“智慧政务”中的文本挖掘应用研究

摘要

随着网络的日益普及，互联网在我国民众的政治、经济和社会生活中扮演着日益重要的角色，成为我国公民行使知情权、参与权、表达权和监督权的重要渠道。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。

针对于问题一，为建立关于一级标签的分类模型，首先使用**改进的 TextRank 算法**对较长的留言详情进行关键句提取，采用同义词替换、回译等方法对样本进行数据增强以及预处理。通过 **Word2Vec 预训练模型**对留言详情词向量化，建立于一级标签分类的**双向 LSTM 神经网络**模型。结果表明，模型在训练集上的准确率为 99.27%，F1 分数为 0.99，在验证集上的准确率为 91.57%，F1 分数为 0.92，说明双向 LSTM 模型能够很好地对一级标签进行分类，并且具有不错的泛化能力。

针对于问题二，为提取特点地点或特定人群反映的热点问题，利用**TF-IDF** 对经过预处理的文本数据进行词向量化，然后用**DBSCAN算法**对留言主题进行聚类，并且利用留言详情关键词对聚类结果进行修正，得到排名前5的热点问题分别为“A市A2区丽发新城小区附近搅拌站灰尘、噪音扰民”“A市伊景园滨河苑开发商捆绑销售停车位”“市民咨询A市人才新政购房补贴问题”“A市A5区魅力之城小区临街餐饮店油烟噪音扰民”以及“A市A5区五矿万镜K9县存在房屋质量及物业管理问题”。最后利用***t*分布随机邻域嵌入算法**把文本向量映射到二维平面，对热点问题的聚类结果进行可视化，并给出相应的词云图。

针对于问题三，对留言的答复意见质量给出一套评价方案，建立了基于线性规划的**WMD模型**，用WMD距离衡量答复意见和留言详情的相关程度，然后结合答复意见长度和留言详情长度构建了关于答复意见质量的评价指标。结果表明，该指标能够较好地对答复意见质量进行评价：质量好的答复意见具有WMD距离小和答复留言对数长度比在1附近的特点，质量差的答复意见具有WMD距离大和答复留言对数长度比偏低的特点。

关键词：文本挖掘 双向LSTM DBSCAN *t*分布随机邻域嵌入算法 WMD模型

目录

一、问题重述.....2
1.1 问题的背景.....2
1.2 题目所给的信息和参数.....2
1.3 所要解决的问题.....2
二、问题分析.....3
2.1 问题一的分析.....3
2.2 问题二的分析.....3
2.3 问题三的分析.....3
三、模型假设.....4
四、模型的建立与求解.....4
4.1 问题一的解决.....4
4.2 问题二的解决.....13
4.3 问题三的解决.....21
五、模型评价.....27
5.1 模型优点.....27
5.2 模型缺点.....27
六、参考文献.....28

一、问题重述

1.1 问题的背景

随着网络的日益普及，互联网在我国民众的政治、经济和社会生活中扮演着日益重要的角色，成为我国公民行使知情权、参与权、表达权和监督权的重要渠道。

近年来，随着市长信箱、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，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。

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。

1.2 题目所给的信息和参数

附件 1 提供了一种关于内容分类的三级标签体系。

附件 2、附件 3、附件 4 均为群众在互联网上的问政留言记录。附件 2 主要有留言主题、留言详情和内容分类等信息；附件 3 主要有留言主题、留言相亲、点赞数和反对数等信息；附件 4 留言主题、留言详情和答复意见等信息。

1.3 所要解决的问题

问题一 群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

问题二 热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。请根据附件 3 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，定义合理的热度评价指标，并给出评价结果。

问题三 答复意见评价

针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

二、问题分析

2.1 问题一的分析

问题一要求建立关于留言内容的一级标签分类模型，附件 2 中的数据主要有留言主题、留言详情及其一级标签，首先要对文本数据进行长度统计、类别统计，去除文本中的一些无用字符，如空格、换行符等，对于文本长度过长的留言，用 TextRank 算法对其进行关键句提取，从而转化成短文本，以减少文本表示的维度；如果类别数量不均衡，则要对样本进行数据增强，常用的数据增强方法是欠采样和过采样。欠采样会丢失部分样本信息，而过采样只是单纯地复制样本。针对于文本数据的特殊性，基于过采样，可采用同义词替换、随机删除部分词语、对长文本重提取和回译等方法进行数据增强。然后对文本进行分词、去停用词以及词向量化。对于文本词向量化，采用开源的语料库可直接生成词向量，作为分类模型的输入，建立关于留言内容一级标签的双向 LSTM 神经网络分类模型。最后根据精确率、召回率和 F1-score 等指标评价模型的分类效果。

2.2 问题二的分析

问题二要求根据附件 3 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，定义合理的热度评价指标，并给出评价结果，给出排名前 5 的热点问题。附件 3 给出了留言时间、留言主题和留言详情等数据，一般来说，留言主题简洁精炼，能够很好地概括留言详情，包含了地点、人物、问题等信息，因此对留言主题进行聚类分析。聚类之前需要对文本进行分词和去停用词，在分词时不能把数据中“A 市”“A1 区”等地点分成“A，市”“A1，区”，否则会影响热点问题的提取。使用 TF-IDF 方法和词袋模型对留言主题进行词向量化。利用 DBSCAN 算法对留言主题进行聚类分析，用轮廓系数评价聚类效果，提取每个类别的关键词。如果样本没有被聚类到它本应在的类中，则需要结合留言详情进行修正。根据各类中的留言数量、点赞数和反对数构建热度指标，筛选出热点问题。最后利用 tSNE 算法对文本数据降维并对聚类结果进行可视化。

2.3 问题三的分析

问题三要求从答复的相关性、完整性、可解释性等角度对附件 4 相关部门对留言的答复意见质量给出一套评价方案。对于评价答复意见质量，可使用 WMD 算法计算答复意见和留言详情之间的相关程度，然后结合答复意见长度和留言详情长度构建关于答复意见质量的评价指标，对每条留言的答复意见进行评价。

三、模型假设

- 假设附件 2 中的数据都是正确分类的，类别有且仅有所给出的类别数量。
- 对于每一条留言，用户只能从点赞和反对点一个，或者不点，且每个用户只能点赞/反对一次。

四、模型的建立与求解

4.1 问题一的解决

问题一要求根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

4.1.1 数据预处理

- **数据描述性统计**

对附件 2 给出的留言主题、留言详情和一级标签进行描述性统计。数据中一共有 9210 条样本，一级标签有七个类别，每个类别在样本中的分布如表 1 所示。

表 1 一级标签分类统计

一级标签	数量	比例(%)
城乡建设	2009	21.81
劳动和社会保障	1969	21.38
教育文体	1589	17.25
商贸旅游	1215	13.19
环境保护	938	10.18
卫生计生	877	9.52
交通运输	613	6.66
合计	9210	100

从表 1 可以看出，大部分数据的一级标签为城乡建设、劳动和社会保障、教育文体，比例占了 50%以上，交通运输最少，仅占 6.66%。在数据不平衡的情况下建立模型对样本进行分类，在一定程度上会影响分类效果。

在留言主题中，最小长度为 2 个字符，最长长度为 48 个字符，平均长度为 19.49 个字符，样本分布可以近似看成正态分布，如图 1 所示。

在留言详情中存在大量无用字符，如换行符“\n”、空格“\u3000”、替换成“**********”的身份证号码等。去除无用字符后，留言详情的最小长度为 12 个字符，最长长度为 12161 个字符，平均长度为 383 个字符，96%的留言详情长度在 1500 个字符内，样本为右偏分布，如图 2 所示。

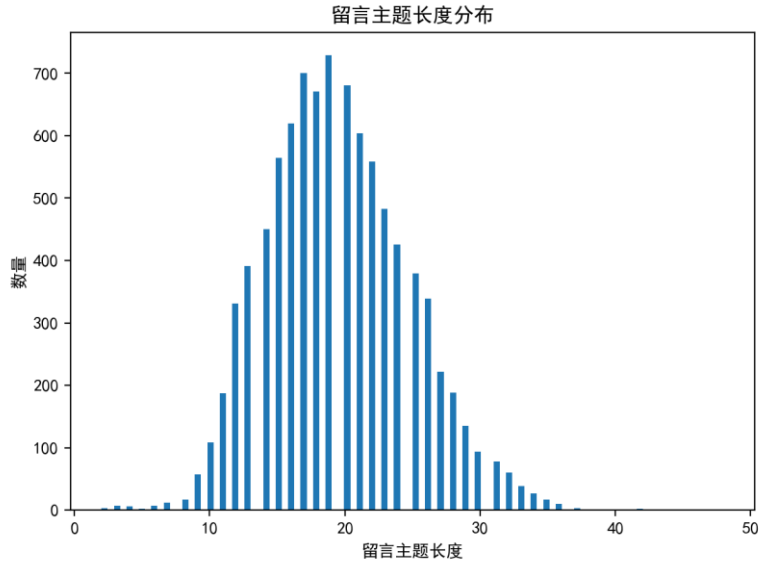

图 1 留言主题长度分布

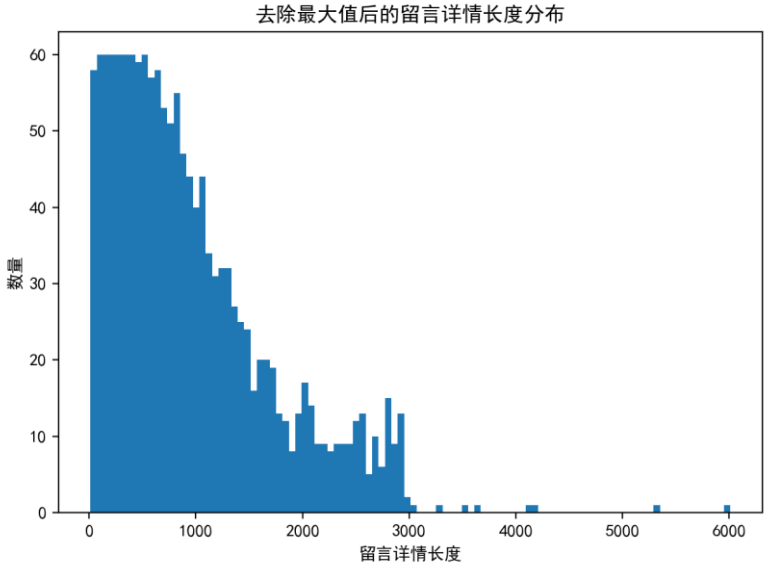

图 2 留言详情长度分布（去除最大值后）

➢ **数据增强**

为了降低一级标签不均衡对模型分类效果的影响，在建立模型之前需要对样本进行数据增强，使得每个类别的数量大致相同。

在数据标签不均衡的情况下，常用的数据增强[1]方法有欠采样和过采样等方法。但是，欠采样会丢失样本信息，而过采样只是单纯的重复了少数类别，对于重复的样本，特征缺少多样性。结合文本数据特性，基于过采样的方法，按照以下四种方法 4:1:3:2 的比例来对样本进行数据增强。

同义词替换：把句子中的一些词语用意思相近的词语替换。

删除词语：随机删除句子中若干个词语。

增加词语：对于较短的留言详情，增加从留言主题或二级分类和三级分类中提取的词语；对于过长的留言详情，提取若干关键句来代替。

回译：把句子先翻译成外语（英文），再翻译回中文。

经过数据增强后的样本数量为 12038，各类标签的数量比较接近，其分布如表 2 所示。

表 2 数据增强后的样本标签分布

一级标签	数量	比例(%)
城乡建设	2009	16.69
劳动和社会保障	1969	16.36
教育文体	1589	13.20
商贸旅游	1586	13.17
环境保护	1669	13.86
卫生计生	1675	13.91
交通运输	1541	12.80
合计	12038	100

- **分词和去停用词**

在建立模型之前对文本进行分词和去除停用词。

4.1.2 模型建立

从图 2 可以看出，留言详情中存在较多字符长度大于 1500 的长文本，为了降低长文本表示的维度，对长文本提取若干关键句转化短文本，这一过程相当于机器学习中的特征提取。本文基于 BM25 对传统的 TextRank 算法进行改进，进而对长文本进行关键句提取，然后建立双向 LSTM 神经网络模型对留言内容的一级标签进行分类。

4.1.2.1 长文本关键句提取

- **TextRank 算法**

TextRank 算法[2]是 Mihalcea 和 Tarau 于 2004 年在自动摘要提取任务中获得的成果，是一种基于图的无监督方法。对于长度大于 1500 字符的文本，本文利用 TextRank 算法提取关键句。

TextRank 算法提取长文本关键句的主要思想是：按照逗号、句号、感叹号和问号等标点符号把长文本分成若干个子句，每一个子句为一个文本单元。以这些文本单元作为节点，节点间是否相似确定边的存在，相似度的数值作为权重，从而构成文本网络图$G=(V,E,W)$，V是节点集合，E是节点间各个边的集合，W是各边权重的集合。用矩阵形式表示向量化后的文本，然后对该矩阵进行迭代计算，根据节点计算值的大小对节点进行排序，取排名靠前的子句组成集合。

根据有向网络图$G=(V,E,W)$，可以得到一个关于子句的n阶相似度矩阵$SM=(w_{ij})_{n\times n}$，矩阵中的每一个元素w_{ij}为句子V_i与句子V_j之间的相似度，其计算公式如下：

$$w_{ij}=\frac{\left|\{w_k|w_k\in S_i \& w_k\in S_j\}\right|}{\ln|S_i|+\ln|S_j|}$$

其中，S_i和S_j分别表示句子V_i和V_j的词语个数，w_k为句子中的词语。

根据G和SM不断地对各个节点的权重进行迭代计算，公式如下：

$$WS(V_i) = (1-d) + d \times \sum_{V_j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

其中，$WS(V_i)$为子句V_i的权重值；d为阻尼系数，取值范围为$(0,1)$，它表示在文本网络图中一个子句跳转到其他任何一个子句的概率大小，一般情况下将d设置为 0.85；$In(V_i)$为指向子句V_i的所有子句集合；$Out(V_j)$为子句V_i所指向的所有子句集合；等式右边的求和部分表示每个相邻的子句对子句V_i的贡献程度。

- **BM25 算法**

在信息检索领域中，BM25[3]是搜索引擎用来估计文档与给定查询的相关性的排名函数，它基于 Stephen E.Robertson、Karen Spärck Jones 等人在 20 世纪 70 年代和 80 年代开发的概率检索框架。

给定查询句子Q，对其进行分词得到词序列$\{q_1, q_2, \dots, q_t\}$。给定文档$d \in D$，计算Q与d之间的相关性，公式如下：

$$Score(Q,d) = \sum_{i=1}^{t} w_i \times R(q_i, d)$$

$$w_i = IDF(q_i) = \ln \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

$$R(q_i, d) = \frac{f_i \cdot (k_1 + 1)}{f_i + K} \times \frac{qf_i \cdot (k_2 + 1)}{qf_i + k_2}$$

$$K = k_1 \cdot (1 - b + b \cdot \frac{dl}{Dl})$$

其中，w_i表示词语q_i的权重，用逆词频 *IDF* 来表示；$R(q_i, d)$为q_i和d的相关性；N表示文本集合D中文本的总数量；$n(q_i)$表示包含词语q_i的文本数量；常数 0.5 用作平滑处理；f_i为词语q_i在文本d中出现的频率，qf_i为词语q_i在句子Q中出现的频率；dl和Dl分别为文本d的长度和文本集合D中所有文本的平均长度。k_1, k_2, b均为可调节的参数，参数b的作用主要是调解文本长度对相关性的影响。在本文中，$k_1 = 1.5, k_2 = 0, b = 0.75$。

从公式可以看出，$Score(Q,d)$就是每个词语q_i和给定文档d的相关性的加权求和。

- **基于 BM25 的 TextRank 算法**

传统的 TextRank 算法进行关键句提取时，仅仅考虑了两个句子共同包含词语的次数，缺乏对文本语义层面的把握。因此，本文用 BM25 算法计算句子之间的相关性，对传统的 TextRank 进行改进，再进行长文本的关键句提取。基于 BM25 的 TextRank 算法提取长文本关键句的步骤如下：

Step1 对长文本T按照逗号、句号、感叹号和问号等标点符号进行分割，即

$$T = \{S_1, S_2, \dots, S_m\}$$

Step2 $\forall S_i \in T$，进行分词和去掉停用词，即

$$S_i = \{t_{i1}, t_{i2}, \dots, t_{im}\}$$

Step3 $\forall S_i, S_j \in T$，计算句子之间的相关性$Score(S_i, S_j)$

Step4 $\forall S_i \in T$，计算每个句子S_i的权重，直至权重不再变化

Step5 对所有句子的权重进行降序排序，选出前k个句子作为长文本T的关键句

文本的预处理和特征提取过程如图 3 所示：

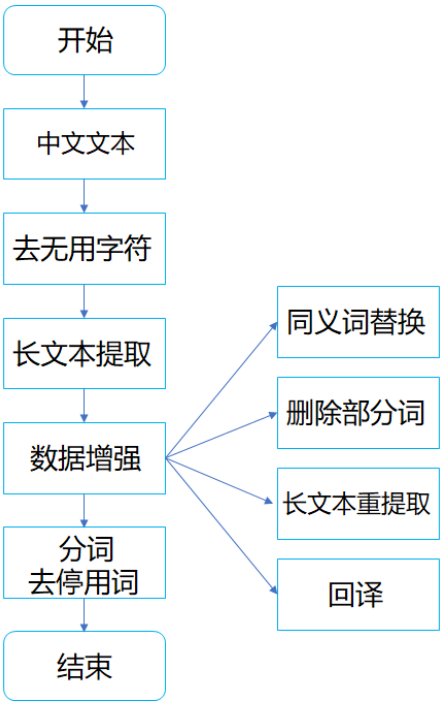

图 3 文本预处理和特征提取过程

4.1.2.2 双向 LSTM 神经网络模型

- **Word2Vec 预训练模型**

对文本数据进行预处理和特征提取之后，需要把文本数据转化成计算机能够计算的数据，这一过程就是文本词向量化。文本词向量化主要有两种实现方式，一是随机生成词向量，与下游任务的模型参数一同训练；二是采用网上开源的预训练模型生成词向量，直接作为下游任务的输入而无需对词向量进行训练。

本文采用北京师范大学中文信息处理研究所等开源的中文 Word2Vec 预训练模型（人民日报新闻语料），对经过预处理的留言详情词向量化，直接作为本文的分类模型的输入。

Word2Vec 模型[4,5]是 Google 公司在 2013 年开源的一种将词语转化为向量表示的模型，它是由神经概率语言模型[6]演化而来，对神经概率语言做了重要改进，提高了计算效率。Word2Vec 模型有两种主要的实现方式：CBOW 模型和 Skip-gram 模型，CBOW 模型是一个三层神经网络，输入已知上下文输出对下个词的预测；Skip-gram 模型也是一个三层神经网络，输入某个词语输出对其上下文词向量的预测。其模型结果如下图 4 所示。

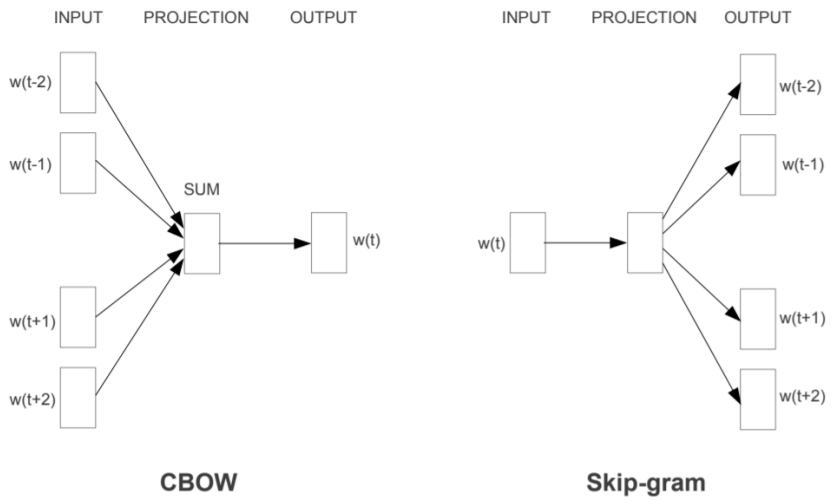

图 4　CBOW 与 Skip-gram 的模型结构

➢ **双向 LSTM 神经网络**

在神经网络中，循环神经网络模型[6](RNN)由于解决了当前神经元与之前神经元之间的相关性，即文本特征向量将很好地被有效表征，因而更合适于文本序列化数据的特征学习，长短期记忆(LSTM)模型[7]作为 RNN 的一种变体，解决了 RNN 梯度消失和梯度爆炸、长期记忆能力不足等问题，但只能沿着一个时间流动方向处理数据，双向 LSTM 模型[8]在 LSTM 模型的基础上增加了一个按时间逆序来传递信息的网络层，能够很好地捕捉上下文信息，从而帮助模型提高性能。

因此，本文基于留言详情，建立关于其一级标签的双向 LSTM 分类模型。双向 LSTM 在时间维度上展开与 LSTM 循环单元结构如图 5 图 6 所示。

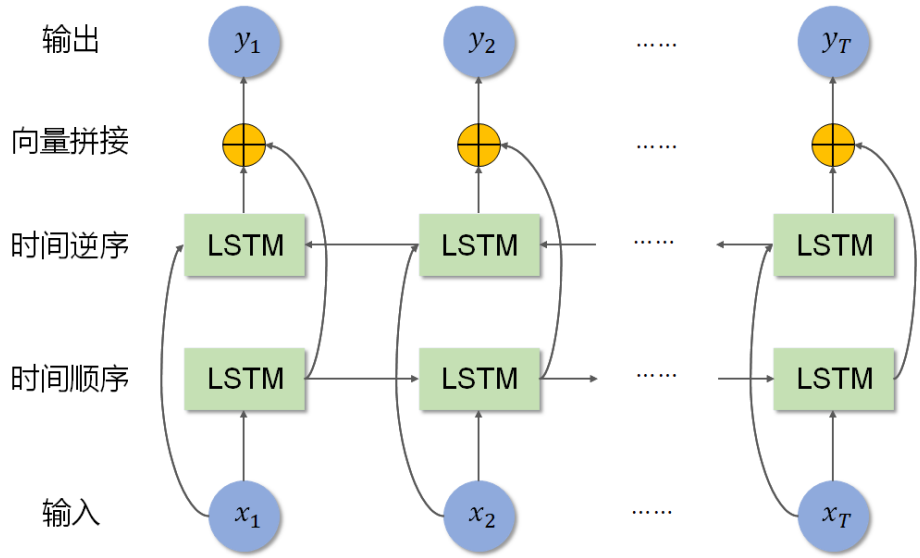

图 5 双向 LSTM 结构

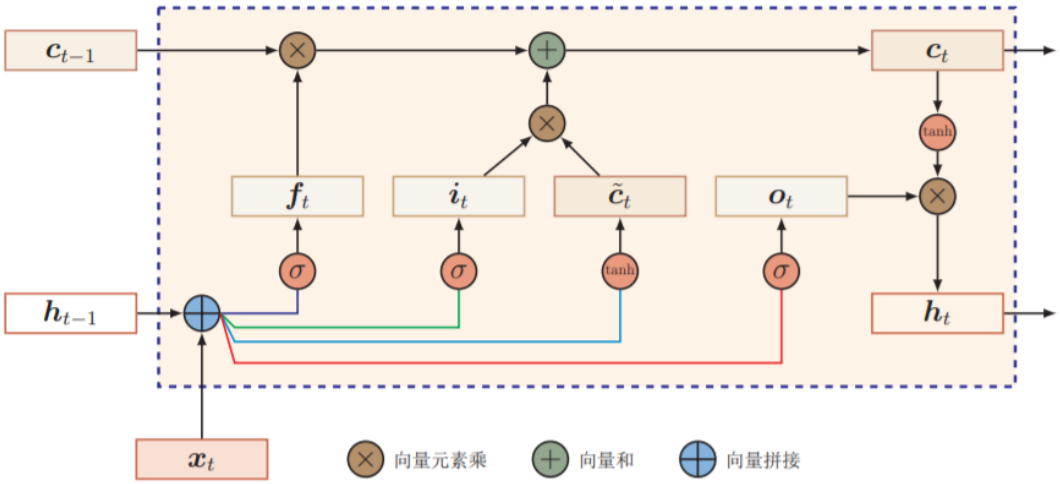

图 6 LSTM 循环神经元结构[l]

LSTM 网络在 RNN 的基础上主要有两个改进：(1)引入一个新的内部状态$\boldsymbol{c}_t$专门进行线性的循环信息传递，同时输出信息给隐藏层的外部状态$\boldsymbol{h}_t$；(2)引入门控机制来控制信息传递的路径，分别为输入门$\boldsymbol{i}_t$、遗忘门$\boldsymbol{f}_t$和输出门$\boldsymbol{o}_t$。

- 遗忘门$\boldsymbol{f}_t$控制上一时刻的内部状态$\boldsymbol{c}_{t-1}$需要遗忘多少信息
- 输入门$\boldsymbol{i}_t$控制当前时刻的候选状态$\tilde{\boldsymbol{c}}_t$有多少信息需要保存
- 输出门$\boldsymbol{o}_t$控制当前时刻的内部状态$\boldsymbol{c}_t$有多少信息需要输出给外部状态$\boldsymbol{h}_t$

在时刻t的输入为$\boldsymbol{x}_t$，LSTM 循环单元结构的计算过程为：，

$$\begin{bmatrix} \tilde{\boldsymbol{c}}_t \\ \boldsymbol{o}_t \\ \boldsymbol{i}_t \\ \boldsymbol{f}_t \end{bmatrix} = \begin{bmatrix} \tanh \\ \sigma \\ \sigma \\ \sigma \end{bmatrix} \left(W \begin{bmatrix} \boldsymbol{x}_t \\ \boldsymbol{h}_{t-1} \end{bmatrix} + b \right)$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t)$$

其中，$\odot$代表向量点乘操作，tanh和σ分别代表非线性激活函数：

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

对于双向 LSTM 模型，第一层按时间顺序，第二层按时间逆序，在时刻t的输入为$\boldsymbol{x}_t$，隐状态为$\boldsymbol{h}_t^{(1)}$和$\boldsymbol{h}_t^{(2)}$均按 LSTM 循环单元计算过程计算，输出为$\boldsymbol{y}_t$，则

$$\boldsymbol{h}_t^{(1)} = f\left(U^{(1)}\boldsymbol{h}_{t-1}^{(1)} + W^{(1)}\boldsymbol{x}_t + \boldsymbol{b}^{(1)}\right)$$

$$\boldsymbol{h}_t^{(2)} = f\left(U^{(2)}\boldsymbol{h}_{t-1}^{(2)} + W^{(2)}\boldsymbol{x}_t + \boldsymbol{b}^{(2)}\right)$$

$$\boldsymbol{h}_t = \boldsymbol{h}_t^{(1)} \oplus \boldsymbol{h}_t^{(2)}$$

$$\boldsymbol{y}_t = V\boldsymbol{h}_t$$

其中，$f(\cdot)$为非线性激活函数，$U^{(1)}$、$U^{(2)}$、$W^{(1)}$、$W^{(2)}$、$\boldsymbol{b}^{(1)}$、$\boldsymbol{b}^{(2)}$和V均为网络参数，$\oplus$表示向量拼接操作。

本文所建立的分类模型以双向 LSTM 模型为基础，采用典型的神经网络模型分层结构框架，自下而上依次为：输入层、词嵌入层、隐藏层、输出层。其中输入层接收经过处理的文本数据传递给词嵌入层；词嵌入层主要用于文本数据的编码化，并转化为维度固定的向量格式；隐藏层有三层，依次为：双向 LSTM 层、LSTM 层和全连接层，神经元个数依次为 256、128 和 32；输出层使用 softmax 激活函数产生输出，模型结构如图 7 所示。

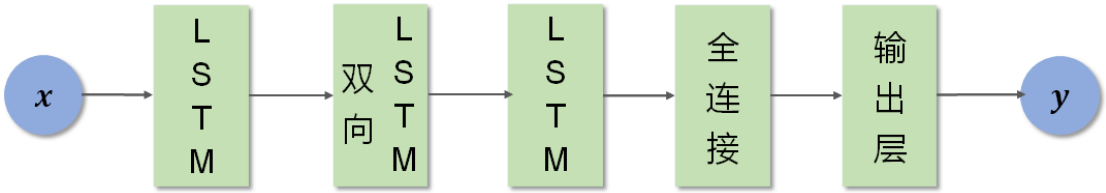

图 8 模型结构

4.1.3 模型求解

4.1.3.1 损失函数

留言内容的一级标签总共有 7 类，对其进行 onehot 编码，每个类别用一个 7 维 0-1 向量表示。选用多分类交叉熵作为网络的损失函数，即

$$loss(\boldsymbol{y},\hat{\boldsymbol{y}}) = -\sum_{i=1}^{n} \boldsymbol{y}_i^{\mathrm{T}} \ln \hat{\boldsymbol{y}}_i$$

其中，$\boldsymbol{y}_i$和$\hat{\boldsymbol{y}}_i$为第i个样本真实标签和预测标签。

4.1.3.2 词嵌入层

采用北京师范大学中文信息处理研究所等开源的中文 Word2Vec 预训练模型（人民日报新闻语料）对文本进行词向量化，嵌入层不参与模型训练，每个词语用维度为 300 的向量表示，使用频率最高的 50000 个词，因此嵌入矩阵的维度为 50000×300。

4.1.3.3 超参数设置

使用 Tensorflow2.0 深度学习框架对模型进行训练，从数据集中随机选出 80%的数据作为训练集，剩余的 20%作为验证集，优化算法使用 Adam 算法，初始学习率为 0.001，批量大小为 64，训练 100 个回合。为了防止在训练过程中发生过拟合，使用提前停止[9]策略，当验证集上的准确率连续 5 个回合都没有提高时，就停止迭代。

4.1.3.4 训练结果

训练过程中训练集和验证集的损失函数和准确率变化图像如图 9 所示。

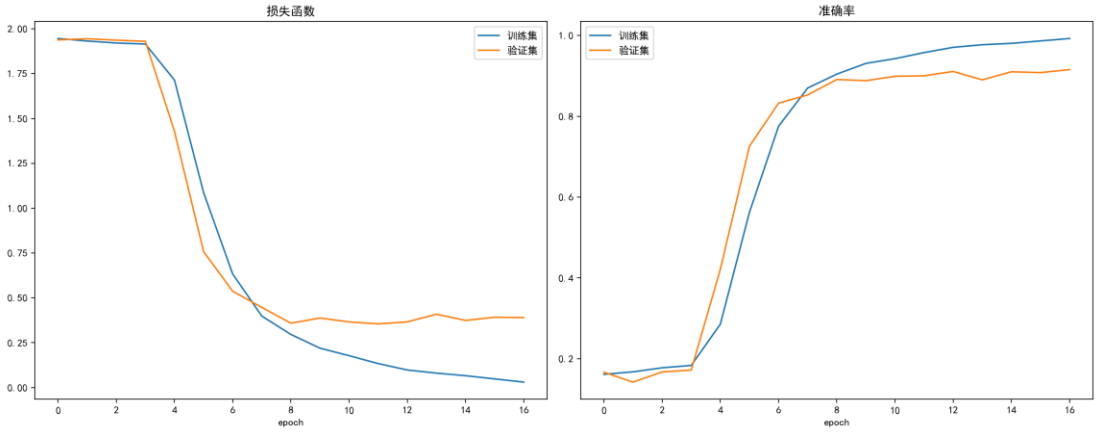

图 9 训练过程

从图 9 可以看出，当训练至第 8 个回合时，模型在训练集上的损失函数的变化率开始减小，模型在验证集上的准确率开始收敛不再有较大的提高；当训练至第 17 个回合时，模型在验证集上的准确率已经连续 5 个回合没有提高，触发停止条件，防止模型继续训练而发生过拟合。

模型最终在训练集上的损失函数为 0.0392，准确率为 99.27%，在验证集上的损失函数为 0.3902，准确率达到 91.57%，说明模型具有很好的泛化能力。

4.1.3.5 模型评价

为衡量模型分类效果[10]好坏，采用查准率(Precision)、查全率(Recall)和 F1 值(F1 Score)对模型进行评价。

给定数据集$\mathcal{T} = \{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(N)}, y^{(N)})\}$，标签$y^{(n)} \in \{1,2,3,4,5,6,7\}$，用训练好的模型对数据集中的每一个样本进行预测，结果为$Y = \hat{y}^{(1)}, \ldots, \hat{y}^{(N)}$，则根据以下公式计算评价指标。

- **查准率**

查准率也叫精确度或精度，类别c的查准率是所有预测为类别c的样本中，预测正确的比例，即：

$$P_c = \frac{TP_c}{TP_c + FP_c}$$

其中，TP_c为真实类别为c并且模型正确地预测为类别c的样本数量，FP_c为真实类别不是c但被模型错误地预测为类别c的样本数量。

- **查全率**

查全率也叫召回率，类别c的查全率是所有真实标签为类别c的样本中，预测正确的比例，即：

$$R_c = \frac{TP_c}{TP_c + FN_c}$$

其中，FN_c为真实类别为c但被模型错误地预测为其它类别的样本数量。

- **F1 分数**

F1 分数是一个综合指标，为查准率和查全率的调和平均，类别c的 F1 分数为：

$$\mathrm{F1}_c = \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}$$

模型在整个数据集上的 F1 分数为：

$$\mathrm{F1} = \frac{1}{7}\sum_{c=1}^{7} \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}$$

分别计算模型在训练集和验证集的评价指标，结果如表 3 和表 4 所示。

表 3 模型在训练集上的各项评价指标

一级标签	精确度	召回率	F1 分数	数量
城乡建设	1.00	1.00	1.00	1207
劳动和社会保障	1.00	0.99	1.00	1578
教育文体	0.99	1.00	1.00	1369
商贸旅游	1.00	0.98	0.99	1268
环境保护	0.98	1.00	0.99	1615
卫生计生	1.00	1.00	1.00	1265
交通运输	0.99	1.00	1.00	1328

表 4 模型在验证集上的各项评价指标

一级标签	精确度	召回率	F1 分数	数量
城乡建设	0.94	0.92	0.93	334
劳动和社会保障	0.94	0.90	0.92	391
教育文体	0.90	0.95	0.92	306
商贸旅游	0.88	0.84	0.86	318
环境保护	0.87	0.89	0.88	394
卫生计生	0.93	0.96	0.94	324
交通运输	0.96	0.97	0.96	341

从表 4 可以看出，模型分类效果强：训练集各类别的 F1 分数几乎都是 1.00，在整个训练集上的 F1 分数为 0.99；验证集各类别的 F1 分数最低为 0.86，为商贸旅游类别，最高为 0.96，为交通运输类别，在整个验证集上的 F1 分数为 0.92。

4.2 问题二的解决

问题二要求根据附件 3 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，定义合理的热度评价指标，并给出评价结果。

4.2.1 数据预处理

对留言主题进行聚类分析之前，对留言主题文本进行分词、去停用词。由于需要根据特定地点或特定人群对留言进行归类，使用 jieba 分词时需要对特定地

点设置更高的频率，使得这类地点词语不被分开，例如：
原句：*A 市A5 区稻田中学午餐伙食太差劲*
一般分词：*['A', '市', 'A5', '区', '稻田', '中学', '午餐', '伙食', '太', '差劲']*
处理后分词：*['A 市', 'A5 区', '稻田', '中学', '午餐', '伙食', '太', '差劲']*

4.2.2 模型建立

4.2.2.1 TF-IDF 词向量化

词袋模型是最基础的文本表示模型，把每个文本看成一袋词语，并忽略每个词语的顺序。具体来说就是将整段文本以词为单位分开，每个文本可以表示成一个长向量，向量中的每一维代表一个词，该维度对应的权重代表这个词在文本中的重要程度，一般用 TF-IDF 计算权重。

TF-IDF(Term Frequency—Inverse Document Frequency)[11]是一种应用广泛的加权级数，常用于进行信息检索和数据挖掘，包含词频 TF 和逆文档频率 IDF 两部分。词频 TF 即词语在文本中出现的频率，逆词频 IDF 则是一个词语普遍关键性的度量。

TF-IDF 的核心思想为：若某个词在一个文本中多次出现，但很少出现在其他文本中，则认为该词具备良好的类别区分性能，TF-IDF 实际上为 TF*IDF。对于特定地点或特定人群问题的留言，其地点、问题关键词等词语在该类问题的文本上应该会出现多次，在其他问题的文本出现的次数应该会很少，从而较好地根据地点、人群和问题关键词区分不同问题的留言。

TF*IDF 计算公式：

$$IDF_j = \log\frac{N}{1+n_j}$$

$$TF * IDF_j = TF_j * \log\frac{N}{1+n_j}$$

其中，N为文本数量，n_j为包含词j的文本数量。

4.2.2.2 DNSCAN 算法

DBSCAN[12]（Density-Based Spatial Clustering of Applications with Noise，具有噪声的基于密度的聚类方法）是一种基于密度的空间聚类算法，采用 DBSCAN 对留言主题向量聚类的好处是无需事先确定形成的簇类数量，并且可以发现任意形状的簇类。

DBSCAN 算法需要提供两个参数$(Eps, MinPts)$，参数Eps描述了每一样本的邻域距离阈值，参数$MinPts$描述了某一样本的距离为Eps的邻域中样本个数的阈值。DBSCAN 实现对留言主题聚类的算法如下：

输入：给定经过预处理的留言主题集合D，邻域半径Eps，邻域密度阈值$MinPts$
输出：基于密度的簇的集合
***Step*1** 标记所有文本为 *unvisited*
***Step*2** 随机选择一个 *unvisited* 对象 *p*
***Step*3** 标记 *p* 为 *visited*
***Step*4** 如果 *p* 的Eps-邻域至少有$MinPts$个对象
　　创建一个新簇 *C*，并把 *p* 添加到 *C*

记 N 为 p 的Eps-邻域中的对象集合
对于 N 中的每个点 q
如果 q 是 *unvisited*
则标记 q 为 *visited*
如果 q 的Eps-邻域至少有$MinPts$个对象，把这些对象添加到 N
如果 q 还不是任何簇的成员，把 q 添加到 C
结束循环
输出 C
否则标记 p 为噪声
直到没有标记为 *unvisited* 的对象

为了确定最佳的聚类参数$(Eps, MinPts)$，引入轮廓系数[13]$s(i)$来评价在不同参数组合下 DBSCAN 对留言主题的聚类效果，选出使得轮廓系数最小参数组合$(Eps, MinPts)$。

对于数据点$i \in C_i$，C_i为数据点i所属的群集，有：

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

$$b(i) = \min_{i \neq j} \frac{1}{|C_i|} \sum_{j \in C_i} d(i,j)$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

其中，$d(i,j)$是群集C_i中的数据点i和数据点j之间的距离，分母$|C_i| - 1$用于去除计算$d(i,i)$的距离，$a(i)$为数据点i和其他数据点的平均距离，可作为衡量数据点i在聚类中的效果，其值越小则聚类效果越好，内聚度越高；$b(i)$为分离度，是数据点i距离其他群集的数据点的平均最小距离，其值越大则分离度越高。

轮廓系数$s(i)$取值范围为$[-1, 1]$，其值越大，说明聚类效果越好。

4.2.2.3 构建热度评价指标

对于同一个问题，同一用户可能会留言多次，从而导致该问题的留言数量增加。例如，反映某个问题的留言数量达到数百条，而这数百条留言均来自同一个用户，则该问题理应不能算作热点问题。一个问题是否为热点问题应该要有很多用户留言反映，定义某个问题的有效留言数量为反映该问题的用户数量。对于某个问题，如果存在用户重复留言，该用户在该问题上获得的总点赞数和总反对数的平均数作为该用户在该问题上获得的点赞数和反对数。

某一问题的有效留言数量越多，即反映该问题的用户数量越多，该问题越有可能成为热点问题。对于在该问题的点赞数和反对数，都反映了围观该问题的用户数量，点赞数越多，说明越多用户关注该问题；反对数越多，说明越多用户认为该问题不值一提。因此，一个问题能否成为热点问题，与该问题的有效留言数和点赞数成正相关，与反对数成负相关。定义热度评价指标如下：

$$h_i = \begin{cases} n_i + \ln(a_i - d_i + 1), a_i > d_i \geq 0 \\ n_i - \ln(d_i - a_i + 1), d_i \geq a_i \geq 0 \end{cases}$$

其中，n_i为问题i的有效留言数了，$a_i \geq 0$为问题i的总点赞数，$d_i \geq 0$为问题i的总反对数。

4.2.2.4 t 分布随机邻域嵌入算法

文本数据基本上都是高维非线性的，无法直接把聚类后的数据分布在二维平面可视化，因此要对文本数据降维再可视化。

t 分布随机邻域嵌入[14](t-distributed Stochastic Neighborhood Embedding, tSNE)算法是一种典型的流形学习方法，它能够将高维空间的点对映射到低维空间，同时保持相互之间分布的概率不变。

记高维数据点对为$X = [x_1, x_2, \dots, x_n] \in R^{m \times n}$，低维空间映射点对为$Y = [y_1, y_2, \dots, y_n] \in R^{d \times n}$，$m \gg d$，本文将文本数据降至二维即$d = 2$。

高维空间中两个数据点(x_i, x_j)的联合概率函数为：

$$p_{ij} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma^2\right)}{\sum_{k \neq l} \exp(-\|x_k - x_l\|^2 / 2\sigma^2)}$$

为避免“拥挤问题[14]”，低维空间中两个数据点(x_i, x_j)的联合概率函数为自由度为 1 的 *t* 分布函数：

$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l}(1 + \|y_k - y_l\|^2)^{-1}}, i \neq j$$

通过联合分布 P 和 Q 之间的 KL 散度来衡量利用低维映射q_{ij}建模高维输入相似度p_{ij}错误率：

$$C = KL(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}, i \neq j$$

采用梯度下降法最小化 KL 散度：

$$\frac{\partial C}{\partial y_j} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)\left(1 + \|y_i - y_j\|^2\right)^{-1}$$

更新规则为：

$$Y^{(t)} = Y^{(t-1)} + \eta \frac{\partial C}{\partial y_j} + \alpha(t)\left(Y^{(t-1)} - Y^{(t-2)}\right)$$

其中，t 为当前迭代次数，η为学习率，$\alpha(t)$为第 *t* 次迭代的动量。

采用 tSNE 算法对文本数据降维表示的过程如下：

输入：文本向量数据$X = [x_1, x_2, \ldots, x_n] \in R^{m \times n}$，最大迭代次数$t = 100$，学习率$\eta = 200$和动量因子$\alpha(t) = 0.8$。
输出：文本向量在二维空间的数据表示。

***Step*1** 计算p_{ij}和q_{ij}

***Step*2** 随机生成初始解$Y^{(0)}$

***Step*3** 从 1 到 *t* 迭代：

(1) 计算q_{ij}

(2) 计算梯度$\frac{\partial C}{\partial y_j}$

(3) 更新$Y^{(t)}$

4.3.3 模型求解

4.3.3.1 DBSCAN 聚类结果

采用网格搜索法搜索最佳的参数$(Eps, MinPts)$，对Eps从 0-1.5 取值，间隔为 0.01，对$MinPts$分别取 2，3，4，不同参数组合对应的轮廓系数如下图 10 所示。

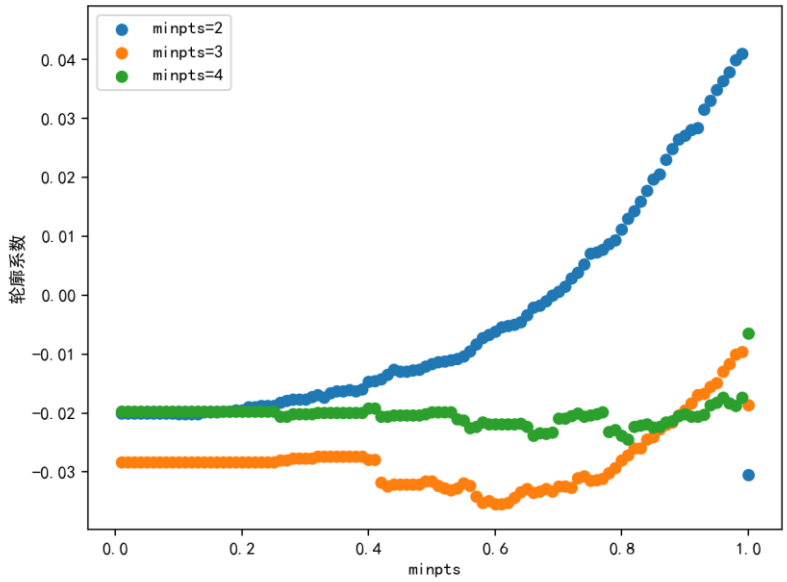

图 10 不同参数下的轮廓系数

从图 10 可以看出，在$MinPts$固定时，Eps越大，轮廓系数越大，当$Eps = 0.99$，$MinPts = 2$，轮廓系数最大，为 0.041；在$MinPts = 2$时，轮廓系数对Eps的变化敏感；当$MinPts = 4$时，轮廓系数对Eps的变化不敏感。

当$Eps = 0.99$，$MinPts = 2$时，一共得到 434 个簇类共 1216 个样本，3065

个样本点为噪声点。经发现，聚类结果中存在一些内容相似但地点不同的留言聚到了同一类，可能是由于留言主题长度较为简短，部分关键词重叠：

A3 区梅溪湖看云路润芳园小区油烟扰民
A5 区劳动东路魅力之城小区油烟扰民
魅力之城小区临街门面油烟直排扰民

根据聚类后的留言主题，提取各类留言出现次数最多的地点和 3 个关键词，并对聚类结果进行修正：若某一样本的关键词和地点与某一个类的关键词和地点相同或相似，则把该样本归到该簇类中。修正后的聚类结果如下表 5 所示：

表 5 各个簇类的样本数

簇类编号	样本数	关键词
8	52	A2 区 丽发 新城 搅拌站 噪音 扰民
16	45	伊景园 滨河苑 捆绑 销售 车位
11	25	人才 购房 补贴
38	21	魅力之城 夜宵 油烟 噪音 扰民
170	10	茶场村五组 拆迁 规划
10	10	星沙 凉塘路 旧城改造
67	10	中心 城市 加快 建设
68	9	经济学院 强制 实习
169	8	转业 士官 异地 安置
30	7	十里天池 玉佩路 交付

4.3.3.2 热点问题

对每个簇类计算热度指标，对其进行排序，选出排名前 5 的问题作为热点问题，结果如下表 6 所示。

表 6 排名前五的热度问题

排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	49.00	2019/11/2 至 2020/01/26	A 市 A2 区 丽发新城小区	小区附近搅拌厂灰尘噪音扰民
2	2	42.00	2019/07/18 至 2019/09/01	A 市伊景园 滨河苑	开发商捆绑销售停车位
3	3	24.69	2018/11/15 至 2019/12/02	A 市	市民咨询关于 A 市人才新政购房补贴问题
4	4	17.00	2019/07/21 至 2019/12/04	A 市 A5 区 魅力之城小区	小区临街餐饮店油烟噪音扰民
5	5	12.71	2019/05/05 至 2019/09/19	A 市 A5 区 五矿万境 K9 县	房屋质量及物业管理问题

4.3.3.3 聚类可视化

➢ **聚类散点图**

通过 tSNE 算法把留言主题向量从高维空间映射到二维平面，并对排名前 5 的热点问题进行可视化，结果如图 11 所示：

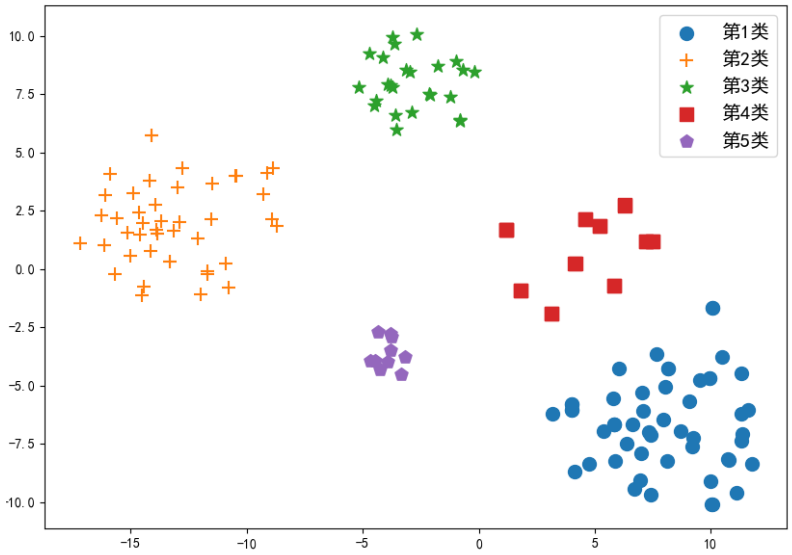

图 11 排名前 5 的热点问题

从图 11 可以看出，留言主题数据不仅能够在原本的词向量空间通过 DBSCAN 算法聚类，而且经过 tSNE 算法从高维的词向量空间映射到二维平面后，仍然保持着类与类之间的界限。

➢ **热点问题词云**

A 市 A2 区丽发新城小区附近搅拌站灰尘、噪音扰民

图 12 热点问题一词云

A 市伊景园滨河苑开发商捆绑销售停车位

图 13 热点问题二词云

市民咨询 A 市人才新政购房补贴问题

图 14 热点问题三词云

A 市 A5 区魅力之城小区临街餐饮店油烟噪音扰民

图 15 热点问题四

A5 区 K9 县万境五矿存在房屋质量及物业管理问题

图 16 热点问题五词云

4. 3 问题三的解决

问题三要求从答复的相关性、完整性、可解释性等角度对附件 4 相关部门对留言的答复意见质量给出一套评价方案。

4.3.1 数据预处理

与前面的问题一样，在建立模型之前需要对文本数据进行预处理，观察附件 4 数据发现部分答复意见存在一定答复格式，例如，“现将网友在 XXX 平台向 XXX 书记留言反映 XXX 问题的调查核实情况向该网友答复如下：您好！……感谢您对我区工作的理解和关心。XXXX 年 XX 月 XX 日”。对于按照格式进行回复的答复意见，其有效的信息为中间部分的文本，因此需要识别出使用格式的答复意见，提取有效的文本信息，然后对其进行分词、去停用词。

留言详情和答复意见长度分布如图 17 和图 18 所示。

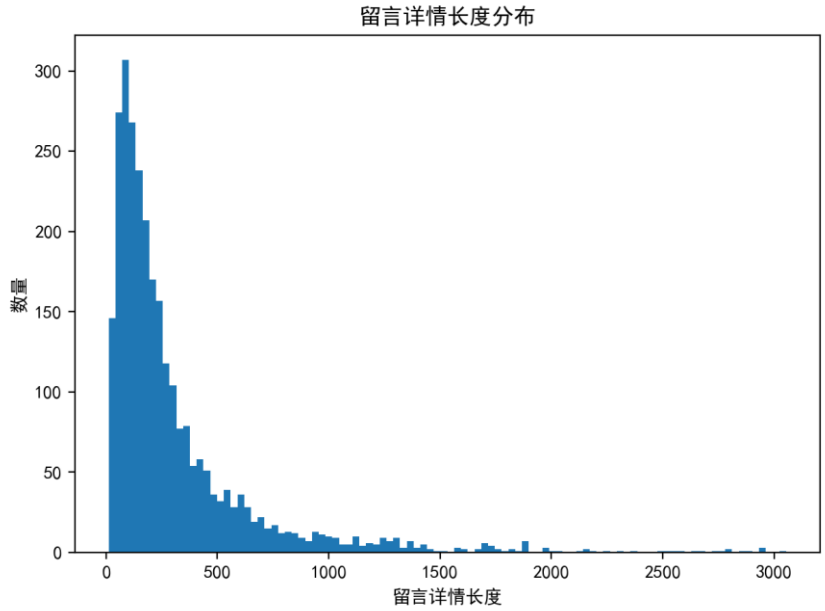

图 17 留言详情长度分布

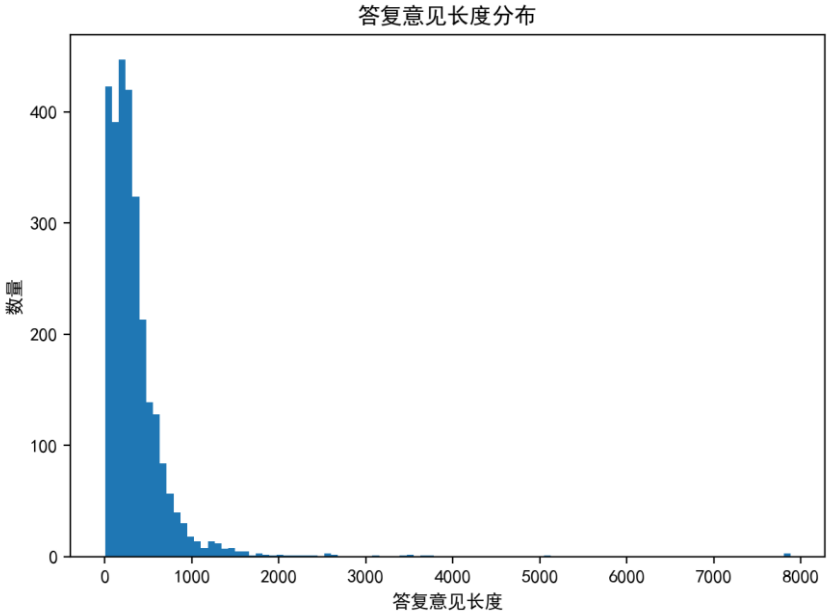

图 18 答复意见长度分布

从图 17 和图 18 可以看到，绝大部分的留言详情和答复意见长度都在 1000 个字符内，因此，对于长度大于 1000 留言详情和答复意见，用问题一中改进的 TextRank 算法进行关键句提取，然后对其进行分词、去停用词。

4.3.2 模型建立

根据前面的分析，对于评价答复意见质量，最为关键的是衡量留言与答复之间的相关程度，然后结合答复意见长度和留言详情长度构建关于答复意见质量的

评价指标，对每条留言的答复意见进行评价。

4.3.2.1 词移距离(Word Move's Distance)

词移距离[15](Word Mover's Distance, WMD)是 EMD[16](Earth Mover's Distance)模型在自然语言处理方向上的应用，与欧氏距离不同，WMD 可以从文本整体上来考虑两个文本之间的相似性，避免了欧氏距离仅在词义级别上考虑相关性的缺陷。

本文基于 Word2Vec 建立 WMD 模型计算答复意见与留言详情之间的相关程度，WMD 本质上是一个数学优化问题，其基本原理[16]如下：

***Step*1** 对文本进行分词、去除停用词，通过 Word2Vec 把文本映射到词向量空间，用词嵌入矩阵$X \in R^{d \times n}$来表示n个词的词嵌入映射表，$x_i \in R^d$表示第 i 个词的词向量；

***Step*2** 使用归一化的词袋模型(nBOW)表示待比较的文字，记为向量$d \in R^n$。如果第 i 个词出现了c_i次，那么$d_i = c_i / \sum_k c_k$；

***Step*3** 计算每对词语之间的欧氏距离$c(i,j) = \left\| x_i - x_j \right\|_2$，$c(i,j)$表示词$i$转移到词$j$的代价；

***Step*4** 记d和d'为用 nBOW 表示的两段文本，允许d中第i个词d_i转移到d'中第j个词，定义$T \in R^{n \times n}$为词移矩阵，矩阵中每个元素$T_{ij} \geq 0$表示有多少个词i从d转移到d'中的词j，最终需要把d中的所有词都转移到d'的所有词上，从而得到两段文本之间的距离即把d中所有词都转移到d'中所有词的最小代价。

将其写成线性规划的形式：

$$\min \sum_{j=1}^{n} \sum_{i=1}^{n} T_{ij}\, c(i,j)$$

$$\text{s.t.} \begin{cases} \sum_{j=1}^{n} T_{ij} = d_i, i = 1, \dots, n \\ \sum_{i=1}^{n} T_{ij} = d'_j, j = 1, \dots, n \end{cases}$$

第一个约束条件表示d中的第i个词的流出总和需要等于d_i，第二个约束条件表示d'中的第j个词的流入总和需要等于d'_j。

利用线性规划[17]中的单纯性法可求得该问题的最优解。

4.3.2.2 答复意见评价方案

在计算答复意见和留言详情的 WMD 距离时，需要把分词后的答复意见的所有词语全部转移到分词后的留言详情的所有词语中，因而 WMD 可以从相关性、可解释性两个方面衡量答复意见和留言详情之间的相关程度。WMD 越大，答复意见与留言详情越不相关，说明答复意见越“答非所问”；WMD 越小，答复意见与留言详情越相关，说明答复意见越“答到点上”。因此，WMD 是评价答复意见

质量的关键。

除了 WMD 距离外，答复意见长度$respon_i$（去除空格符和换行符后的字符个数）也是评价答复意见质量的一个重要方面。一般来说，答复越长，越能显现答复者的认真程度，在一定程度上说明答复意见具有较强的解释性；但并不一定越长越好，质量好的答复应该以最少的字数回答出最好的答复，答复意见过长，有可能答复中存在大量的堆砌文字或引用了相关文件内容；而答复意见越短，则越容易说明答复意见存在问题。

为了尽可能消除答复意见过长的影响，增加留言详情长度$detail_i$（去除空格符和换行符后的字符个数）作为评价答复质量的一个衡量方面。留言详情偏长，可能是网友提出的问题偏多且描述详细，表示网友对这个问题的关注度高，希望得到较好的答复。如果对应的答复意见长度偏短，则表明答复不够完整，解释能力差。留言详情偏短，可能是网友在向政府提建议或者询问一些相对简单的事项。此时如果答复过长，则可能存在答复冗余，导致答复质量低下。

结合答复意见长度$respon_i$和留言详情长度$detail_i$，定义答复留言长度比为$ratio_i = respon_i/detail_i$，质量好的答复意见应该尽可能地控制$ratio_i$在 1 附近，答复留言长度比的分布如图 19 所示。

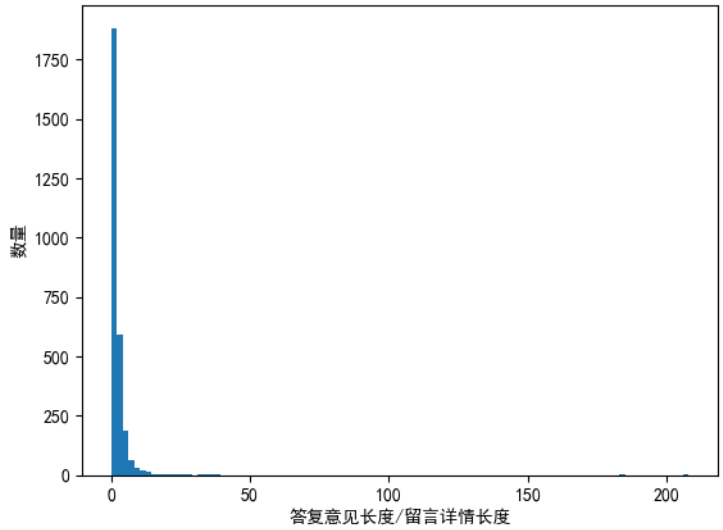

图 19 答复留言长度比的分布

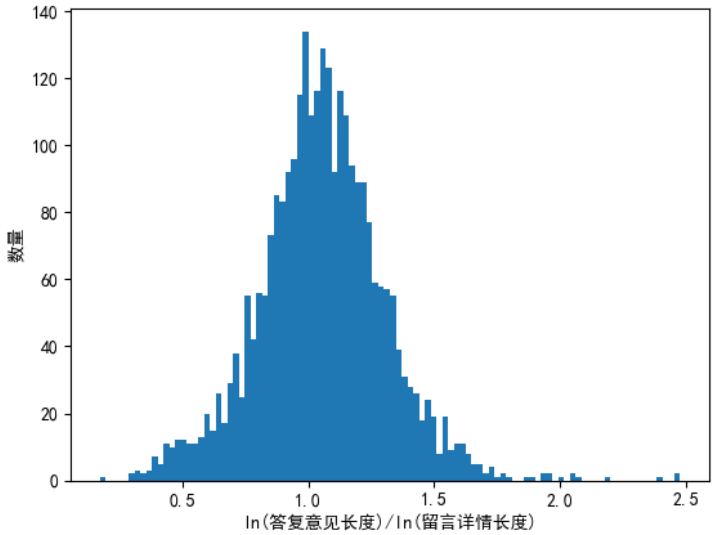

图 20　答复留言对数长度比的分布

从图 19 可以看出，答复留言长度比存在严重的拖尾情况。对$respon_i$和$detail_i$分别取对数后相比，其分布如图 20 所示，不仅可以将比值控制在一个小的范围内，还能使数据近似呈正态分布。因此，重新定义答复留言长度比为：

$$ratio_i = \frac{\ln respon_i}{\ln respon_i}$$

构建答复意见质量的评价指标如下：

$$score_i = \left(\frac{1}{\mathrm{WMD}_i}\right)^{\alpha} \times \left(\frac{1}{k} ratio_i\right)^{\beta}$$

其中，α和β为分别为 WMD 和 *ratio* 的权重,且满足$\alpha + \beta = 1$。$k > 1$为常数，用于把 *ratio* 放缩到$(0, 1)$之间。

$score_i$越大，答复意见质量更高。WMD 与 *ratio* 相结合，能够缓解 *ratio* 的缺陷：即当 *ratio* 越大时，答复意见越有可能被评为质量好；当 *ratio* 越小时，答复意见越有可能被评为质量差。此时 WMD 起到一个“缓解”的作用，缓解 *ratio* 在两个极端的评判倾向。

4.3.3 模型求解

利用 Python 中的 Pulp 库对模型求解，求出每条答复意见和留言详情之间的 WMD 距离，α设为 0.7，β设为 0.3，k设为 3，计算$score_i$，部分结果如下（限于篇幅，仅展示排名靠前和排名靠后的几条内容，解答部分已标红）：。

score ： 0.3741
留言编号：*20893*
留言主题：*咨询下泉塘规划问题*
留言详情：*请问泉塘东四线以西，盼盼路以南，3614 小区以东的土地有什么规划？*
答复意见：*网友“UU008515”您好！经联系区规划局后得知，泉塘东四线以西，盼盼路以南，3614 小区以东的土地已规划为商住用地。感谢您对我们工作的关心、支持和监督！泉塘街道办事处*

score ： 0.3304
留言编号：*6448*
留言主题：*关于A 市实施差别化购房措施的咨询*
留言详情：*领导您好，关于（关于实施差别化购房措施的通知）中，签订拆迁安置协议一年内的属于刚需群体，是否签订拆迁协议一年后还未购房的就不属于刚需群体了。请领导解答，谢谢。*
答复意见：*网友“UU0081211”您好！您的留言已收悉。现将有关情况回复如下：根据《关于实施差别化购房措施的通知》，签订拆迁安置协议一年内的属于刚需群体，签订拆迁协议一年后还未购房的不属于刚需群体。感谢您对我们工作的支持、理解与监督！2018 年 4 月 2 日*

score ： 0.3137
留言编号：*17442*
留言主题：*咨询A 市经开区漓楚路以南地块具体规划*
留言详情：*漓楚路以南，三景国际小区以东，武警后勤基地以西，佳美紫郡以北地块空置了很多年了，请问有什么具体规划*
答复意见：*网友“UU008398”您好！来信收悉。现回复如下：经核查，“漓楚路以南，三景国际小区以东，武警后勤基地以西，佳美紫郡以北”区域有两块地，位于漓楚路以南的地块为商住用地，位于佳美紫郡以北的为居住用地。感谢您对我们工作的理解和支持！2017 年 5 月 22 日*

score ： 0.1012
留言编号：*37482*
留言主题：*建议 B9 市规划一个校车接送计划*
留言详情：*现在 B9 市的交通压力实在太大，上学放学期间更是可怕！各种交通工具，加上道路不够宽，很容易出现更多的交通事故！恳请政府做一个小学生校车接送计划，提高安全保障的同时也能减轻交通压力。请领导慎重考虑！谢谢！*
答复意见：*2018/12/12*

score ： 0.0999
留言编号：*25431*
留言主题：*咨询A7 县星沙镇能否根据房产证开具准迁证*
留言详情：*本人 2014 年在 A7 县泉塘街道购房，现房产证，结婚证，身份证齐全，因为户口是在河南省洛阳市的集体户口，不在自己手中，现在办理落户时无法办理。洛阳方面称无准迁证不能借出户口，而星沙这边无户口不能开具准迁证。本人现居住在星沙，双方行政部门相互推诿，咨询一下能不能由星沙这边先根据房产证开具准迁证？*

表 7 评价结果

留言编号	留言长度	回复长度	对数比	WMD	score	排名
20893	34	88	1.2697	2.8058	0.3753	1
6448	79	122	1.0995	3.1639	0.3304	2
17442	49	125	1.2406	3.5826	0.3141	3
97307	161	310	1.1289	3.4955	0.3106	4
50409	617	622	1.0013	3.8119	0.2820	5
9128	1347	877	0.9404	3.8144	0.2766	6
17904	60	154	1.2302	4.4476	0.2693	7
7625	110	290	1.2062	4.5186	0.2647	8
⋮	⋮	⋮	⋮	⋮	⋮	⋮
25431	147	10	0.4614	11.7946	0.1014	2811
121460	1020	22	0.4462	11.9591	0.0994	2812
25918	289	11	0.4232	11.8696	0.0983	2813
122596	2321	23	0.4046	11.9688	0.0965	2814
114346	646	3	0.1698	8.6709	0.0932	2815
144850	1309	13	0.3574	11.9460	0.0931	2816

从表 7 可以看出，质量较好的答复意见的答复留言对数比基本都在 1 附近，而且 WMD 距离都比较小；而质量差的答复意见的答复留言对数比都在 0.5 以下，而且 WMD 都很大，最大达到了 11.9688。留言编号为 114346 的答复长度仅为 3 个字符，答复留言对数比仅为 0.1698，WMD 距离为 8.6709。由此说明，本文所构建的答复意见质量的评价指标能够较好地对答复意见进行评价。

五、模型评价

5.1 模型优点

- 双向 LSTM 神经网络模型具有强大的非线性拟合能力，同时能够捕捉文本中的上下文信息，非常适用于文本分类问题
- DBSCAN 算法聚类速度快且能有有效处理噪声点和发现任意形状的簇类
- 采用 tSNE 算法把高维文本数据映射到二维平面对聚类结果进行可视化，为评估聚类效果提供了直接参考
- WMD 算法能够从整体上来计算两个文本的相似程度，并且能够求出最优解

5.2 模型缺点

- DBSCAN 算法需要给定两个参数(*Eps*, *MinPts*)，当空间聚类的密度不均匀、聚类间距差相差很大时，聚类质量较差
- WMD 算法时间复杂度为$O(p^3 \log p)$，p为两个文本中去重后的词语数量

六、参考文献

[1]李新叶,龙慎鹏,朱婧.基于深度神经网络的少样本学习综述[J/OL].计算机应用研究:1-8[2020-05-08].https://doi.org/10.19734/j.issn.1001-3695.2019.03.0036.

[2]陈志泊,李钰曼,许福,冯国明,师栋瑜,崔晓晖.基于 TextRank 和簇过滤的林业文本关键信息抽取研究[J/OL].农业机械学报:1-11[2020-05-08].http://kns.cnki.net/kcms/detail/11.1964.S.20200309.1813.024.html.

[3]陈乐乐,黄松,孙金磊,惠战伟,吴开舜.基于 BM25 算法的问题报告质量检测方法[J/OL].清华大学学报(自然科学版):1-8[2020-05-08].https://doi.org/10.16511/j.cnki.qhdxxb.2020.25.002.

[4]毛郁欣,邱智学.基于 Word2Vec 模型和 K-Means 算法的信息技术文档聚类研究[J].中国信息技术教育,2020(08):99-101

[5]阮光册,谢凡,涂世文.基于 Word2vec 的图书馆推荐系统多样性问题应用研究[J].图书馆杂志,2020,39(03):124-132.

[6]杨仲江,马俊彦,王昊.序列结构的 RNN 模型在闪电预警中的应用[J].灾害学,2020,35(02):90-96.

[7]Pingyang Lyu,Ning Chen,Shanjun Mao,Mei Li. LSTM based encoder-decoder for short-term predictions of gas concentration using multi-sensor fusion[J]. Elsevier B.V.,2020,137.

[8]刘帅,王磊,丁旭涛.基于 Bi-LSTM 的脑电情绪识别研究[J/OL].山东大学学报(工学版):1-6[2020-05-08].http://kns.cnki.net/kcms/detail/37.1391.t.20200429.2139.002.html.

[9]Effland Alexander,Kobler Erich,Kunisch Karl,Pock Thomas. Variational Networks: An Optimal Control Approach to Early Stopping Variational Methods for Image Restoration.[J]. Pubmed,2020,62(3).

[10]Ibrahim Bounhas,Nadia Soudani,Yahya Slimani. Building a morpho-semantic knowledge graph for Arabic information retrieval[J]. Elsevier Ltd,2019.

[11]张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.

[12]张美玉,王洋洋,吴良武,秦绪佳.结合 DBSCAN 聚类与互信息的图像拼接算法[J].小型微型计算机系统,2020,41(04):825-829.

[13]孙石磊,王超,赵元棣.基于轮廓系数的参数无关空中交通轨迹聚类方法[J].计算机应用,2019,39(11):3293-3297.

[14]朱献超. 基于梯度下降和自适应学习的高维生物数据降维可视化方法研究[D].华中师范大学,2018.

[15]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.

[16]黄栋,徐博,许侃,林鸿飞,杨志豪.基于词向量和 EMD 距离的短文本聚类[J].山东大学学报(理学版),2017,52(07):66-72.

[17]程明. 基于超像素分布与 EMD 度量的快速手势识别算法[D].湖南科技大学,2016.

[18]李鹏,杨元维,高贤君,杜李慧,周意,蒋梦月,张净波.基于双向循环神经网络的汉语语音识别[J/OL]. 应用声学,2020(03):1-8[2020-05-08].http://kns.cnki.net/kcms/detail/11.2121.o4.20200506.1009.022.html.

[19]韩皓,谢天.基于注意力 Seq2Seq 网络的高速公路交织区车辆变道轨迹预测[J/OL].中国公路学报:1-17[2020-05-08].http://kns.cnki.net/kcms/detail/61.1313.U.20200507.1521.016.html.

[20]王太勇,王廷虎,王鹏,乔卉卉,徐明达.基于注意力机制 BiLSTM 的设备智能故障诊断方法[J].天津大学学报(自然科学与工程技术版),2020,53(06):601-608.

[21]石凤贵.基于百度网页的中文自动问答应用研究[J].现代计算机,2020(08):104-108.

[22]黄晨晨,索朗拉姆,拉姆卓嘎,群诺.基于 SVM 的藏文微博文本情感分析研究与实现[J].高原科学研究,2020,4(01):92-96.

[23]陈俊芬,张明,赵佳成.复杂高维数据的密度峰值快速搜索聚类算法[J].计算机科学,2020,47(03):79-86.

[24]Rudolf Scitovski,Kristian Sabo. A combination of k -means and DBSCAN algorithm for solving the multiple generalized circle detection problem[J]. Springer Berlin Heidelberg,2020(prepublish).

[25]薛嬰. 基于文本相似度的主观题自动评分系统的设计与实现[D].北京邮电大学,2019.

[26]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.

[27]吴德亮. 基于降维与聚类的单细胞 RNA 测序数据分析[D].哈尔滨工业大学,2018.

[28]赵明月. 基于词性和关键词的英文短文本测量方法[D].河南大学,2018.

[29]刘拼拼. 领域问答系统中问句相似度计算方法研究[D].哈尔滨工业大学,2018.

[30]吴欣辉. 基于中英文主题向量空间的文本分类算法[D].中国科学技术大学,2018.

[31]刘海军. 基于时空流形学习的人体动作识别[D].电子科技大学,2014.

[32]黄维. 合成纤维系缆非线性动力特性及绷紧式系泊系统响应研究[D].天津大学,2012.