它偏向长文本。

$$TF_{t_i}=\frac{Time_{t_i}}{Sum_{D_j}} \quad (2)$$

其中，$Time_{t_i}$表示特征词t_i在文本D_j中的出现次数，Sum_{D_j}表示文本D_j中所有特征词的个数。

需要注意的是，一些通用的词语对于主题并没有太大的作用，反而是一些出现频率较少的词更贴合文章主题，所以单纯使用 TF 作为权重是不合适的。权重的分配标准应当满足：一个词预测主题的能力越强，权重越大；反之，权重越小。IDF 就是在完成这样的工作。

(2) 计算逆文本频率

逆文本频率的主要思想是：如果包含特征词t_i的文档越少，IDF 越大，则说明特征词t_i具有很好的类别区分能力。

$$IDF_{t_i} = \log\left(\frac{|D|}{|D_{inc(t_i)}|+1}\right) \quad (3)$$

其中，$|D|$表示语库中的文本总数，$|D_{inc(t_i)}|$表示包含特征词t_i的文本数。（分母加 1 是为了防止除 0 错误）

某一特定文件内的高词语频率，以及该词语在整个文本集合中的低文本频率，可以产生值较高的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。[3]

(3) 计算 TF-IDF

$$TF-IDF_{t_i} = TF_{t_i} \times IDF_{t_i} \quad (4)$$

TF-IDF 值越大，则这个词成为一个关键词的概率就越大。

在实现中，我们主要用到了 Scikit-Learn 中的 Tfidf-Vectorizer 类。首先，根据训练集数据构建一个 TF-IDF 词向量空间，并且得到了词向量空间坐标 *vocabulary* 和 TF-IDF 权重矩阵tdm_{train}（$tdm[i][j]$表示第 j 个词在第 i 个类中的 IF-IDF 值）。然后，把测试集数据也映射到这个词向量空间中，也就是说，测试集和训练集处在同一个词向量空间（*vocabulary* 相同），但测试集也有自己的权重矩阵tdm_{test}。

2.1.5 分类器

现在，我们已经得到了训练集和测试集同处的 TF-IDF 词向量空间，接下来需要选择合适的分类算法对其进行分类测试。我们调用了 sklearn 库，共测试了 4 种分类器：基于朴素贝叶斯算法的 MultinomialNB、基于随机森林算法的 RandomForest、基于逻辑回归算法的 LogisticRegression 及 LogisticRegressionCV。分类器先获取训练集的权重矩阵和标签，进行训练，然后获取测试集的权重矩阵，给出预测标签。下面分别对它们进行介绍。

(1) MultinomialNB

朴素贝叶斯模型这个概念来自于古典数学理论，它独立的确定每一维度

特征被分类后的条件概率，然后综合这些概率对其所在的特征向量做出分类预测，即“假设被分类的条件概率在各个维度上的特征之间是相互独立的”[4]，该假设使得模型预测需要估计的参数规模从指数数量级减少到线性数量级，极大地节约了计算时间和空间。

但是，该模型在训练时没考虑各个特征之间的联系，对于数据特征关联性较强的分类任务表现不好，因此不是本题的最优选择。

(2) RandomForest

随机森林是多重决策树的组合，而不只是一棵决策树。随机森林算法下决策树的数量越多，泛化的结果更好。在实现中，该模型训练周期较长，调参过程复杂，时间成本高，且每次得到的结果不同。但由于训练数据中存在噪音，随机森林中的数据集会出现过拟合的现象。

(3) LogisticRegression

Logistic Regression 是线性回归的思想，但用作分类器：它从样本集中学习拟合参数，将目标值拟合到$[0,1]$之间，然后对目标值进行离散化，实现分类。它使用了 Logisitic 函数（又称为 Sigmoid 函数），这个函数将分类任务的真实标记和线性回归模型的预测值联系起来。Sigmoid 函数具体的计算公式如下。

$$\sigma(z)=\frac{1}{1+e^{-z}} \tag{5}$$

结合两种坐标尺度下的 Sigmoid 函数图，如图 3 所示。可以看到，当横坐标的尺度足够大时，在 x=0 处 Sigmoid 函数看起来很像单位阶跃函数。而这种类似于阶跃函数的效果正是我们想要的，考虑二分类任务，其输出标记为 0 和 1，而 Sigmoid 函数将 z 值转化为一个接近 0 或 1 的 y 值，并且其输出值在 z = 0 附近变化很陡。

将 Sigmoid 函数的输入记为z，暂且由下面公式表出：

$$z = w_1x_1 + w_2x_2 + ... + w_nx_n + b \tag{6}$$

其中，x_i表示示例x在属性i上面的取值。因此，为了实现 Logistic 回归分类器，我们可以在每个特征上都乘以一个回归系数，然后把所有的结果值相加，将这个总和代入 Sigmoid 函数中，进而得到一个$[0,1]$之间的数值。任

何大于 0.5 的数据被分入 1 类，小于 0.5 即被归入 0 类。所以，Logistic 回归也可以被看成是一种概率估计。

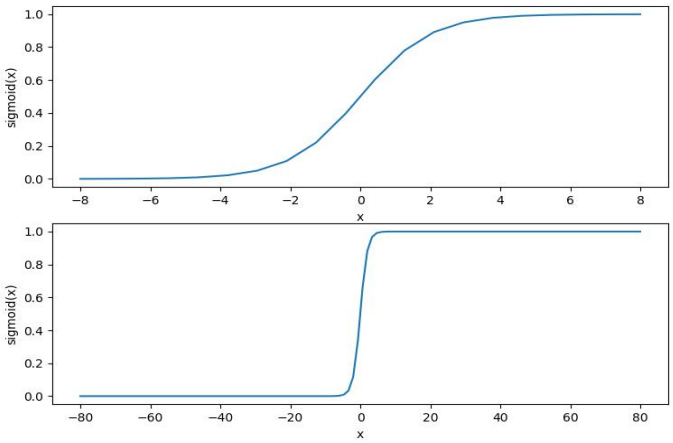

图 3 两种坐标尺度下的 Sigmoid 函数图

为了使得分类器尽可能地精确，我们需要找到最佳参数（系数）。给定包含 n 个示例的数据集 $D=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$，其中，$x_i=(x_{i1};x_{i2};...;x_{id})$，$x_{ij}$ 表示 x_i 第 j 个属性上的取值，$y_i\in[0,1]$。线性模型试图习得一个通过属性的线性组合来进行预测的函数，即

$$f(x_i)=\omega_1x_{i1}+\omega_2x_{i2}+...+\omega_dx_{id}+b \tag{7}$$

一般用向量形式表示：

$$f(x_i)=\omega^Tx_i+b \tag{8}$$

其中，$\omega=(\omega_1;\omega_2;...;\omega_d)$。为了便于讨论，我们把 ω 和 b 吸收入向量形式 $\hat{\omega}=(b;\omega)$，得到 $\hat{\omega}=(\omega_0;\omega_1;...;\omega_d)$ 的形式。因此有

$$f(x_i)=\omega_0\times1+\omega_1\times x_{i1}+...+\omega_d\times x_{id}=\hat{\omega}^Tx_i \tag{9}$$

为兼顾计算的准确性和便捷性，我们使用均方误差 L（亦称为平方损失）来衡量 $f(x)$ 和 y 之间的差别，从而确定系数 ω 和常数项 b。

$$L=\frac{1}{2}\sum_{i=1}^{n}[f(x_i)-y_i]^2 \tag{10}$$

此公式是二次方程，当 L 取得最小值时，所对应的 $\hat{\omega}$ 就是最佳拟合参数。求解 $\hat{\omega}$ 使 L 最小化的过程，称为线性回归模型的最小二乘“参数估计”。[5]

接下来，我们使用梯度下降法推导$\hat{\omega}$的迭代公式。梯度下降法基于的思想是：要找到某函数的最小值，最好的方式就是沿着该函数的梯度方向的反方向搜寻。其步骤是，先随机给$\hat{\omega}$赋值，然后沿着公式一阶偏导的反方向计算下降量值，多次重复，最终会让公式收敛到一个极小值。用向量作为表示形式，梯度下降法的迭代公式如下：

$$\hat{\omega} = \hat{\omega} + \Delta\hat{\omega} = \hat{\omega} - \alpha \frac{\partial L}{\partial \hat{\omega}} \quad (11)$$

其中，α是步长，即每次迭代的移动量。

由于涉及到矩阵的计算，过程较为复杂，详细步骤见附件。这里给出以梯度下降法计算最优$\hat{\omega}$的最终迭代公式：

$$\hat{\omega} = \hat{\omega} + \alpha * X^T(Y - X\hat{\omega}) \quad (12)$$

(4) LogisticRegressionCV

LogisticRegressionCV 和(3)的主要区别是，LogisticRegressionCV 使用了交叉验证来选择正则化系数 C，而 LogisticRegression 需要自己每次指定一个正则化系数。除了交叉验证，以及选择正则化系数 C 外， LogisticRegression 和 LogisticRegressionCV 的使用方法基本相同。

2. 1. 6 模型评价

四种模型的表现情况如表 2 所示。

Precision 体现了模型对负样本的区分能力，Precision 越高，模型对负样本的区分能力越强；Recall 体现了模型对正样本的识别能力，Recall 越高，模型对正样本的识别能力越强。F1-score 是两者的综合，F1-score 越高，说明模型越稳健。

表 2 四种分类器的对比

分类器	Precision	Recall	F-score
MultinomialNB	0.826	0.818	0.813
RandomForest	0.774	0.761	0.757
LogisticRegression	0.859	0.852	0.851
LogisticRegressionCV	0.904	0.904	0.903

通过结果来看，表现最好的分类器是 LogisticRegressionCV。需要指出的是，在参数选择和调优上，我们令 class_weight='balanced'，使类库根据训练样本量来

计算权重，也就是说，某种类型样本量越多，则权重越低，样本量越少，则权重越高。从而有效解决误分类的代价较高和样本比例失衡的问题。另外，多元逻辑回归模型与我们的问题较为贴合，因此令 multi_class='multinomial'，solver='newton-cg'。newton-cg 是牛顿法家族的一种，利用损失函数二阶导数矩阵（即海森矩阵）来迭代优化损失函数。

2.1.7 问题小结

本节通过设计中文文本分类系统，给出了一种关于留言内容的一级标签分类模型，具有较为理想的准确性和稳健性，可以有效应用于网络平台留言分类，较好地解决人工处理工作量大、效率低、差错率高等问题。同时也为问题二的解决提供了思路、奠定了基础。

2.2 问题二：热点问题挖掘

首先我们需要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，然后根据我们定义的热度评价指标，给出相应的评价结果，见图 4。

2.2.1 数据预处理

这里我们主要进行语料采集、分词、去停用词及词向量空间构造。大体思路与实现和 2.1 中的内容相似，只有一点不同：为了更好地突出某个问题的聚集程度，我们在 TF-IDF 权重矩阵的基础上，增加了词性对权重的影响，主要是提高了名词和动词的权重比，同时也对一些虚词进行了二次过滤。我们先利用 2.1 中构建的子系统对语料库进行了初步分类，得到的部分结果如表 3 所示。

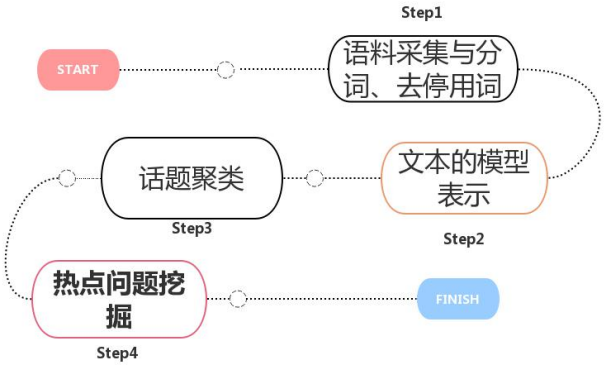

图 4 热度模型流程图

表 3 问题二预处理结果示例

留言文本编号	留言详细内容	留言预测类别
qs2_corpus_seg/unknown/1003.txt	...A 市 生殖 医学 医院 以 做 男孩 为 诱饵 ， 生出 基因 缺陷 女婴...	卫生计生
qs2_corpus_seg/unknown/1051.txt	...投诉 丽发 新城 小区 附近 违建 搅拌站 噪音 扰民...	城乡建设
qs2_corpus_seg/unknown/1196.txt	...西地省 高考 学生 少数 民加分 名单 何时 公布...	教育文体

2.2.2 文本聚类

有了预处理后的语料库，再分别在每个类别下进行文本聚类、筛选特定问题就有了一定基础。聚类算法属于常见的无监督分类算法，在很多场景下都有应用，如用户聚类，文本聚类等。常见的聚类算法可以分成两类，如表 4 所示。为解决以上两类算法的问题，我们选择了 DBSCAN 聚类算法。DBSCAN 是一类基于密度的算法，它的基本假设是一个集群的密度要显著高于噪声点的密度。因此，其主要思想是对于集群中的每一个点，在给定的半径范围 *eps* 内，相邻点的数量必须超过预先设定的某个阈值 *min_ samples* 。

表 4 聚类算法说明

聚类算法	代表	缺点
基于分区的算法	K-means	需要事先确定聚类的个数
		只适用于具有凸形状的簇
		对内存的占用资源较大
基于层次划分的算法	层次聚类	需要确定停止分裂的条件
		计算速度慢

DBSCAN 算法中包含两个重要的参数：

eps：聚类类别中样本的相似度衡量，与类别内样本相似度成反比。可以理解为同一个类别当中，对两个样本之间距离的最大值限定。

min_samples：每个聚类类别中的最小样本数。会对未分类样本数量造成影响，与未分类样本数量成正比。当相似样本数量少于该参数时，不会形成聚类。

DBSCAN 算法基本定义如表 5 所示。

表 5 DBSCAN 基本概念

名词	定义
核心点 *p*	距离该点 *eps* 范围内的其他点个数不少于 *min_samples* 个。
可到达点	在核心点 *p* 的 *eps* 范围内的点（可到达点可通过与其相邻的核心点，到达距离超过 *eps* 的其他核心点）。
异类点	无法到达核心点的点。
簇	以核心点为中心，每个簇至少包含一个核心点，相连的核心点形成簇，非核心点可以成为簇的一部分。

聚类从随机节点开始，保证每个节点被访问一次且仅访问一次。对每个节点，计算其范围 *eps* 内节点的个数。如果节点个数超过预定的点数 *min_samples*，则该节点被标记为核心点，否则被标记为噪音点。核心点及其范围 *min_samples* 内的点形成簇。查找簇的过程在访问遍历所有点后完成。伪代码描述见图 5。

```
while(exist_any_point(data_set))          //数据集中存在未被访问的点
{
    p=get_point(data_set);                //从数据集中抽取一个未被访问的点p
    if (count_of_eps(p)>=min_samples)     //p周围eps范围内点个数不小于min_samples
        set_core(p);                      //说明点p为核心点
    else
        set_noise(p);                     //说明点p为噪音点
    if(if_core(p))                        //如果p是一个核心点
        make_cluster(p);                  //找出所有可达的点，形成一个簇
    else
        set_marginal(p);                  //说明p是一个边缘点
}
```

图 5 DBSCAN 算法伪代码描述

通过对文本进行聚类，我们可以得到在每个主题下的多个子问题，以及每个子问题下的多个留言文件。如图 6 所示。

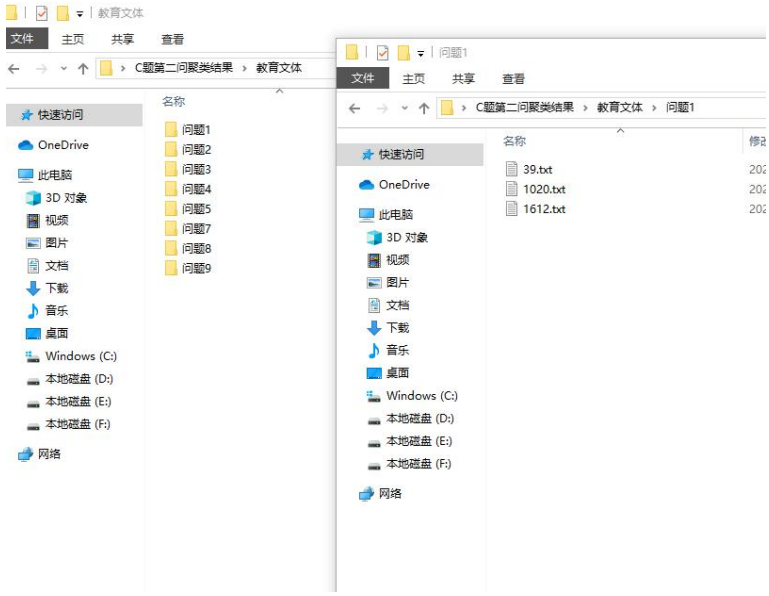

图 6 聚类结果格式示例

2. 2. 3 热度模型的建立

从数据集提取出某些特定问题后，需要建立模型来对其热度进行分析和评价。

2.2.3.1 基本原理

关于问题热度指数评价模型的确立，我们参考了类似的概念，如新闻热点的热度模型，舆情热点的模型等，并根据所给数据的特征设计了适合本题的热度指数模型。需要了解的是，热度算法也是需要不断优化去完善的。[6]

当群众反映的问题被收入库中，我们应该根据一定的评判标准对留言进行分类，对于不同类别的问题，本身具有的热度值也有所不同；群众不断地建议以及点赞、反对等行为会增加这类问题的热度，而问题热度一定会随着时间而流失。由此得出热度模型建立的基本公式为：

$$Score = S_0 + S_{users} - S_{Time} \tag{13}$$

即：问题热度分 = 初始热度分 + 用户交互产生的热度分 – 随时间衰减的热度分

2.2.3.2 初始热度的设置

初始热度，指的是单独看待某一特定问题自身时的热度值。最初，我们选择

的是一个静止的初始热度值，但在测试后发现效果不佳。结合现实考虑，如卫生计生、教育问题自被提出就较商贸旅游热度高得多；再如某些特殊事件发生期间（如留学生特权问题、新冠肺炎疫情等），一些相关问题的重要性发生变化，进而导致它们的关注度变化，而此时如果还是给每类问题同样的初始热度显然偏离实际，因此，我们将初始热度修正为变量。详细说明如下。

(1) 按照问题所属类别赋给问题不同的初始热度，如图 7 所示。

(2) 按照问题所在地区/所涉人群赋给问题不同的初始热度

我们先对所有问题聚类分析，将它们分类好之后统计出问题频发的地区或者人群，对这种问题赋予更高的热度初始值。

图 7 按照类别设置初始热度

2.2.3.3 用户行为分规则

首先我们要知道大众的哪些行为会增加事件的热度值，然后对这些行为设立相应的得分规则。[7]例如，对于单个问题，用户可以选择点赞（support）或反对（against）。得到问题的实时用户行为分可以表示为为：

$$S_{user}=\frac{\text{反映问题数}*1+(\text{点赞数}+\text{反对数})*5}{\text{时间跨度}} \tag{14}$$

这里对不同行为我们选择赋予的权重为 1.021、4.995，当然，这个值不应该是一成不变的：当数据集规模小的时候，各项事件都相对较少，此时需要提高每个居民所发生行为的权重来提升用户行为的影响力；当数据集规模变大时，权重

也应该慢慢降低。通过分析本题中数据集的特点，我们选择较小规模人群所对应的行为分，通过多个各种小型群体的模型分析，我们选取了较为固定的行为分。特别需要注意的是，由于点赞和反对的行为都会使问题的热度变高，甚至在同时存在时会产生争议，进而使问题整体的热度再蹿高。因此，当点赞数和反对数大致相等的区间内，我们会整体增加问题的热度指数。

2.2.3.4 热度随时间变化的衰减规则

由于问题的时效性，问题热度值应当随着时间流逝而减小，并且趋势应该是先平稳然后在迅速下降，直至趋于零。换句话说，如果某个问题的热度要一直处于很高的位置，随着时间的推移，它需要越来越多的关注度来维持他的热度。

我们假设群众的注意力是有一定期限的，理论上讲，衰减算法必须保证问题在群众注意力转移后的热度一定会衰减到很低，如果是线性衰减，当某些问题突然有大量群众反映，获得很高的热度分时，可能会使问题热度一直过高，导致后面短时间内矛盾比较尖锐的问题得不到有效的重视。如图 8 所示。

参考牛顿冷却定律，时间衰减因子应该是一个类似于指数函数的函数：

$$T(time) = e^{(k*(T_1 - T_0))} \tag{15}$$

再结合热度公式的基本原理，我们就可以得到最终热度指标的基本模型：

$$Score = S_0 + \frac{amount*1 + 5*(suggest + against)}{timespan} + e^{0.01*timespan} \tag{16}$$

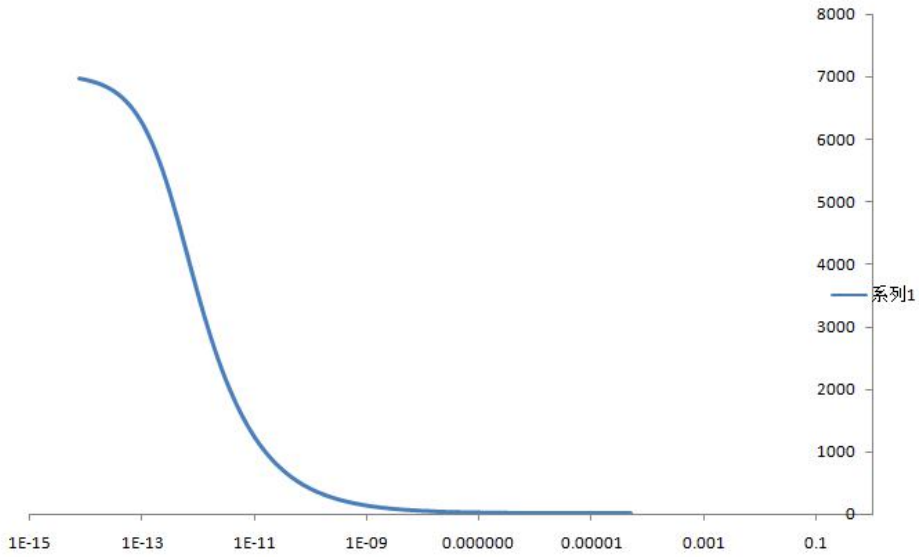

图 8 热度随时间衰减规律示意图

2.2.4 问题小结

本节通过 DBSCAN 算法进行了文本聚类分析，构建了热度模型来衡量某一特定问题的热度指数。在事件的一维尺度基础上增加了时间维度，考虑了时间对问题热度的影响。经测试，模型能够较好地发现热点问题，形成可读性较高的结果，有助于相关部门及时进行针对性处理，提升服务效率。

2.3 问题三：答复意见的评价

首先，我们需要设计关于答复意见质量的评价方案。评分机制主要由四个方面组成：相关性、完整性、可解释性和时效性。

其次，我们要进行科学合理的权重分配。指标权重分配的方法很多，归纳起来可分为三大类，分别是主观赋权法、客观赋权法以及主客观综合赋权法。主观赋权法多是依赖于专家的知识经验进行主观判断来确定指标权重，客观赋权法主要是依靠样本数据分析计算出权重，而主客观综合赋权法是基于主、客观赋权法各自的不足和优势，将两者所得的权重综合集成。在这个问题中我们采用的是最后一个方法，也就是主客观结合的方法。

主观部分的判断，通过查阅资料发现，新华网等国家官方网站上对于政务服务系统的要求主要有透明性、时效性、服务到位、解决问题等，其中着重强调了切实解决问题，而这个要求对应在本问题中就是相关性和完整性，因此，在我们的权重划分中，相关性和完整性的权重就要相对高一些。

至于客观性的判断，对于提出问题的群众来说，最重要的就是要解决这个问题，也就是说，回复的内容要切实解决群众的疑问，这也与上文说到的主观部分判断相一致。但是在相关性和完整性中，前者是群众更在意的，同时也是后者的前提和基础，所以相关性和完整性之间是递进的关系。

综合以上分析，最终形成的评价方案如表 6 所示。

表 6 答复意见质量评分规则

评价指标	所占权重	评价指标	所占权重
相关性	40%	可解释性	20%
完整性	30%	时效性	10%

2.3.1 数据预处理

通过观察附件 4 中的数据，留言用户这一列与我们的分析关联度较小，可以暂时删除，并用留言编号抽象代替每一条留言。

针对答复意见这一列，可以做进一步的精简。通过分析答复意见，总结出所有答复意见的结构均为：

阐述问题标题 + “答复如下（回复如下）”+ 回复内容

由于我们分析的侧重点在于回复内容，所以我们对“回复如下”之前的文本进行了批量处理，精简数据集，提高分析的针对性和准确度。

2. 3. 2 相关性分析

相关性主要考量的是留言答复与留言内容的相关程度。我们这里采用 NLP 相似度分析。NLP 中关于计算文本相似度计算的算法很多，我们组主要选择了两个算法进行了实验和比较，分别是基于词向量的余弦相似度计算和基本词嵌入模型的 doc2vec 相似度计算。

(1) 基于词向量的余弦相似度计算

原理及公式参见 2.1.3 中的公式（1）。流程如图 9 所示。

图 9 余弦相似度计算流程图

首先，我们使用 jieba 进行分词。然后，利用 BERT 预训练模型进行分词编码及向量化。最后，使用余弦距离计算两段文本的相似度。对附件 4 中留言和回复计算的文本相似度如下表所示（仅展示前 5 行，详见附录）。

表 7 余弦相似度计算结果示例

留言编号	相似度
2549	20.488
2554	17.965
2555	15.973
2557	16.726
2574	20.401

根据表格数据绘制柱状图如图 10 所示。

(2) doc2vec 相似度计算

其实在 doc2vec 之前，2.1.4 中介绍 TF-IDF 方法也可以计算相似度，但是该方法的局限性在于，没有考虑到文字背后的语义关联，例如两个文本共同出现的单词很少甚至没有相同的单词，但其语义是相似的。这时，doc2vec 的语义分析功能就很好地解决了以上问题。

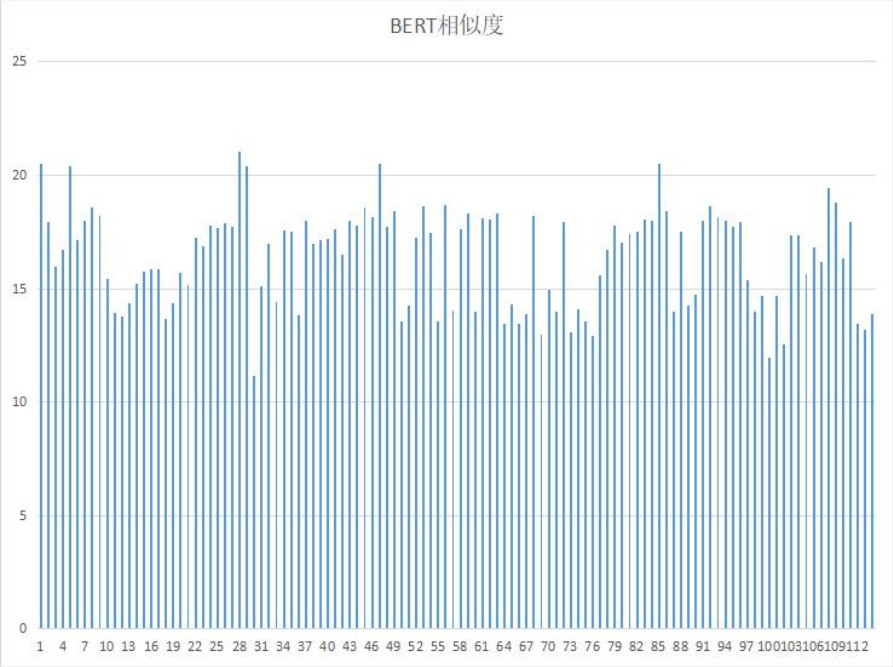

图 10 BERT 相似度结果柱状图

我们使用 PV-DM 方法训练 doc2vec 模型。PV-DM 方法框架如图 11 所示。

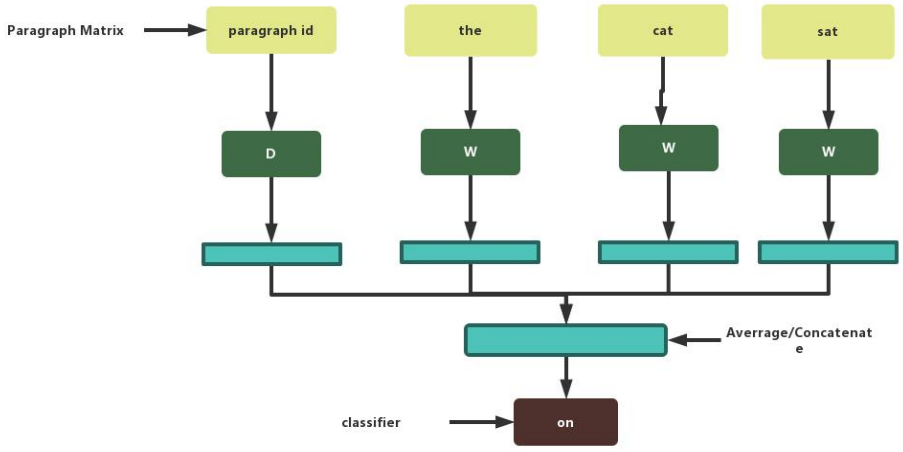

图 11 PV-DM 方法训练 doc2vec 模型流程

我们在模型训练过程中增加了 paragraph id 这个属性，即语料库中的每个句子都有一个唯一的 id 标示。词向量和输出层 softmax 的参数保持训练阶段得到的参数不变，重新利用梯度下降算法训练该句。

paragraph id 是先映射成一个向量，即 paragraph vector 这一点和普通的 word 一样，收敛后即可得到待预测句的 paragraphvector。

使用 doc2vec 计算句向量相似度的流程如图 12 所示。

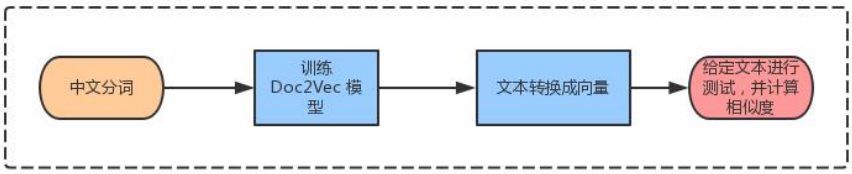

图 12 doc2vec 计算句向量相似度流程

最终对附件 4 中留言和回复计算的 doc2vec 相似度如表 8 所示。

表 8 doc2vec 相似度计算结果示例

留言编号	doc2vec 相似度
2549	23.488
2554	19.963
2555	10.973
2557	-2.726
2754	18.401

根据表格数据绘制柱状图如图 13 所示。

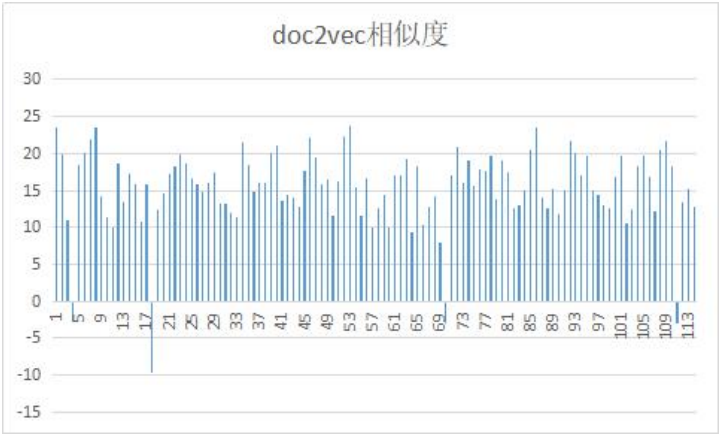

图 13 doc2vec 相似度柱状图

比较以上两个模型，不难看出，BERT 得到的结果更为准确，也更符合直观判断。对于 doc2vec 中得到的结果存在负数等错误值的情况，通过分析，我们认为 doc2vec 不能得到准确结果的原因在于其对短文本的计算效果不佳，以及我们模型训练的不到位。因此，针对相关性分析我们最终采用了 BERT 模型。

2. 3. 3 完整性分析

完整性包括对留言中提到的问题的作答个数，以及每个问题下进行相关答复的程度。我们首先采用关键词提取技术对留言和回复信息进行处理。如果直接将留言和回复进行整体的语义分析与比较，虽然实现难度比较小，但是误差相对较大。因为我们的模型训练的是小规模数据集，我们可以先对留言以及回复中的关键信息进行提取，将文本进行微元，把整个文本的比较转换为文本关键词的比较，从而提高模型的效率和准确率。

2.3.3.1 关键词提取

目前，用于文本关键词提取的主要方法有四种：基于 TF-IDF 的关键词抽取、基于 TextRank 的关键词抽取、基于 Word2Vec 词聚类的关键词抽取，以及多种算法相融合的关键词抽取。我们主要使用 TF-IDF 和 TextRank 关键词抽取进行对比。

(1) TF-IDF 关键信息提取

TF-IDF 的主要思想和公式参见 2.1.4。遍历 TF-IDF 权重矩阵，得到 $topk$ 个我们需要的关键词。通过多组实验发现，当 $topk=6$ 时结果较为理想。部分留言以及回复的关键词如图 14 所示。

留言编号：	2549				
小区	收入	物业公司	业主	方式	业委会
业委会	业主	同意	停车	工作	问题
留言编号：	2554				
方便	很大	有时候	一直	老百姓	部门
土质	道路	渠道	排水	项目	大道
留言编号：	2555				
工作	教师	压力	没交	大力	国家
待遇	民办	教师	依法	学前教育	方面
留言编号：	2557				
研究生	公寓	您好	人才	毕业	购买
购房	补贴	管理中心	房屋交易	购买	平台

图 14 TF-IDF 提取关键词

(2) TextRank 文本关键词抽取

基于 TextRank 的文本关键词抽取的基本思想是：利用局部词汇关系，即共现窗口，对候选关键词进行排序，该方法的步骤如下。

① 对于给定的文本 D 进行分词、词性标注和去除停用词等预处理操作

② 构建候选关键词图 G=(V,E)

③ 根据公式迭代计算各节点的权重

④ 对节点权重进行倒序排列

对文本 D 进行分词后，保留'n','nz','v','vd','vn','l','a','d'这几个词性的词语，最终得到 n 个候选关键词，即 $D = D(t_1, t_2, ..., t_n)$。节点集 V 由候选关键词组成，并采用共现关系构造任两点之间的边，两个节点之间当且仅当它们对应的词汇在长度为 K 的窗口中共现时存在边，K 表示窗口大小，即最多共现 K 个词汇。得到排名前 $TopN$ 个词汇作为文本关键词，部分结果如图 15 所示。

留言编号:	2549				
小区	投票	物业公司	业主	方式	业委会
业委会	业主	业主大会	停车	工作	业委会
留言编号:	2554				
方便	很大	有时候	一直	生意	部门
施工	道路	渠道	排水	项目	大道
留言编号:	2555				
工作	教师	压力	没交	大力	国家
待遇	民办	教师	依法	学前教育	方面
留言编号:	2557				
研究生	公寓	您好	人才	毕业	购买
购房	补贴	管理中心	房屋交易	购买	平台

图 15 TextRank 提取关键词

观察两种方法得到的结果，TextRank 产生的关键词更能体现文本的真正含义。通过查阅资料，我们将其原因归结于对单文档直接应用 TF-IDF 方法时，选择低频率词本身就是不准确的，因此计算得到的低频率词语也不一定是关键词，导致用这种方法得到的结果不理想；而 TextRank 方法是基于图模型的排序算法，在单文档关键词抽取方面有较为稳定的效果。所以在本题中我们最终选择 TextRank 方法进行关键词提取。

2.3.3.2 完整性定量计算

提取了关键词之后，我们需要计算回复与留言的完整度。这里我们使用word2vec 模型，word2vec 包含两种结构，一种是 skip-gram 结构，一种是 CBOW 结构。skip-gram 结构是利用中间词预测邻近词，CBOW 模型是利用上下文词预测中间词。[8]CBOW 一般用于数据数据量较小的情况，所以本题中采用 CBOW 结构，见图 16。

先将上述模型得到的两个关键词列表进行逐一的单词间相似度计算，再进行结果的卷积，从而得到一个句子的完整性定量结果。

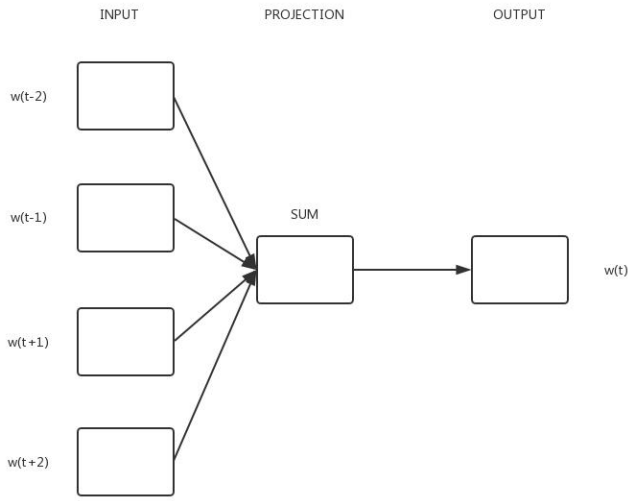

图 16 CBOW 结构图

2. 3. 4 可解释性评定

我们将可解释性定义为，回复的语言态度以及回答的内容力度。针对语言态度，可以使用 NLP 的情感分析进行处理，得到某一回复的态度是偏积极还是消极，是否存在对群众的问题不耐烦的现象等。而回答力度主要指某一回复对于解决群众的问题的实用性与清晰度。

通过观察附件 4 的数据我们发现，有的回答是直接明了的，对解决问题具有较大帮助，如图 17 所示。

有的回答较为模糊，不能对问题解决起到直接作用，但是也给予了一定意义上的回应，如图 18 所示。

还有一些回答明显属于“套话”，不具备任何实际意义，如图 19 所示。

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，关于您在平台栏目给胡华衡书记留言，反映“A2区景蓉花苑物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下：经调查了解，针对来信所反映的“小区停车收费问题”，景蓉华苑业委会于2019年4月10日至4月27日以“意见收集方式”召开了业主大会，经业委会统计，超过三分之二的业主同意收取停车管理费，在业主大会结束后业委会也对业主提出的意见和建议进行了认真疏理归纳并进行了反馈，业委会制定的停车收费标准不高于周边小区价格。针对来信所反映的“物业公司去留问题”，5月5日下午辖区桂花坪街道牵头组织社区、物业公司、业主委员会、业主代表的会议，区住房和城乡建设局也参加了会议。在综合各方面的意见后，辖区桂花坪街道、区住房和城乡建设局已要求业委会依法依规召开业主大会，根据业主大会的表决结果再执行相应的程序。再次感谢您对我区工作的理解和关心。2019年5月9日

图 17 可解释性较好的回复示例

网友“A00077538”：您好！针对您反映A3区含浦镇马路卫生很差的问题,A3区学士街道、含浦街道高度重视，现回复如下：您留言中反映的含浦镇在2013年已经析出两个街道，分别是学士街道和含浦街道，鉴于您问题中没有说明卫生较差的具体路段，也没有相应的参照物，同时您也未留下联系方式，请您看到回复后，致电学士街道0731-0000-00000000或者含浦街道0731-0000-00000000反映相关问题。感谢您对我们工作的关心、监督与支持。2019年4月24日

图 18 可解释性不好的回复示例（1）

您好，你所反映的问题已转交相关单位调查处置。

图 19 可解释性不好的回复示例（2）

由此，我们设立的针对可解释性的打分规则如下：

可解释性满分 20 分，情感分析和力度分析各占 10 分，评判结果越好则分数越高。

2.3.4.1 情感分析

我们使用的是 Word2vec 模型，选用的词典来源于 BosonNLP 数据下载的情感词典[9]，该词典源于社交媒体文本，适用于处理媒体内容的情感分析问题。流程如图所示。

2.3.4.2 力度分析

对于回答的力度判定，我们选择了基于负面词典打分的方法，因为直接分析语义去判定力度对模型的依赖性过强。由于力度较小的回复中多会出现“你所反映的问题已转交相关单位调查处置”，“建议您至 XXXX 部门咨询”等语句，我们将附件 4 中类似的短语人工提取出来，形成力度判定的字典。然后将回复文本作为输入，与负面词典中的词向量进行比较，根据相似度得出回复力度的负面指数，最终转化为力度得分。

2. 3. 5 时效性判定

对于时效性的定义较为简单，即回复与留言之间的时间差。

首先我们计算出了附件 4 中的每一组回复与留言之间的时间差（单位：天），得到的数据如表 9 所示（仅展示前 5 组，详见附录）。

根据表格数据绘制柱状图如图 20 所示。

表 9 回复与留言时间差示例

留言编号	时间差（单位：天）
2549	15
2554	15
2555	15
2557	15
2754	16

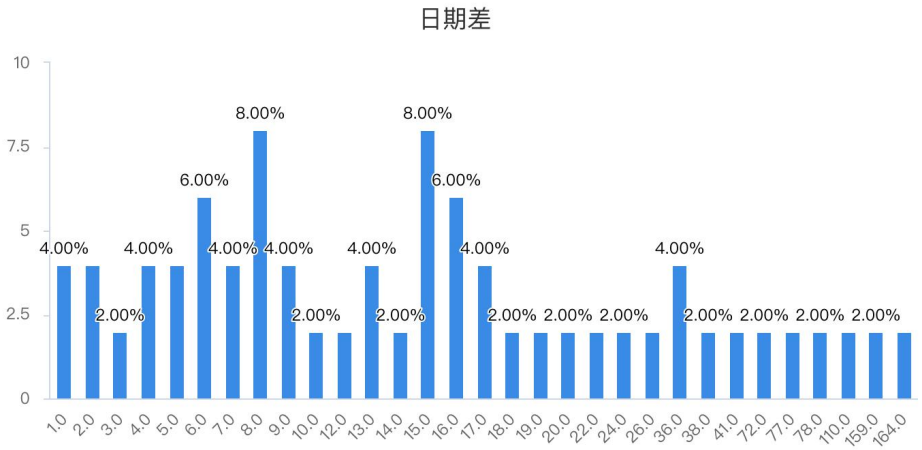

图 20 留言与回复日期差

从图中可以看出大部分留言与回复的时间差都集中在 15 天以内，由此我们制定的时效性评判标准如图 21 所示。

```
//Time_score : 时效性得分
//Delta_Time : 留言与回复时间差

if delta_time in [13,15] :
    Time_score = 5
if delta_time in [1,13) :
    Time_score = 5 + (13-delta_time)/13*5
if delta_time in (15,+∞) :
    Time_score = 5 - (delta_time-15)/delta_time*5
```

图 21 时效性评判标准

得到的时效性得分结果如表 10 所示。

表 10 时效性得分结果示例

留言编号	时效性得分
2549	5.000
2554	5.000
2555	5.000
2557	5.000
2754	4.686

根据表格数据绘制柱状图如图 22 所示。

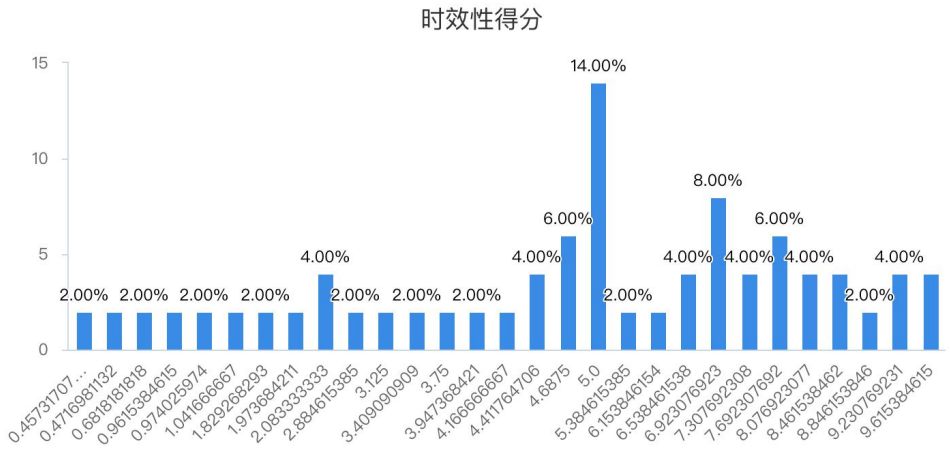

图 22 时效性得分柱状图

2.3.6 问题小结

本节设计实现了一套答复意见质量评价系统，根据相关性、完整性、可解释性和时效性四个方面分别对答复进行打分，通过对以上每个方面评判规则的建模

以及计算结果，最终得到了每个留言答复的总成绩以及各小项的得分情况。在实际应用中，可以作为评价员工绩效的参考。

最终完整流程如图 23 所示。

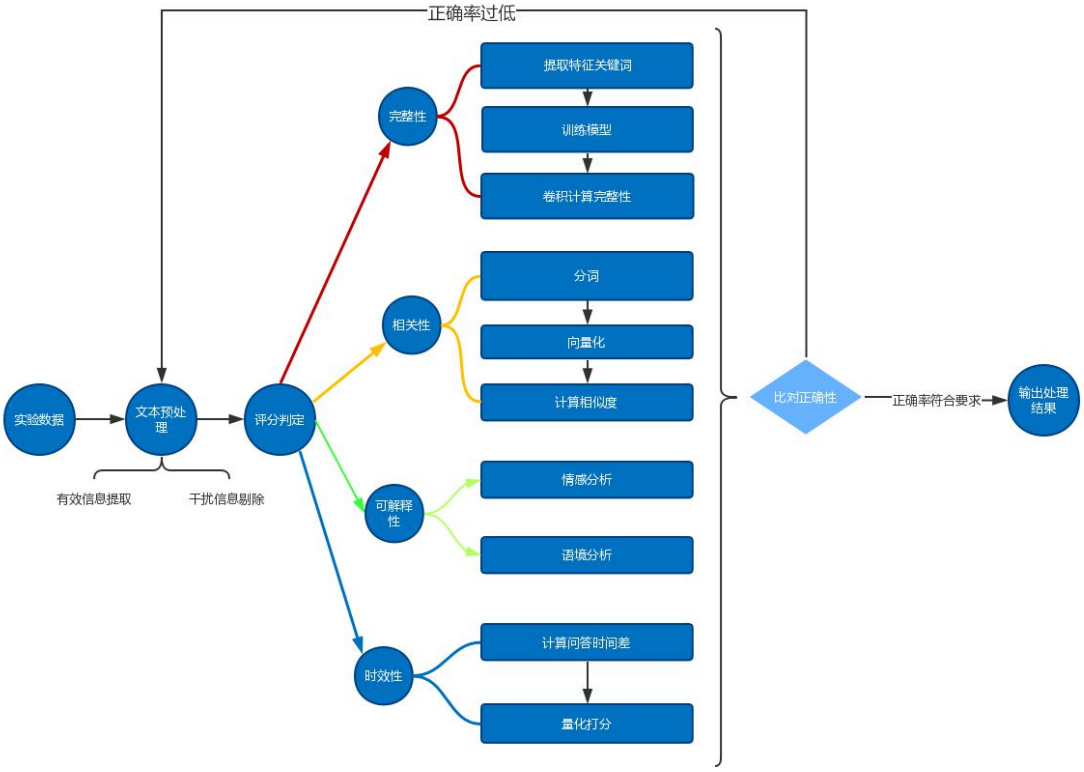

图 23 问题三完整流程图

3 结果分析

3.1 问题一结果展示与分析

通过上述分析与实现，基于题目给出的附件 2，我们建立了关于留言内容的一级标签分类模型，得到了测试集的预测结果。详表请见附件，这里展示部分结果如下。

留言文件编号	留言详细内容	留言实际类别	留言预测类别
2949	...请 加强 对 中小学生 交通 知识 的 宣传教育...	交通运输	交通运输
2973	...交通拥堵 A 市 大城市 病之一...	交通运输	城乡建设
3587	...请 解决 A4 区辰北 三角洲 片区 教育资源 不足 问题...	教育文体	教育文体
3647	M6 州雅师 长期 补课 ，该 由 谁 管 ？...	教育文体	教育文体
2923	...严惩 “ 保护伞 ” 和 “ 官商勾结 ” 行为...	环境保护	劳动和社会保障
7271	...B 市 B9 市 中医院 和 B9 市楚东 医院 垄断 药品...	商贸旅游	卫生计生

图 24 问题一结果示例

对于我们设计的一级标签分类模型，F-score 评价达到了 0.903，预测结果与实际类别相同的不再赘述，重点分析一下不同的情况。原因主要有二：

(1) 附件 2 中给出的一级标签正确，模型本身存在一定局限，导致某些文本预测结果与实际有出入。如上表中 2973 号文件。说明仅通过词语权重来标定文本所属类别是不够准确的，应当更多地考虑句意与主旨间的关系。

(2) 附件 2 中给出的一级标签错误，而模型有较好的容错率，能够在一定程度上排除噪声数据的干扰，得到的预测类别反而与留言文本内容更相符。如上表中 2923 号、7271 号文件。说明面对大量原始语料时，人工处理的正确度可能会随着连续工作时长的增加而下降，包括工作人员的主观判断、工作态度等也会对工作效果产生一定影响。而利用本系统有助于减轻政府工作人员负担、合理调度部门资源，更重要的是提高办公效率、更好地服务人民。

3. 2 问题二结果展示与分析

通过上述热度模型的建立与测试，我们得出了排名前 5 的热点问题及对应的留言信息，部分结果如表 11、表 12 所示。详表请见附件“热点问题表.xlsx”和“热点问题留言明细表.xlsx”。

通过分析得知，群众反映的热点问题主要有以下特点：

(1) 话题上，集中在与群众日常生活息息相关的民生问题、教育问题方面。

(2) 时长上，大多在一个月左右，说明相关部门发现问题、解决问题的周期较长，对于问题初现的重视度有待加强。

(3) 地点/人群上，留言群众具有一定的共性，侧面体现了该问题的真实性和严重性。

表 11 热点问题表示例

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	29	6.606	2019/9/5 至 2019/9/25	A5 区烧烤摊	临街门面油烟扰民
2	9	6.106	2019/8/23 至 2019/9/6	A4 区绿地海外滩二期	施工噪音扰民

表 12 热点问题留言明细表示例

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
29	246598	A00054842	A5 区劳动东路魅力之城小区临街门面烧烤夜宵摊	2019/9/25 0:31:33	A5 区劳动东路魅力之城小区临街夜宵摊、烧烤摊 24 小时经营，油烟扰民……	0	1
29	195095	A00039089	魅力之城小区临街门面油烟直排扰民	2019/9/5 12:29:01	魅力之城小区楼下烧烤摊、快餐店无证经营，长期油烟烧烤熏死人……	0	3

3. 3 问题三结果展示与分析

通过我们设计的答复质量评价系统，得到每条留言对应的答复意见总分如表 13 所示，详见附件 source.xsl。

四项具体的分数如图 25 所示，详见附件 source.json。

各留言答复的综合成绩分布区间如图 26 所示。

表 13 答复综合得分示例

留言编号	综合得分
2549	87.508
2554	87.418
2555	84.936
2557	87.008
2574	86.650


```
[{"留言编号": 2549.0, "相关性": 39.30020523071289, "完整性": 20.0, "可解释性": 24.723606555335937, "时效性": 5},
{"留言编号": 2554.0, "相关性": 39.376808166503906, "完整性": 20.0, "可解释性": 25.236380044907797, "时效性": 5},
{"留言编号": 2555.0, "相关性": 35.54915237426758, "完整性": 20.0, "可解释性": 23.028726463755227, "时效性": 5},
{"留言编号": 2557.0, "相关性": 37.244876861572266, "完整性": 20.0, "可解释性": 24.471561565530465, "时效性": 5},
{"留言编号": 2574.0, "相关性": 38.66112518310547, "完整性": 20.0, "可解释性": 23.99879874612853, "时效性": 4.6875},
{"留言编号": 2759.0, "相关性": 38.09440612792969, "完整性": 20.0, "可解释性": 23.13538426266595, "时效性": 4.6875},
{"留言编号": 2849.0, "相关性": 38.63417053222656, "完整性": 20.0, "可解释性": 24.64652905462929, "时效性": 1.829268292682
{"留言编号": 33970.0, "相关性": 37.40629196166992, "完整性": 20.0, "可解释性": 23.584258211114456, "时效性": 6.9230769230
```


图 25 答复各项得分示例

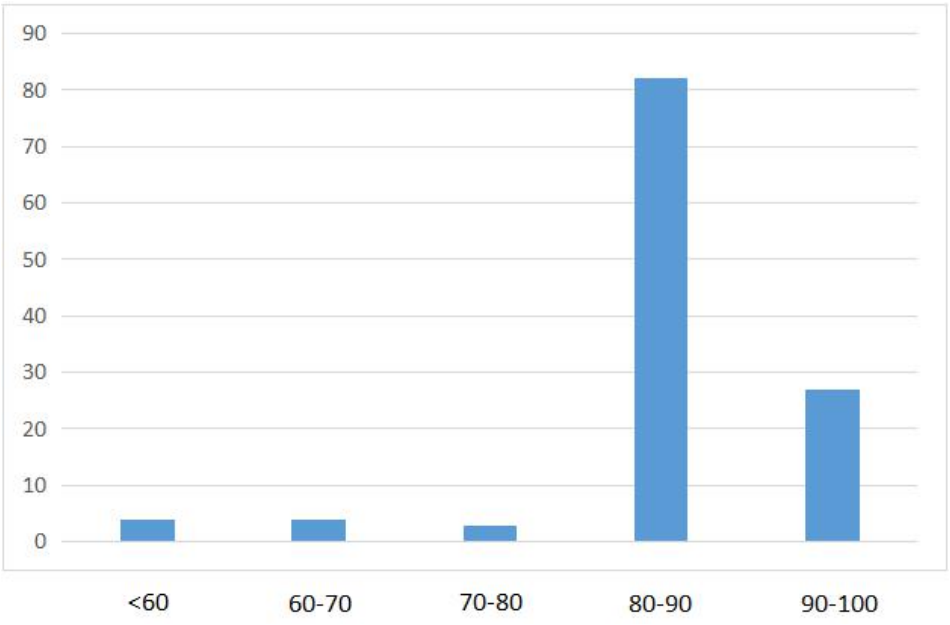

图 26 答复成绩分布柱状图

可以看到，大部分答复得分在 80-90 之间，说明大多数情况下，群众反映的问题能够得到及时有效的处理。但是仍然存在个别回复在 60 分以下，于是我们对得分比较低的数据进行了人工比对，验证打分系统的正确率。

综合得分低于 60 的留言答复共有 4 条，如表 14 所示。详情如表 15 所示。

表 14 不及格答复的得分情况

留言编号	综合得分
172654	46.009
37459	50.543
174574	54.700
37482	58.922

通过分析得知，有些留言几乎没有实际意义，只有问候开头或落款日期；而对于 174574 号留言，尽管答复中给出了一定的方法，但由于和留言中所反映的问题相关性较低，且切实帮助群众解决问题的力度不大，因此综合得分不高。

整体上看，本套答复意见质量评价系统的设计较为合理，结果准确率及公平性较高。

表 15 不及格留言答复详情表

留言编号	答复意见	答复时间
172654	尊敬的网友：您好！您所反映的事情已收悉，现回复如下： 2019 年 1 月 14 日	2015/1/6 10:21:18
37459		2019/1/14 16:06:08
174574	网友：您好，您反映的问题已收悉，现将有关情况回复如下：开户口迁移证，需持迁入地县级公安局开具的准予迁入证明、迁移人户口本原件，到出迁出地派出所办理。专此回复。 2019 年 11 月 18 日	2019/11/19 9:26:27
37482	2018 年 12 月 12 日	2018/12/13 18:53:19

4 总结与推广

4.1 模型优缺点分析

1. 优点
   (1) 在中文文本分类系统中有较好的容错率，能够在一定程度上排除噪声数据的干扰。
   (2) 答复质量评价系统考虑了多种情况，在题目的要求上进行了适度的拓展。
2. 缺点
   (1) 关于热度模型的构建，各因素之间的相关性有待加强，且灵活性不高，尤其是对于用户行为的评分规则考虑情况较少且分值略微主观。
   (2) 在使用词典分析的模型中，由于中文的博大精深，词性的多变成为了影响模型准确度的重要原因。

4.2 结论及推广

在本问题中，我们通过使用自然语言处理、权重分析法等方法，基于提供的表格数据，在已有通用模型的基础上，设计实现了中文文本分类系统、热点问题挖掘系统以及答复质量评价系统。经过不断的测试与改进，最终得到了相对人工分析正确率较高的结果，较为完整地解决了给出的问题。

本套“智慧政务”系统不仅适用于对政务留言回复的数字化智能化处理，对于目前发展火热的电商平台也有一定的使用价值。通过对客户的留言和评论进行智能化处理，可以及时准确地了解消费者的需求。一方面，依据不同的产品进行用户锁定与挖掘，另一方面，也能依据用户的喜好进行个性化推荐，从而帮助电商平台制定更为科学的策略。

参考文献

[1] 李伟. 基于决策树的网页敏感词过滤系统设计[D].西北农林科技大学.2018-05-01

[2] 王小平. 语义分析的一些方法(中篇)[J].http://www.sciencenet.cn/.2015-05-12

[3] 周诗咏. Web 环境下基于语义模式匹配的实体关系提取方法的研究[D].东北大学. 2009-06-30

[4] 兰见春. 基于 Spark 的犯罪预警分析系统的设计与实现[D].江西财经大学.2018-06-01

[5] 强保华. 基于线性回归和属性集成的分类算法[J].2017-01-17

[6] 卢争超. 产品经理需要了解的算法—热度算法和个性化推荐[J].2018-07-24

[7] 金石. 基于运营商管道大数据的智能电商推荐系统[D].南京邮电大学.2018-04-01

[8] 吴振华. 基于 GPU 计算连续分布式词向量的方法[C].2014-11-06

[9] 徐威. 一种文本分类方法在诈骗短信识别中的应用[D].暨南大学.2017-06-01