“智慧政务”中的文本挖掘应用

摘　要

互联网问政平台社会认知度越来越高，网络问政平台逐步成为政府了解民意等重要渠道。社情民意相关的文本数据量不断攀升，给相关部门的工作带来了极大挑战。建立一个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。

本次数据挖掘的目的是根据题目所给的收集来自互联网公开来源的群众问政留言记录，以及各对应部门对群众留言的答复意见，解决群众留言分类、热点问题挖掘和答复意见质量评价三个问题。

针对问题一：本文首先将附件 2 中的数据进行去重去噪，中文分词，停用词过滤等预处理，采用 word2vec 模型对数据进行预训练使其变成维度为 100 的词向量序列。将一级标签编码，并赋给相对应的留言详情，将处理后的训练集输入 CNN 进行训练，并通过可视化观察模型的 accuracy 值与 loss 值调整模型参数，最后得到七个一级标签平均 F1 值为 0.84。

针对问题二：本文首先将附件 2 中的数据进行去重，使用 FoolNLTK 对留言主题进行命名实体识别获得名称，将名称在留言详情中进行频率统计，剔除频率少的词汇，构建字典。基于字典再通过一系列的预处理得到文本集。将文本集通过 TF-IDF 构建 TF-IDF 矩阵来对字典中地名或人群名进行加权处理，得到文本向量矩阵，再通过 OPTICS 聚类对得出的进行筛选，计算热度，通过热度值排序，得出排名前五的热点问题。

针对问题三：本文将附件 4 中的数据进行预处理，再将数据使用 word2vec 提取词向量，然后使用 TF-IDF 的权值矩阵对 word2vec 进行加权平均，再通过留言与答复句向量之间余弦相似度得出答复质量的评价分数。

关键词：Word2vec　TF-IDF　CNN　OPTICS 算法　文本聚类　命名实体识别

Text Mining Application in "Smart Government"

Abstract

Internet political platform social awareness is getting higher and higher, the Internet political platform gradually become the government to understand public opinion and other important channels. The increasing volume of text data related to social and public opinion has brought great challenges to the work of relevant departments. The establishment of a intelligent government system based on natural language processing technology is already a new trend of innovation and development of social governance, which has a great effect on improving the level of government management and the efficiency of governance.

The purpose of this data mining is to collect the records of public questions from the Internet from public sources according to the topic, as well as the corresponding departments to reply to the people's message, to solve the mass message classification, hot issue mining and reply to the quality evaluation of opinions three questions.

For question one: This paper first of all, the data in annex 2 to re-noise, Chinese word-sharing, de-stop word filtering and other pre-processing, the use of word2vec model of pre-training the data to make it into a dimension of 100 word vector sequence. The first-level label is encoded and assigned to the corresponding message details, the processed training set is entered into CNN for training, and the model parameters are adjusted by visually observing the accuracy and loss value of the model, and the average F1 value of the seven first-level labels is 0.84.

For question two: This article first of all, the data in Annex 2 to re-emphasis, using FoolNLTK to name the message topic to identify the name, the name in the message details of the frequency statistics, remove the less frequent vocabulary, build a dictionary. Based on the dictionary, the text set is obtained through a series of pre-processing. The text set is constructed by TF-IDF TF-IDF matrix to weight the place name or population name in the dictionary, the text vector matrix is obtained, and then the resulting is filtered by OPTICS clustering, the heat is calculated, and the top five hot issues are obtained by the heat value sorting.

For question three: This paper pre-processes the data in Annex 4, then uses word2vec to extract word vectors, then uses TF-IDF's weight matrix to weight the word2vec, and then draws the evaluation score of the response quality by leaving a message and the cosine similarity between the reply sentence vector.

Key words: Word2vec TF-IDF CNN OPTICS text clustering named entity recognition

目　录

一、挖掘目标 1
二、问题分析 1
2.1 问题 1 分析 1
2.2 问题 2 分析 1
2.3 问题 3 分析 1
三、总体流程 2
四、文本预处理 3
4.1 文本去重 3
4.2 命名实体识别 3
4.3 中文分词 3
4.4 停用词过滤 4
4.5 正则表达式 4
五、算法介绍 4
5.1 词袋模型 4
5.2 Word2vec 模型 4
5.3 卷积神经网络 CNN 5
5.4 TF-IDF 7
5.5 LDA 主题聚类 8
5.6 Dbscan 聚类模型 9
5.7 Optics 聚类模型 9
六、问题求解 11
6.1 群众留言分类 11
6.1.1 文本数据处理 11
6.1.2 群众留言分类情况 11
6.1.3 分类模型评价 13
6.2 热点留言挖掘 15
6.2.1 热点问题 15
6.2.2 热点问题划分 15
6.2.3 热度评价及排名 17
6.3 留言答复意见评价方案 17
6.3.1 答复意见介绍 17
6.3.2 数据处理 17
6.3.3 答复评价模型构建 18
七、总结 20
八、参考文献 21

一、挖掘目标

本次数据挖掘的目的是根据题目所给的收集来自互联网公开来源的群众问政留言记录，以及各对应部门对群众留言的答复意见，通过文本去重、实体命名、中文分词、停用词过滤、正则化等文本预处理后，使用相应的文本挖掘方法实现以下三个目标：

（1）根据留言内容的数据特点建立卷积神经网络 CNN 模型，实现对群众留言内容进行分类。

（2）利用特征提取和文本聚类并结合热点问题自身定义内容，实现对热点问题的挖掘。

（3）针对留言内容的答复意见的相关性、完整性、可解释性等制订一套评价方案。

二、问题分析

2.1 问题 1 分析

问题 1 是针对附件 2 数据及一级标签进行多分类。日常人工分类都是通过关键字词去判断句子的类别，故该题可以通过关键词特征去进行机械学习，使模型能通过句子中的词判断类别。该问题可以将数据去重去噪、停词等预处理减少干扰词汇。进行中文分词及 Word2vec 向量化，提取文本特征词与表示文本。建立卷积神经网络 CNN 文本分类模型对数据进行学习，进行一级标签分类并根据题目所给的 F-Score 值对分类方法评价。

2.2 问题 2 分析

问题 2 是针对附件 3 数据的热点问题挖掘。热点问题是一个地点短时间内获得多关注的问题。该题通过文本数目进行文本聚类得出预选问题。我们可以先对数据进行命名实体识别提取出地名，人群名等词汇加入词典，经过数据预处理，通过 TF-IDF 提取文本特征，以及对提取出来的命名实体进行加权，然后用聚类模型得到预选文本，通过热度计算，将热度值前五的文本输出作为热点问题。

2.3 问题 3 分析

对留言问题的分类还不是能解决社会群众问题的关键，最重要的是将群众留言问题分派到相应的各个机构部门后，相关部门能够给出合理的答复意见与实际行动。而答复的质量至关重要，为此，我们提出了一个以 TF-IDF 为权重，对 Word2vec 做加权平均的余弦相似度模型，通过得到的结果作为答复意见的质量

的评价。

三、总体流程

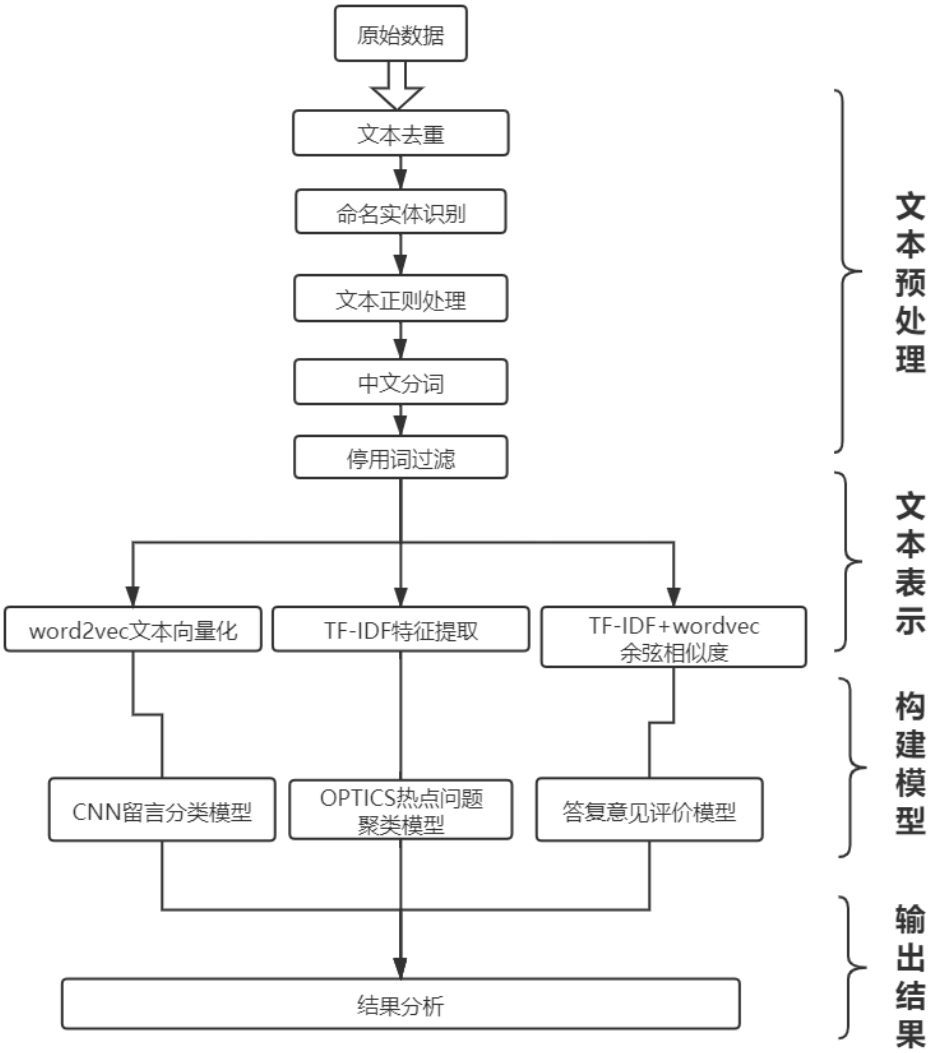


图 1 整体流程图

本文的整体思路主要有以下四个步骤：

步骤一：文本预处理。在原始数据中有许多干扰后期模型建立的因素，必须先对它们进行预处理操作，具体依次为文本去重、命名实体识别、文本正则处理、中文分词、停用词过滤，便于接下来的工作。

步骤二：文本表示。针对所建模型的特点，本文采用不同的方法对文本内容进行处理，使之成为计算机能够识别的形式，为模型的建立做准备。

步骤三：构建模型。为了处理群众留言分类、热点问题挖掘和答复意见评价，分别搭建了 CNN 留言分类模型、Optics 热点问题聚类模型和答复意见评价模型这三个模型。

步骤四：输出结果。每个模型将会输出相应的结果，对结果进行分析得出结

论。

四、文本预处理

下面我们将针对附件 2.xlsx、附件 3.xlsx 及附件 4.xlsx 中的留言主题、留言详情和答复意见内容进行文本预处理，去除带来干扰的数据，为后面的模型建立做准备，以免影响模型效果。

4.1 文本去重

政务留言常常会出现留言用户多次留言、使用他人用户留言、纠正留言等行为，故文本会出现文本数据重复的情况，故要进行文本去重。通过我们初期对各个附件的反复筛查，发现了文本中无空值，但群众问政留言记录中多次出现了两条甚至两条以上的相同的留言信息（包括留言用户、留言主题、留言详情）。

这些重复的留言信息对后期解决问题没有太大意义，反而会增加解题负担，所以，我们对这些重复了的留言信息选择一条保留，去掉重复的信息。

名称	初始数目	处理后数目
附件 1	9210	8956
附件 2	2816	4207
附件 3	2788	2788

表 1 文本去重对比表

4.2 命名实体识别

命名实体识别指的是从文本中识别出人名、地名、时间等具有特定意义的实体。命名实体识别主要是针对问题 2 中对热点问题的挖掘。热点问题主体由时间地点事件三个部分组成，故对地名等的识别显得尤为重要。

由于留言详情中含有大量的噪音词，会导致命名实体识别的不准确以及速度慢。基于观察到留言主题一般拥有地名等命名实体，故本文采取 FoolNLTK 模块对留言主题逐条进行命名实体识别。FoolNLTK 模块基于 BiLSTM 模型训练而成，虽然速度慢，但是命名实体识别准确率高。

对识别出来的实体仍有噪音及分得不准确的词语，如：（本人系春华镇金鼎村七里组），人工切分串联一起的词语以及删除部分不合理的词语，构造字典文本。再将字典文本进行处理，去重、同义词替换，如：（街道—街，路口—路），最后将字典文本与留言详情比对，将出现频率较小的词去掉，构成本文的字典。

4.3 中文分词

中文分词(Chinese Word Segmentation) ，即将连续的字序列按照一定的规则切分成一个一个单独的词，重新组合成词序列的过程，是文本挖掘的基础，属

于自然语言处理范畴。词，是可以独立活动且具有意义的最小的语言成分，在英文文本中，词与词之间是以空格作为分界符分开的；而中文文本是以字为书写单位，词与词之间没有明显的分界符，所以，在自然语言处理技术中，对于一段中文文本，成功准确地对它进行中文分词，是十分重要的。本文采用 jieba 分词结合添加自定义词典的方式对留言内容进行分词处理，得到分词结果。

4.4 停用词过滤

对留言信息进行中文分词，将文本划分为词之后，会发现划分完的这些词条中含有大量的单一独立字或词，如“个”、“的”、“就”、“了”、“啊”等，这些单一独立字或词不但携带的文本信息量极少、对文本的标识和表达无意义，而且还会在我们特征选取或者对其他有价值的词进行分析时，起到一定的抑制作用，降低了分类、聚类系统的准确度和处理效率。这些应被剔除的字或词称为停用词(Stop Words)。本文采用停用词表的停用词过滤方法，将文本中文分词后的结果与所选的停用词表进行文本匹配，如若匹配成功，则对该字或词进行删除。

4.5 正则表达式

除了通过停用词过滤的符号和字词外，留言内容中出现了大量像”0000-00000000”、” https://baidu.com/”的无效数字串和网站地址等，这些与我们对留言内容的分析无关，并会影响对数据的处理效果，故我们采用正则表达式对它们进行删除操作。

五、算法介绍

5.1 词袋模型

BOW 词袋模型原理是不考虑词的语法和语序问题，将所有的词都装进一个袋子里，然后对词语做映射，每个句子以向量的形式被表示出来，向量中的数表示对应的词在这个句子中出现的次数。在答复意见评价问题中采用的是 Sklearn 模块下的 CountVectorier 来构建词袋模型，由于词袋模型仅只单单考虑了词的量，当语料库过大时会产生维度灾难，以及常用词在句子中出现的数量也是很多导致对句子关键意思的影响。所以，本文为解决这些不良影响，在使用词袋模型时，会引入 TF-IDF 算法，该算法会在后文介绍。

5.2 Word2vec 模型

Word2vec 是一种由谷歌开源出的词嵌入工具。在 Word2vec 被提出之前，词向量自然语言处理通常是采用 one-hot representation，这种文本向量化方式

虽然简单易得，无需复杂的操作，但是它有着较大的缺点：（1）可能会由于词汇表长度过长，导致维数灾难的发生；（2）无法很好地体现词与词之间的的关系。

而 Word2vec 则解决了 one-hot representation 的缺陷，其作用是将文本中的字词转为计算机可以理解的稠密向量（Dense Vector），它的实现过程经过输入层、隐藏层和输出层三层。

输入层(Input Layer)：输入以 one-hot 表示的向量。

隐藏层(Hidden Layer)：此层的神经元数目等于输入的词向量长度。从输入层来到隐藏层的所有 one-hot 行向量会与该矩阵相乘，最终通过隐藏层。

输出层(Output Layer)：输出层的维度与输入层的维度一样，利用 softmax 回归，得到归一化之后的词向量。

文本向量化训练过程中，可以通过调整 Word2vec 中的 size、min_mount 等参数来提高训练时的速度与实现词向量的优质化。

在本文中我们采取的是 word2vec 的 Skip-gram 算法，其原理是通过给定当前中心词，最大化上下文单词的概率，在我们语料库相对较小有不错的效果，下图是其原理过程：

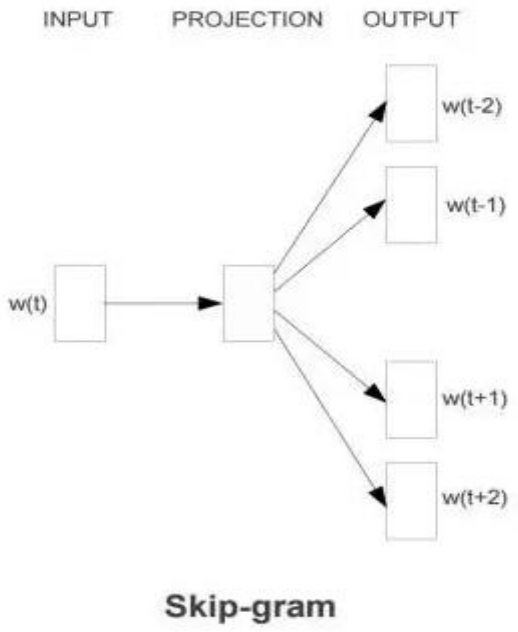


图 2 Skip-gram 原理图

5.3 卷积神经网络 CNN

CNN，是一种由 Yoon Kim 提出的利用卷积神经网络对文本进行分类的算法，可实现特征提取和分类同时进行，在图像分类领域取得了显著的成效。近几年来，卷积神经网络已逐步成为了研究热点之一，它主要是采用多层卷积、池化等过程对输入的文本或图像进行特征抽取与降维，再经过全连接层采用 softmax 算法实现分类的过程。

优点：

（1）不需要手动选取特征，只需训练好权重，就可得到好的特征分类效果。

（2）卷积核可共享，处理维数高的数据无压力。

（3）使用参数少，却有出色性能。

（4）支持特征提取和分类同时进行。

缺点：

（1）在进行模型训练时，样本量要大。

（2）要对参数进行选择调整。

本文将选取卷积神经网络对群众留言内容进行一级标签分类。卷积神经网络 CNN 的模型结构一般包括下面几层：

第一层：输入层

顾名思义，输入层即用于数据的输入，输入的是一个经向量化的文本的矩阵。假设该文本由 n 个词向量组成，且每个词向量维度为 d，则输入的矩阵就应是 n×d 的 n 行 d 列矩阵，由于卷积层的输入需要每个词向量的维度需要一致，这里我们把 d 设置为 100，并把向量矩阵归一化。

第二层：卷积层

卷积层，是卷积神经网络中的一个核心层次，在每个卷积层中含有若干个卷积单元。在卷积层，若是对图片处理，则是通过卷积核对图片中某一块像素区域进行操作；而对自然语言处理，主要是在词向量上进行操作，在这里我们我们分别进行卷积核大小为 3、4、5 的卷积操作，通过从上往下多次卷积运算实现对特征的提取，然后通过 ReLU 非线性激活函数与池化层连接，这里的特征就是所谓的词向量。其操作如下图：

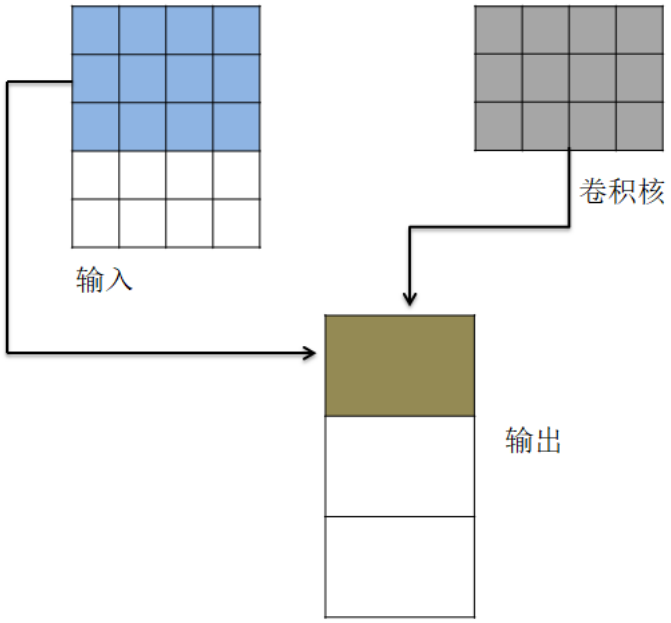


图 3 卷积层操作示意图

第三层：池化层

池化层，也叫做下采样层。将上一层次卷积层中提取的特征输入到该层，通过池化操作去掉一些不重要的样本，从而实现降低数据维度，避免过拟合。常见的池化方法有均值池化、随机池化和 Max-Pooling 最大池化法，本文是采用该最大池化法对特征进行池化操作。这种方法不被输入的句子长度所限制，原理是输出各个特征文本的最大值，作为采样后的样本值。

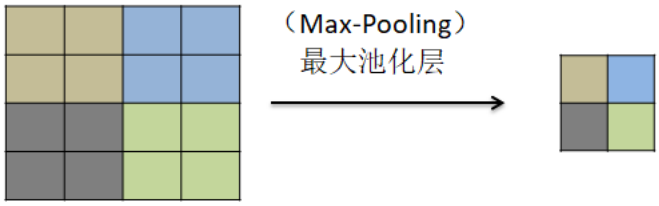


图 4 最大池化法示意图

第四层：全连接层

在全连接层主要是通过 softmax 函数实现对图片或文本的分类，得到分类模型，它输出的是一个 N 维向量，维数 N 就是分类类别数目，由于题目给我们的数据只有七个类别，为此，我们采取的是根据一级标签的七分类模型。

第五层：输出层

5.4 TF-IDF

TF-IDF 由词频(TF)和逆文档频率(IDF)两部分构成：

词频((Term Frequency，TF)表示某一个词语或关键字在文档中出现的频率，其计算公式如下：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中，分子 $n_{i,j}$ 是特征词 t_i 在文档 d_j 中出现的次数，分母 $\sum_k n_{k,j}$ 是文档 d_j 中所有词语的个数。

逆文档频率(Inverse Document Frequency，IDF)反映特征词在文档集中的重要程度，其计算公式为：

$$idf_i = \lg \frac{|D|}{|\{j : t_i \in d_j\}|}$$

其中，分子 $|D|$ 表示所有文档的数目，分母 $|\{j : t_i \in d_j\}|$ 是指包含特征词 t_i 的文档总数（即 $n_i, y \neq 0$ 的文档数），可能会出现文档数为 0 的情况，会使得整个分式无意义，故一般情况下分母使用 $1+|\{j : t_i \in d_j\}|$。即 idf_i 是文本数量与某一个特征词 t_i 在文本集中出现的次数比值。

综上，TF-IDF 计算公式如下，其结果即为权重：

$$tf - idf_{i,j} = tf_{i,j} \times idf_i$$

其基本思想是一个词在某一文档中出现的频率(即 TF)越高，则其表示文档

内容的能力就越强；而如果一个词在其他的所有文档中出现的频率越高，所包含的文档信息量越少，则该特征词区分不同文档的能力越弱。

5.5 LDA 主题聚类

LDA(Latent Dirichlet Allocation)是 Blei 在 2003 年提出的基于贝叶斯的文档主题生成模型，又称为三层贝叶斯概率模型，它从上到下包含文档、主题、词这三层结构。它采用了词袋的方法，将每篇文档视为一个词频向量，把文本化为易于建模的形式，使用狄利克雷分布求解最终的概率分布，确定潜在主题。

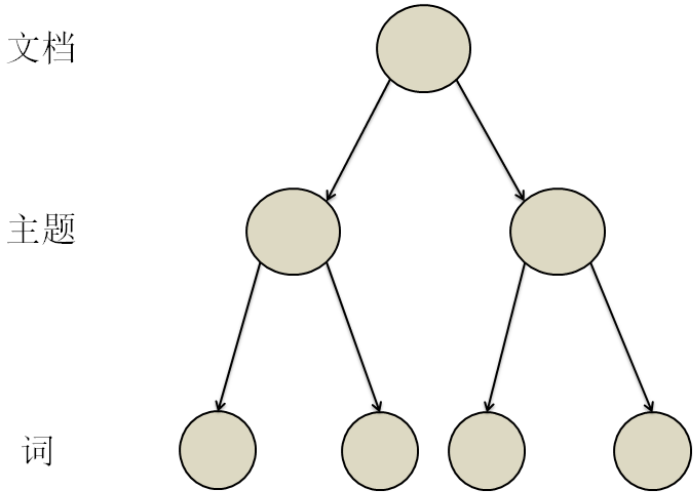


图 5 LDA 结构图

其整体算法步骤是：

（1）先定义文档集合D，话题集合T，D中所有不同的单词组合成一个大集合 VOC。

（2）对文档集合D中的每篇文档d进行预处理，得到语料集合$W=\{w_1,w_2,...,w_k\}$。

（3）对每篇文档d对应到集合 T 中不同话题的概率$\theta_d<P_{t1},...,P_{tk}>$，其中$P_{ti}$ i 表示d对应T中第i个话题的概率，公式为：

$$P_{ti}=n_{ti}/n$$

（4）对每个话题集合 T 的t生成不同单词的概率$\varphi_t<P_{w1},...,P_{wm}>$，其中$P_{wi}$表示$t$生成 VOC 中第$i$个单词的概率，公式为：

$$P_{wi} = N_{wi} / N$$

（5）LDA 的核心公式如下：

$$p(\ \psi\ )d= \quad (p\ \ |w)\ t \quad ($$

5.6 Dbscan 聚类模型

Dbscan (Density-Based Spatial Clustering of Applications with Noise) 是基于一组邻域来描述样本集的紧密程度的的算法，它有两个参数邻域距离阈值 Eps 和距离为 Eps 的邻域中样本个数的阈值 $Minpts$ ， $(Eps, Minpts)$ 可用来描述邻域的样本分布紧密程度。对于基于划分的 K-means 聚类算法和基于层次的 BIRCH 聚类算法一般只适用于对凸样本集的聚类，而 Dbscan 既适用于凸样本集，也适用于非凸样本集。该算法优点是能发现任意形状的空间聚类、可有效处理噪声点。

算法原理简介：

第一步，先输入数据库，半径 Eps 和最少数目 $Minpts$ ；

第二步，选择任意一个未被访问的点，看看它在半径 Eps 的领域内的点的个数，若大于 $Minpts$ ，则该点与这个领域中的点可构成一个簇；若小于 $Minpts$ ，则该点暂时被标记为噪声点。

第三步，一直循环第二步，直至所有点都被访问。

Dbscan 算法有个缺点就是对参数的选择极其敏感，微小的差别都可能导致分类效果很大的变化。

5.7 Optics 聚类模型

Optics 聚类算法是一种基于密度的聚类算法，可视为 Dbscan 算法的一种改进升级版算法，它改进了 DBSCAN 算法中对输入参数敏感的问题，而且可以获得不同密度的聚类。基于密度的算法与其它算法的区别是：它不是基于距离的，而是基于密度的，改善了其他算法中只能发现“类圆形”聚类的缺点。它的指导思想是，只要一个区域中的某个点密度大于某个阈值，就把这个点加到它附近的聚类中去。

Optics 聚类算法中几个基本概念介绍：

（1） Eps 邻域：数据集 D 中除 x_i 以外，与 x_i 距离小于等于半径 Eps 的样本点的集合。可记为： $N_{Eps}(x_j) = \{x_j \in D \mid dist(x_i, x_j) \le Eps\}$

（2） $Minpts$ ：给定的 Eps 邻域中最小的样本量数。

（3）核心对象/核心点：对于 $\forall x_i \in D$，如果其 Eps 邻域中样本点数大于 $Minpts$，则样本点 x_j 是核心点，也叫核心对象。

（4）直接密度可达：如果 x_i 是核心点，x_i 在 x_i 的 Eps 邻域中，则称 x_i 由 x_i 直接密度可达。

（5）密度可达：若存在样本序列 $p_i, p_2, \ldots, p_n$，p_{i+1} 由 p_i 直接密度可达且 p1 的核心对象是 x_i，p_n 的核心对象是 x_i，则称 x_i 由 x_j 密度可达。

（6）密度相连：若 $x_i, x_j, x_k \in D$，x_i 和 x_j 均由 x_k 密度可达，则称 x_i 和 x_j 密度相连。

（7）核心距离：设 $x_j \in D$，在给定参数 Eps 和 $Minpts$ 的情况下，使得 x_j 成为核心对象的最小半径叫做 x_j 的核心距离。数学表达式如下：

$$cd\left(x_j\right)=\begin{cases} undefined & ,\left|N_{Eps}(x)<Minpts\right| \\ d\left(x_j, N_{Eps}^{Minpts}\left(x_j\right)\right) & ,\left|N_{Eps}(x)\geq Minpts\right| \end{cases}$$

（8）可达距离：设 $x_i, x_j \in D$，在给定参数 Eps 和 $Minpts$ 的情况下，xj 关于 xi 的可达距离可用如下数学表达式计算：

$$rd\left(x_j, x_i\right)=\begin{cases} undefined, & \left|N_{Eps}(x)<Minpts\right| \\ \max\{cd(x_i), d(x_i, x_j)\}, & \left|N_{Eps}(x)\geq Minpts\right| \end{cases}$$

Optics 聚类运算流程：

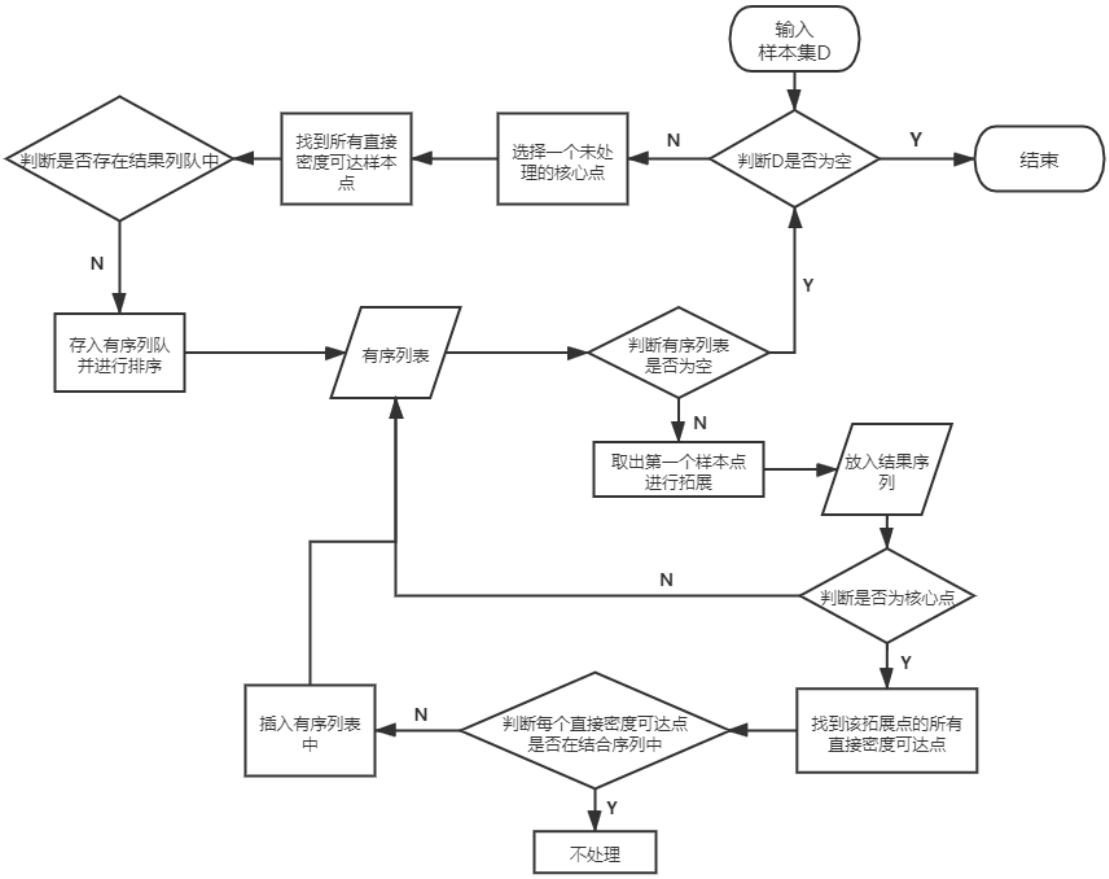


图 6 Optics 原理流程图

运算过程结束，则输出结果列队中具有可达距离信息的有序样本点。

六、问题求解

6.1 群众留言分类

6.1.1 文本数据处理

读取附件 2.xlsx 文本，对其进行预处理后得到文本集，将文本集按 8:2 的比例划分训练集和验证集，将文本集作为语料库放入 Word2vec 模型中，并使用其中的 Skip-gram 算法提取词向量，得到词向量矩阵。并对训练好的词向量做归一化完成了预训练过程，为卷积神经网络提供了训练数据。

6.1.2 群众留言分类情况

根据附件 1 中的一级分类标签对附件 2 的所有留言内容进行分类，分类结果如下图，其中城乡建设类留言共计 1989 条记录，劳动和社会保障类共计 1961 条，教育文化类共计 1566 条，商贸旅游类共计 1159 条，环境保护类共 921 条，卫生计生类总计 874 条，交通运输类共 598 条。

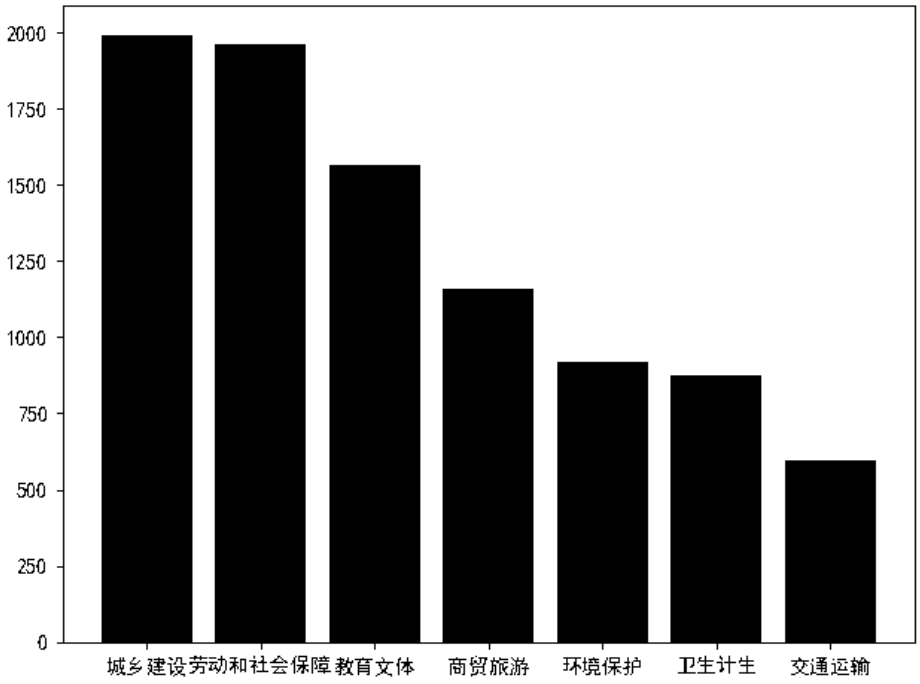
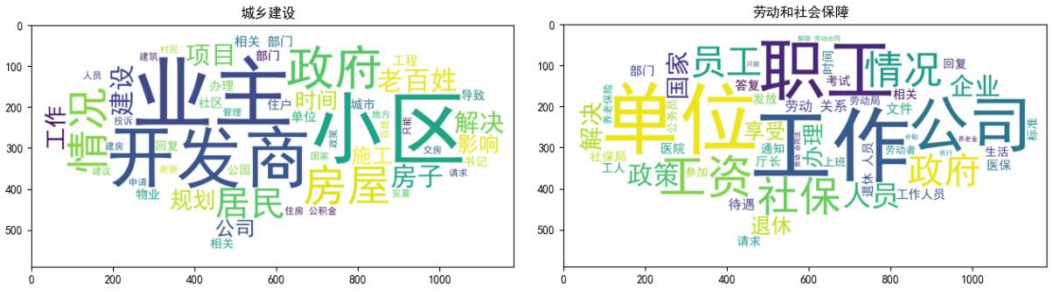


图 7 留言分类汇总图

一级标签	总计
城乡建设	1989
劳动和社会保障	1961
教育文化	1566
商贸旅游	1159
环境保护	921
卫生计生	874
交通运输	598

表 2 各一级标签总计表

为了方便对每个留言类别中重点内容的获取，本文对每个留言类别的词频都做了词云图，图中词越大，则代表该词在留言中出现的频率越高，反映留言内容的能力越强。



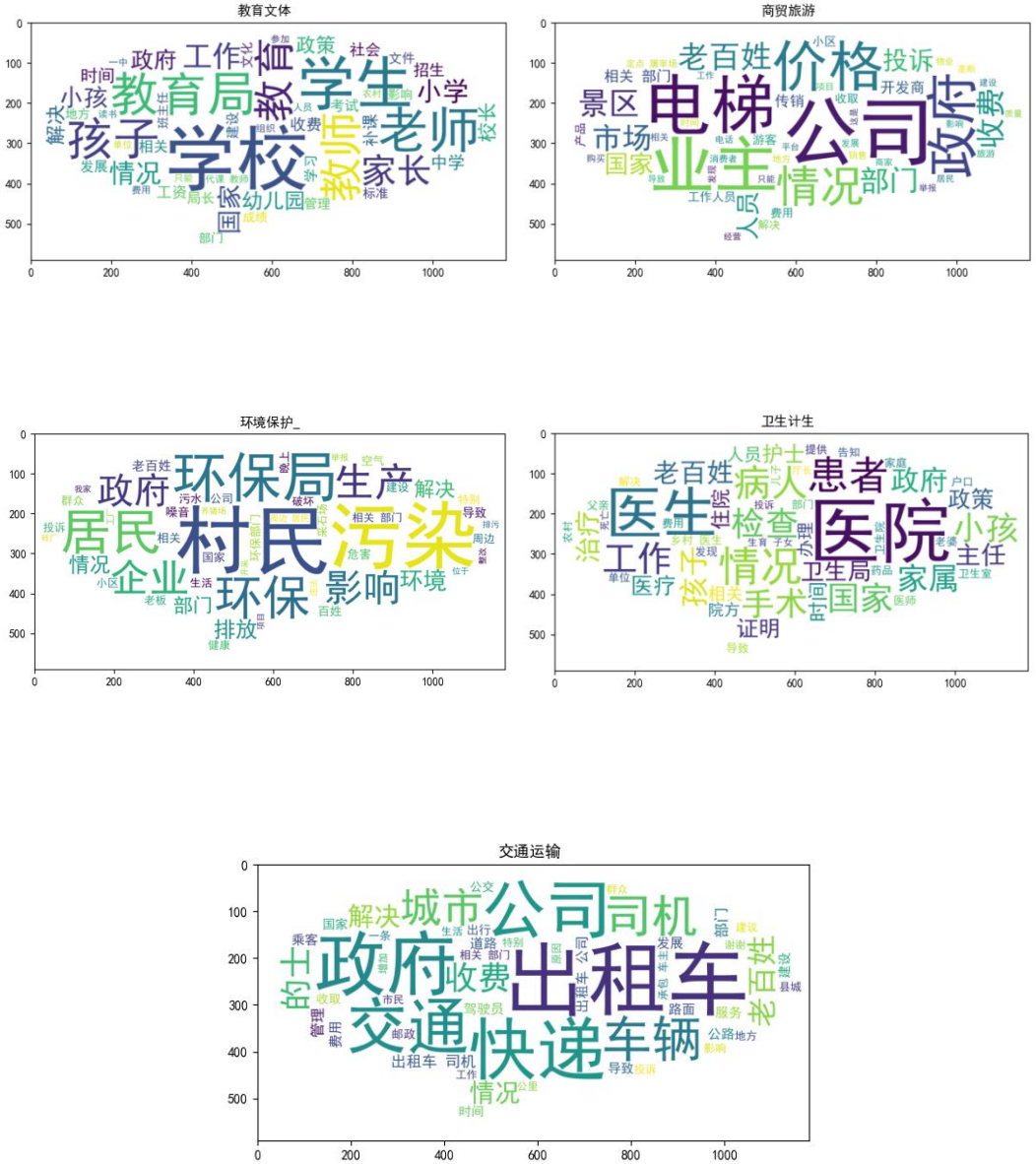


图 8 七大留言分类词云图

6.1.3 分类模型评价

（一）模型结果

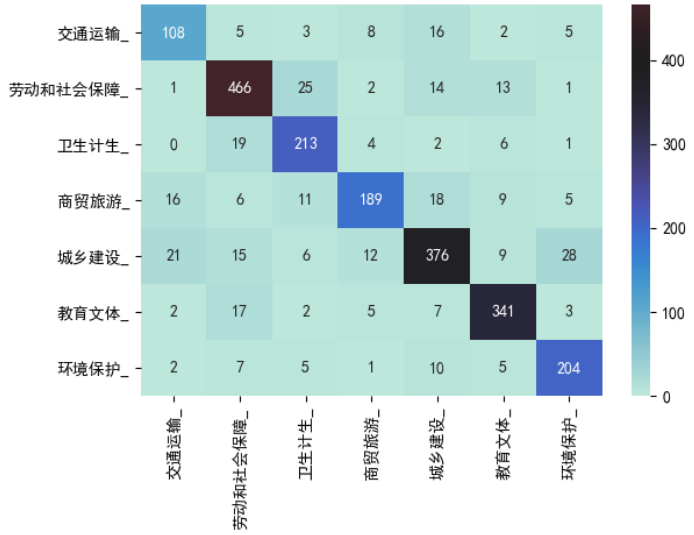


图 9

通过卷积神经网络的迭代训练之后得到的分类效果如上图所示，结果表明，大量数据都分布到斜对角线当中，而其他地方的数据量极少，说明我们大部分的数据都分类对了，只有少量的数据分错，达到了我们分类模型的预期效果。

（二）模型评价

通过对学习率为 0.01、卷积核数量为 200，卷积核大小为 3 的卷积神经网络 20 次迭代训练之后我们得到如下结果：

	precision	recall	f1-score	support
交通运输_	0.75	0.63	0.69	134
劳动和社会保障_	0.85	0.90	0.87	502
卫生计生_	0.84	0.84	0.84	267
商贸旅游_	0.84	0.77	0.80	279
城乡建设_	0.78	0.84	0.81	443
教育文体_	0.90	0.89	0.90	398
环境保护_	0.91	0.85	0.88	227
accuracy			0.84	2250
macro avg	0.84	0.82	0.83	2250
eighted avg	0.84	0.84	0.84	2250

图 10 CNN 迭代 20 次结果图

模型的准确度达到了 84%，数据结果显示，除交通运输的准确度为 69%之外，其他分类都达到了 80%的准确度，可见该卷积神经网络的训练得到的分类模型效果还不错。

如下是 20 次迭代过程的 accuracy 和 loss 的变化过程：

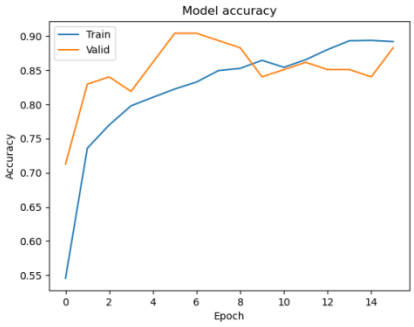


图 11

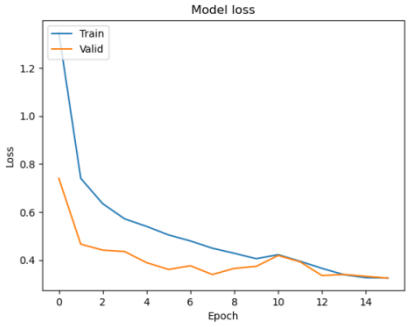


图 12

由上图可知模型的准确度在不断的提升，并且损失值在下降，同时训练集和验证集上的准确度和损失值相差较小，说明模型拟合效果不错，不存在过拟合的情况。

6.2 热点留言挖掘

6.2.1 热点问题

（一）热点问题定义

热点问题即是在某一个时间范围内、某一片地区、某一群社会人群集体反映的某一问题。通过我们的研究分析发现，热点问题具有以下特点：

（1）影响范围大，有许多人持续关注并发表看法。

（2）普遍性，在人们的社会经济生活中普遍存在。

（3）动态性，会随着该事件的发展和社会人群关注点的变化而发生改变。

（4）敏感性，对于一些事态较严重的事件，可能大众群体会对其相当敏感。

（二）文本数据处理

在文本聚类前，先进行如下处理，对附件 2 中数据进行命名实体识别得到词典，并将词典加入分词库中，再根据新的分词库进行分词，由于文本类别数未知，并且噪音过大，故采用 TF-IDF 进行特征提取，且对地名、人群名等实体乘以 1.2 进行加权处理，使文本等特征更加突出。

6.2.2 热点问题划分

（一）特征提取

在对热点问题留言进行划分前，必须要先对留言文本进行特征词提取。虽然在数据预处理中，已经对文本进行去重、实体命名、去停用词等操作，减少了无意义字词的干扰；但留言文本中仍然包含大量的字词，这将不利于文本聚类的精确度。所以，文本采用了传统的词频-逆向文档频率(TF-IDF)方法进行特征词提取，采用 TF-IDF 是原因是操作简单、速度快且已经足以满足我们对文本的处理需求。通过特征词的提取，不仅使得能体现留言核心内容的词语更加突出，而且降低了向量空间维数。本文将文本每一个句子通过构建 TF-IDF 矩阵表示向量，并且对命名实体识别中的词汇进行加权处理，使得同一地点或同一人群可以更好的凝聚到一起。

（二）文本聚类

文本聚类是一种无监督的、不需要经过训练、不需要先对文本进行手工标记类别，主要依据聚类假设：同一类文本相似度高，不同类文本相似度低的计算机

学习方法。使用相似度计算，将具有相同特征或者相似特征的文本聚类在一起。

本文依次采用了 LDA 主题聚类、DBSCAN 密度聚类和 Optics 模型聚类对热点问题留言进行归类，聚类如下：

1.LDA 主题模型

LDA 主题模型，本文使用参数 K 为 65 进行聚类，提取热点话题，并进行可视化观察。

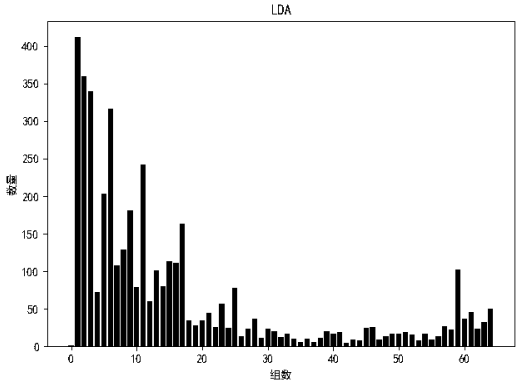


图 13　LDA 组别分布图

由 LDA 组别分布图可看到，组数小于 20 的组别，数量较高，而组别大于 20 的组别，数量较低，各组别间的差异大、分类不平均。

模型评价：LDA 聚类结果不理想，数据拥有大量的噪音项，LDA 主题模型无筛除的性质，会让噪音文本也加进聚类分组，而且会出现万能选项，影响分类结果，并且 LDA 主题模型的 K 值难以调整。

2.DBCAN 模型

由于 LDA 模型的聚类效果差，本文接着使用了 DBSCAN 模型进行聚类，DBSCAN 算法是基于密度的聚类算法，可以有效的筛除离群点，本文使用 Eps 为 1，Minpts 为 5 进行聚类，聚类效果如下：

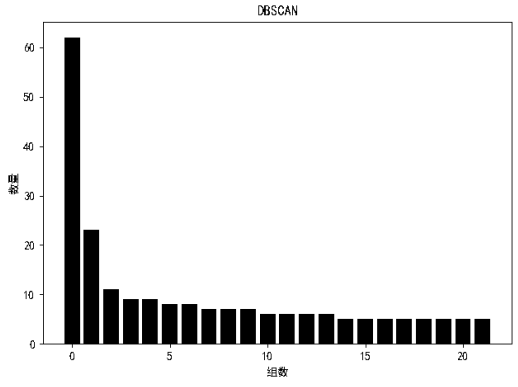


图 14　DBSCAN 组别分布图

DBSCAN 的组别分布图可以看到，聚类结果分布不平均，但能去除离群数据。

模型评价：DBSCAN 模型对参数 Eps 与 $Minpts$ 过于敏感，并且训练时间过长，而且也容易出现文本聚类不均匀的情况。

3.OPTICS 算法

由于 DBSCAN 对参数敏感的的缺点，本文采用了 OPTICS 算法对文本进行聚类，可以通过观察数据可达距离调整 EPS 的值，实验参数为，Eps=1.22，Minpts=12

进行聚类。

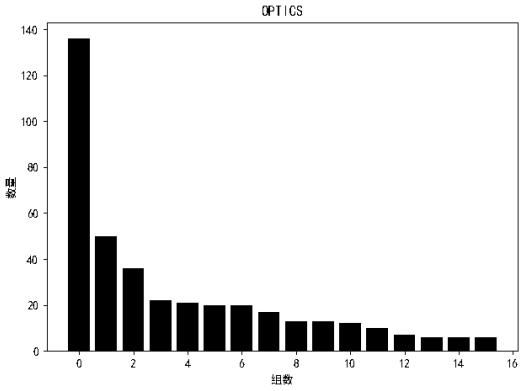


图 15 OPTICS 组别分布图

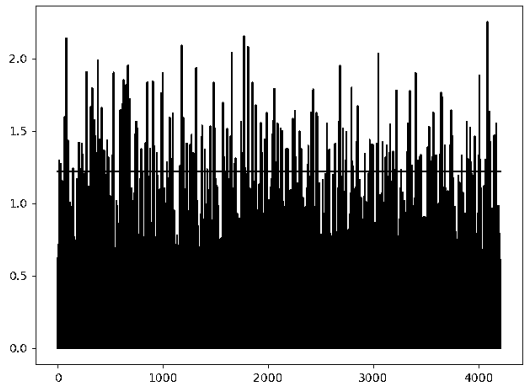


图 16 数据可达距离图

从 OPTICS 的组别分布图可以看到 OPTICS 比 dbscan 分组更加符合概率规律，但是仍是会出现分组不均匀的现象，对输出的文本进行人工的筛选，并将问题输出，作为候选问题。

模型评价：OPTICS 也会出现分布不平均的情况，但是仍达到了聚类的效果，相比于其他模型，OPTICS 对参数值没那么敏感，并且可以有效筛除离群点。

6.2.3 热度评价及排名

由于留言热度由留言数量 n，时间差 t，点赞数 z 与反对数 f 按照比例构成，参考清博大数据平台评价微博传播指数的新媒体指数(BCI Micro-blog Communication index)方式，确定每一类留言的热度值，热度计算公式如下所示：

$$heat = 0.5\times n + 0.4\times \ln(t) + 0.1\times \ln(z) - 0.1\times \ln(f)$$

再将每类留言内容根据其对应的热度指数进行排名，热度指数越高则排名越前，根据附件 3 的留言内容，已对其中的热点问题针对时间范围、地点人群等进行整合。

热度排名	地点/人群	问题描述
1	A 市伊景园滨河苑	小区车位捆绑销售
2	A 市丽发新城小区	小区附近搅拌站烟尘噪音扰民
3	A 市	住房公积金贷款新政不灵活
4	A 市 58 车贷关注者	案件多月无进展
5	A 市 A1 区辉煌国际城居民	小区楼下违法开饭店

表 3 排名前五的热点问题表

详情见“热点问题表.xls”。

6.3 留言答复意见评价方案

6.3.1 答复意见介绍

线上留言处理问题已经应用非常广泛，而答复的质量显得非常重要，据观察发现答复的意见也是千奇百怪，分析发现关于答复的相关性、完整性、可解释性等是判断答复质量的重要指标，并根据这些性质提出答复质量的评价方案。

6.3.2 数据处理

在数据处理之前统计留言和答复的总字符长度和留言、答复最长的字符长度，通过数据预处理后得到相对干净的数据集：

类别	长度
留言详情总字符	969259
答复意见总字符	996127
留言详情最长字符	3083
答复意见最长字符	7883

表 4 留言详情和答复意见字符长度表

6.3.3 答复评价模型构建

首先我们引入了一个关于 Word2vec 的余弦相似度方法，通过计算句子向量的余弦值来作为答复的质量评价指标：

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

由于 Word2vec 的相似度计算是采用平均值的，但是词语在句子中所占的分量是不同的，单靠 Word2vec 的余弦相似度计算准确度显然不够。为此我们提出了另一种评价方法，计算句子的加权平均值。我们对每个词都乘以一个权重值再来求解，通过采用 TF-IDF 计算的到的权值作为我们的权重结合 Word2vec 的词向量，通过加权平均的余弦相似度计算作为我们的答复质量评价指标。

加权平均公式：

$$S=\frac{\sum_{i=1}^{n}W_i\times I_i}{n}$$

加权平均后得到了 sentence embedding ，然后通过余弦相似度的计算公式计算得到结果即为我们的答复质量评价指标：

$$Q=\frac{dot(s_1,s_2)}{norm(s_1)\times norm(s_2)}$$

其中 s_1 、s_2 为加权平均值，dot 是 numpy 中的乘法函数，norm 为 scipy 模块中的函数，通过调用这两个函数来计算矩阵的余弦相似度。

属性	符号
TF-IDF 权值	W
Word2vec 词向量	I
加权平均	S
答复质量	Q

表 5 公式符号说明表

首先我们通过预处理之后的文本集来创建我们的语料库，对数据进行 8：2 的训练集和验证集切分，然后通过语料库来训练我们的 TF-IDF 权值矩阵，接下来使用我们的语料库进行 word2vec 模型训练，得到我们的词向量矩阵，然后通过权值矩阵和词向量矩阵的加权平均，再进一步做余弦相似度计算即可得到我们的答复质量评价分数。

其中 word2vec 的训练默认参数更改如下表：

参数	参数值
Skip-gram	1
维度	80
最少词数	1
并行训练数	4

表 6 word2vec 参数更改表

本文通过多次检验得出，当答复质量值大于 0.8 时，该答复质量可定义为高；当答复质量值大于 0.5、小于 0.8 时，该答复质量可定义为一般；当答复质量值大于 0.5 时，该答复质量可定义为差。

下面从附件 4 中随机抽取举例出一条留言详情、答复意见并对其作答复意见评价。

留言详情：

欢迎领导来A市泥泞不堪的小含浦镇滚泥巴，这个小镇的蓝天保卫战的奇葩战术:没有一台吸尘车，停留在用两台破洒水车把泥巴冲到左边，然后再把泥巴冲到右边，越是上下班，几台破机器越是这样！把整个整个含浦搞得泥泞不堪，没有一台干净车，居民民愤极大!!!全国最脏的小镇!职能部门严重造成水资源的浪费!

答复意见：

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

结果如下：

```
留言与答复的匹配程度为： 0.5515268243930987
该答复质量一般
```

由结果显示答复的质量一般，从人工去判断也比较符合实际。

七、总结

全文采用基于 Word2vec 的卷积神经网络 CNN 实现对群众留言内容进行一级标签分类；通过三种文本聚类模型的聚类结果对比，择优选取了 Optics 密度聚类算法将相似的热点问题划分为一类，并结合留言数量、热点问题时间差、点赞数与反对数设计热度指数算法对每一类热点问题进行热度计算与排名，分析得出排名前五的热点问题；最后针对意见答复内容，我们提出了一个以 TF-IDF 为权重，对 Word2vec 做加权平均的余弦相似度模型，通过得到的结果作为答复意见的质量的评价。

与此同时，我们的文本挖掘也存在着一些不足之处，在对热点问题进行划分类别时，并不能完全准确无误地将其进行划分，这也显示出了中文文本挖掘模型中的不足，接下来，我们也会继续进一步深入地对中文文本型数据进行研究。

八、参考文献

[1]曾凡锋,李玉珂,肖珂.基于卷积神经网络的语句级新闻分类算法[J].计算机工程与设计,2020,41(04):978-982.

[2]李海磊,杨文忠,李东昊,温杰彬,钱芸芸.基于特征融合的 K-means 微博话题发现模型[J].电子技术应用,2020,46(04):24-28+33.

[3]曹鲁慧,邓玉香,陈通,李钊.一种基于深度学习的中文文本特征提取与分类方法[J].山东科学,2019,32(06):106-111.

[4]张旭,孙玉伟,成颖.不同特征对文本聚类效果的比较研究——以新闻文本为例[J].情报理论与实践,2020,43(01):169-176.

[5]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.

[6]李彦冬,郝宗波,雷航.卷积神经网络研究综述[J].计算机应用,2016,36(09):2508-2515+2565.

[7]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.

[8]周练.Word2vec 的工作原理及应用探究[J].科技情报开发与经济,2015,25(02):145-148.

[9]曾依灵,许洪波,白硕.改进的 OPTICS 算法及其在文本聚类中的应用[J].中文信息学报,2008(01):51-55+60.

[10]荣秋生,颜君彪,郭国强.基于 DBSCAN 聚类算法的研究与实现[J].计算机应用,2004(04):45-46+61.

[11]孙学刚,陈群秀,马亮.基于主题的 Web 文档聚类研究[J].中文信息学报,2003(03):21-26.

[12]郑洪英. 数据挖掘聚类算法的分析和应用研究[D].重庆大学,2002.

[13]https://blog.csdn.net/Yellow_python/article/details/90034067

[14]https://blog.csdn.net/ACM_hades/article/details/90776525