智慧政务开足“码”力，文本挖掘为“虎”傅翼
——基于自然语言处理技术的网络问政应用研究

摘　要

无论是传统社会还是互联网时代，有价值的信息和及时准确的传播都是社会运行所必需的重要基础。近年来，随着媒体融合和大数据时代的发展，互联网改变了信息传播的方式、内容，网络问政日渐成为一种新型的问政形式。但互联网大数据也造成了浩如烟海的信息过载和信息垃圾现象，如何把握大数据、云计算、人工智能等技术带来的崭新机遇，建立基于自然语言处理技术的智慧政务系统，在第一时间准确地把握舆论和信息脉搏，挖掘热点信息，做出及时、准确、有效的回复，无疑是当代中国网络问政发展的新使命。

对群众留言进行分类(问题 1)：群众留言数据共 9210 条，存在较为明显的类别不均衡情况，为有效避免欠拟合状态的出现，本文采用 EDA 文本增强技术进行模型性能的提升。在此基础上，考虑到分类处理的准确性、全面性，本文对于留言详情和留言主题采用不同的分类方法进行处理。首先，对于留言详情文本，利用维基百科中文语料库训练 Word2Vec 词向量模型，并最终利用 XGBoost 算法实现分类，F1score 达到 89.76%；其次，对于留言主题文本，通过对每条留言详情中的每个词汇进行 word embedding 映射，完成词向量化过程，进而利用 TextCNN 算法实现留言主题分类，F1score 达到 95.28%。

对热点问题进行挖掘(问题 2)：考虑到 Word2Vec 非常擅长在命名实体识别中找出相似性，本文首先基于 Word2Vec 计算文本相似度，并利用热度概率选取潜在热点标题 250 条。其次，利用 Word Mover’s Distance 对潜在热点标题再次归类后获得 13 个类中心。尽管热点事件的类型复杂多变，但对于固定小区的问题，往往时间集中度更明显，因此本文进一步地针对类中心引入时间变量，通过改良 K-means 聚类，剔除错误分类数据。最后，计算热度指数以实现热点问题的挖掘，热点问题和热点问题留言明细如表 13、14 所示。

对答复意见进行评价(问题 3)：为了全方位评估“智慧政务”中相关部门对留言的答复意见(反馈意见)的质量，本文针对附件 4 中相关部门对留言的答复意见，设立了一套答复意见评价指标体系(包含相关性、完整性、可解释性、时效性四个一级指标和相关程度、信息量、行文逻辑、受理效率等 10 个二级指标)，并利用变异系数权重法进行指标赋权，进一步地对网络问政平台中的“输出”即答复意见进行全方位的评估，最终得到答复意见评价评分如表 14 所示。

关键词：智慧政务 文本挖掘 Word2Vec XGBoost TextCNN WMD 指标体系

目　录

表目录

图目录

第 1 章 绪论

1.1 研究背景

随着智能手机的普及和互联网的快速发展，“人人都是通讯社、个个都有麦克风”的时代俨然到来[1]。在这个全媒体时代，网络问政已成为各级领导干部了解群众所思所想，了解群众诉求以及他们的呼声的重要手段。网络问政把网民的利益诉求反映给了政府，也迫使政府应对网民的压力做出反应和姿态，促使政府解决问题，“倒逼改革”[2]。2017 年 10 月 18 日，习近平书记在党的十九大报告中指出要“全面增强执政本领”，明确提出要“善于运用互联网技术和信息化手段开展工作[3]。”

近年来，网络问政在各级政府的重视下，在广大网民的支持下，取得了显著成效。但随着互联网技术的快速发展、网民参政需求的不断提升、新媒体和网络问政的快速结合，回复慢、答非所问、互相推诿、资源浪费等问题频繁出现。如何把握大数据、云计算、人工智能等技术带来的崭新机遇，建立基于自然语言处理技术的智慧政务系统，摒弃“鸵鸟心态”，防范“制度空转”，实现网络问政的高质量发展，成为新时代社会治理创新发展的新课题。

1.2 问题重述

基于上述背景，如何把握大数据、云计算、人工智能等技术带来的崭新机遇，建立基于自然语言处理技术的智慧政务系统，是一个具有重要而深远意义的研究项目。就本文而言，需根据收集自互联网公开来源的群众问政留言记录，完成以下任务：

问题 1：对群众留言进行分类

网络问政平台的群众留言数量繁多、种类复杂，对留言按照一定的体系标准进行分类，是后续将留言分派至相应职能部门进行回复处理的必要之举。囿于当前网络问政数据技术的局限性，现如今，大部分电子政务系统还是依靠人工处理，存在工作量大、效率低下的问题，开发群众留言分类系统已成为当务之急。根据附件 2 的数据内容，完成关于留言内容的一级标签分类模型，并利用 F-score 对分类方法进行评价是本文要解决的首要问题。

问题 2：对热点问题进行挖掘

所谓热点问题,一般是指在一段时间内群众普遍关注、最感兴趣、具有切实意义的问题。使问题成为热点，并引起群众的普遍关心，往往是有关问题方面发现不及时、解决不得力造成的。及时发现热点问题，有助于相关责任部门进行针对性处理，明确分工，责任到位。在此问，我们需要通过合理的手段，对一段时间内反映特定地点或特定人群的问题进行分类，并定理合理的热度评价指标，确定热点问题。

问题 3：对答复意见进行评价

在网络问政中，群众留言是前提条件，政府反馈才是关键环节，如果没有反馈环节，则只是徒增问政形式。张哲也指出公众参与和政府回应是民主政治研究问题的两个方面，及时有效的政府回应是公众实质性参与的先决条件和重要考虑因素[4]。那么政府对公众的回应情况究竟如何，回应是否及时有效？答复是否相关完整？多角度出发对答复意见的质量给出一套评价方案，是此问需要解决的重点问题。

第 2 章 群众留言分类

2.1 数据预处理

2.1.1 文本数据增强——EDA

文本数据共 9210 条，存在较为明显的类别不均衡情况，这容易导致模型对于小样本类别往往处于欠拟合状态，在实际预测时，几乎不会对这一类别给予太高的概率。

近年来，文本数据增强领域快速发展，基于回译方法的文本数据增强成为了质量高又几乎无技术门槛的通用文本增强技术，但其对长文本的支持较弱。此外还有非核心词替换、基于上下文信息的文本增强、基于语言生成模型的文本增强等等技术方法。

本文一方面考虑到在小样本场景下，采用 EDA 文本增强技术进行模型性能的提升，简单而有效；另一方面考虑到与其他方法相比，EDA 方法实现代价较低，结合初次调试 XGBoost 和 TextCNN 得到的结果，本文对环境保护、商贸旅游以及卫生计生三类利用 EDA 进行文本增强，模型的 F1-score 均提升了约 5%。

2.1.1 文本数据增强——EDA

2.1.1.1 EDA 算法原理

EDA，即简单数据增强(easy data augmentation)，包括了四种方法：同义词替换、随机插入、随机交换、随机删除，这四种方法的具体操作如下所示[5]：

同义词替换(Synonym Replacement, SR)：从句子中随机选取 n 个不属于停用词集的单词，并随机选择其同义词替换它们；

随机插入(Random Insertion, RI)：随机的找出句中某个不属于停用词集的词，并求出其随机的同义词，将该同义词插入句子的一个随机位置。重复 n 次；

随机交换(Random Swap, RS)：随机的选择句中两个单词并交换它们的位置。重复 n 次；

随机删除(Random Deletion, RD)：以 p 的概率，随机的移除句中的每个单词。

运用 EDA 方法可以生成类似于原始数据的增强数据，引入一定程度的噪声，有助于防止过拟合；此外，EDA 可以通过同义词替换和随机插入操作引入新的词汇，允许模型泛化到那些在测试集中但不在训练集中的单词。基于以上两点原

因，EDA 可以有效地提高文本分类的效果。

2.1.1.2 文本增强操作

针对交通运输类别数据，随机抽取 40%左右的文本利用 EDA 算法进行文本增强，并且随机选择一条生成的文本(四个方法都会生成文本)；针对商贸旅游类别数据，随机抽取 10%；针对卫生计生类别数据，随机抽取 15%，进行相同处理。具体处理结果如下所示：

表 1　EDA 处理设定

EDA	同义词替换	随机插入	随机删除	随机替换
代码参数	alpha_sr	alpha_ri	alpha_rs	alpha_rd
概率	25%	25%	25%	25%
新文本	12			

表 2　文本增强处理结果

类别	城乡建设	环境保护	交通运输	教育文体	劳动和社会保障	商贸旅游	卫生计生
Class	0	1	2	3	4	5	6
个数	2009	938	613	1589	1969	1215	877
文本增强	2009	938	870	1589	1969	1452	1054

2.2 基于留言详情的分类——XGBoost

本文首先选择留言详情作为分类依据，建立关于留言详情的一级标签分类模型。基于留言详情的分类，本文将采用陈天奇等人开发的开源机器学习项目 XGBoost 来进行处理[6]。

2.2.1 词向量技术

由于文本数据的规范性不足，本文对获取到的文本数据进行规范化处理，排除掉句子中包含的错误信息、不规范信息等等问题。本文首先去除数字、字母和特殊符号等，并用 Python 的中文分词库 jieba 分词。进一步地建立停用词集合，对已经进行分词处理后的文本词语集合进行停用词处理。

文本预处理结束后产生的文本信息，仍需要将得到的语料转换成数值型向量，才能够转变成机器学习的分类器能够识别的结构化向量数据。本文运用的是 Word2Vec 词向量技术。

Word2Vec 词向量技术具有两个方法来产生词向量[7,8]：连续词袋模型(CBOW)和 Skip-gram 模型。

1.连续词袋模型(CBOW)

连续词袋模型的主要思路就是利用当前词的上下文表示C_{ij}，预测当前词的表示W_{ij}，对于给定的文本信息D，使用 CBOW 模型主要是最大化文本信息的后验概率，计算方式如下所示[9]：

$$\arg\max_{\theta}\prod_{j}^{D}[\prod_{ij=1}^{T_j}p(w_{ij}\mid C_{ij},\theta)] \tag{1}$$

其中，T_j代表文本信息D中的第j个句子，w_{ij}代表第j个句子中的第i个词，C_{ij}代表为w_{ij}的上下文表示，θ代表后验概率参数。计算条件概率首先需要把文本信息包含的词汇转换为词向量矩阵$M\in R^{N\times d}$，M代表文本信息每一行中的词向量，N代表文本信息包含的词汇表，d代表词向量的维度，然后对于给定当前词w_{ij}的上下文表示C_{ij}，利用神经网络模型将当前词的条件概率$p(w_{ij}\mid C_{ij},\theta)$进行最大化处理。

对于后验概率参数θ的计算主要有 Hierarchical Softmax 和 Negative Sampling 两种方式，本文主要介绍 Hierarchical Softmax 方法，模型结构如图 1 所示，其网络结构主要包括三层，分别为输入层，投影层和输出层。

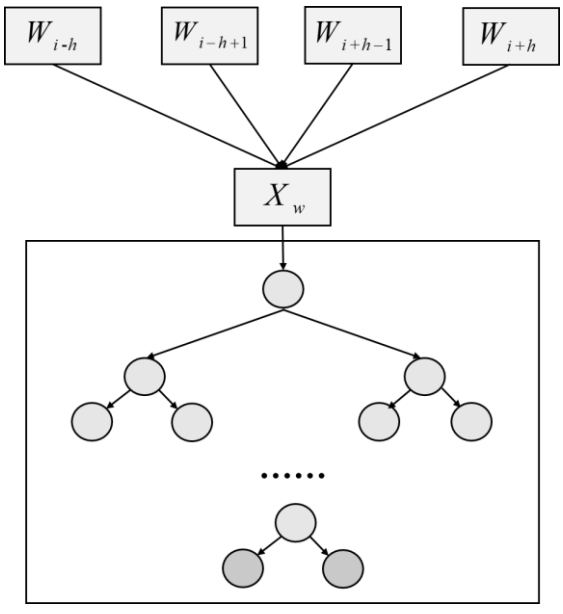

图 1　CBOW Hierarchical Softmax 模型结构图

(1) 输入层：输入层主要通过利用词向量矩阵寻找当前词的上下文向量表示 $[w_{i-h}, w_{i-h+1}, \cdots, w_{i-1}, w_{i+1}, \cdots, w_{i+h-1}, w_{i+h}] \in R^{2h \times d}$，进一步地将其作为输入向量。

(2)投影层：投影层的主要作用是将输入层的向量做求和处理，得到维度为 d 的中间层向量。

(3)输出层：输出层对应一颗二叉树，它是以预料中出现过的词当叶子结点，以各词在语料中出现的次数当权值构造出来的 Huffman 树。因此，当前词的条件概率为：

$$P_H(w_{ij} \mid C_{ij}; \theta) = \prod_{k_{ij}=1}^{K_{ij}} (\sigma(q_{k_{ij}} \cdot C_{ij})^{1-d_{k_{ij}}} \cdot (1-\sigma(q_{k_{ij}} \cdot C_{ij}))^{d_{k_{ij}}}) \tag{2}$$

其中，$d_{k_{ij}}$ 代表叶子节点在非叶子节点中的位置，若是在非叶子节点的右子树中则代表 1，若是在非叶子节点的左子树中则代表 0；$q_{k_{ij}}$ 代表从根结点到叶子节点所经过路径下的第 k 个非叶子节点向量，σ 代表 softmax 函数，利用随机梯度下降法可以进行参数更新。

2. Skip-gram 模型

如上所述，CBOW 是已知当前词的上下文，来预测当前词，而 Skip-gram 则恰恰相反，是在已知当前词的情况下，预测其上下文，最大化文本信息的后验概率的计算方式为：

$$\arg\max_{\theta} \prod_{w_{ij}}^{D} [\prod_{c \in C_{ij}} p(c \mid w_{ij}, \theta)] \tag{3}$$

与 CBOW 相同，有 Hierarchical Softmax 和 Negative Sampling 两种方式提高训练结果的准确率和加速训练过程，本文同样主要介绍 Hierarchical Softmax 方法。模型结构如图 2 所示，该模型不再考虑投影层，而是直接将输入层与输出层的二叉树非叶结点相连接。输出层的概率分布转变成为 Huffman 树，每个叶子节点表示词表中的一个词，非叶子节点是将词向量分类到具体的孩子节点上。

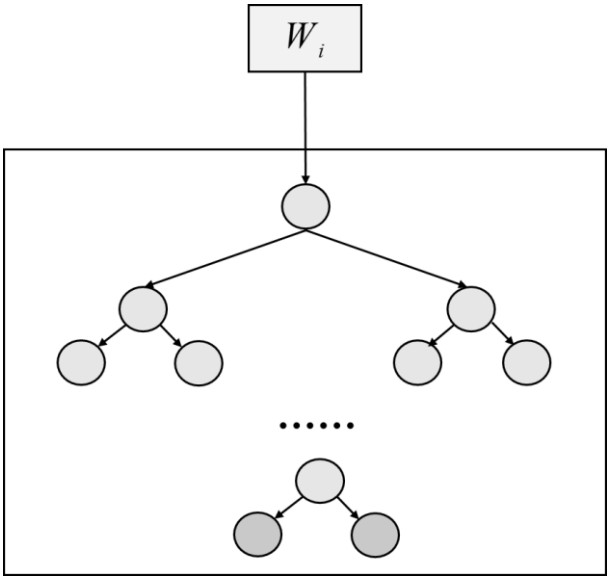

图 2　Skip-gram Hierarchical Softmax 模型结构图

概率计算公式如下所示：

$$P_H(c \mid w_{ij};\theta)=\prod_{k_{ij}=1}^{K_{ij}}(\sigma(q_{k_{ij}}\cdot c)^{1-d_{k_{ij}}}\cdot(1-\sigma(q_{k_{ij}}\cdot c)^{d_{k_{ij}}}) \tag{4}$$

上述公式中，w_{ij} 代表当前词，c 代表上下文信息中的词。

维基百科的中文语料是目前公认的大型中文语料库，因此，本节经过对比尝试，选择基于维基百科中文语料库的利用 Skip-gram 训练 400 维的 Word2Vec 词向量模型，并从生成的词向量中抽取待研究留言的特征词向量。

2.2.2 XGBoost 分类

2.2.2.1 XGBoost 原理

极端梯度提升(extreme gradient boosting，XGBoost)是众多 boosting 算法的其中一种[10]，最初是由采用陈天奇等人开展的一项研究项目，在 Kaggle 的希格斯自信号识别竞赛中崭露头角。

XGBoost 本质上是一种提升树模型，该算法汇集多个分类树模型，得到一个很强的分类器。XGBoost 的基本思想主要是通过对特征进行分类，进而生长出一棵树，并且不断添加树的数量。添加一棵树的过程，就是对一个新的函数进行学习，并将学习到的结果用来对上次预测的残差值进行拟合的过程。

Xgboost 模型的目标函数用数学函数表示如下[11]：

$$L(\theta)=\sum_i l(\tilde{y},y_i)+\sum_k \Omega(f_k) \tag{5}$$

其中

$$\Omega(f)=\Upsilon T+\frac{1}{2}\lambda\|w\|^2 \tag{6}$$

$L(\theta)$为损失函数，Ω代表模型复杂程度的惩罚项，Υ为L_1正则化的系数，λ为L_2正则化的系数。目标函数可化为：

$$L(\theta)=\sum_{j=1}^{T}[(\sum_{i\in I_j}g_i)w_j+\frac{1}{2}(\sum_{i\in I_j}h_i+\lambda)w_j^2]+\gamma T \tag{7}$$

令：

$$G_i=\sum_{i\in I_j}g_i,\ \ H_i=\sum_{i\in I_j}h_i \tag{8}$$

则目标函数可变换为：

$$L(\theta)=\sum_{j=1}^{T}[G_jw_j+\frac{1}{2}(H_j+\lambda)w_j^2]+\gamma T \tag{9}$$

此时，可以将目标函数改写成关于叶子结点分数w的一个一元二次函数，可以用顶点公式求解最优w和最优的目标函数值，最优权重为：

$$w_j=-\frac{G_j}{H_j+\lambda} \tag{10}$$

则目标函数为：

$$L(\theta)=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j+\lambda}+\Upsilon T \tag{11}$$

2.2.2.2 XGBoost 分类

在本文具体实现中，首先选择基于维基百科中文语料库的利用 Skip-gram 训练 400 维的 Word2Vec 词向量模型，在此基础上，利用 XGBoost 算法实现基于留言详情的文本分类。考虑到 Softmax 回归适合处理一个样本尽可能属于一种类别的多分类问题，因此本文采用 softmax 目标函数处理多分类问题。

1.混淆矩阵

在机器学习领域，混淆矩阵(confusion matrix)，又称为可能性表格或是错误矩阵。它是一种特定的矩阵用来呈现算法性能的可视化效果，通常是监督学习(非监督学习，通常用匹配矩阵：matching matrix)。其每一列代表预测值，每一行代表的是实际的类别。这个名字来源于它可以非常容易的表明多个类别是否有混淆

(也就是一个 class 被预测成另一个 class)。

XGBoost 分类最终得到的混淆矩阵如下所示：

$$\begin{Bmatrix} 378 & 0 & 0 & 0 & 0 & 16 & 3 \\ 7 & 304 & 9 & 7 & 7 & 18 & 35 \\ 3 & 21 & 163 & 0 & 0 & 0 & 7 \\ 3 & 5 & 0 & 283 & 0 & 0 & 5 \\ 4 & 13 & 0 & 0 & 265 & 15 & 9 \\ 22 & 11 & 1 & 4 & 6 & 343 & 2 \\ 4 & 8 & 0 & 3 & 4 & 4 & 526 \end{Bmatrix}$$

从上述结果可以看出，对于**城乡建设类别文本**，实际上应该有 378 条城乡建设主题留言，但是系统错误的将其中 16 条预测成了商贸旅游类别文本，3 条预测成了卫生计生类别文本。类似地，对于**环境保护类别文本**，实际上应该有 304 条城乡建设主题留言，但是系统错误的将其中 7 条留言预测成了城乡建设类别文本，9 条留言预测成了交通运输类别文本，7 条预测成了教育文体类别文本，7 条预测成了劳动和社会保障类别文本，18 条预测成了商贸旅游类别文本,35 条预测成了卫生计生类别文本。其余几类以此类推，本文不再赘述。

2.F-score

分类结果的好坏一般通过准确度和效率两个指标来判断，准确度主要用来判断分类器的分类能力[1]。效率主要从分类器在分类过程中所消耗的资源和时间来考虑，但由于现阶段计算机配置等因素的影响，目前分类结果评价中暂不考虑效率这个指标的影响，主要从准确度指标的高低来体现。目前，衡量准确度的优劣一般用查准率，查全率和 F1 值三种评估指标，其中查准率主要是判断分类器把其他类别中的文本划分到研究类别中的程度；查全率则是判断分类器把研究类别中的文本划分到其他类别中的程度；F1 值则是两个比值的中和。为了更容易的描述三种指标的公式，我们建立如下分类结果表，如表 3 所示：

表 3　人工判定与系统分类结果对比表

类别归属	人工判定属于此类	人工判定不属于此类
分类系统判定属于此类	X	Y
分类系统判定不属于此类	M	N

其中 X、N 表示分类系统与人工判定结果一致的情况，Y、M 表示分类系统与人工判定结果不一致的情况。

查准率 P 的公式如下所示：

$$P=\frac{X}{X+Y} \tag{12}$$

查全率 R 的公式如下所示

$$R=\frac{X}{X+M} \tag{13}$$

F1 值是将查准率 P 和查全率 R 综合考虑，公式如下所示：

$$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{14}$$

本文应七级分类，共七个模型，F1 值在 81.18%-95.45%之间，分类准确率如表 4 所示。

表 4　XGBoost 分类结果准确率

类别	准确率	类别	F1 值
城乡建设	92.43%	劳动和社会保障	90.14%
环境保护	81.17%	商贸旅游	87.39%
交通运输	88.83%	卫生计生	92.61%
教育文体	95.45%	**平均**	**89.76%**

2.2 基于留言主题的分类——TextCNN

为了确保分类结果的准确性、完整性，本文在选择留言详情作为分类依据的基础上，进一步地，还将选择留言主题作为分类依据，建立关于留言内容的一级标签分类模型。基于留言主题的分类，本文将采用 Yoon Kim 于 2014 年在 CNN 输入层改进基础上提出的文本分类模型 TextCNN。

2.2.1 划分数据集

基于 EDA 文本数据增强后的 9873 条数据集，本文按照 8:1:1 的比例随机划分为训练集、验证集、测试集，数据内容为标题+类别。数据集划分结果如下表 5 所示：

表 5　数据集划分

数据集	训练集	验证集	测试集
数据量	7896	981	996

2.2.2 TextCNN 分类过程梳理

2.2.2.1 词表 vocab 构建

在进行 NLP 相关编码工作时，将文本进行序列化编码是一个必要的环节。本文对留言主题以字为单位构建词表 vocab，通过短填长切(句子大于设定参数会进行自动的剪切，如果小于便会自动填充)，将所有句子处理成一个长度。本文在实际处理时，将参数设定为 32，一次训练所选取的样本数设定为 128。

2.2.2.2 Word Embedding 构建词向量

Chinese-Word-Vectors 是由北京师范大学和人民大学研究者开源出来的 100 多个中文预训练词向量，其中所有向量都是在 word2vec 和 skip-gram 上训练出来的。

本文运用 Chinese-Word-Vectors 词向量集合技术中的预处理词向量进行处理。该词向量是通过上文所述的 word2vec 方式处理的、以搜狗新闻为数据背景训练出来的维数为 300 维的预处理词向量。进一步地，将词表 vocab 中的词(字)与预处理词向量的词比对，找到本题词表对应的 300 维词向量，拼成 4672*300 的矩阵，作为最初的输入。此外，需要指出的是，由于本文一次训练选取 128 个样本数，因此每次输入为 128*300。

2.2.2.3 Convolution 卷积

卷积一词最开始出现在信号与线性系统中，信号与线性系统中讨论的主要就是信号经过一个线性系统以后会产生的变化。从物理意义上可以理解为系统某一时刻的输出是由多个输入共同作用(叠加)的结果。

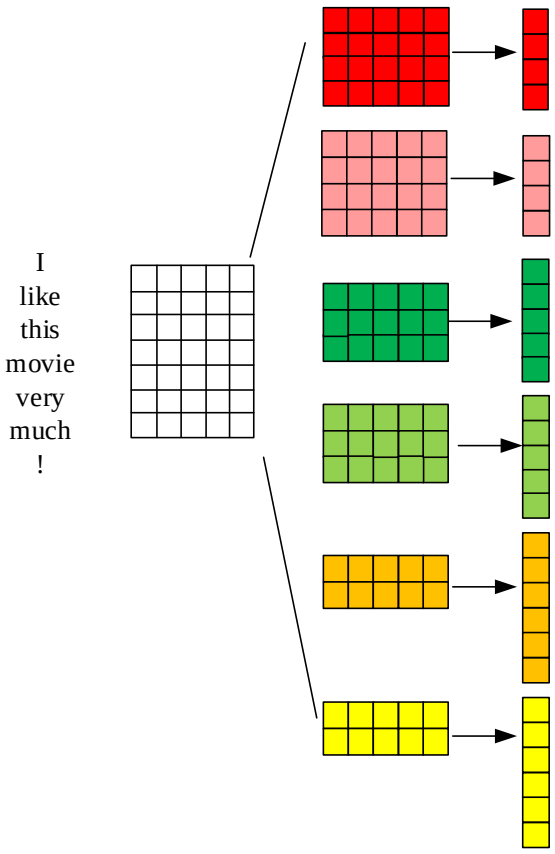

图 3　卷积过程示例图

在 TextCNN 分类中，通过卷积操作，将输入的矩阵映射成一个更小的矩阵，feature_map 是卷积之后的输出。

在计算机视觉上，我们的过滤器(卷积)在图像的局部区域上滑动，但是在 NLP 中，我们通常使用滑过整行矩阵(单词)的过滤器。因此，我们的滤波器的“宽度”通常与输入矩阵的宽度相同。对于矩阵的宽度其实就是一个单词转化为的词向量的长度，然后这个滤波器相当于一次遍历多个词向量，而且可以有不同的滤波器。高度或区域大小可能会有所不同，通常使用一次滑动 2-5 个字的窗口。每个过滤器(即卷积)对句子矩阵执行卷积并生成(可变长度)特征映射。

对示例图图 3 的卷积分析如下所示：

图 3 中，输入为(7,5)，7 个字，每个字特征 5 维。卷积核尺寸(高度)为 2,3,4，卷积核尺寸控制了窗口的字数量，卷积核宽度一般与输入的维度一致。

若卷积核尺寸为 2，步长为 1，如上图所示，先看 I like，再看 like this，再滑动到 this movie；类似地，若卷积核尺寸为 3，步长为 1，如上图所示，先看 I like this，再看 like this movie，再滑动到 this movie very。

图中不同尺寸的卷积核个数各有 2 个，因此图中中间层可以看到有 2 个 4*5

矩阵(红色)、2 个 3*5 矩阵(绿色)，2 个 2*5 矩阵(黄色)。假设这次选择的卷积核尺寸为 2，则 7 个字需要用到(I like)(like this)(this movie)(movie very)(very much)(much !)6 次滑动，一次滑动等于(I like)对应的(2*5)矩阵与卷积(2*5)做卷积操作，那么经过 6 次操作就变成了 6*1 维的 feature map，此外，由于每次有 2 个尺寸为 2 的卷积核，因此输出为 2 个 6*1 的 feature map(即第三层黄色所示)。以此类推，假设这次选择的卷积核尺寸为 3，则输出 2 个 5*1 的 feature map(即第三层绿色所示)；假设这次选择的卷积核尺寸为 4，则输出 2 个 4*1 的 feature map(即第三层红色所示)。因此选择不同的尺寸的卷积核会有不同的 feature map 输出。

在示例图卷积分析的基础上，对应于本题的处理内容如下所示：

本题输入为(128,300)，同样选择了尺寸为 2,3,4 的卷积核，也就是说卷积核高度为 2 或 3 或 4，但宽度为 300(与词向量维度统一)(上图中宽度为 d=5)，另外每个尺寸的卷积核有 256 个。因此，本题中间层可以看到 256 个 4*300 矩阵(对应红色)，256 个 3*300 矩阵(对应绿色)，256 个 2*300 矩阵(对应黄色)。在第三层，本题可以看到 256 个 125*1 的 feature map(对应红色)，256 个 126*1 的 feature map(对应绿色)，256 个 127*1 的 feature map(对应黄色)。

2.2.2.4 max-pooling 最大池化

池化(Pooling)是卷积神经网络中另一个重要的概念，它实际上是一种形式的降采样。有多种不同形式的非线性池化函数，而其中“最大池化(Max pooling)”是最为常见的。它是将输入的图像划分为若干个矩形区域，对每个子区域输出最大值。池化操作一方面可以使模型更加关注是否存在某些特征而不是特征具体的位置。另一方面，由于池化层会不断地减小数据的空间大小，因此参数的数量和计算量也会下降，这在一定程度上也控制了过拟合。

池化层通常会分别作用于每个输入的特征并减小其大小。当前最常用形式的池化层是每隔 2 个元素从图像划分出 2*2 的区块，然后对每个区块中的 4 个数取最大值，如图 4 所示。

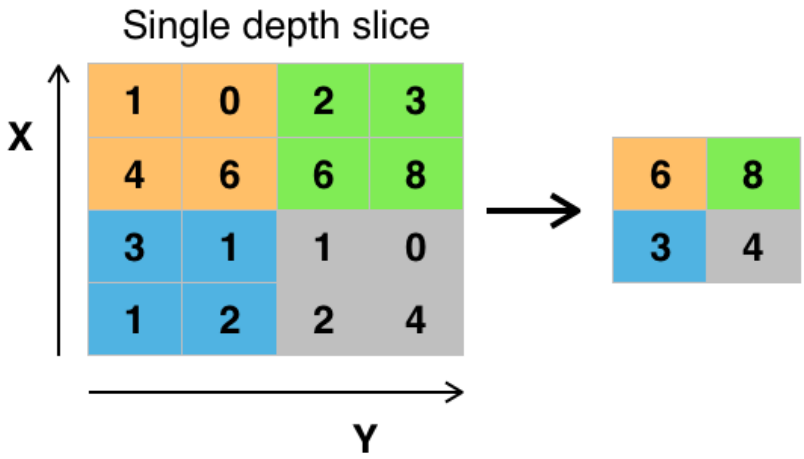

图 4　池化原理图

同样地，对示例图图 5 的最大池化过程分析如下所示：

经过卷积操作后生成(可变长度)特征映射。然后在每个特征图上执行 1-max 池化，即记录来自每个特征图的最大数。因此，从六个特征图生成一元特征向量，并且这六个特征被连接以形成倒数第二层(6,1)(图中最右层)的特征向量。

在最大池化过程分析的基础上，对应于本题的处理内容如下所示：

本题对卷积后的 768 个特征图(256*3)做全局最大池化，从 768 个不同的特征图生成 768 个一元特征向量，连接这 768 个特征形成(768,1)的特征向量。

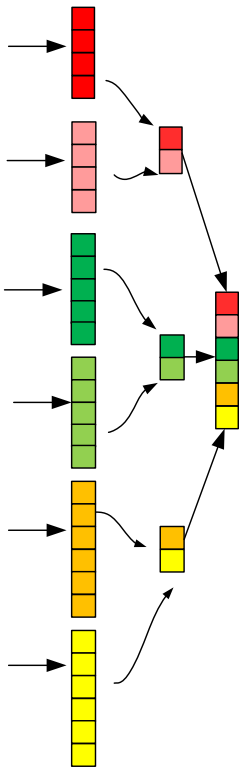

图 5　最大池化过程

2.2.2.5 softmax k 分类

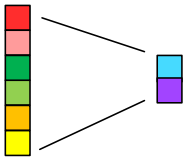

图 6　Softmax 分类过程

最后的 softmax 层接收上一层的特征向量作为输入，并用它来分类句子。

对示例图图 6 的 Softmax 分类过程如下所示：

图 6 是 2 分类，我们将 max-pooling 的结果拼接起来，送入到 softmax 当中，得到各个类别比如 label 为 1 的概率以及 label 为-1 的概率。预测处理的情况下，到这里整个 textCNN 的流程便告一段落。训练的情况下，此时便会根据预测 label 以及实际 label 来计算损失函数，计算出 softmax 函数，max-pooling 函数，激活函数以及卷积核函数四个函数当中参数需要更新的梯度，来依次更新这四个函数中的参数，完成一轮训练 。

在 Softmax 分类分析的基础上，对应于本题的处理内容如下所示：

本题设置了 7 分类，输入为经过最大池化的 768 个特征，输出为 7 个分类。

2.2.2.6 TextCNN 流程综述

根据上述内容总结出的 TextCNN 流程完整图例如下所示：

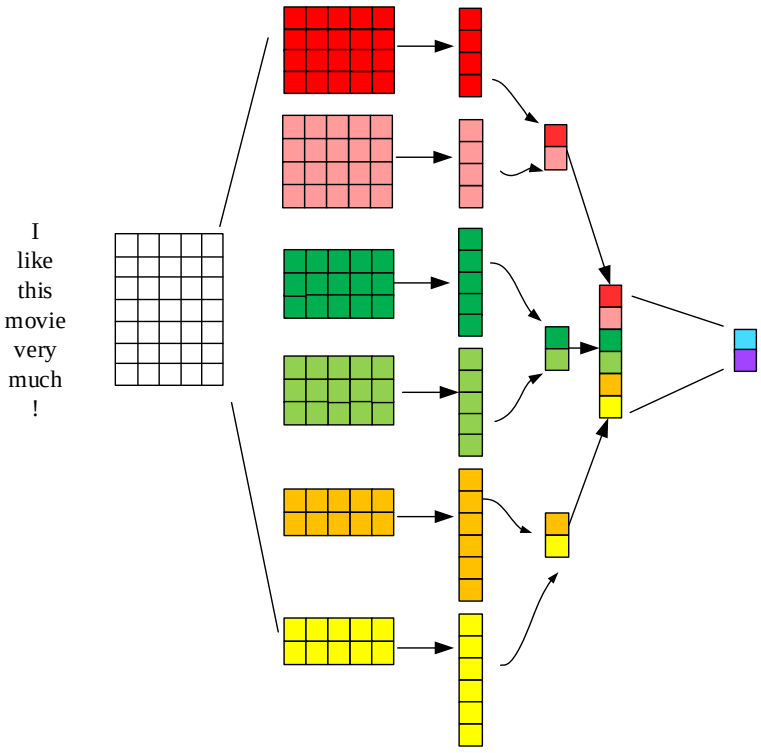

图 7　TextCNN 流程图

相对应地，本文运行过程中的部分数据如图 8 所示：

```
(embedding): Embedding(4762, 300)
(convs): ModuleList(
  (0): Conv2d(1, 256, kernel_size=(2, 300), stride=(1, 1))
  (1): Conv2d(1, 256, kernel_size=(3, 300), stride=(1, 1))
  (2): Conv2d(1, 256, kernel_size=(4, 300), stride=(1, 1))
)
(dropout): Dropout(p=0.5, inplace=False)
(fc): Linear(in_features=768, out_features=7, bias=True)
```

图 8　TextCNN 运行数据

2.2.3 TextCNN 分类结果展示

将信息增强后的数据集训练 30 代(Epoch=30)结果如下所示，Test Acc 达到 95.28%。

```
Loading data...
Vocab size: 4762
7896it [00:00, 41045.81it/s]
981it [00:00, 40899.50it/s]
996it [00:00, 70614.04it/s]
Time usage: 0:00:00
<bound method Module.parameters of Model(
  (embedding): Embedding(4762, 300)
  (convs): ModuleList(
    (0): Conv2d(1, 256, kernel_size=(2, 300), stride=(1, 1))
    (1): Conv2d(1, 256, kernel_size=(3, 300), stride=(1, 1))
    (2): Conv2d(1, 256, kernel_size=(4, 300), stride=(1, 1))
  )
  (dropout): Dropout(p=0.5, inplace=False)
  (fc): Linear(in_features=768, out_features=7, bias=True)
```

```
Epoch [1/30]
Iter:      0,  Train Loss:   2.1,  Train Acc:  0.00%,  Val Loss:   3.5,  Val Acc: 18.45%,  Time: 0:00:02 *
Epoch [2/30]
Iter:    100,  Train Loss:   2.8,  Train Acc:  0.00%,  Val Loss:   2.2,  Val Acc: 16.41%,  Time: 0:01:06 *
Epoch [3/30]
Epoch [4/30]
Iter:    200,  Train Loss:   1.9,  Train Acc: 25.78%,  Val Loss:   1.6,  Val Acc: 39.35%,  Time: 0:02:10 *
Epoch [5/30]
Iter:    300,  Train Loss:   1.9,  Train Acc:  7.03%,  Val Loss:   1.4,  Val Acc: 52.91%,  Time: 0:03:13 *
Epoch [6/30]
Epoch [7/30]
Iter:    400,  Train Loss:  0.67,  Train Acc: 79.69%,  Val Loss:   1.0,  Val Acc: 66.16%,  Time: 0:04:17 *
Epoch [8/30]
Epoch [9/30]
Iter:    500,  Train Loss:   1.2,  Train Acc: 53.12%,  Val Loss:  0.86,  Val Acc: 65.24%,  Time: 0:05:25 *
Epoch [10/30]
Iter:    600,  Train Loss:   0.2,  Train Acc: 92.97%,  Val Loss:   1.8,  Val Acc: 54.64%,  Time: 0:06:34
Epoch [11/30]
Epoch [12/30]
Iter:    700,  Train Loss:   1.1,  Train Acc: 64.84%,  Val Loss:   1.1,  Val Acc: 65.24%,  Time: 0:07:39
Epoch [13/30]
Iter:    800,  Train Loss:   1.4,  Train Acc: 54.69%,  Val Loss:  0.66,  Val Acc: 76.96%,  Time: 0:08:41 *
Epoch [14/30]
Epoch [15/30]
Iter:    900,  Train Loss:  0.43,  Train Acc: 85.94%,  Val Loss:  0.42,  Val Acc: 86.44%,  Time: 0:09:41 *
Epoch [16/30]
Epoch [17/30]
Iter:   1000,  Train Loss:  0.56,  Train Acc: 81.25%,  Val Loss:  0.33,  Val Acc: 89.50%,  Time: 0:10:41 *
Epoch [18/30]
Iter:   1100,  Train Loss:  0.45,  Train Acc: 86.72%,  Val Loss:  0.32,  Val Acc: 91.13%,  Time: 0:11:42 *
Epoch [19/30]
Epoch [20/30]
Iter:   1200,  Train Loss:  0.24,  Train Acc: 92.19%,  Val Loss:  0.28,  Val Acc: 92.25%,  Time: 0:12:43 *
Epoch [21/30]
Iter:   1300,  Train Loss:   0.3,  Train Acc: 89.06%,  Val Loss:  0.23,  Val Acc: 93.58%,  Time: 0:13:44 *
Epoch [22/30]
Epoch [23/30]
Iter:   1400,  Train Loss:  0.24,  Train Acc: 92.19%,  Val Loss:  0.23,  Val Acc: 93.88%,  Time: 0:14:44
Epoch [24/30]
Epoch [25/30]
Iter:   1500,  Train Loss:  0.16,  Train Acc: 94.53%,  Val Loss:   0.2,  Val Acc: 94.29%,  Time: 0:15:44 *
Epoch [26/30]
Iter:   1600,  Train Loss:  0.25,  Train Acc: 91.41%,  Val Loss:  0.21,  Val Acc: 94.90%,  Time: 0:16:45
```

```
Epoch [26/30]
Iter:   1600,  Train Loss:  0.25,  Train Acc: 91.41%,  Val Loss:  0.21,  Val Acc: 94.90%,  Time: 0:16:45
Epoch [27/30]
Epoch [28/30]
Iter:   1700,  Train Loss: 0.063,  Train Acc: 98.44%,  Val Loss:   0.2,  Val Acc: 95.31%,  Time: 0:17:45 *
Epoch [29/30]
Epoch [30/30]
Iter:   1800,  Train Loss:  0.25,  Train Acc: 95.31%,  Val Loss:  0.17,  Val Acc: 96.13%,  Time: 0:18:46 *
Test Loss:  0.17,  Test Acc: 95.28%
Precision, Recall and F1-Score...
               precision    recall  f1-score   support

  Construction     0.9632    0.9385    0.9506       195
   Environment     0.9320    0.9796    0.9552        98
Transportation     0.9444    0.9551    0.9497        89
     Education     0.9497    0.9869    0.9679       153
         Labor     0.9528    0.9528    0.9528       212
       Tourism     0.9493    0.9493    0.9493       138
        Health     0.9712    0.9099    0.9395       111

      accuracy                         0.9528       996
     macro avg     0.9518    0.9532    0.9522       996
  weighted avg     0.9531    0.9528    0.9527       996

Confusion Matrix...
[[183   5   2   1   1   3   0]
 [  1  96   0   0   1   0   0]
 [  1   0  85   0   2   1   0]
 [  1   0   0 151   0   1   0]
 [  2   1   0   3 202   1   3]
 [  1   0   3   1   2 131   0]
 [  1   1   0   3   4   1 101]]
Time usage: 0:00:01
```

图 9　TextCNN 分类结果

TextCNN 分类最终得到的混淆矩阵如下所示：

$$\begin{Bmatrix} 183 & 5 & 2 & 1 & 1 & 3 & 0 \\ 1 & 96 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 85 & 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 151 & 0 & 1 & 0 \\ 2 & 1 & 0 & 3 & 202 & 1 & 3 \\ 1 & 0 & 3 & 1 & 2 & 131 & 0 \\ 1 & 1 & 0 & 3 & 4 & 1 & 101 \end{Bmatrix}$$

从上述结果可以看出，对于**城乡建设类别文本**，实际上应该有 195 条城乡建设主题留言，但是系统错误的将其中 5 条留言预测成了环境保护类别文本，2 条留言预测成了交通运输类别文本，1 条预测成了教育文体类别文本，1 条预测成了劳动和社会保障类别文本，3 条预测成了商贸旅游类别文本。类似地，对于**环境保护类别文本**，系统错误的将其中 1 条留言预测成了城乡建设类别文本，1 条预测成了劳动和社会保障类别文本。其余几类以此类推，本文不再赘述。

与 XGBoost 处理类似，TextCNN 应七级分类，共七个模型，F1 值在 93.95%-96.79%之间，分类准确率如表 6 所示。

表 6　TextCNN 分类结果准确率

类别	准确率	类别	F1 值
城乡建设	95.06%	劳动和社会保障	95.28%
环境保护	95.52%	商贸旅游	94.93%
交通运输	94.97%	卫生计生	93.95%
教育文体	96.79%	**平均**	**95.28%**

第 3 章 热点问题挖掘

Word2Vec 一方面可以有效识别单词之间的重要关系，这使得它在许多 NLP 项目案例中非常有用；另一方面，Word2Vec 非常擅长在命名实体识别中找出相似性，因此名称实体识别也可以使用 Word2Vec，通过将所有类似的实体聚集在一起获得更好的结果。基于上述优点，本文首先基于 Word2Vec 计算文本相似度，并利用热度概率选取潜在热点标题。其次，利用 Word Mover's Distance 对潜在热点标题再次归类后获得类中心，进一步地针对类中心引入时间变量，通过改良 K-means 聚类，剔除错误分类数据。最后，计算热度指数以实现热点问题的挖掘。具体操作流程如图 10 所示：

图 10　热点问题挖掘操作流程图

3.1 WMD 算法简介

Word Mover's Distance(WMD)算法是 Matt 等人于 2015 年提出的计算文本文档距离的方法，是将 Earth Mover's Distance (EMD)和词嵌入结合起来，通过计算一篇文本中的词语完全转移到另一篇文本所需要的最短距离来衡量文本间的差异性的一种方法。

WMD 算法是在 EMD 算法基础上改进得来的，本文首先对 EMD 算法做一个简单的介绍[12]。EMD 算法主要应用于图像和语音信号处理领域，同时 EMD 算法可以简化为线性规划中运输问题的最优解问题。假设某运输问题有 P 和 Q 两个地方，需要将货物从 P 地的工厂运输到 Q 地的仓库，P 地有 $P_1,P_2,\cdots,P_m$ 共 m 座工厂，工厂 P_i 有重量为 w_{P_i} 的货物；Q 地有 $Q_1,Q_2,\cdots,Q_n$ 共 n 个仓库，仓库 Q_j 的容量为 w_{Q_j}。该运输问题的优化目标是采取何种方式可以尽可能高效地将所有货物从 P 地运输到 Q 地。定义货物从工厂 P_i 运输到仓库 Q_j 距离为 d_{ij}，运送货物的重量为 e_{ij}，则一次运输的工作量为 $d_{ij}\cdot e_{ij}$。不难看出当运送距离越远时运输工作量就越大，工作量总和的最小值 W 为[13]：

$$W=\sum_i^m\sum_j^n d_{ij}e_{ij}\rightarrow\min \tag{15}$$

同时需要满足以下条件：

$$\begin{cases} e_{ij}\geq 0,\ (1\leq i\leq m,1\leq j\leq n) \\ \sum_{i=1}^{n} e_{ij}\leq w_{A_i},\ 1\leq i\leq m \\ \sum_{i=1}^{m} e_{ij}\geq w_{B_j},\ 1\leq j\leq n \\ \sum_{i=1}^{m}\sum_{j=1}^{n} e_{ij}=\min(\sum_{i}^{m} w_{A_i},\sum_{i}^{m} w_{B_j}) \end{cases} \tag{16}$$

通过训练词向量得到文本中词语的低维分布向量后，进一步地，计算向量间的余弦距离或欧式距离即 d_{ij}，便可将 EMD 距离引入自然语言处理领域。

WMD 距离是依靠 Word2Vec 模型生成的高质量和大规模的数据集中的 word embedding 工具实现的。由于文本语言是由词组成的，因此 Word2Vec 将每一个词表示成一定维度的向量，如果该词在第三个位置出现，那么就将第三个位置的值设为 1，其余设为 0，基于上述设置就可以对所有样本进行神经网络的训练直到收敛。收敛后可以获得权重，进一步地将这些权重作为每一个词的

向量。此外，在 Word2Vec 中使用了 Huffman 树，这样的话就可以根据上下文来推测这个词的概率。WMD 算法的具体计算过程如下所示：

(1)单词向量化

使用 word2vec 的向量化矩阵 $X \in R^{d \times n}$ 来表示一个有 n 个词语的词汇表，第 i 列表示在 d 维空间中第 i 个单词的向量。首先运用归一化的词袋模型将文档表示为归一化的 n 维的词袋向量 d[14]。一个单词 i 的词频权重为：

$$D_i = \frac{c_i}{\sum_{j=1}^{n} c_j} \tag{17}$$

其中 c_i 表示词语 c_i 在文档中出现的次数。

(2)文档距离

由于两个语义相似的文本因为所用词汇完全不同而导致文档向量的非零部分分布在不同部分的情况的存在，例如，“特朗普会见媒体”和“总统对话记者”。将上述两个句子进行分词处理后，两个词袋向量不存在共有的不为零维度，并且向量间距离最大，但是实际来看，两者距离却很小。考虑到一些句子即使没有相同的词汇，也传达着几乎相同的信息，这样会产生独立词距离难以表示的问题。为了将独立词的语义有效地融合在文档距离矩阵中，可以采用欧式距离计算每个 word2vec 词向量之间的距离，公式如下所示：

$$c(i, j) = \left\| Vec_i - Vec_j \right\|_2 \tag{18}$$

其中 $c(i, j)$ 代表一个词语转移成另一个词语所花费的代价。

在得到每一个单词到单词之间距离的基础上，就可以得到整个文档 P 到文档 Q 之间的距离：

$$\sum_{i,j} T_{ij} c(i, j)$$

将累积代价最小化，有以下公式：

$$\min_{T \geq 0} \sum_{i,j=1}^{n} T_{i,j} c(i, j) \tag{19}$$

满足：

$$\sum_{j=1}^{n} T_{ij} = d_i \quad \forall i \in \{1,\cdots,n\}$$
$$\sum_{i=1}^{n} T_{ij} = d_j^{'} \quad \forall j \in \{1,\cdots,n\} \tag{20}$$

WMD 距离利用 word2vec 中的语义信息，实现高度语义共现精确度，同时可以挖掘出独立词之间的语义相关性。但同时 WMD 也存在一些不足，比如 WMD 算法只是简单的对所有词随机赋予一个权重，并不考虑词在句子中的重要程度，这容易造成对句子的分类错误等等。

3.2 WMD 处理过程

3.2.1 数据预处理

首先对附件 2 中的留言标题删除符号以及空格，然后在导入自建词库的基础上，利用 jieba 对留言标题进行分词，并删去 Word2vec 无法识别的 oov（out of vocabulary），得到最终结果。

3.2.2 重新训练 Word2Vec

1.在对群众留言详情进行分类时，本文利用的是通过维基百科中文语料库训练的 Word2Vec 词向量模型，并从生成的词向量中抽取待研究留言的特征词向量。但考虑到维基百科中文语料库中一些词汇的缺乏，比如群众留言中涉及到的地区名称、小区名称、学校名称等等，本文重新训练 Word2Vec，利用正则化建立了小区等词库，希望可以在正确识别小区、学校等的基础上提高计算文本相似度的正确性，更新词库示例如表 7 所示。

表 7　更新词库示例

序号	名称	标签数	例 1	例 2	例 3
1	地区	62	A 市	A1 区	A2 区
			K3 县	M 市	L 市
2	小区	385	丽发新城	伊景园	魅力之城
			中建嘉和城	广福园	时代年华
3	学校	188	商学院	外贸学院	物流学院
			梅溪湖中学	周南中学	博才小学
4	医院	45	康乃馨医院	163 医院	南湖医院
			武警医院	市人民医院	儿童医院

2.添加上述词库，利用 jieba 对维基百科文本(400 多 mb)和附件 2、附件 3 的留言标题以及留言详情进行分词，重新训练一个 100 维度的词向量。新的词向量可以对小区、学校、医院等当地信息有更强的表示效果，比如利用 most similar 函数查询到的“伊景园”的相似词如表 8 所示：

表 8　伊景园的相似词

相似词	概率
滨河苑	0.803
万润	0.688
铁路职工	0.668
广铁集团	0.669
金茂府	0.665

3.2.3 计算热度概率

为了优化算法运行时间，提高热点挖掘准确度，本文提出利用热度概率寻找潜在热点事件。首先基于词向量利用 n_similarity 函数对全部留言标题计算文本相似度，热度概率选出排名前 250 名的留言标题。热度概率的计算公式如下所示：

$$R_j = \sum_{i=1}^{n} sim_i + 0.1 \cdot (pro_j - against_j) \quad i = 1,2,..,n; \quad j = 1,2,..,N \tag{21}$$

其中，R_j 为热度概率，pro 为点赞数，against 为反对数，sim 为相似度，n 为与标题 j 相似度(sim)高于 0.85 的标题总数，N 为附件 2 所有的标题总数。

在这里需要指出的是，利用 WMD 算法计算文本相似度效果更佳，但出于运行速度及运行内存的考量，本文首先利用 n_similarity 进行初步的筛选，筛选结果如表 9 所示：

表 9　热度概率排名前 250 名的留言

排名	时间	标题	热度概率
0	2019/8/19	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	211.5752
1	2019/4/11	反映 A 市金毛湾配套入学的问题	176.7
2	2019/2/21	请书记关注 A 市 A4 区 58 车贷案	83.95137
3	2019/2/25	严惩 A 市 58 车贷特大集资诈骗案保护伞	82.59803
	…	…	…

247	2019/2/26	A 市万润滨江天著忽悠业主延迟网签	11.47013
248	2019/2/21	A7 县海德公园楼盘没有配套公办幼儿园或普惠性幼儿园吗	11.46758
249	2019/10/10	A2 区银桂苑小区西侧工地严重扰民	11.4553

3.2.4 确定类中心

对上述确定的 250 条留言标题再次利用相似度进行归类，首先将 250 条留言的初始类别定为他们本身，并将词移距离(WMD)小于 1 的标题归为一类的方式进行迭代，类别仍为本身的即为类中心。

表 10　类中心表

排名	时间	标题	热点概率
0	2019/8/19	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	211.5752065
1	2019/4/11	反映 A 市金毛湾配套入学的问题	176.7
2	2019/2/21	请书记关注 A 市 A4 区 58 车贷案	83.95137141
3	2019/9/5	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	67.9
4	2019/12/10	投诉丽发新城小区附近违建搅拌站噪音扰民	61.08729804
	…	…	…
10	2019/1/8	西地省展星投资有限公司涉嫌诈骗	14.34128863
11	2019/3/26	关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	12.49461095
12	2019/2/21	A7 县海德公园楼盘没有配套公办幼儿园或普惠性幼儿园吗	11.46757925

3.2.5 根据类中心匹配文本相似度

根据上文得到的 13 个类中心，再次利用 WMD 算法对全部标题进行匹配计算文本相似度，将词移距离小于 0.8 的归为一类，结果如表 11 所示：

表 11　同类别个数汇总表

类中心	个数
投诉丽发新城小区附近违建搅拌站噪音扰民	105
A 市地铁 3 号线星沙大道站地铁出入口设置极不合理	66
投诉 A 市伊景园滨河苑开发商违法捆绑销售无产权车位	48
咨询 A 市购房政策的社保缴纳相关问题	42
西地省展星投资有限公司涉嫌诈骗	41

A7 县海德公园楼盘没有配套公办幼儿园或普惠性幼儿园吗	24
A5 区劳动东路魅力之城小区临街门面烧烤夜宵摊	17
A7 县泉塘街道漓楚东路与东六路交汇处是否可以架设人行天桥	10
关于 A6 区月亮岛路沿线架设 110kv 高压线杆的投诉	9
A 市 A5 区汇金路五矿万境 K9 县存在一系列问题	7
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	4
请书记关注 A 市 A4 区 58 车贷案	3
反映 A 市金毛湾配套入学的问题	1

3.2.6 加入时间变量，剔除错误事件

热点事件的类型复杂多变，比如有针对新政策的咨询，有针对正在建设的交通干线的建议征集，有针对某小区发生的特定问题的投诉等等。但我们发现，“噪音扰民”等类型事件频发，但对于固定小区，往往时间集中度更明显，所以对“丽发小区违建搅拌厂扰民”以及“伊景园强制捆绑车位”事件，利用时间和词移距离进行改良 K-means 聚类，剔除被错误放入的事件。

其中，时间的处理为

$$T_i' = T_i - T_{\min} \qquad i = 1, 2, ..., n \tag{22}$$

T_i 为第 i 条留言标题的发布时间，n 为该事件总留言数。

由于 K-means 算法需要人为事先确定聚类中心的个数，并且在之后的迭代过程中聚类中心的个数也保持不变，所以如何正确的选取合适的 k 值便显得尤为重要。本文为选取 k 值做了如下准备工作：

(1)手肘法：手肘法因图形酷似手肘得名，其主要的判断依据是误差平方和(SSE)。

$$SSE = \sum_{i=1}^{k} \sum_{p \in C_i} \left| p - m_i \right|^2 \tag{23}$$

其中，C_i 是第 i 个“自然小类”，p 是 C_i 中的点，m_i 是 C_i 的聚类中心。

随着聚类中心的增多，数据的分类会越来越准确，直到 k 值达到真实的聚类中心个数。与此同时，误差平方和的下降趋势会变缓，以“丽发小区违建搅拌厂扰民”事件为例，聚类的手肘图如图 11 所示，可以发现，k=2、4 时，趋势减缓明显。

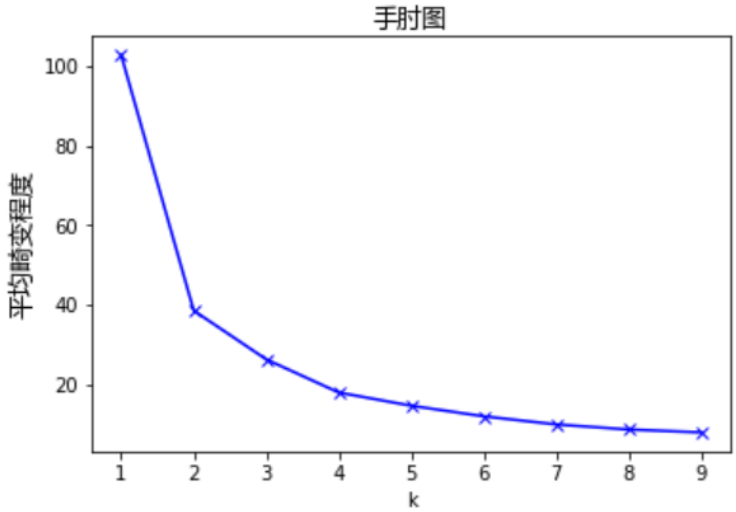

图 11　聚类分析手肘图

(2)轮廓系数法：轮廓系数法的核心逻辑就是利用样本点到某个类的平均距离作为衡量这个样本点应该被归属到一个类的指标。

$$G = \frac{b-a}{\max(a,b)} \tag{24}$$

其中，G 为某点的轮廓系数，a 是该点与同类其他样本点的平均距离，b 是该点到最近的类中所有点的平均距离。按照定义可以看出，平均轮廓系数最大的 k 就是最优聚类数。以“丽发小区违建搅拌厂扰民”事件为例，聚类的轮廓系数图如图 12 所示，由图可以得到，k=2,4 是峰点。

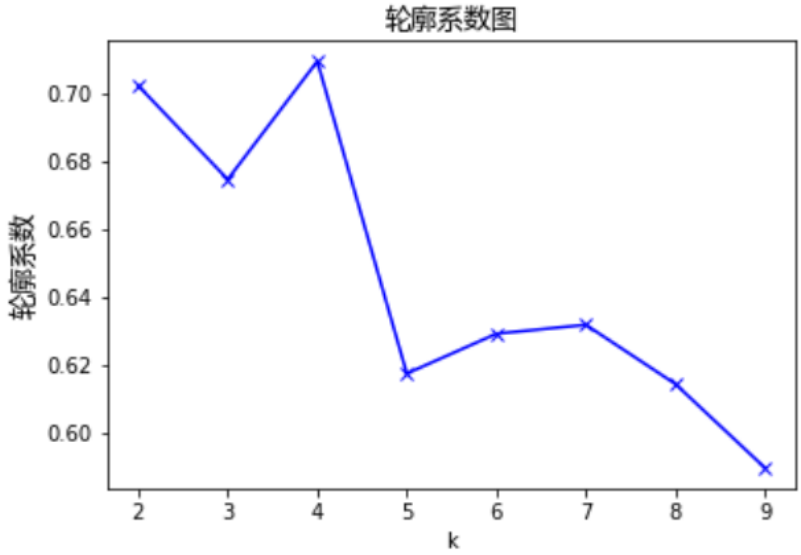

图 12　聚类分析轮廓系数图

综合手肘图和轮廓系数图的结果，本次聚类最佳的 k 值应该选取为 4。对聚类后的数据，取类中心所在簇作为最终结果，即词移距离为 0 的文本数据所在簇。

最终获得的分类效果图如图 13 所示：

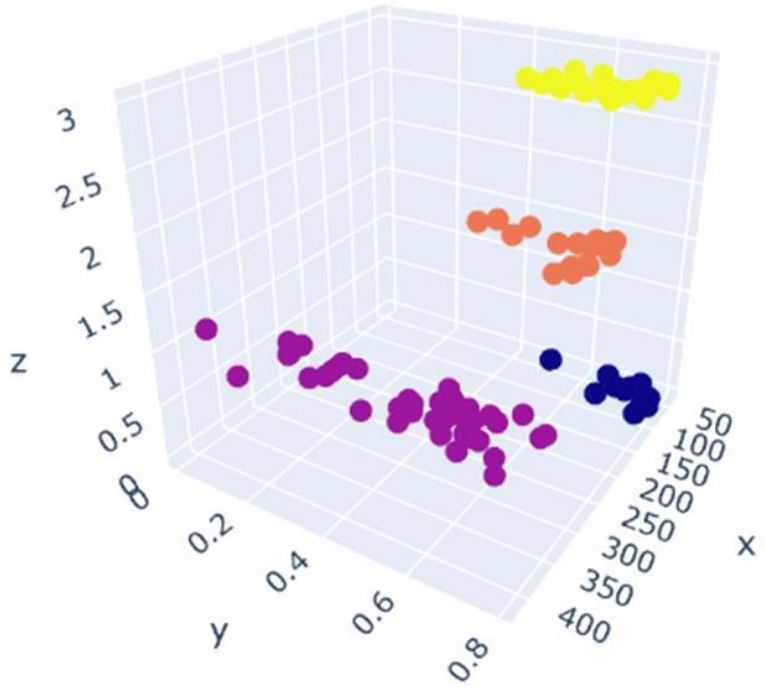

图 13　分类效果图

分类效果图中，x 轴为 T，y 轴为词移距离，z 轴为类别。分类效果的部分结果如下表所示：

表 12　分类效果部分展示表

留言标题	类别
投诉小区附近搅拌站噪音扰民	1
投诉小区附近搅拌站噪音扰民	1
A 市丽发新城三期违建是否会影响交房	2
A2 区丽发新城附近修建搅拌站污染环境影响生活	1
A5 区劳动东路魅力之城小区油烟扰民	2
A 市西二环夜晚重型货车噪音严重扰民	2
A7 县 M9 县城三期小区麻将馆扰民	0
魅力之城小区临街门面油烟直排扰民	2
A2 区丽发新城小区内垃圾站散发严重臭味	2

处理“伊景园强制捆绑车位”事件步骤相同，本文不再赘述。经过改良 K-means 算法处理后，“丽发小区违建搅拌厂扰民”以及“伊景园强制捆绑车位”事件留言总量分别为 47、46 条。

3.2.7 计算热度指数，确定热点问题

基于上述文本相似度的处理，本文定义热度评价指标的计算方法如公式(25)所示：

$$H_i=\sum_{j=1}^{n}\left[0.7n+0.5(1-dis_j)+0.2(pro_j-against_j)\right]\cdot e^{-0.01(T_j-T_{\min})} \tag{25}$$

$$i=1,2,...,N \quad j=1,2,..,n$$

其中，N 为热点事件的总个数(类中心的总个数)，n 为第 i 个热点事件的总留言条数，H_i 为热度指数。$e^{-0.01(T_j-T_{\min})}$ 为时间衰减函数，T_j 表示该条留言出现时间，$T_{\min}$ 表示该问题最早出现的留言时间。

借助热度指数，本文确定出群众留言中的热点问题，即某一时间内群众集中反映的某一问题，排名前 5 的热点问题如表 13 所示，热点问题留言详情表如表 14 所示。

表 13　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	1058.368699	2019/11/02 至 2020/01/26	A 市 A2 区丽发新城小区	小区附近违建搅拌站噪音扰民
2	2	892.7331578	2019/06/24 至 2019/12/31	A 市伊景园滨河苑	小区开发商违法捆绑销售无产权车位
3	3	463.9341903	2018/10/27 至 2020/01/07	A 市地铁	A 市地铁线路规划问题
4	4	414.5063167	2019/01/10 至 2019/12/25	A 市购房政策	咨询购房政策的社保缴纳相关问题
5	5	352.6	2019/4/11	A 市金毛湾	小区配套入学问题

热点问题的问题描述以及相关留言详情在表 13、表 14 中标注的较为清晰，在此不做赘述。需要特别指出的是排名第五位的“A 市金毛湾——小区配套入学问题”。尽管留言内容仅有一条且点赞数在热度评价指标中赋予权重较小，但由于其点赞数过高，达到 1762，而反对数仅为 5，过高的点赞与反对数之差助力其

热度指数排名第五，成为群众广泛关注的热点问题。

表 14　热点问题留言详情表

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
1	208285	A909205	投诉小区附近搅拌站噪音扰民	2019/12/15 12:32:10	尊敬的领导，我是 A 市暮云街道丽发新城的一名业主，最近遇到了意见特别烦心的事情，我是做小区安保的，有白天上班也有晚上班。小区边上建了个大型搅拌厂，白天晚上噪音很大，不能正常休息，关闭门窗还是有很大噪音 ……	0	24
1	261072	A909207	投诉小区附近搅拌站噪音扰民	2019/11/23 23:12:22	投诉 A 市暮云街道丽发新城附近大型搅拌站水泥厂噪音严重扰民，扬尘污染环境，希望有关部门回复……	2	9
1	208714	A00042015	A2 区丽发新城附近修建搅拌站污染环境影响生活	2020/1/2 0:00:00	尊敬的领导：您好！作为一名居住在 A2 区丽发新城的业主，和小区内的每一位业主一样，最近我们正面临一个十分头痛的问题：我们小区附近百米范围内修建了搅拌厂，严重污染环境，小区内空气质量和声环境质量急剧下降……	0	4
…	…	…	…	…	…	…	…
1	287331	A909217	A2 区李丽发新城附近无资质混凝土搅拌站为何禁而不止	2020/1/15 15:56:21	举报 A2 区丽发新城小区附近建了一个无资质的搅拌站！这家搅拌站环境糟糕，粉尘到处飞扬，噪声隆隆，典型的脏、乱、差，严重影响附近居民生活！请领导早日处理！	0	0
2	191001	A909171	A 市伊景园滨河苑协商要求购房同时必须购买车位	2019/8/16 9:21:33	商品房伊景园滨河苑项目是由 A 市政府办牵头为广铁集团铁路职工定向销售的楼盘，作为集团的一名退休员工，我深深感觉到政府对铁路员工的关怀，在广铁辛苦几十年，住的是职工原来配置的福利房，遇到这样一个利好消息十分激动，现在到了缴纳认购款的时候，却被告知除了要缴纳 18.5 万的认购款还要缴纳 12 万元的钱买车位，不然就取消购房资格……	1	12
2	213584	A909172	投诉 A 市伊景园滨河苑定向限价商品房违规涨价	2019/7/28 13:09:08	投诉 A 市伊景园滨河苑定向限价商品房项目，广州铁路集团公司与开发商联合,违规提高房价,违规收取认购款 18.5 万，违法强行捆绑销售车位 12 万/户，无视法律法规，无视民生，坑员工血汗钱……	0	0
2	224767	A909176	伊景园滨河苑车位捆绑	2019/7/30 14:20:07	伊景园滨河苑车位捆绑销售！广铁集团做个人吧！我辛辛苦苦攒钱买个房这么	0	0

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
			销售广铁集团做个人吧		多事！先是强行收我首期认购金，还不给我合同，说什么预购不用！……		
…	…	…	…	…	…	…	…
2	196264	A00095080	投诉A市伊景园滨河苑捆绑车位销售	2019/8/7 19:52:14	A 市伊景园·滨河苑现强制要求购房者捆绑购买 12 万车位一个，不买就取消购房资格，国家三令五申禁止捆绑车位销售，为何还敢顶风操作，2 千多购房者苦不堪言，无处申冤，现垦求领导帮忙讨回公道！	0	0
3	267630	A000100648	反映A市地铁3号线松雅湖站点附近地下通道问题	2019/5/22 23:37:38	沈书记：您好！请求您百忙中关注下居住在地铁3号线松雅西地省站西北方向10 万民众方向将来乘降地铁出行安全风险问题。上周地铁3号线成功穿越京港高铁，突破施工最大难点，官宣预计地铁3号线有望试运行，居住在3号线的居民们欢欣鼓舞。近日又有媒体公布了3号号线各站点出入口设置方案，我们在仔细研读3号线松雅西地省站出入口设置方案后发现设计不合理 ……	0	42
3	193286	A000103197	居住在地铁3号线A7县松雅西地省站西北方向10万民众的心声	2019/4/17 11:13:12	沈书记：您好！欢迎您来 A7 县工作！请求您百忙中关注下居住在地铁3号线松雅西地省站西北方向 10 万民众方向将来乘降地铁出行安全风险问题。上周地铁3号线成功穿越京港高铁，突破施工最大难点，官宣预计地铁3号线有望试运行，居住在3号线的居民们欢欣鼓舞。近日又有媒体公布了3号号线各站点出入口设置方案，我们在仔细研读 3号线松雅西地省站出入口设置方案后发现设计不合理	0	32
3	228796	A00053304	强烈建议将地铁7号线南延至A市生态动物园	2019/3/6 14:20:16	尊敬的A市委领导、A市规划局领导：长株潭一体化提出已9年，暮云作为融城“核心”区域，目前进入发展的关健时刻，在A市大力发展地铁等轨道交通的大背景下，迫切希望暮云组团能赶上地铁这躺“快车”：一、暮云轨道交通现状：暮云片区无地铁……二、建议……	0	31
…	…	…	…	…	…	…	…
3	203187	A00024716	咨询A9市高铁站选址的问题	2019/8/1 13:48:57	尊敬的A市委书记，听说A9市高铁站选址已经很久了，一直都没有准确的定下来，许多人也特别关心，可是选址渺	53	10

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
					无音讯，现在听说备选有两个，关口和平水，再次我发表一下我的建议……		
4	204245	A00024579	关于A市高级技师购房补贴的疑问	2019/3/1 10:58:43	尊敬的胡书记：《A市人才购房及购房补贴实话办法(试行)》颁布后，本人于2018年2月份通过政府网站公布的窗口电话0000-00000000咨询高级技师购房补贴政策，窗口工作人员答复是“只要是2017年6月21号后，也就是办法颁布后取得证书及购房就可以申请购房补贴，至于是先买房还是先取得证书没有规定”……	0	14
4	213772	A00074436	关于修改A市住房限购限售政策的建议	2019/4/30 15:00:28	尊敬的市委领导您好，A市实行住房限购限售已有一段时间，在落实房住不炒方面取得了很大成效，非常感谢政府作出的重大努力。但是政策中有一条规定，要求已购住房需要拿到房产证四年后才能出售一直在困扰着我……	0	12
4	225657	A00051791	关于A市人才购房补贴的疑问	2019/2/25 14:43:15	根据《A市人才购房及购房补贴实话办法(试行)》中第八条规定“A市域内企业2017年6月21日后新引进或新获得高级技师职业资格证的人员在A市首次购房后，可申请3万元购房补贴”的规定，本人于2017年9月在A市购房，为响应政策的号召，于2018年获得高级技师职业资格，但在申请购房补贴时，却因“购房时间应在取得高级技师之后”为由被拒，对此存在以下疑惑……	2	6
…	…	…	…	…	…	…	…
4	189733	A0009754	A市限卖房产政策一刀切	2019/10/10 17:09:47	A市的限购政策有效的控制了房价疯涨的局面，也得到了大多数老百姓的支持，但政策出台这么久也没有完善细节。政策规定房产需在产权证满4年后才能出售，这种方式太粗暴，侵犯了公民处置财产的正当权益……	1	0
5	223297	A00087522	反映A市金毛湾配套入学的问题	2019/4/11 21:02:44	书记先生：您好！我是梅溪湖金毛湾的一名业主，和其他业主一样因为当初金毛湾的承若学校都是金毛建的，果断买了梅溪湖金毛湾。楼盘当时承若小学配套周南小学或实验三小，中学配套周南中学或西雅中学。2018年4月16号学区划分没有金毛湾，诸多业主进行了维	5	1762

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
					权，最终由市教育局缪副局长做出了相关解释……明确于2018年7月31日前将金毛湾纳入周南梅溪湖中学或西雅中学配套入学范畴。但是，时至今日仍然没有相关配套入学文件出台……		

第 4 章 答复意见评价

4.1 指标体系设置

群众向网络政务平台“输入”利益需求时，不管是否合理，作为代理人民行使权力的政务平台有义务向他们做出相应的回应，每一次“输出”都是检视政务平台是否把人民利益放在最高位置的机会[15]。网络政务平台需要根据一套固定的制度和决策程序进行分析判断，讨论协商，并向群众进行“输出”合理的反馈结果。

为了全方位评估“智慧政务”中相关部门对留言答复意见(反馈意见)的质量，本文将设立一套答复意见评价指标体系，以相关部门的答复内容为评估对象，对网络问政平台中的“输出”即答复意见的相关性、完整性、可解释性、时效性进行全方位的评估。答复意见评价指标体系的架构如下表所示：

表 15　指标体系设置

一级指标	二级指标	评估方式
相关性	相关程度	基于 word2vec 的相似度计算
完整性	信息量	实词数量
	话题熵	LDA
	问候语	问候语数量
可解释性	行文逻辑	逻辑词数量
	政策引用	max{书名号数量,中括号数量}
时效性	受理效率	提问与答复的时间间隔

4.1.1 相关性

相关性是评估留言答复意见(反馈意见)质量最为重要的指标，一方面是为了确保答复工作人员掌握一定的语言艺术和表达技巧，在工作中正确表达客观事物和主观思想，传递信息，交流思想，切勿答非所问[16]；另一方面更是为了有效防止部分工作人员处于某些目的对“输入”问题故意回避、大事化小、导致“输出”的结果言之无物，据此，本文设置相关程度指标对该项内容进行评价。具体操作中，本文利用的是 n_similarity 函数计算留言详情以及答复意见之间的相似度。

4.1.2 完整性

考察答复意见的相关性是为了确保不出现答非所问的现象，那么完整性更多

关注的是答复人员应具备的素质和能力。首先，答复意见中要包含足够多的信息，确保留言群众能够获得其想要了解的内容，因此本文设置信息量这一指标对其进行评估；其次，答复意见要全面准确、由浅入深，与留言内容相关联的政策、措施尽可能的完善，确保答复内容完整，本文设置话题熵这一指标进行评估；此外，工作人员的态度也至关重要，适当的礼貌用语能让留言群众产生信任感和亲近感，拉近距离，消除隔阂，相应的，本文设置问候语这一指标对其进行评估，指标体系具体设置内容如下所示。

1.信息量

文本所含的信息量通常用文本的长度，或者文本中包含的词的数量来衡量。汉语与英语不同，英语的每个单词都有其明显的含义，并且词与词之间通过空格断开，即存在明显的分隔符。比较来说，中文的词与词之间不存在空格，并且中文连续多个字才能形成具有明确含义的词汇。因此中文分词的目的是在于将一个连贯的句子按照一定的分词标准将其分成有一个一个具有独立含义的词汇[17]。基于此，本文选择答复文本中所包含的实词的数量作为信息量的量化指标[18]。

由于相关部门的对留言的答复意见属于文本形式，即属于非结构化数据，是计算无法直接分析的数据，所以对于经过预处理的数据，本文首先对其进行分词处理，将非结构数据转换为结构化数据，以便计算机进一步地识别与处理。对于汉语，由于词与词之间的界定模糊不清，并没有明确的分割标志，本文采用 Python 中的 jieba 库进行处理。

在使用 jieba 库对于答复意见文本进行分词后，并不是所有的词都是有实际意义的，像的、啊等的一些没有实际意义的虚词，本文通过引入自定义停用词库，在进行分析前首先将其剔除，只保留有意义的实词。最后通过计算实词的数量来确定答复意见的信息量。

2.话题熵(主题多样性)

LDA 模型是 Blei 提出的一种对离散数据集建模的概率主题模型[19]。借助于超参数，LDA 可以将“文档—词汇”的高维空间映射到“文档—主题”和“主题—词汇”的低维空间[1]。该模型具有三层生成式贝叶斯网络结构[20]，基于前提假设：文档是由若干个隐含主题构成，而这些主题是由文本中若干个特定词汇构成，忽略文档中的句法结构和词语出现的先后顺序[21]。对于每一篇文档，文档内的词

语的计算公式见公式(26)：

$$P(\text{词语}|\text{文档}) = \sum_{\text{主题}} P(\text{词语}|\text{主题}) * (\text{主题}|\text{文档}) \tag{26}$$

其中，P(词语|文档)表示词语在选定文档中出现的概率，该概率值即为每个单词的词频，故公式(26)的左边是已知的。P(词语|主题)表示单词出现在每个主题中的概率，P(主题|文档)表示每个主题出现在每篇文档中的概率，这两类概率值都是未知的，LDA 的目的便是通过公式(26)左边已知的 P(词语|文档)值推出右边的两个概率值[22]。它有两个常用假设，第一个是与实际情况基本相符的假设，即文档集合中的文本包含很多个主题，并且是根据一定比例混合组成的；另外一个假设是“词袋假设”，即一篇文档是由大量无序的词构成的，不同的词之间是可以交换顺序的[24]。

本文应用 LDA 计算出了描述文本中所覆盖的主题，并给出每一个描述文本所述主题的概率大小。最初由 Claude Shannon 提出来的香农熵，可以用来测量文本的不确定性[23]。近年来，熵更多的被应用于衡量文本中潜在主题的多样性。字符串中字母出现的概率越均衡，不确定性就越大，即熵就越大。类似地，答复意见文本中所属的主题概率越均衡，熵越大，表示文本包含的主体数目越多。综上，答复意见文本的话题熵计算公式如下：

$$E(d) = -\sum_{i=1}^{n} p_i \log(p_i) \tag{27}$$

其中，$E(d)$ 代表答复意见文本 d 的主题熵，n 是基于 LDA 确定出的对应的留言详情文本所包含的主题数量，p_i 是答复意见文本 d 中属于话题 i 的概率。

本文从答复意见文本对应的留言详情文本中识别主题，在探索尝试后，对每个文本选择确定两个主题。

3.问候语

问候语是人们生活中最常用的重要交际口语，是指能够拉近人与人之间的距离的语言，表示自己对别人的尊重。随着社会服务意识的增强，许多单位对接听电话如何打招呼皆有严格的规定，网络政务平台作为电子政务的一部分更是不能例外。“您好”，“感谢您对我们工作的关心、监督与支持。”这种必备的问候语自然也是答复意见评价方案中的重要考察指标。本文通过对问候语进行遍历计数，通过问候语的出现频率对这一指标进行考察。问候语的具体设定及赋值示例如表

16、17 所示。

表 16　问候语示例

名称	标签数	例 1	例 2	例 3	例 4	例 5
问候语	17	你好	尊敬	特此回复	感谢	祝福

表 17　问候语赋值示例

<table>
<tr><th>留言编号</th><th colspan="7">答复意见</th></tr>
<tr><td>2574</td><td colspan="7">网友“A0009233”，您好，您的留言已收悉，现将具体内容答复如下：关于来信人建议“白竹坡路口”更名为“马坡岭小学”，原“马坡岭小学”取消，保留“马坡岭”的问题。公交站点的设置需要方便周边的市民出行，现有公交线路均使用该三处公交站站名，市民均已熟知，因此不宜变更。感谢来信人对我市公共交通的支持与关心。2019 年 5 月 5 日</td></tr>
<tr><td rowspan="2">问候语</td><td>总计</td><td>你好</td><td>您好</td><td>收悉</td><td>答复如下</td><td>...</td><td>感谢</td></tr>
<tr><td>4</td><td>0</td><td>1</td><td>1</td><td>1</td><td>...</td><td>1</td></tr>
<tr><td>179601</td><td colspan="7">“UU008609”你好，你反映的问题已转交相关部门调查处置。9 月 30 日尊敬的网友：您好！就你反映的问题，先回复如下：身份证挂失：挂失者需持本人相关证件(如：户口本，社保卡等，能够证明自己身份的有效证件)，户籍不限，到 H1 区公安人口与出入境管理大队办证大厅进行现场挂失。不能委托他人代办，因挂失期间需要采集本人人像和指纹。感谢您对公安工作的支持和关注。2019 年 10 月 28 日</td></tr>
<tr><td rowspan="2">问候语</td><td>总计</td><td>你好</td><td>您好</td><td>收悉</td><td>答复如下</td><td>...</td><td>感谢</td></tr>
<tr><td>5</td><td>1</td><td>1</td><td>0</td><td>0</td><td>...</td><td>1</td></tr>
</table>

4.1.3 可解释性

可解释性一词并没有一个整体概念，因此关于可解释性的任何观点都应该首先为“可解释性”确定一个特定的定义。本文在这里所说的“可解释性”，指作为答复意见的阅读者(留言者)，在正常的阅读条件之下能否正确、容易地理解作为解释对象(答复意见)的文本，答复意见文本如果不具备必要的条件，便不能达到传达意义的目的，那么网络问政便会失去实质性的意义。本文主要从三个角度对可解释性进行分析，首先是“行文逻辑”，答复意见要根据一定的顺序和构段方式使其有条理性、清晰性，便于理解；其次是“政策引用”，网络问政平台是政府听取民声、汇聚民智、推进民主决策和依法治国的重要渠道，回复网民的意见和建议更需要引“法”据典，提高依法行政水平。

1.行文逻辑

语言是答复工作人员的重要工具，提高答复意见的可解释性，有效表达观点、阐述事实、宣传政策、释疑解惑和进行疏导和劝解等都离不开语言[16]。语言表达能力中行文逻辑更是重中之重。通常来讲，行文逻辑是指在把握了主题的基础上，准确选词，句子与句子之间意思连贯，段落与段落之间逻辑严密，通篇内容分层次分明、文理贯通。

行文逻辑作为回复内容的着眼点，语句间意思的连贯、段落间逻辑的严密及通篇回复内容的畅通，都离不开过渡词语。据此，本文通过对逻辑词进行遍历计数，通过逻辑词的出现频率对这一指标进行考察。逻辑词的具体设定及赋值示例如表 18、19 所示。

表 18　逻辑词示例

名称	标签数	例 1	例 2	例 3	例 4	例 5
逻辑词	136	同时	根据	鉴于	下一步	初步

表 19　逻辑词赋值示例

留言编号	答复意见						
3684	网友“UU008687”您好！您的留言已收悉。现将有关情况回复如下：您所反映的地点为洋湖新城片区和顺路两厢，北至先导路，南至建兴路，全长 301 米，道路两侧绿化带宽 20 米。目前，沿行车道两侧栽种了行道树，其余部分也按规划要求完成了建设，其中西边绿化带面积约 6000 平方米，由于整体为洋湖壹号小区配建，权属为全体业主，目前按 6 大块造型建设，绿化标准符合原先规划设计要求；东边绿化带约 6000m2 按园林景观路标准建设。感谢您对我们工作的支持、理解与监督！						
逻辑词	总计	已	如下	如	初步	...	目前
	10	1	1	1	0	...	1
3700	网友“UU00877”您好！您的留言已收悉。现将有关情况回复如下：经查，A4 区政府已责成区市政局牵头，区城乡建设局、区规划分局配合进行具体选址，招标(邀标)进行方案设计等，尽快启动万国城小区人行天桥建设并投入使用，方便出行。感谢您对我们工作的支持、理解与监督！2019 年 1 月 8 日						
逻辑词	总计	已	如下	如	初步	...	目前
	8	1	1	1	0		0

2.政策引用

答复意见人员作为政府公务人员，遇到问题依法办事，引导群众依法维权，分析群众诉求是否合法，决不以权压法、以言代法，努力推动形成依法办事、遇事找法、解决问题运用法、化解矛盾靠法的良好法治环境乃是基本的工作准则。

在对群众留言的问题进行回复处理时，要使群众切实感受到问题得到了有效解决，权益受到了公平对待，利益得到了有效维护，回复内容说的明白、解释得通，就必须牢固树立法治在化解社会矛盾中的主体地位，树立法律在化解社会矛盾中的权威地位。考虑到“依法是原则，法治是底线”的工作要求，本文将对于相关政策的引用程度也纳入答复意见的评价指标体系中，希望相关工作人员能够切实坚持法治思维和法治方式，对群众各方面的留言诉求认真梳理、仔细分析，严格区分不同情况，切实做到依法按政策解决。由于大部分情况下，引用政策需要书名号或中括号的存在，本文通过对书名号和中括号进行遍历计数，通过书名号及中括号出现的最大频率对这一指标进行考察。具体操作中，政策引用的赋值示例如表 20 所示。

表 20　政策引用赋值示例

<table>
<tr><th>留言编号</th><th colspan="3">答复意见</th></tr>
<tr><td>3980</td><td colspan="3">网友“UU008173”您好！您的留言已收悉。现将有关情况回复如下：目前，市规划局正在编制《A 市轨道交通线网规划修编规划》，对于您的意见和建议，将充分论证研究。感谢您对我们工作的支持、理解与监督！
2018 年 12 月 29 日</td></tr>
<tr><td rowspan="2">政策引用</td><td>总计</td><td>书引号《》</td><td>中括号[]</td></tr>
<tr><td>1</td><td>1</td><td>0</td></tr>
<tr><td>4043</td><td colspan="3">网友“UU008187”您好！您的留言已收悉。现将有关情况回复如下：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：“公益性岗位开发实行申报制度。对政府投资开发的公益性岗位，用人单位填报《A 市政府投资开发的公益性岗位申报表》，在每年 7-8 月向同级机构编制部门申报，由机构编制部门负责审核。”2013 年 7 月由 A2 区交警大队申报，经 A2 区编办审核确认、A2 区财政局审核，对 9 名交通劝导员进行了公益性岗位认定，A2 区人社局按程序审核对交警大队 2013 年 7 月至 2015 年 3 月(共 21 个月)申报的公益性岗位拨付了足额的岗位补贴。根据《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：“企业(单位)在与招用的就业困难人员签订劳动合同之后，应按照社会保险缴费级次到市、区、县(市)公共就业服务机构办理社会保险补贴申报手续。”在接到 A2 区交警大队的申报后，A2 区人社局于 2012 年 1 月至 2014 年 12 月(共 36 个</td></tr>
</table>

	月)对持有**《就业失业登记证》**且进行困难认定的 11 名文明交通劝导员审核并拨付了公益性岗位社保补贴。以上公益性岗位岗位补贴及社保补贴，A2 区人社局均严格按照相关文件依法依规进行了审核及办理，申请和发放都是实名制，且均未超过三年，不存在套取国家再就业资金。具体发放数据，您可至 A2 区人社局进行现场查实。感谢您对我们工作的支持、理解与监督！2018 年 12 月 21 日		
政策引用	总计	书引号《》	中括号[]
	4	4	3

需要指出的是，在统计政策引用时，带书名号的除了政策文件外，还有平台以及帖子。如：您在《问政西地省》专栏发表了题为《呼吁 C5 市领导制止违法违规建房行为》的网帖已知悉.....。将类似“《问政西地省》”的 155 个书引号通过 excel 替换删去；将类似“《问政西地”的 15 个书引号通过 excel 替换删去；将 5 个关于“网帖”的书引号及 4 个关于“帖子”的书引号人工删去。

4.1.4 时效性

答复意见的时效性是指答复意见仅在一定时间段内对留言问题具有价值的属性。答复意见的时效性很大程度上制约着解决问题的客观效果，就是说同一答复内容在不同的时间进行回答具有很大的性质上的差异，本文将这个差异性称为答复意见的时效性。

通常来说，时间间隔越短，答复信息越及时，答复内容的可使用程度越高，时效性越强。此外，信息的时效性更是影响着一些决策的生效时间，可以说是答复意见的时效性决定了答复内容在哪些时间内有效。这一指标的设置，主要目的是防止相关工作人员思想上对其不够重视，处理留言问题较随意，致使出现懈怠拖延的情况，切实督促相关工作人员紧紧把握答复时效性特点，充分发挥其时效性的功能。

本文通过受理效率这一指标对时效性进行评价，计算提出问题与收到回复的时间间隔，受理时间越短，说明处理效率越高，时效性越强，相应的指标得分就越高。以天为时间单位，所有留言的最短受理时间为 1 天，最长受理时间为 1161 天。由于时间间隔为负向指标，因此对其分值做正向处理：

$$x\_score = \frac{\max}{\mathrm{x}} \tag{28}$$

其中 max 表示最大时间（即 1161），x 表示当前时间。具体赋值示例如表 21 所示。

表 21　时效性赋值示例

留言编号	留言时间	答复时间	时间间隔 (单位：天)	时间指标分
10694	2016/1/8 10:34:19	2016/1/8 15:22:06	1	1161
3704	2018/12/28 17:18:45	2019/1/3 14:03:07	6	193.5
3720	2018/12/27 15:18:07	2019/3/6 10:26:14	69	16.826087
19133	2015/2/7 12:59:42	2015/8/24 15:19:53	199	9.7563025
159285	2014/6/14 22:59:34	2017/8/18 9:27:48	1161	1

需要指出的是，具体操作中，时间上对两个部分的格式进行了一定的修改处理，确保所有时间格式统一。修改的第一处内容是附件 4 第 7 行的留言时间(秒数未显示，通过修改单元格格式进行调整)；修改的第二处内容是第 1141 行的答复时间(缺少时分秒，填补上 00:00:00)。

4.2 指标赋权——变异系数权重法

变异系数法是直接利用各项指标所包含的信息计算得出系统各指标变化程度的方法，是一种客观赋权的方法。在评价指标体系中，指标取值差异越大的指标权重越大，指标取值差异越小的指标权重越小。具体原理如下所示：

1.首先将目标向量和各年份指标向量构造矩阵 M=(指标 1，指标 2，……，指标 n)=($A_1, A_2, \cdots, A_n$)。

2.计算第 i 项评价指标的标准差。

$$D = \sqrt{\frac{\sum (x_i - \bar{x})}{n-1}} \tag{29}$$

3.计算第 i 项评价指标的变异系数。

$$CV_i = \frac{D}{\bar{x}_i} \tag{30}$$

4.将变异系数进行归一化处理，进而得到各评价指标的变异系数权重 w_i 。

$$w_i = \frac{CV_i}{\sum_{i=1}^{n} CV_i} \tag{31}$$

最终的各个二级指标权重如表 22 所示：

表 22　二级指标权重

二级指标	变异系数下的权重
相关程度	0.027819
信息量	0.181932
话题熵	0.082143
问候语	0.071211
行文逻辑	0.088143
政策引用	0.352641
受理效率	0.196112

4.3 答复意见评价展示

针对附件 4 相关部门对留言的答复意见，基于本文设置的答复意见评价指标体系，根据相关性、完整性、可解释性和时效性四个一级指标以及相关程度、信息量、行文逻辑和受理效率等 11 个二级指标，本文利用变异系数权重法进行指标赋权，最终得到答复意见评价评分如下表所示。由于文本的空间限制，文本中仅展示排名前五位和排名最后五位的留言以答复的相关信息。

表 23　答复意见得分表

排名	留言编号	得分	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间
1	88359	100	UU0082390	咨询 J9 县的油茶种植政策	2019/1/5 15:52:14	A1	A2	2019/1/7 17:00:37
2	96757	96.84	UU0082390	咨询 J9 县的油茶种植政策	2019/1/5 15:56:23	B1	B2	2019/1/16 8:38:52
3	96762	96.39	UU0082390	咨询 J9 县农产品品牌建设扶持政策的相关问题	2018/12/15 23:28:40	C1	C2	2019/1/3 10:53:46
4	99213	66.1	UU0082352	K8 县三中危房问题 7 年无处理结果，盼望领导给	2018/5/25 16:56:36	D1	D2	2018/6/4 21:13:23

排名	留言编号	得分	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间
				于关注解决！				
5	19133	44.5	UU008720	万明村征收问题回复的质疑	2015/2/7 12:59:42	E1	E2	2015/8/24 15:19:53
…	…	…	…	…	…	…	…	
2812	75000	0.1	UU081539	请求恢复F7县三联卫生院	2018/3/25 23:20:25	F1	您的留言已收悉。关于您反映的问题，已转F7县调查处理。	2018/6/29 16:07:22
2813	12451	0.09	UU081019	请求易书记增加A市交通辅警工资或待遇	2014/3/11 17:02:28	G1	网友：您好！留言已收悉	2014/4/28 16:06:58
2814	74851	0.08	UU008665	咨询F2区社区基层干部的基本保障问题	2018/7/1210:22:54	H1	您的留言已收悉。关于您反映的问题，已转F2区调查处理。	2018/10/31 16:09:57
2815	12452	0.08	UU081171	关于A市旅游发展的看法和建议	2014/3/1111:56:22	I1	网友：您好！留言已收悉	2014/4/28 16:05:42
2816	159285	0	UU081173	咨询低保、残疾人补助的相关问题	2014/6/1422:59:34	J1	你好。请向当地民政部门询问。2017年8月18日	2017/8/18 9:27:48

由于表格展示空间有限，排名前五位的留言详情及答复意见所含内容较多，排名后五位的留言详情所含内容较多，在表格中暂用字母替代，并于答复意见评价Excel表中进行较为完整的展示。

参考文献

[1]李少温.基于网络问政平台大数据挖掘的公众参与和政府回应问题研究[D]. 华中科技大学, 2019.

[2]沈国麟,李良荣.网络理政:中国的挑战、目标和理念[J].新闻大学,2018(03):107-113+151.

[3]袁颖. 服务型政府视域下政务新媒体网络问政存在的问题与治理策略研究[D].华南理工大学,2019.

[4]张哲.政民互动平台中公民参与的行为逻辑研究[D].兰州大学,2016.

[5]Wei J W, Zou K. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks[J]. 2019.

[6]杜世民.基于分类模型的电商用户复购行为预测研究[D].杭州师范大学,2019.

[7] Rong X. Word2vec Parameter Learning Explained [J]. Eprint Arxiv, 2016.

[8]胡西祥. 基于深度学习的微博评论情感倾向性分析[D].哈尔滨工业大学, 2017.

[9]梁宁.基于注意力机制及深度学习的文本情感分析研究[D].华北电力大学,2019.

[10]段立,徐鸿宇,王懿,赵莉,刘冲,郭娇.基于 word2vec 和 XGBoost 相结合的国网95598 客服投诉工单分类[J].电力大数据,2019,22(12):50-57.

[11]杨颖.基于 xgboost 分类算法的企业债发行主体违约风险研究[D].浙江大学,2019.

[12]赵明月.基于词性和关键词的短文本相似度计算方法[J].计算机时代,2018(05):66-70+73.

[13]徐鑫鑫. 基于 WMD 距离的文本相似度算法研究[D].太原理工大学,2019.

[14]王子璇,乐小虬,何远标.基于 WMD 语义相似度的 TextRank 改进算法识别论文核心主题句研究[J].数据分析与知识发现,2017,1(04):1-8.

[15]蔺代标. 网上信访：一种新的非暴力利益表达形式[D].湖南师范大学,2018.

[16]李峰.浅谈基层信访工作人员应具备的素质和能力[J].现代国企研究,2016(04):266-267.

[17] 张启宇，朱玲，张雅萍.中文分词算法研究综述[[J].情报探索,2008(11):53-56.

[18] 张乐. 共享短租中房客信任影响因素及信任传递研究[D].北京邮电大学,2019.

[19] Blei D, Ng A, Jordan M. Latent Dirichlet allocation[J]. Journal of Machine Learning Research,2003,3:993.

[20] Friedman N, Geiger D, Goldszmidt M. Bayesian Network Classifiers[J]. Machine Learning, 997,2:131.

[21]姚全珠,宋志理,彭程.基于 LDA 模型的文本分类研究[J].计算机工程与应用,2011,47(13):150-153.

[22]李文峰.基于主题模型的用户建模研[D].北京邮电大学，2013.

[23]Shannon, C. E. A mathematical theory of communication[J]. The Bell System Technical Journal, 1948, 27: 379-423 and 623-656.