“智慧政务”中的文本挖掘应用

摘要：随着“互联网+政务服务”模式的深入推进，大量政务服务网站、微博、微信、APP 等成为获取社情民意的重要途径，与此相应的留言数据的数量迅速增长，过去处理留言的人工方式越来越无法满足速度和效率的要求。大数据技术为这个问题带来了新的解决方式，通过数据挖掘等技术可以大大减少留言处理的过程需要的人力资源。本文基于深度学习、聚类等方法实现了政府网站留言的自动分类、留言热点问题的挖掘和留言回复自动评价。本文文章结构如下：

第一章简要介绍问题研究背景和研究目标。

第二章主要介绍整体技术路线和文本挖掘的理论知识，包括 word2vector 词向量模型、用于文本分类的长短时记忆网络（LSTM）和卷积神经网络（CNN）、用于热点问题挖掘的 Sing-pass 聚类算法。

第三章介绍文本挖掘任务的实现过程。本文使用腾讯 AI Lab 开源大规模高质量中文词向量数据实现词向量化，采用一个单层 LSTM 网络和一个单层 CNN 网络组成的文本分类模型来实现留言分类；通过使用 Sing-pass 聚类算法并定义了一个合理的留言热度计算公式来实现热点问题挖掘，并通过留言的词性分析提取留言中的位置/人群描述和问题描述；以国务院有关文件中的要求为依据，定义了留言回复的及时性、相关性、完整性指标，实现对留言回复内容的评价。

第四章介绍实验结果：本文实现的留言分类模型，训练完成后在验证集上 F-score 得分可以达到 0.9，实现了较好的分类效果；留言热点挖掘模型可以较为合理的获取热度指标为前五的热点问题，并对热点问题分布进行了简要分析；留言回复评价模型可以依据及时性、相关性、完整性对留言内容进行评价打分，并可以对留言回复情况进行简要的分析。

第五章对本次文本挖掘任务进行了总结。

关键字：Sing-pass 聚类；LSTM+CNN;留言热度计算；留言回复评价

Text Mining Application in "Smart Government Affairs"

Abstract: With the in-depth advancement of Internet + government services, a large number of government service websites, Weibo, WeChat, APP, etc. have become an important way to obtain social conditions and public opinion. The amount of message data corresponding to this has grown rapidly. In the past, manual methods for processing messages were increasingly unable to meet the requirements of speed and efficiency. Big data technology has brought a new solution to this problem, and data mining and other technologies can greatly reduce the human resource requirements of the message processing process. In this paper, through deep learning, clustering and other methods to achieve the automatic classification of government website messages, the mining of message hotspot issues and the automatic evaluation of message replies. The article structure is as follows:

The first chapter briefly introduces the problem research background and research objectives.

The second chapter mainly introduces the overall technical route and the theoretical knowledge of text mining, including word2vector word vector model, long and short time memory network (LSTM) and convolutional neural network (CNN) for text classification, and Sing -pass clustering algorithm for hot topic mining.

The third chapter describes the implementation process of text mining tasks. This article uses Tencent AI Lab open source large-scale high-quality Chinese word vector data to achieve word vectorization, and uses a single-layer LSTM network and a single-layer CNN network text classification model to achieve message classification; A reasonable formula for calculating the popularity of messages is defined to realize the mining of hot issues, and the location / crowd description and problem descriptions in the messages are extracted through the part-of-speech analysis of the messages; Based on the requirements in the relevant documents of the State Council, the timeliness, relevance and completeness indexes of the message reply are defined to realize the evaluation of the content of the message reply.

The fourth chapter introduces the experimental results: the message classification model implemented in this paper, after the training is completed, the F-score on the verification set can reach 0.9, achieves a better classification effect; the message hotspot mining model can reasonably obtain the top five hotspot issues, and a brief analysis of the hotspot problem distribution； the message reply evaluation system can score the content of the message based on timeliness, relevance, and completeness, and can briefly analyze the message reply.

Chapter 5 summarizes this text mining task。

Keywords: Sing-pass clustering; LSTM + CNN; message popularity calculation; message reply evaluation system

目录

1. 研究背景及目标

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。然而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。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和数据挖掘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。

我们基于竞赛官网给出的政务网站群众问政留言记录以及相关部门对部分群众留言的答复意见，利用自然语言处理（NLP）和大数据文本挖掘技术，实现：

1、政务网站群众留言自动分类

2、政务网站群众留言 Top5 热点问题挖掘

3、政务网站群众留言答复意见评价。

2. 研究方法

2.1. 整体技术路线

本项目整体技术路线如图 1 所示：

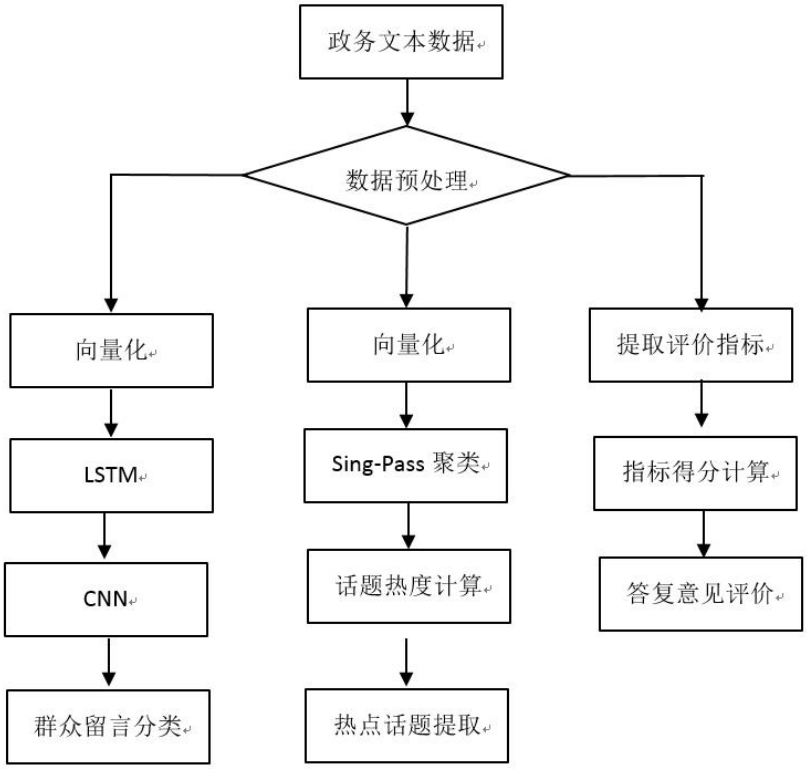

图 1. 整体技术路线图

其中，问题一政务留言分类使用一个由单层 LSTM 和单层 CNN 组成的文本分类模型实现，问题二热点问题挖掘通过对留言主题使用 Sing-Pass 聚类算法进行聚类后计算自定义的话题热度来得到，问题三留言答复意见评价主要通过及时性、相关性、完整性三项指标对留言回复内容进行打分评价。

2.2. 理论介绍

2.2.1. word2vec 词向量模型

word2vec 是由 Mikolov 等人 2013 年提出的词向量表示模型，利用该模型可以通过神经网络，将一个词映射到一个较短的向量，能够较好的表达词语的语义特征。word2vec 模型有两种，分别是 CBOW 模型(图 2)以及 Skip-gram 模型(见图 3)。CBOW 模型利用词 w(t)前后各 c(这里 c=2)个词去预测当前词；而 Skip-gram

模型则相反，它利用词 w(t)去预测它前后各 c(c=2)个词[1]。

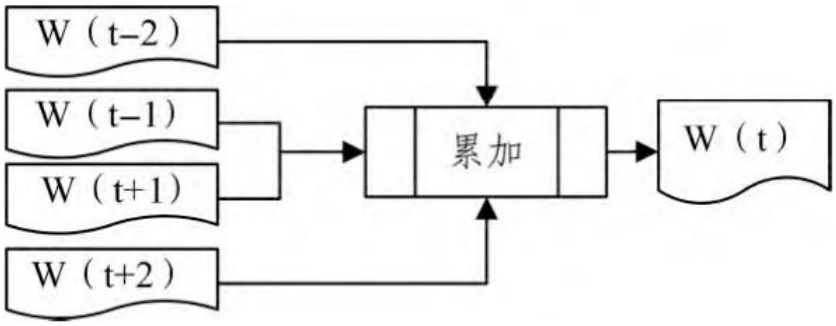

图 2.　CBOW 模型图

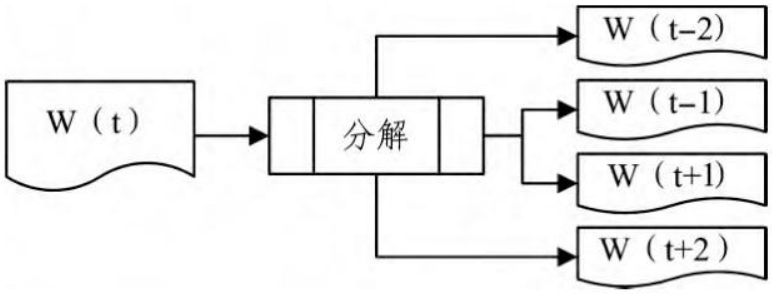

图 3.　Skip-gram 模型图

使用预训练的词向量已经在多项研究中证明是提高实验效果的较好选择，词向量的生成有两种方式，一是应用特定任务的文本语料自行训练词向量，二是使用已训练好的开源词向量。本次实验通过效果比较选择使用腾讯 AI Lab 开源大规模高质量中文词向量数据和自行训练的词向量中任务效果更好的。

2.2.2. 文本分类模型

（1）卷积神经网络

卷积神经网络（Convolutional Neural Networks, CNN）是一类包含卷积计算且具有深度结构的前馈神经网络（Feedforward Neural Networks），是深度学习（deep learning）的代表算法之一，在图像识别有较多成功的应用。常用的 CNN 网络一般由一个输入层、一个或多个卷积层、一个或多个池化层、全连接层及输出层构成[2]，如图 4 所示。

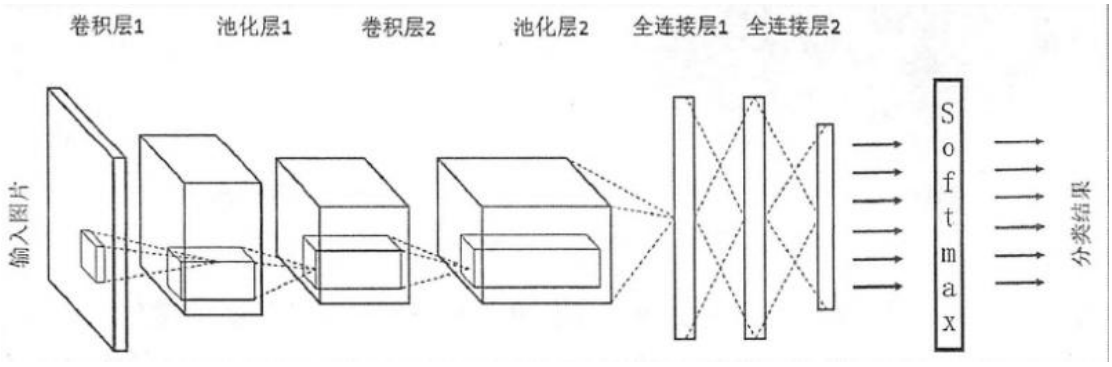

图 4. 通用卷积神经网络结构

卷积神经网络的输入层可以处理多维数据，因此经过预处理的文本数据能够作为输入送入卷积神经网络。

卷积层的功能是对输入数据进行特征提取，其内部包含多个卷积核，组成卷积核的每个元素都对应一个权重系数和一个偏差量，类似于一个前馈神经网络的神经元。卷积层通过卷积核作为局部感受视野每次提取输入的一部分特征，在局部视野内共享同一个权重，因此大大减少了运算量。一般卷积后有激活函数进行计算，将卷积后的结果进行非线性映射，常用的激活函数有 Sigmoid 函数、Relu 函数等。

池化层的功能是对卷积层输出的特征图进行特征选择和信息过滤，目的是缩小卷积后的特征图的大小，以减少训练参数。根据使用局部范围内最大值或者平均值可分为最大池化层和平均池化层。

全连接层的作用是对提取的特征进行非线性组合以得到输出，通过与上一层的神经元的全连接，进行 Softmax 全连接操作，以完成分类或者回归。

卷积神经网络通过卷积层的局部感受和权值共享，相较于一般的全连接人工神经网络，可以大幅度减少运算量，通过池化层的池化作用可以有效的提取特征，具有运算速度快、可以有效提取特征的优点，但是也因为池化作用会导致丢失一部分特征信息。

CNN 网络常用于图像领域，自 Yoon Kim 于 2014 年发表经典之作《Convolutional Neural Networks for Sentence Classification》，将 CNN 网络使用于文本分类任务，并在文中提出了一种经典的 CNN 模型[3]，CNN 在文本分类中应用逐步广泛了起来。

（2）LSTM 神经网络

长短期记忆网络（Long Short Term Memory, LSTM）为循环神经网络

（Recurrent Neural Network, RNN）的变形结构，基于经典 RNN 的基础上，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[4]，图 5 是典型 LSTM 网络结构。

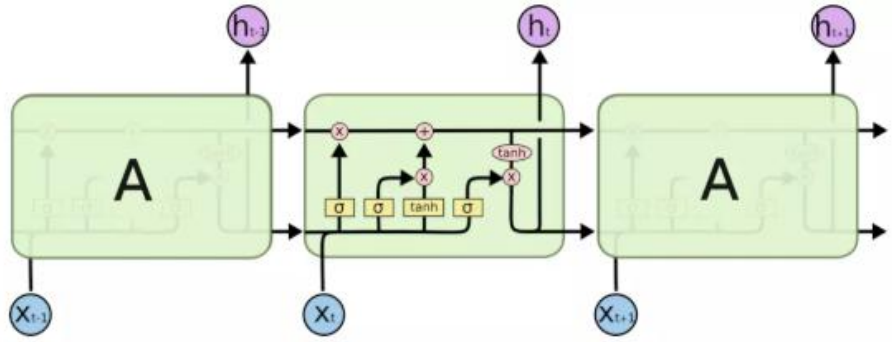

图 5. 典型 LSTM 网络结构

LSTM 应用于自然语言领域可以较好的保留语言的上下文相关性，但是因为保留了上下文信息，导致运算时间较长。

2.2.3. 话题识别模型

（1）Single-pass 聚类算法

Signal-pass 算法一种增量聚类算法，是简单的文本聚类算法，其优点是比较灵活，不需要事先确定聚类数目，对于聚类数目不确定的任务具有天然的优势。其基本思想是将文本特征向量做相似度比较，相似度值大于阈值的文本归为一类文本，从而使主题更容易被发现，使计算更精准[5]。该算法的主要思想是，依次输入一个文本，判断当前文本与已有簇的相似程度，如果相似度大于阈值，则把当前文本归入到该簇，反之则创建新的簇。相似度一般通过两个向量之间的距离进行衡量，常用的指标有欧氏距离、马氏距离、余弦距离等，本文中使用余弦距离度量相似度，计算方法用公式（1）表示。

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad (1)$$

（2）TF-IDF 算法

本文采用 TF-IDF（term frequency－inverse document frequency，词频-

逆向文件频率）算法作为衡量文本中每个词对文本的重要性程度。TF-IDF 是一种统计方法，用以评估某个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。TF-IDF 算法的核心思想是：如果某个词或短语在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，对该类的贡献更大。

设文档集合 C，N 表示 C 中全部文档的数目。在给定的某个文档 d，利用 TF-IDF 方法计算给定词 i 的权重公式，其公式如下[6-8]：

$$w_{ij} = tf \times idf = f_{ij} \times \log\left(\frac{N}{N_i}\right) \quad (2)$$

其中指f_{ij}词语 i 在文档d_j中出现的频率，N_i表示文档中出现词语 i 的文档数。

3. 实现流程

3.1. 数据预处理

对数据挖掘任务，数据预处理已被证明是非常重要的环节，数据挖掘的目的是在大量的、潜在有用的数据中挖掘出有用的模式或信息，挖掘的效果直接受到源数据质量的影响。因此高质量的数据是进行有效挖掘的前提，高质量的结果必须建立在高质量的数据上。

针对本次文本挖掘任务，数据预处理的内容包括以下方面，本文三个任务都要经过如下预处理过程：

1、转码。原始数据中存在大量编码格式为 GBK 的数据，而 Python 适合处理的格式 UTF-8，需要进行转码；

2、处理非法字符。因编码的不同及数据本身的来源的问题，文本中存在较多无法处理的非法字符，如网页换行符等，需要进行剔除；

3、去掉无意义符号。文本中存在的各种计算机语言的空格、换行、制表符等对于挖掘没有实际的作用，需要剔除。

4、格式转换。原始数据中的日期和时间格式不统一，比如分隔符不一致，单元格时间数据类型不一致等，需要统一转换为同一种日期和时间格式。

5、去停用词。原始数据中的字母、数字、标点符号、语气词、助词等一些没有实际意义的内容对文本挖掘任务没有帮助，反而会带来噪声，需要建立停用词典去除。

6、词序列化。因文本分类 CNN 模型中无法接受直接输入的文本数据，需要将挣个文本按照词在词典中的顺序将词变成一串数字序列，模型在接受序列后按照序列转换为向量化进行后续计算。

3.2. 群众留言分类

3.2.1. 分类模型简介

在上文中已经提到 CNN 与 LSTM 模型的各自优缺点，结合两个模型的优点，考虑首先通过 LSTM 层保留文本的上下文相关性，然后通过 CNN 层提取特征和加快运算速度，来提高模型的分类效果同时兼顾运算效率。模型结构如图所示。

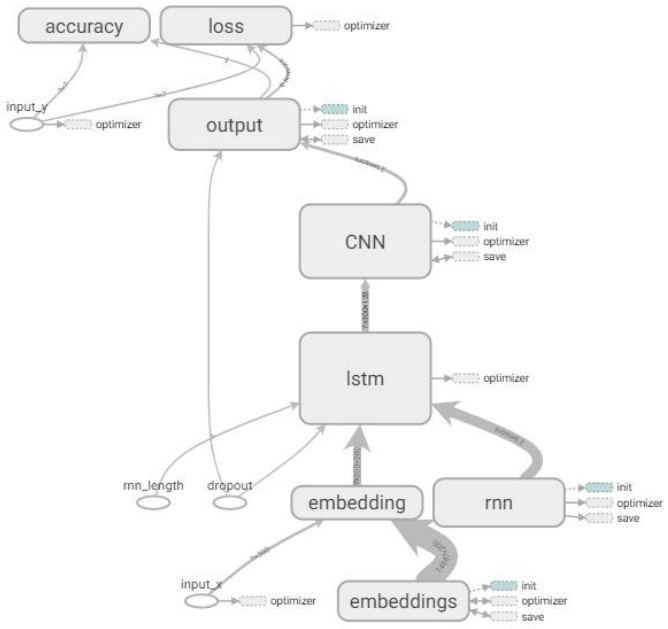

图 6. 本文留言分类模型

其中输入 embedding 为词向量，随后通过一个单层的 lstm 层和一个单层的 CNN 网络，在各层使用一个 Dropout 层减少过拟合，Dropout 率为 0.5，CNN 网络包括 1 个卷积层，1 个最大池化层和 1 个全连接层，CNN 的 3 个卷积核尺寸分别 2,3,4。

3.2.2. 算法实现过程

对附件 2 中原始数据经上文预处理过程后提取留言主题和留言详情，按照一定的比例划分训练集、测试集、验证集。因中文词与词之间没有明显分隔符，中文类文本挖掘任务首先要经过分词，分词后的结果去除停用词后保留。

（1）词向量化

中文文本在输入模型进行分类之前要先映射到向量空间，将词转换为向量进行文本分类任务。本文在留言分类任务中使用了通过 word2vec 基于 Skip-Gram 训练自行训练的 30000 词的词向量，每个词向量 200 维。

将文本的词向量数据序列化之后输入模型进行训练，在训练过程中通过测试集不断测试训练过程中保存的模型，观察训练过程的精度和损失函数曲线，当训练精度在较长时间内没有提高时结束训练，保存效果最好的模型。

（2）测试

使用保存的模型在验证集上进行预测，通过预测结果与数据标签的比较计算查准率（Precision）和查全率（Recall），计算 F-Score 指标。

3.3. 热点问题挖掘

3.3.1. 热点问题挖掘流程

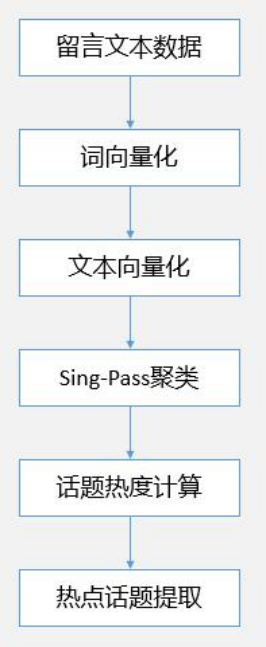

图 7. 留言热点挖掘流程

3.3.2. 算法实现过程

预处理和分词与留言分类任务相同，但是与留言分类任务不同的是，本次提供的留言的主题具有高度概括性，对于表达问题已经足够，而详情中则带有较多的噪音，使用留言主题进行聚类反而能提高精度，因此本次只使用留言主题进行聚类。

（1） 文本向量化

本文热点问题挖掘中使用了腾讯 AILab2018 年 10 月发布的使用 word2vec 模型训练的开源词向量-腾讯 AI Lab 开源大规模高质量中文词向量数据，该开源数据基于 Skip-Gram 训练，包含 800 多万中文词汇，其中每个词对应一个 200 维的向量，，相比自行训练的词向量，覆盖率、新鲜度、准确性的表现更为优秀。为解决在训练过程中在 800 万词语的词典中查找效率低的问题，通过对本次留言数据的词频进行统计，选择词频最高的词，建立一个小规模词典，后续词的向量从小规模词典中查找，可以大大加快查找效率。

文本是由一个个词组成的，对每个词的词向量通过 tfidf 加权得到文本向量，计算方法见公式（3）

$$SentVector = \sum_{i=1}^{i=n} TfIdf_i * WordVector_i \quad （3）$$

其中 $WordVector_i$ 代表组成文本的每个词的词向量，$TfIdf_i$ 代表每个词的 TfIdf 权值。

（2）Sing-Pass 聚类

使用 Sing-Pass 算法进行聚类的算法过程如下：

输入：D 篇新闻文档向量，相似度阈值 s

输出：问题集合 T

具体步骤：

1：首先输入文档中第一篇文档 d1，作为第一个新话题；

2：输入文档 (d2...di...dn),将 di 与已有的类簇中所有样本计算相似度。如果最大的相似度大于阈值 s ，则加入该问题集，如果都不大于 s，则再次

建立一个新问题。

3：重复步骤 2，直至最后一个文档输入完，输出所有问题集 T。

其中相似度阈值 s 的确定要进行反复试验，相似度阈值取值较高的时候聚成的类簇的数目越多，类内相似度越高，但是可能会遗漏部分表达方式不一样的文本，影响精度。相似度阈值越低，聚成的类簇数目越少，会尽可能包含各种表达方式的文本，但是可能类内会聚集实际语义关联不大的文本。因此相似度阈值 s 的选择要综合考虑聚类的精度和聚类完整性通过反复实验进行选择。

（3）问题热度计算

问题热度的计算首先要定义问题热度的指标，结合现实生活应用与政务网站的功能，秉持不人为制造热点、不人为抹消热点，不漏掉热点、不放大热点的原则，可以通过以下几方面进行考量：

1）类内文本数量越多的类应具有更高的热度

类内文本数量是最重要的衡量指标，代表某个问题被反映的频数，并且有详细的问题描述，可信度是最高的。

2）类内文本的点赞和反对数对类热度有影响，但不超过文本数量的影响

相对于详细的问题描述，点赞和反对的操作要相对简单很多，可信度要下降许多。有些网站需要登录才能点赞或者反对，可信度稍高，有些网站不登录就可以点赞或者反对，可信度更低。同时为了避免人为刻意点赞制造热点和人为刻意反对抹消热点，点赞数和反对数只能在一定程度上影响热点，影响程度可视具体情况斟酌确定。

3）某个问题被同一用户反复反映，应给予一定的权重

同一用户反复反映的问题，代表该问题长期未得到解决，对反映人的影响较大，应给予一定的权重。

4）不同用户反映的问题应比同一用户多次反应的问题具有更高的热度

不同用户反映的问题代表影响群体较大的问题，表示该问题出现的广度，应比同一用户重复反映的问题热度更高

综合以上考量得到话题热度评价公式：

$$\text{Hotwei} = (\text{Totals} - \sum_{i=1}^{i=n} SameU * \alpha)(1 + Thumwei)(1 - Agawei) \quad (4)$$

其中 $Totals$ 代表类内样本总数，SameU 代表类内同一用户数量，$(Totals-\sum_{i=1}^{i=n}SameU+n)$代表类内相同用户重复反映的次数，α 为重复留言剔除系数，Thumwei 代表点赞指数，Agawei 代表反对指数。Thumwei 计算公式为：

$$\text{Thumwei}=\begin{cases}\mathcal{E},Cthum>\mathcal{E}Totalthum\\ \frac{Cthum}{Totalthum},Cthum<\mathcal{E}Totalthum\end{cases}\quad(5)$$

其中 Cthum 为类内点赞总数，Totalthum 为所有样本点赞总数，ε代表点赞数目对热度的贡献上限，可根据实际情况设定。

Agawei 代表反对指数，计算公式为：

$$\text{Agawei}=\begin{cases}\mathcal{E},Caga>\mathcal{E}Totalaga\\ \frac{Caga}{Totaltaga},Caga<\mathcal{E}Totalaga\end{cases}\quad(6)$$

其中 Caga 代表类内反对总数，Totalaga 代表所有样本反对总数。ε代表反对数目对热度的抵消上限，可根据实际情况设定。

（4）热点话题提取

在聚类完成之后，因聚类过程的时间复杂度较高 O（n^2），考虑后续计算效率，可以计算出话题热度 Top5 之后，聚类结果只保留 Top5，将 Top5 类内信息写入热点明细表，从类内选择其中一条内容作为整个话题的代表写入热点问题表。热点问题的选择应具有代表性，能反映出整个问题集合的内容，因此在每个类内选择与类内其他文本具有最高相似度和或平均最高相似度的文本作为话题代表，平均相似度计算公式如下：

$$\text{Sim}_i=\frac{1}{n-1}\sum_{j=1}^{j=n-1}\cos(i,j)\quad(7)$$

其中 n 为类内文本数量，j 代表类内其他文本，cos(i,j)代表两条文本的余弦距离。

（5）热点问题地点/人群和问题描述提取

因本次竞赛热点信息表需要写入地点/人群和问题描述，因此提取出热点问题代表后还需要将文本分解为地点/人群和问题描述。通过观察留言主题，可以发现本次竞赛留言主题的语法结构具有高度的相似性，主要可分为两类：

1、以地点或人群开始，问题描述在后。如“A 市伊景园滨河苑协商强制购买车位”、“A3 区西湖街道茶场村拆迁问题”等，可以提取动词之前连续的名

词作为地点/人群，以动词开始的部分作为问题描述。

2、以动词开始，如“请问”、“咨询”、“反映”等，地点/人群位于两个动词之间。如“反映 A3 区西湖街道茶场村拆迁问题”，可以提取两个动词之间的部分作为地点/人群，其余部分作为问题描述。

通过使用具有词性分析功能的分词工具匹配词性，按照以上两种结构建立两种分解模式，分别提取地点/人群和问题描述后写入热点问题表。

3.4. 答复意见的评价

3.4.1. 评价指标设计

我国的政府网站建设可以 2017 年为界分为两个时期，在 2017 年以前各地的网站建设属于粗放型，由各地政府部门独立建设，网站建设也没有统一规划，风格、栏目各不相同，重复建设、网站重建设轻维护的现象比较严重。自 2017 年开始，国务院办公厅印发《国务院办公厅关于印发政府网站发展指引的通知国办发〔2017〕47 号》文件，对政府网站的发展给出了方向，要求各地推进集约化建设、推进一网通办、加强网站维护，并对网站的风格、栏目、管理等给出了指导意见[9]。此后全国清理了大量网站，形成了目前各级政府独立的门户网站和统一的政务服务网的格局。因此目前的政务网站的评价原则多来自于国务院的有关文件以及在此基础上各省、自治区、直辖市形成的具体指标。

在 2015 年开展的第一次全国网站普查的基础上，从 2017 年开始每季度进行全国网站抽查，国务院发布抽查情况通报，给出了明确的扣分项。各省、自治区、直辖市依据国务院的通报进行本省的抽查通报，并依通报的问题形成本省、自治区、直辖市的更为详细的要求。我们依据《政府网站发展指引》和国务院《2017 年第一季度全国政府网站抽查情况的通报》[10]等有关文件，设计了及时性、相关性、完整性三个指标。

1）及时性指标

及时性是个相当重要的指标，根据《国务院办公厅关于开展第一次全国政府网站普查的通知 国办发〔2015〕15 号》中的指标介绍，有超过 3 个月未回复的留言，单项否决[11]，因此本文取 3 个月为留言回复最长期限。各省、直辖市、自治区的有关通报如《云南省人民政府办公厅关于 2017 年一季度全省政府网站抽

查情况的通报》、《河南省人民政府办公厅关于2017年一季度全省政府网站抽查情况的通报》等，对于及时回复的指标有1日或2日各不相同，我们本次确定为5日。

具体而言：留言回复期限为5日以内的，及时性指标满分，超过5日至3个月期间递减，直到扣完。

2）相关性指标

相关性指标代表留言回复与留言内容的相关性，以回复的内容解答了留言为准，避免答非所问[10]，可以通过留言回复和留言内容的相似度乘以相关性指标分值来衡量。

3）完整性指标

完整性代表留言回复必须包含指定的要素，本次确定两个要素。一是在当前依法治国的背景下，无论解答的留言是咨询、投诉或是建议都应当有相应的法律、法规、政策、文件等依据，因此要求答复内容须包含有关法律、法规、政策、文件依据。二是根据政府网站发展指引的要求，对于留言除文字答复外，应提供电话等其他渠道，即便不属于本单位职责范围的也应当提供电话等其他咨询渠道，因此要求答复内容应包含办公电话，通过正则表达式匹配文本中的电话格式串和文本中的书名号和特定字眼来提取电话和依据。

3.4.2. 评价体系

结合以上指标得出留言答复评价指标，如表1所示。

表1. 政府网站留言回复评价表

评价指标	及时性指标	相关性指标	完整性指标
指标分值	20	60	20

评分标准	1、答复时间在 5 天内，20 分 2、答复时间超过 90 天，0 分 3、答复时间在 5-90 天之间，按比例递减。	相关性*60（相关性以文本相似度衡量）	1、答复中出现法律、法规、政策、文件，得 10 分 2、答复中出现电话，得 10 分

3.4.3. 算法实现过程

留言答复评价算法过程如图 8 所示，针对每一项指标分别提取相应的内容，逐项指标计算得分后，将三项指标得分合计后作为整体得分，按照留言编号保存为表格文件。

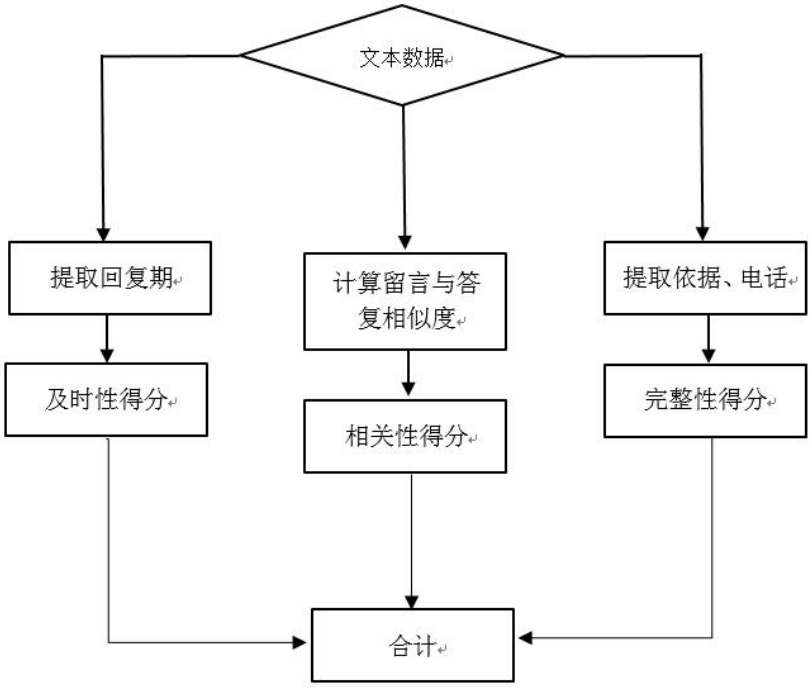

图 8. 留言评价得分计算流程

4. 实验结果

本次实验采用 python 3.7，使用 Pycharm 2019 软件平台在 windows 7 环境下进行开发。

4.1. 群众留言分类

4.1.1. 模型训练情况

本次实验使用附件 2 的数据，使用 jieba 分词工具进行分词，按照划分的训练集、测试集进行模型训练，划分比例按照训练集、测试集、验证集 7:1.5:1.5 进行划分。在模型训练过程中使用 tensorboard 工具监控训练过程中精度的变化情况如图 9 所示，可以看到随着训练逐步稳定，精度在达到在 0.9-1 之间有小范围的震荡，最终在迭代次数或者测试过程中长时间精度没有提高后结束训练。

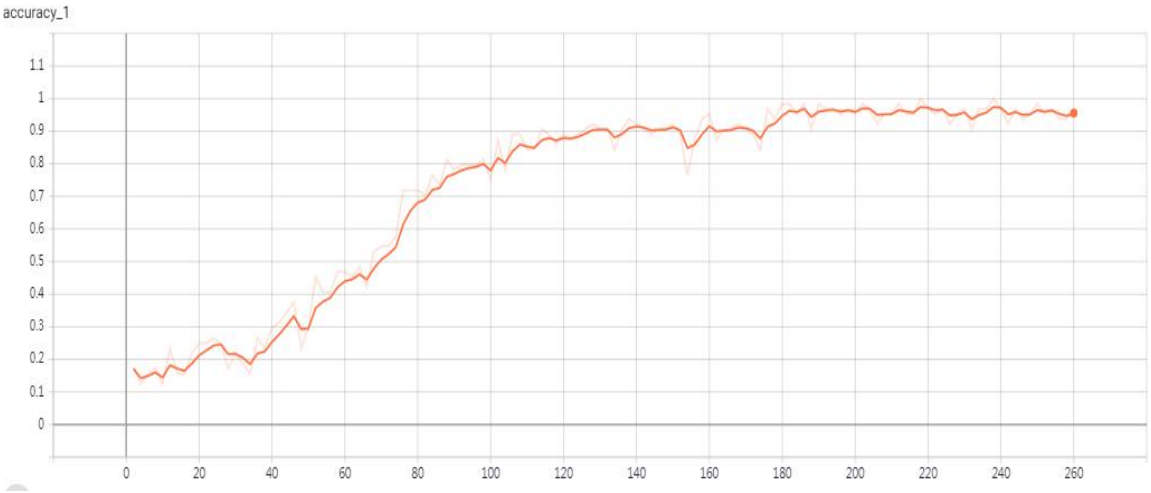

图 9. 留言分类模型训练过程精度变化图

训练过程中损失函数的变化情况如图 10 所示，随着训练的进行，损失函数的 loss 值在 0.05 左右小范围震荡一直到训练结束。

图 10. 留言分类模型训练过程损失函数变化图

4.1.2. 测试情况

使用划分的验证集进行分类测试，输出每个类的查准率和查全率计算F-Score，可以看到各个类别的查准率，查全率以及F-score得分，总体平均查准率为0.91，平均查全率0.90，平均F-score达到0.90，该模型具有较为优秀的分类能力。

```
Loss:   0.35, Acc:  90.97%
Precision, Recall and F1-Score:
              precision    recall  f1-score   support

        城乡建设       0.89      0.92      0.91       287
        卫生计生       0.89      0.91      0.90       133
        商贸旅游       0.89      0.87      0.88       195
        交通运输       0.90      0.84      0.87        89
     劳动和社会保障       0.90      0.95      0.92       286
        教育文体       0.96      0.91      0.93       233
        环境保护       0.92      0.91      0.91       158

    accuracy                           0.91      1381
   macro avg       0.91      0.90      0.90      1381
weighted avg       0.91      0.91      0.91      1381
```

图11. 留言分类模型验证集上查准率、查全率和F-score指标

4.2. 热点话题挖掘

4.2.1. 热点话题挖掘

在附件3数据上进行热点话题挖掘，将相似度阈值取为0.93，点赞指数和反对指数的上限确定为0.5，同一用户重复反映的剔除系数定为0.5，输出的热点话题表如图12所示，从图中可以看到，排名在前五的热点问题分别是“滨河苑小区车位问题”、“A市住房公积金问题”、“西湖街道茶场村的拆迁问题”、“魅力之城小区油烟问题”，热点的地点/人群和问题描述具有较好的可理解性。因本次数据中隐藏了实际的地区信息，对分词造成了较大的干扰，分词工具无法将“A市A7区”识别为一个词语，从而影响了精度，在实际应用中使用真实的地址，使用分词工具的地理信息词典或者调用地图接口效果可以进一步更高。在本次实验中为通过词性分离地址/人群和问题描述，在提取位置/人群和问题描述时使用了hanlp分词工具进行词性分析。

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	27.99	20190417至20190901	A市伊景园滨河苑	伊景园滨河协商要求购房同时必须购买车位
2	2	11	20190116至20191225	A市住房公积金贷款问题	咨询住房公积金贷款问题
3	3	11	20190106至20190912	A3区西湖街道茶场村五组	咨询西湖街道茶场的拆迁规划
4	4	10.99	20191118至20200125	小区搅拌站噪音	投诉小附近搅拌站噪音扰民
5	5	9.74	20190721至20190925	A5区	劳动东路魅力之城小一楼的夜宵摊严重污染附近的空气，急需处理！

图 12.　Top5 热点问题图

热点问题明细表如图所示，因明细表较长，折叠后只显示其中部分内容，从表中内容看，第一类都集中在“滨河苑车位问题”，第二类集中在“公积金咨询”，第三类集中在“西湖街道茶厂村的拆迁问题”，第四类集中在“丽发新城搅拌站噪音问题”，第五类集中在“魅力之城小区油烟污染问题”，可以看出该方法具有较好的聚类问题的能力，结合热点问题表，可以看出热点问题的选择也具有较好的代表性，综合来看具有较好的实际应用效果。但是从聚类内容上也可以看出，数据本身的规范性也有待提高，数据中对于同一个位置的描述出现了多种描述，有“市/区/小区”格式，也有“市/小区”，也有直接使用小区名称的，这对于数据挖掘会造成比较大的干扰。在本次实验中使用了停用词典对部分词语进行过滤，以便降低干扰。这一类问题对原始数据提出了更高的要求，可以在收集留言时在界面通过下拉单等固定格式或者通过调用互联网上地图接口的方式形成规范的地理位置描述，对于提高热点问题挖掘的精准度将会有较大的帮助。

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	190337	A00090519	关于伊景园滨河苑捆绑销售车位的维权投诉	43700.51528	投诉伊景园.滨河苑开发商捆绑销售车位！A市房产新城片区下的伊景园.滨河苑是广铁集团铁路职工的定向商品房，之前已经统一交了18.5万的认购款，但没有正规的合同，现在集团下发文件，强制要求职工再交12万的车位费，不交就取消购房资格，捆绑销售车位！请问相关部门领导，强制捆绑购买车位不属于侵犯消费者权益吗？还请相关部门切实维护职工权益，取消捆绑销售要求！	0	0
2	202249	A00092267	投诉A8总住房公积金无法办理组合贷款	2019/5/10 15:25:47	长金管委（2018）8号、长金管发（2019）1号、长金管发（2019）3号都明确表示，自2019年2月1日起，鼓励因取货调查公积金贷款额度不够部分，可以采取公积金商贷组合贷款，可A市地区A8总市因房产部门的原因几个月都执行不了，请问A8总住保局和A8总公积金管理部是怎么回事？能不能为百姓贷款买房干点实事？	0	0
3	189093	A00051608	A3区西湖街道茶场村五组什么时候能启动征地拆迁	2019/2/21 12:02:17	请问胡书记，A3区西湖街道茶场村六组已经因为A3区山景区修建西大门而拆迁了，，，据悉，西湖街道茶场村五组已经前两年被A3区山大学科技城报批了用地红线手续，，，那么换着的五组什么时候能启动征地拆迁呢？？	0	0
4	189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	43782.47247	我是A2区丽发新城小区的一名业主，我要投诉同发投资有限公司在未经小区业主同意的情况下，在离小区不到百米的地方建搅拌站，可想而知，一个大型搅拌站每天的噪音输出有多严重！还有扬尘污染，极度影响了我们的生活，请尽快关闭或搬迁该搅拌站！	0	0
5	236798	A00039089	A5区劳动东路魅力之城小区油烟扰民	2019/07/28 12:49:18	尊敬的政府：A5区劳动东路魅力之城小区临街门面长期油烟直排，长期投诉无果，由于各部门无权力强关，目前油烟机清洗也没有，每天油烟直排，黑死树木，对环境污染巨大，并且经常深夜经营，各种噪音扰民，无法生活，请求市政府出来管理一下，影响开封无到不在，	0	4

图 13.　热点问题明细表

4.2.2. 热点问题聚类情况分析

对聚类的问题按照各种问题包含的文本数量进行简单的可视化分析可以发现，大部分问题的反映次数在 2 次以内，绝大部分的问题反映次数都在 5 次以内，反映次数 5 次以上的只占有很小的比例，说明反映的问题整体上很分散，没有在较大范围内形成热点。这可能是实际问题本身较为分散，也是由于网站的覆盖面不够大，留言更新不够及时造成的。近些年随着微博、微信、政务 APP 的推广，网站的用户活跃程度相对于移动端已经大幅度下降，热点问题数据可以考虑从多个渠道进行收集，更加便捷的留言收集方式可以鼓励用户更加广泛的参与，留言数据量的增大，留言数据的及时更新，对于数据挖掘效果也有较大的帮助。

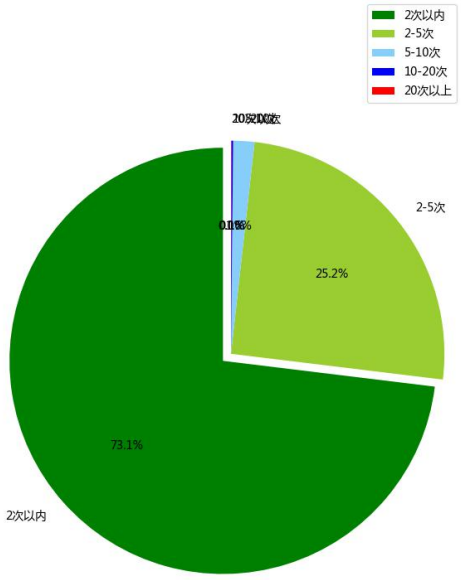

图 14.　热点问题聚类情况图

4.3. 答复意见的评价

4.3.1. 留言回复评价表

在附件 4 的数据上运行评价算法，得到各条留言回复的评价表格如图所示，第一列为留言编号，第二列为及时性得分，第三列为相关性得分，第四列完整性得分，第五列为得分合计。

编号	及时性得分	相关性得分	完整性得分	合计
2549	15.56	57.48	10	83.03
2554	15.56	54.6	0	70.15
2555	15.56	57.91	10	83.47
2557	15.56	57.44	20	93
2574	15.33	58.05	0	73.39
2849	9.78	58.25	20	88.03
3681	12.44	57.16	20	89.61
3683	15.33	57.31	0	72.64
3684	15.33	56.49	0	71.83
3685	3.11	56.87	0	59.98
3692	12	57.32	10	79.32
3700	15.33	55.1	0	70.44
3704	17.56	50.73	0	68.29
3713	15.11	55.45	0	70.57
3720	3.56	57.06	10	70.61
3727	17.33	53.82	0	71.15
3733	14.67	59.41	0	74.08
3747	15.78	56.78	10	82.55

图 15.　留言评价得分情况图

4.3.2.　留言评价结果分析

对每项评价指标进行简单可视化分析，留言回复期限分布情况见图 16，对及时性指标主要关注留言回复期限超过 3 个月的较差情况和回复期间在 5 天以内较好的情况。从图上可以看出，本次提供的数据的回复期限在 5 天以内和超过 3 个月的都较少，主要集中在 5 天到 3 个月之间。随着国家放管服改革的进行，对于各类业务的承办期要求逐步提高，各地都在推行“358 制度”、“710 制度”等，不断提高及时性要求，5 天到 3 个月的期限针对留言答复来说仍然过长，留言回复的及时性仍然有待提高。

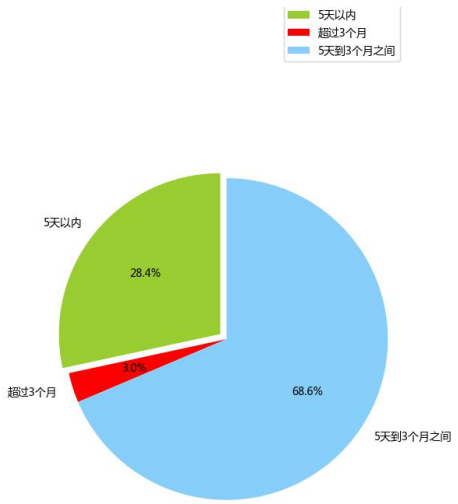

图 16.　留言回复期限分布

留言回复完整性情况见图 17，从图中可以看到本次提供的数据中回复的完整性较差，较多的留言回复中既没有有关依据也没有电话出现，留言的完整性方面有待加强。

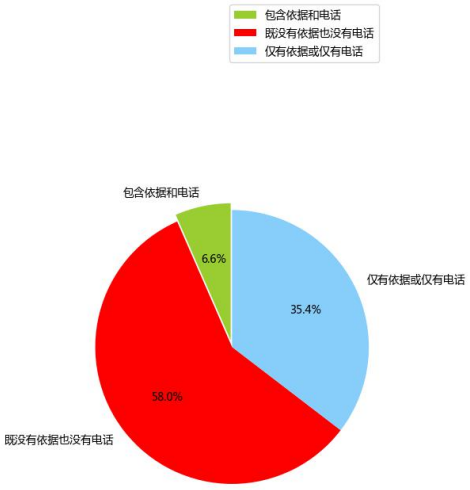

图 17.　留言回复完整性情况

留言回复的相关性情况如图 18 所示，相关性低于 50%为很差，相关性介于 50%到 70%为较差，相关性介于 70%到 90%为较好，相关性高于 90%为很好。从图中可以看出，本次提供的留言回复数据的相关性都基本处于较好和很好的水平，相关性整体情况较好。

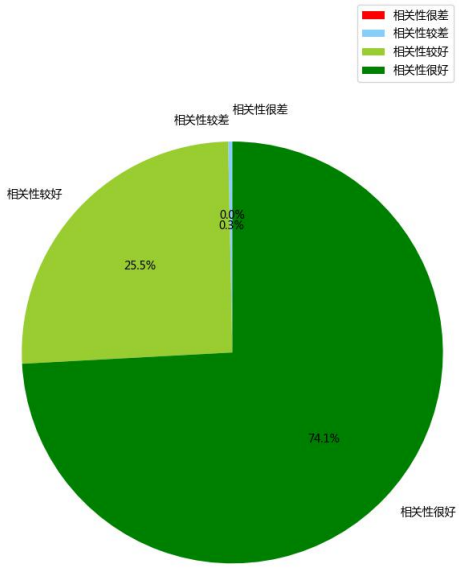

图 18.　留言回复相关性情况

5. 总结和展望

本文通过构建文本分类模型、热点话题识别算法和留言回复评价体系，实现了政府网站留言的自动分类、热点话题识别和留言回复自动评价，在实际留言数据上实现了较好的效果，并可以对问题分布和留言回复状况进行简单分析，具有实际应用的价值。但是也要看到，因为留言数据比较分散，提取的热点问题的热度值并不很高，这是源于留言数据来源比较单一，仅有来自网站的内容，缺乏微博、微信等用户活跃程度更高的来源，未能及时提取到较大范围内的留言数据，而且因留言地址表述不一致、不规范也对挖掘效果造成了一定的干扰，可以通过过优化留言界面的方式改善数据质量，将能够更加精准的提取到热点问题。相信随着微博、微信等移动端产品加入政务应用，数据数量、数据质量可以得到进一步提升，数据挖掘的效果可以进一步得到提升，政务文本的挖掘也可以不仅局限于留言，可以进一步对政务业务、信息公开等进行更广范围的数据挖掘，数据挖掘将可以在政务服务领域发挥更大的作用。

参考文献

[1] 唐明，朱磊，邹显春．基于 Word2Vec 的一种文档向量表示[J]. 计算机科学,2016,43(06):214-217+269.

[2] 周飞燕，金林鹏，董军．卷积神经网络研究综述[J]. 计算机学报,2017,40(06):1229-1251.

[3] Yoon Kim. Convolutional Neural Networks for Sentence Classification.

[4] 胡新辰. 基于 LSTM 的语义关系分类研究[D].哈尔滨工业大学,2015.

[5] 黄建一,李建江,王铮,方明哲.基于上下文相似度矩阵的 Single -Pass 短文本聚类[J]. 计算机科学,2019,46(04):50-56.

[6] https://baike.baidu.com/item/tf-idf/8816134?fr=aladdin.

[7] 张瑾．基于改进 TF-IDF 算法的情报关键词提取方法[J]. 情报杂志,2014,33(04):153-155.

[8] Gerard Salton, Christopher Buckley.Term-weighting Approaches in Automatic Text Retrieval[J].Information Processing & Management, 1988, 24 (5) :513-523.

[9] 《国务院办公厅关于印发政府网站发展指引的通知 国办发〔2017〕47 号》

[10] 《国务院办公厅秘书局关于 2017 年第一季度全国政府网站抽查情况的通报》

[11] 《国务院办公厅关于开展第一次全国政府网站普查的通知 国办发〔2015〕15 号》