“智慧政务”中的文本挖掘应用

摘要

随着技术和时代的发展，政务系统的广泛应用已经是社会治理创新发展的新趋势，因此，运用网络文本分析和数据挖掘技术对网络问政信息的研究具有重大的意义，同时对提升政府的管理水平和施政效率具有极大的推动作用。

针对问题一：通过正则表达式，对群众留言信息表的留言内容进行提取，使用停用词典等工具去除群众留言信息中的停用词和其他噪声词。利用 jieba 中文分词工具对群众留言信息进行分词，将群众留言信息、一级分类标签转为 id 向量，构建具有映射关系的字典，并结合运用腾讯词向量构建的字典用于组成构建神经网络中词嵌入层的词向量矩阵。将已训练过的腾讯词向量相关字典放入 LSTM 神经网络模型，实现关于群众留言内容一级标签的分类和文本数据的分类测试评估。最终选择 7 次模拟训练，学习速率为 0.001，得到最终为 0.89 的评估分数值并实现其准确率和损失值可视化。

针对问题二：需要拆分两个部分进行文本数据处理：第一，通过留言 id 对留言主题信息、留言详情信息进行去重，得到不重复的群众留言反馈信息。同样利用 jieba 中文分词工具对留言信息进行分词，并通过 TF-IDF 算法提取留言的关键词。再利用 TF-IDF 算法得到各个留言问题描述的 TF-IDF 权重向量，采用 DBSCAN 聚类对权重向量进行聚类，得到 26 类热点问题。剔除噪声项，进行排序后，得到热点问题留言明细表。第二，基于热点问题留言明细表中的留言主题，通过搭建对人民日报语料库进行训练的 BERT+LSTM+CRF 网络模型对地点提取，并拼接通过 Hanlp 的维特比分词器提取的人群。定义热度评价标准，根据对相关热点的影响程度，赋予一定的权重，进而构建热度指数进行计算。按照热度指数的大小，对每一类留言进行排序，根据热度排名，得到热度排名前五的五类热点问题。同时根据得到的热点问题的 ID，选择每一类留言中点赞数与反对数最高的一条留言中涉及的地点/人群作为代表，放入热点问题表中，构建最终的热点问题表。

针对问题三：通过留言答复意见表的原始文本数据进行去缺失值、去重、去短句，以及运用 jieba 中文分词，建立词典，将分词列表集转换成稀疏向量。再运用 TF-IDF 算法与余弦相似度求出答复意见与留言主题及详情、答复意见模板的相似度。分组统计后得出分析结果图，再确定答复意见的相关性、完整性的指标及权重，构建答复意见评价得分公式，分别计算得分，将数据可视化得其数值分布图，其分布符合原始答复意见质量的大体趋势。

关键词：jieba 中文分词 TF-IDF 算法 DBSCAN 聚类 命名实体识别

Text mining application in "intelligent government affairs"

Summary

With the development of technology and times, the wide application of government affairs system has become a new trend in the innovative development of social governance. Therefore, the application of network text analysis and data mining technology is of great significance to the research of network political information, and plays a great role in promoting the government's management level and administration efficiency.

For question 1:Through regular expressions, extract the message content of the mass message table, and use stop dictionary and other tools to remove the stop words and other noise words in the mass message.Use the jieba Chinese word segmentation tool to segment the message of the masses, convert the mass message and the first-level classification tags to id vectors, construct a dictionary with a mapping relationship, and combine the dictionary built with Tencent word vectors to form words in the neural network The word vector matrix of the embedded layer.Put the trained Tencent word vector related dictionary into the LSTM neural network model to realize the classification of the first-level tags of the message content of the masses and the classification test evaluation of the text data.Finally, 7 simulation trainings were selected, the learning rate was 0.001, and the final evaluation score value of 0.89 was obtained and the accuracy and loss value were visualized.

For Question 2:Two parts need to be split for text data processing: first, the message subject information and message detail information are deduplicated by message id to obtain non-repeated feedback messages from the masses.Similarly, jieba Chinese word segmentation tool is used to segment message information, and TF-IDF algorithm is used to extract message keywords. Then, TF-IDF algorithm is used to obtain the TF-IDF weight vector of each message problem description, and DBSCAN Ju Lei is used to cluster the weight vector to obtain 26 types of hot issues.Remove the noise items and sort them to obtain the hot issue message list.Second, based on the subject of the message in the questionnaire list of hot issues, the location is extracted by building a BERT + LSTM + CRF network model that trains the People ' s Daily corpus, and the people extracted through Hanlp ' s Viterbi tokenizer are stitched together.Define the heat evaluation standard, give certain weight according to the degree of influence on related hot spots, and then construct the heat index to calculate.Sort each type of messages according to the heat index, and get the top five hot issues according to the heat ranking. At the same time, according to the obtained hotissue ID, select the location/crowd involved in the message with the highest number of likes and opponents in each type of message as the representative, and put it into the hotissue table, build the final hotspot problem table.

For Question 3:Through the original text data of the message reply opinion table to remove missing values, deduplication, and short sentences, and using jieba Chinese word segmentation, establish a dictionary, convert the word segmentation list set into sparse vectors.Then use the TF-IDF algorithm and cosine similarity to find the similarity between the reply opinion and the subject and details of the message, and the reply opinion template.The analysis results were obtained after grouping statistics, and the correlation and integrity indexes and weights of the replies were determined. The evaluation score formula of the replies was constructed, and the scores were calculated respectively. The data were visualized to obtain its numerical distribution map, which was consistent with the general trend of the quality of the original replies.

Key words: jieba Chinese word segmentation TF-IDFalgorithm DBSCAN clustering Named entity identification

目录

1. 挖掘目标

本次建模目标是利用网络问政平台系统公开来源的问政留言及相关部门的答复意见记录数据，利用 jieba 中文分词工具对群众留言进行分词、腾讯词向量对 LSTM 神经网络词嵌入层进行预处理、DBSCAN 聚类的方法及 TF-IDF 算法、BERT+LSTM+CRF 网络训练人民日报语料库、Hanlp 的维持分词器提取相关词性词语对文本数据处理后，达到以下四个目标：

（1）利用文本分词和构建词向量矩阵的方法对非结构化的数据进行文本挖掘，根据神经网络模型并结合一级指标内容进行文本分类。

（2）结合群众留言主题及详情的文本数据，采用相关算法把文本信息转换为权重向量，根据聚类结果，归纳出热点问题并构建热点问题明细表。

（3）根据热点问题的文本数据，运用模型对地点、人群进行命名实体识别并提取相关信息，定义热度评价指标，构建热度指数排名前五的热点问题表。

（4）针对相关部门对留言的答复意见的原始文本数据，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进行评价。

2. 分析方法与过程

2.1. 总体流程

本部分使用一个总体流程图描述建模方法及过程，并对各部分进行简要说明。

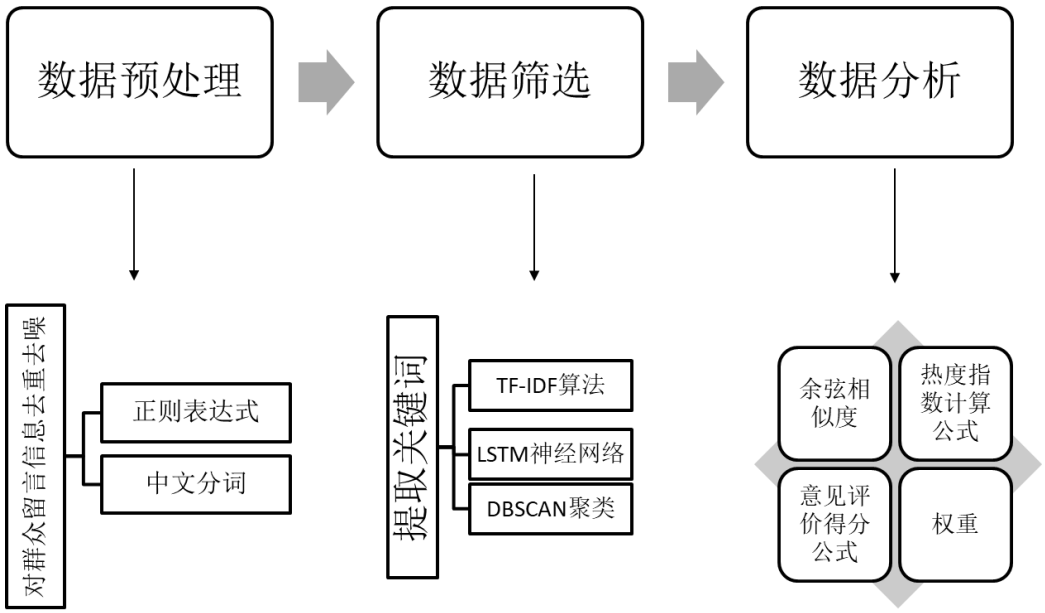

图 1 总体流程图

本用例主要包括以下步骤：

步骤一：数据预处理，在题目给出的数据中，出现了很多重复的群众留言信息数据，在原始的数据上进行去重、去噪处理，在此基础上通过正则表达式以及进行中文分词。

步骤二：数据筛选，统计相关文本数据，分类筛选汇总，在对群众留言信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法，找出每个留言的关键词，把群众留言信息转换为权重向量。采用 LSTM 神经网络模型及 DBSCAN 聚类方法。

步骤三：在对数据进行分析时，构建热度指数计算公式和意见评价得分公式，运用余弦相似度及权重进行情况分析评价。

2.2. 具体步骤

2.2.1. 问题 1 的具体步骤

2.2.1.1.流程图

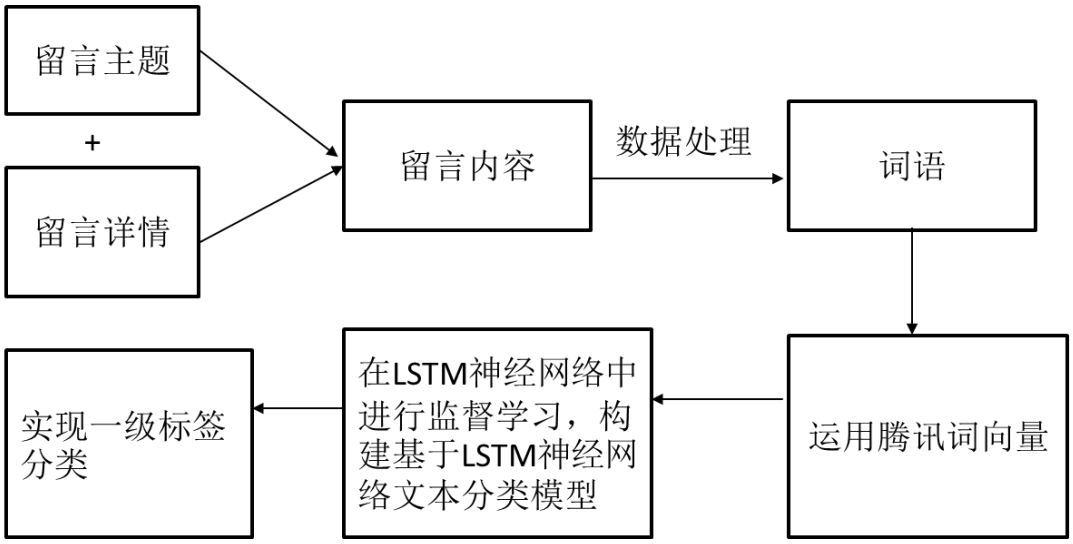

图 2 问题 1 流程图

2.2.1.2.数据预处理

2.2.1.2.1. 对群众留言内容的提取

通过正则表达式，提取出留言内容中的中文和标点符号，由此可以去掉英文和一些其他的字符。

2.2.1.2.2. 对群众留言的划分体系表进行中文分词

在群众留言信息进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。在附件划分体系表中，以中文文本的方式给出了数据，为了便于转换，先要对这些群众留言信息进行中文分词。这里采用 python 的中文分词包 jieba 进行分词。jieba 采用了基于前缀词典实现高效的词语和图形扫描，生成句子中汉字所有可能构成词语情况的有向无环图 (DAG)，同时采用了动态规划查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用了基于汉字构成词语能力的 HMM 模型，使得能更好的实现中文分词效果，这样速度快，精度较高。

2.2.1.2.3. 去除群众留言信息中的停用词和其他噪声词

在附件给出的数据，出现了像“的”、“把”等停用词会影响分类效果，因此我们通过使用停用词典，将把停用词剔除。另外，因为这是市民写给市领导的留言，很多留言中存在例如：“您好”、“尊敬的”、“请问”、“市长”等词，这些会成为留言分类时的噪声，也将其剔除。

2.2.1.2.4. 将群众留言内容转为 id 向量

将留言内容中处理后的词语，进行去重，构建词典 all_dict，并构建具有映射关系的 {word：id} 字典和 {id：word} 字典，用于补充长度的词‘PAD’标记为 0，用于表示未识别词‘OOV’标记为 1，通过映射关系，可以对留言内容中的词语进行标记，进行标记的目的是为了可以使用 tensorflow.keras.preprocessing.sequence 的 pad_sequences 方法来统一句子长度。

将一级分类标签转为 id 向量：通过构建 {标签：id} 字典，通过映射关系，将标签转为 id。

2.2.1.2.5. 拆分训练集和测试集：

通过 train_test_split 方法，将数据 20%分为测试集，80%分为训练集。

2.2.1.3. 处理输入数据

统一句子长度：因为放入神经网络中的数据需要具有相同的结构，因此需要统一句子

长度。通过 pad_sequences 方法统一句子长度为 100，此时的文本是用 id 向量表示的。

2.2.1.4. 处理神经网络词嵌入层

2.2.1.4.1.构建 {word：vector} 字典。

通过 gension 加载二百万的腾讯词向量，利用 all_dict，构建 {word：vector} 字典，word 是 all_aict 中所有已经登录在二百万腾讯词向量的词，一个词对应一个 200 维的腾讯词向量。

2.2.1.4.2.构建{id:vector}字典：

由 {word：id} 和 {word：vector} 可以构造出 {id：vector} 字典，用于构建神经网络中词嵌入层的词向量矩阵。

2.2.1.4.3.构建词向量矩阵：

词向量矩阵的大小是：len(all_dict)*vector_size,，第一行表示的是 id 为 0 的词向量，第二行表示的是 id 为 2 的词向量，以此类推。这里用 200 维的零向量表示用于填充句子长度的字符‘PAD’，未登录词则将 200 维的零向量其中一个 0 改为 1。

2.2.1.4.4.腾讯词向量

腾讯人工智能实验室嵌入中文词语语料库，该语料库为超过 800 万个中文单词和短语提供了 200 维向量表示，即嵌入。这些单词和短语是在大规模高质量数据上预先训练的，向量捕获了中文单词和短语的语义，可以广泛应用于许多下游中文处理任务(例如命名实体识别和文本分类)以及进一步的研究。

之所以选择腾讯词向量，是因为与现有的中文嵌入语料库相比，该语料库包含了大量特定领域的词汇、含最近出现或流行的新词，入能够更好地反映汉语单词或短语的语义。

2.2.1.5.搭建 LSTM 神经网络

embedding 层使用上面处理的词向量矩阵，因为腾讯词向量已经是训练过的，因此这里的 embedding 层不参与训练时的调整，整个模型结构如下：

Model: "sequential_2"

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, None, 200)	15788600
bidirectional_2 (Bidirection	(None, 400)	641600
dropout_4 (Dropout)	(None, 400)	0
dense_4 (Dense)	(None, 200)	80200
dropout_5 (Dropout)	(None, 200)	0
dense_5 (Dense)	(None, 8)	1608

Total params: 16,512,008
Trainable params: 723,408
Non-trainable params: 15,788,600

图 3 模型结构图

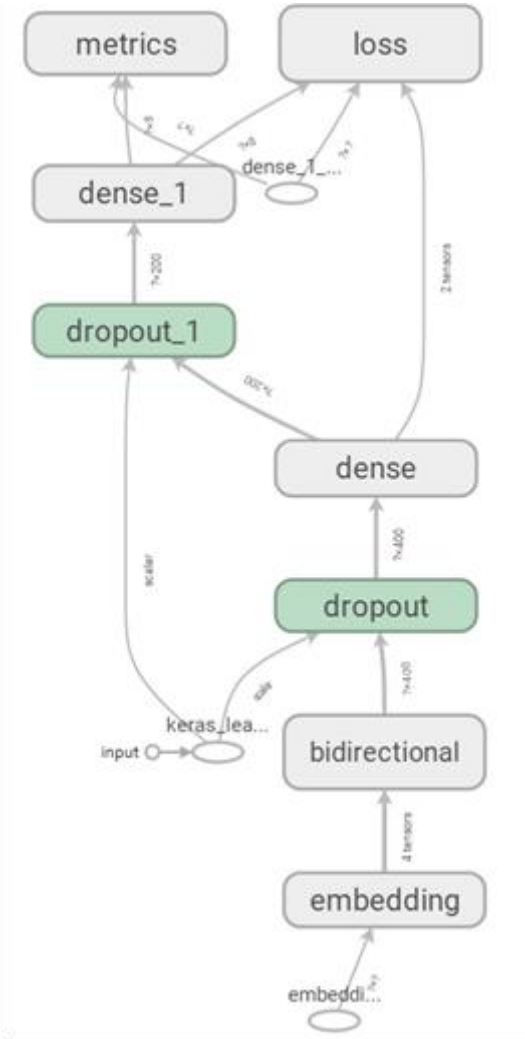

图 4 模型图

2.2.1.6. 长短时记忆网络（LSTM）算法

长短期记忆模型（long-short term memory）是一种特殊的 RNN 模型，于 1997 年由 Hochreiter 提出。

2.2.1.6.1. 递归神经网络原理

在结构上，递归神经网络模型由输入层、隐藏层、输出层构成，通过将上一时刻状态信息传递到当前状态，从而将时间序列展开为一系列具有相互性的神经元，其结构图如图 4 所示。如图 4 所示，在每一时间点上的神经元用节点表示。输入层与隐藏层之间的权重连接记为“U”隐藏层；与神经元之间的权重连接记为“W”；隐藏层与输出 层之间的权重连接记为“V”。RNN 模型的不足之处在 于，随着时间的累加，传递过程中不可避免地出现信息损失，从而导致初始信息的能力逐渐弱化，随之出现梯 度消失等问题，也就丧失了对时间序列的长时间处理能力。

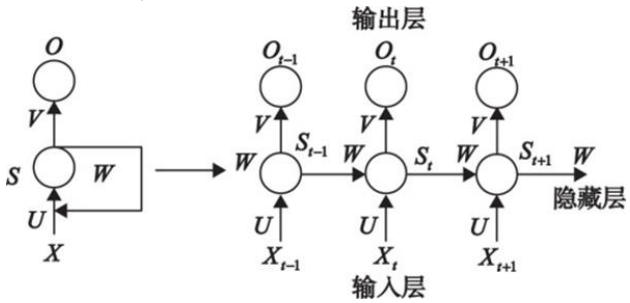

图 5 神经网络结构图

2.2.1.6.2. LSTM 原理

针对 RNN 训练过程中出现的梯度消失问题，LSTM 网络被提出并得到完善。LSTM 的主要理论是设计记忆模块，该模块可受控于多个控制门，从而实现对长时间 信息的记忆功能。

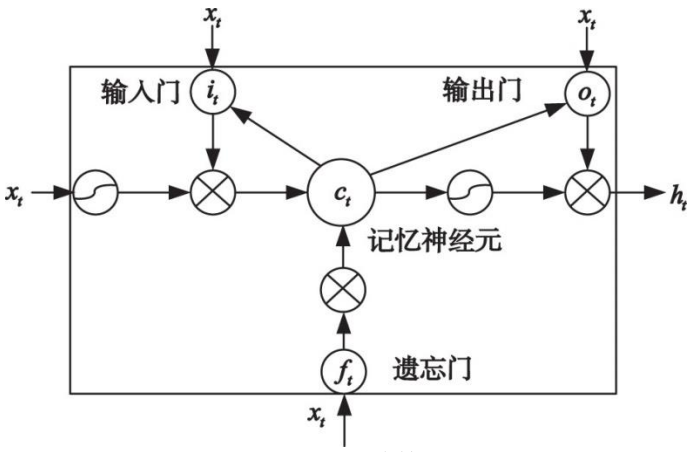

图 6 LSTM 结构图

LSTM 网络更新为:

设 h 为 LSTM 单元输出，c 为 LSTM 记忆模块值，x 为单元输入，W_{xc}、W_{hc} 分别为输入数据和上一时刻输出 的权值，时刻 t 的候选记忆单元值为:

$$\tilde{c_t} = \tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right)$$

输入门可控制当前数据输入对记忆模块状态的影响，可表示为:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i)$$

遗忘门可控制历史学习对当前记忆模块状态值的影响，可表示为:

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f)$$

当前时刻记忆单元状态值 c_t 为(符号⊙表示逐点乘积):

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c_t}$$

输出门可控制记忆模块状态值的输出，可表示为:

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_{t-1} + b_o)$$

LSTM 网络单元输出可表示为:

$$h_t = o_t \tanh\left(c_t\right)$$

2.2.1.7.1. 模型训练

进行模型训练，同时调整训练次数和学习速率，最终选择训练 7 次，学习速率为 0.001。

```
Train on 7368 samples, validate on 1842 samples
Epoch 1/7
7368/7368 - 121s - loss: 1.0571 - accuracy: 0.7036 - val_loss: 0.7044 - val_accuracy: 0.8252
Epoch 2/7
7368/7368 - 105s - loss: 0.6670 - accuracy: 0.8143 - val_loss: 0.6366 - val_accuracy: 0.8208
Epoch 3/7
7368/7368 - 106s - loss: 0.4789 - accuracy: 0.8629 - val_loss: 0.5019 - val_accuracy: 0.8561
Epoch 4/7
7368/7368 - 105s - loss: 0.4059 - accuracy: 0.8802 - val_loss: 0.4850 - val_accuracy: 0.8588
Epoch 5/7
7368/7368 - 105s - loss: 0.3521 - accuracy: 0.8966 - val_loss: 0.4069 - val_accuracy: 0.8833
Epoch 6/7
7368/7368 - 105s - loss: 0.2868 - accuracy: 0.9191 - val_loss: 0.4572 - val_accuracy: 0.8735
Epoch 7/7
7368/7368 - 103s - loss: 0.2606 - accuracy: 0.9245 - val_loss: 0.4038 - val_accuracy: 0.8865
```

图 7 训练过程图

2.2.1.7.2.模型预测

进行预测，得到测试集的预测值。

2.2.1.8. 模型评估

通过 classification_report 方法对模型进行评估，如图，f1 分数为 0.89。

	precision	recall	f1-score	support
0	0.89	0.73	0.80	123
1	0.90	0.94	0.92	320
2	0.93	0.93	0.93	398
3	0.83	0.83	0.83	241
4	0.86	0.94	0.90	159
5	0.89	0.82	0.85	407
6	0.85	0.97	0.91	194
accuracy			0.89	1842
macro avg	0.88	0.88	0.88	1842
weighted avg	0.89	0.89	0.89	1842

图 8 模型评估图

其准确率和损失值可视化为

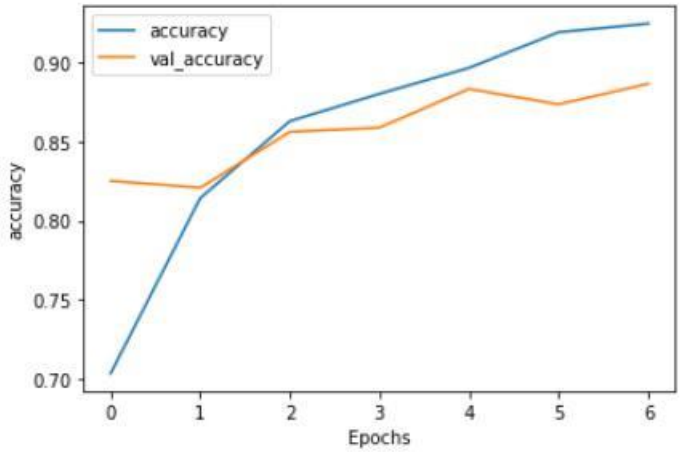

图 9 可视化图 1

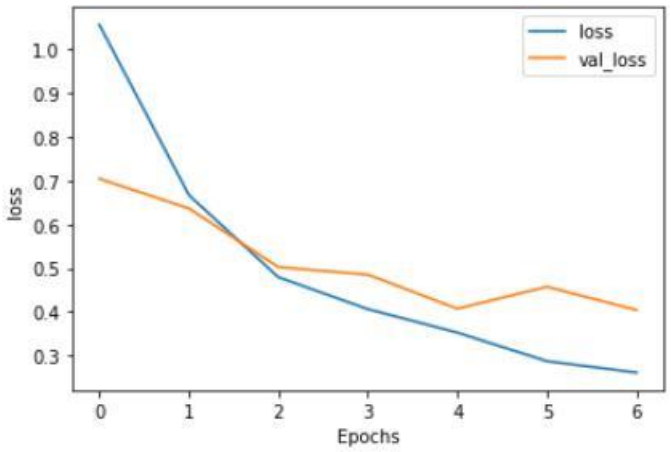

图 10 可视化图 2

2.2.2. 问题 2 的具体步骤

2.2.2.1. 流程图

把问题 2 拆分成两个流程进行

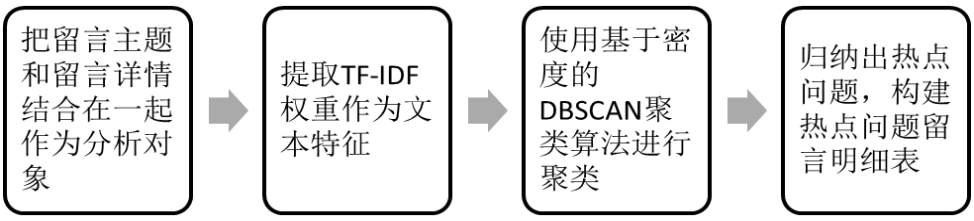

图 11 问题 2 第一部分流程图

图 12 问题 2 第二部分流程图

2.2.2.2. 归类出热点问题，构建热点问题留言明细表

2.2.2.2.1. 数据预处理

2.2.2.2.1.1. 群众留言信息的去重

在附件给出的信息数据中，出现了重复的留言信息，例如留言用户多次重复留言。考虑到相关部门划分、整理相关留言时，可能每天都会对要处理的问题进行更新，因此在去重的时候应该取更新时间最晚的记录，去掉历史记录。考虑到 python 中的字典在保存数据时，key 相同的内容，value 取值为最后更新的值。因此在读取数据时，按时间升序把留言用户作为 key， 把整个留言信息作为 value 保存在 value 中。最后再将字典中的内容写入文本即可。

2.2.2.2.1.2. 对留言主题、留言详情的提取

通过正则表达式提取中文字符和标点符号、分词。

2.2.2.2.1.3. 去除停用词、部分噪声的词

2.2.2.2.2. TF-IDF 算法

在对群众留言信息分词后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法，把群众留言信息转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即TF 权重（Term Frequency）。

词频（TF）=某个词在文本中出现的次数 （1）

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即:

$$词频（TF）= \frac{某个词在文本中的出现次数}{文本的总词数} \quad （2）$$

或

$$词频（TF）= \frac{某个词在文本中的出现次数}{该文本出现次数最多的词的出现次数} \quad （3）$$

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$逆文档频率（IDF）=\log(\frac{语料库的文本总数}{包含该词的文本数+1}) \quad （4）$$

第三步，计算TF-IDF 值（Term Frequency Document Frequency）。

TF - IDF = 词频（TF）× 逆文档频率（IDF） （5）

实际分析得出TF-IDF 值与一个词在留言主题出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的热点问题。

2.2.2.2.3. 生成TF-IDF 向量

生成TF-IDF 向量的具体步骤如下：

（1）使用 TF-IDF 算法，找出每个留言主题描述的关键词；

（2）对每个留言主题提取的关键词，合并成一个集合，计算每个留言主题对于这个集合中词的词频，如果没有则记为 0；

（3）生成各个热点问题的 TF-IDF 权重向量，计算公式如下：

TF-IDF=词频（TF）×逆文档频率（IDF） （6）

2.2.2.2.4. 文本特征的提取

通过调用 sklearn 库中的 CountVectorizer,与 TfidfTransformer 方法，计算每条留言的 TF-IDF 权重，并转为权重矩阵。

2.2.2.2.5. DBSCAN聚类

由于留言中有很多不能作为热点问题，需要通过该聚类方法将留言问题过滤，因此，把上面的TF-IDF权重矩阵作为输入，进行DBSCAN聚类，观察聚类结果，不断调整聚类参数eps和min_samples，最后确定eps=1.1，min_samples=10，得到12类热点问题。

2.2.2.2.6. DBSCAN聚类分类算法

DBSCAN是一个比较有代表性的基于密度的聚类算法。与划分和层次聚类方法不同，它将簇定义为密度相连的点的最大集合，能够把具有足够高密度的区域划分为簇，并可在噪声的空间数据库中发现任意形状的聚类。

DBSCAN中的几个定义：

（1） E 邻域：给定对象半径为 E 内的区域称为该对象的 E 邻域；

（2）核心对象：如果给定对象 E 邻域内的样本点数大于等于min_samples，

则称该对象为核心对象；

（3）直接密度可达：对于样本集合D，如果样本点q在p的Ε邻域内，并且p为核心对象，那么对象q从对象p直接密度可达。

（4）密度可达：对于样本集合D，给定一串样本点p1,p2….pn，p= p1,q= pn,假如对象pi从pi-1直接密度可达，那么对象q从对象p密度可达。

（5）密度相连：存在样本集合D中的一点o，如果对象o到对象p和对象q都是密度可达的，那么p和q密度相联。

DBSCAN中两个重要参数eps ,min_samples

(1)eps：越大类别数越少——参数越大的话，多个簇和大部分对象会归并到同一个簇中

(2)需要对eps和min_samples手动设置，eps默认值是0.5,表示的是半径

如果 附近点的数量 ≥min_samples，则当前点与其附近点形成一个簇，并且出发点被标记为已访问(visited),如果 附近点的数量 <min_samples,则该点暂时被标记作为噪声点。

(3)扫描半径 (eps)和最小包含点数(min_samples)，给定点在邻域内成为核心对象的最小邻域点数：min_samples

(4)min_samples设置为1是不合理的，因为设置为1，则每个独立点都是一个簇，min_samples≤2时，与层次距离最近邻域结果相同，因此，min_samples必须选择大于等于3的值。若该值选取过小，则稀疏簇中结果由于密度小于min_samples,从而被认为是边界点儿不被用于在类的进一步扩展；若该值过大，则密度较大的两个邻近簇可能被合并为同一簇。因此，该值是否设置适当会对聚类结果造成较大影响，如果min_samples不变，Eps取得值过大，会导致大多数点都聚到同一个簇中，Eps过小，会导致一个簇的分裂；如果Eps不变，min_samples的值取得过大，会导致同一个簇中点被标记为离群点，min_samples过小，会导致发现大量的核心点。

2.2.2.2.7. 构建热点问题留言明细表

DBSCAN聚类的结果中，-1类是噪声项，需要剔除，因此将聚类结果作为“问题ID”插入附件3后，将“问题ID”等于-1的样本剔除，进行排序后，得到热点问题留言明细表，部分数据如下。

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
0	228423	A00051608	A3区西湖街道茶场村五组不属于拆迁部分的村民该何去何从	2019/1/9 14:19:17	\n\t\t\t\t\t\n\t\t\t\t\t各位领导，你们好！我是A3区西湖街道茶场村五...	0	0
0	214447	A00051608	A3区西湖街道茶场村五组是如何规划的?	2019/4/1 17:07:36	\n\t\t\t\t\t\n\t\t\t\t\t请问局长，（A3区山后山）的A3区西湖街道茶...	0	0
0	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	\n\t\t\t\t\t\n\t\t\t\t\t请问胡书记，A3区西湖街道茶场村六组已经因为...	0	0
0	189739	A00051608	请问A3区西湖街道茶场村五组是如何规划的	2019/9/12 8:30:47	\n\t\t\t\t\t\n\t\t\t\t\t请问领导，政府对于A3区西湖街道茶场村五组是...	0	0
0	240551	A00051608	A3区西湖街道茶场村五组什么时候能拆迁	2019/7/9 17:09:32	\n\t\t\t\t\t\n\t\t\t\t\t请问领导，A市A3区西湖街道茶场村五组何时能...	0	0
...	...	...	...	...	...	...	...
11	260007	A00036889	A6区二中要求学生提前到校补课	2019/8/22 7:25:15	\n\t\t\t\t\t\n\t\t\t\t\tA6区二中老师在家长群通知：8.24号可能提...	0	0
11	256328	A000100973	A市长郡楚府中学强制学生周六补课	2019/10/22 15:40:17	\n\t\t\t\t\t\n\t\t\t\t\t我是A市长郡楚府中学初三的学生，我们学校强制...	0	0
11	278791	A00024813	A市雅礼中学强制高二学生周六补课	2019/3/1 15:01:54	\n\t\t\t\t\t\n\t\t\t\t\tA市雅礼中学违反教育局规定，强制高二学生在周...	0	0
11	249309	A00015335	A市同升湖学校寒假对学生实施有偿补课	2019/1/8 9:59:37	\n\t\t\t\t\t\n\t\t\t\t\tA市同升湖学校在今年寒假对学生实施有偿补课：...	0	1
11	276340	A00030644	A7县第一中学违规补课收费	2019/11/30 22:12:32	\n\t\t\t\t\t\n\t\t\t\t\tA7县第一中学违规补课收费将于12月1日进行...	1	3

图13热点问题留言明细表部分数据

2.2.2.3.定义热度评价指标，构建热度指数排名前五的热点问题表

2.2.2.3.1．地点、人群命名实体识别

具体步骤如下：

（1）基于热点问题留言明细表中的留言主题，我们将其中的地点、人群提取出来。我们将留言主题中的“请求”、“关于”、“建议”等干扰词去掉。

(2）构建Bert+LSTM+CRF网络，用于训练人民日报语料库。模型结构如下：

Model: "model_3"

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, None)	0	
input_2 (InputLayer)	(None, None)	0	
model_2 (Model)	multiple	101322240	input_1[0][0] input_2[0][0]
bidirectional_1 (Bidirectional)	(None, None, 128)	426496	model_2[1][0]
crf_1 (CRF)	(None, None, 7)	966	bidirectional_1[0][0]

Total params: 101,749,702
Trainable params: 101,749,702
Non-trainable params: 0

图14 Bert+LSTM+CRF网络模型结构

训练好模型后，对留言主题进行命名实体识别，部分数据如下：

0	中海国际社区三期与四期
1	A7县开元路深业睿
2	号线保利麓谷林语段
3	万家丽马王汽堆配城地铁5号线
4	保利麓谷林语小区
5	
6	市四医院滨水星城院区
7	才市场站
8	窑岭站
9	五家岭天健城
10	

图15 模型训练后部分数据

从中发现，对于“A市”开头的文本，该模型无法提取到A市，只能提取到后面的地点，因此通过正则表达式，将格式如“字母+市”的地点提取出来。

（3）人群的提取，选择通过Hanlp的维特比分词器，提取出词性为“nnt”的词语，作为该留言中人群。

（4）将地点与人群拼接起来，插入到热点问题留言明细表中，由此得到每一条热点问题留言所涉及地点、人群。

2.2.2.3.1.1. BERT 模型

BERT 模型（Bidirectional Encoder Representation from Transformers）是谷歌 AI 团队发布于 2018 年 10 月的 BERT 预训练模型，被认为是 NLP 领域的极大突破，刷新了 11 个 NLP 任务的当前最优结果。

其对 GPT 语言模型做了进一步的改进，通过左、右两侧上下文来预测当前词和通过当前句子预测下一个句子，预训练的 BERT 表征可以仅用一个额外的输出层进行微调，在不对任务特定架构做出大量修改条件下，就可以为很多任务创建当前最优模型。

2.2.2.3.1.2. 条件随机场 CRF

条件随机场（Conditional random field，CRF）是给定一组输入随机变量

条件下另一组输出随机变量的条件概率分布模型，其特点是假设输出随机变量构成马尔可夫随机场。条件随机场常用于序列标注问题，比如命名实体识别等。

CRF 的建模公式如下：

$$\mathbf{P(I|O)} = \frac{\mathbf{1}}{\boldsymbol{Z(O)}}\prod_{i}\boldsymbol{\psi_i(I_i|O)} = \frac{\mathbf{1}}{\boldsymbol{Z(O)}}\prod_{i} e^{\sum_i^T\sum_k^M \lambda_k f_k(O,I_{i-1},I_i i)}$$

下标 i 表示当前所在的节点（token）位置，下标 k 表示我这是第几个特征函数，并且每个特征函数都附属一个权重 $\boldsymbol{\lambda_k}$ ，$token_i$ 有 M 个特征，每个特征执行一定的限定作用，然后建模时我为每个特征函数加权求和；$\boldsymbol{Z(O)}$ 是用来归一化生成概率值的；

$\mathbf{P(I|O)}$表示了在给定的一条观测序列 O=(o_1, …, o_i)条件下，用 CRF 所求出来的隐状态序列 I=(i_1, …, i_i)的概率。

其特征函数：

$$\text{Score} = \sum_i^T\sum_k^M \lambda_k f_k(O,I_{i-1},I_i,i)$$

为 $token_i$ 打分，满足条件的就有所贡献。最后将所得的分数进行 log 线性表示，求和后归一化，即可得到概率值。

2.2.2.3.1.3. Hanlp 的维特比分词器

HanLP 是由一系列模型与算法组成的 Java 工具包，目标是普及自然语言处理在生产环境中的应用。HanLP 具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。

HanLP 功能有：中文分词、词性标注、命名实体识别、依存句法分析、关键词提取新词发现、短语提取、自动摘要、文本分类、拼音简繁。

2.2.2.3.2. **提取代表地点/人群**

将每一类热点问题中点赞数和反对数之和最高的留言所涉及的地点/人群，作为代表地点/人群。

问题ID	地点/人群	点赞数	反对数
0	A市西湖街道茶场村五组	0	0
1	碧云街道丽发新城小区	9	2
2	联合广铁集团	20	2
3	A市辉煌国际城二期居民楼	4	1
4	A市梦想枫林湾建筑工	0	0
5	西地省中交驾校	3	0
6	长房云时代小区幼儿	22	1
7	西地省直公积金管理中心	26	0
8	A市高级技师	14	2
9	A市陵王公园	9	0
10	A市西站	9	2
11	楚府中学学生	3	2

图16每个类型热点问题的点赞和反对数图

2.2.2.3.3.计算热度

首先定义热度评价指标：选择留言的持续天数、每一类的留言总数和每一类留言平均每条留言的点赞数与反对数之和作为衡量指标，并根据其对热点的影响程度，赋予一定的权重。其中热点是某一时期内的引人注目的问题，因此持续天

数对热度的影响较小，权重较小，设为 0.8。一条留言能引起的社会关注度比反映某一个问题的人数更重要，因此留言数的权重要小于留言平均参与量的权重，其中留言平均参与量即为每一类的留言总数和每一类留言平均每条留言的点赞数与反对数之和，由此构建的热度计算公式：

热度指数=持续天数 × 0.8 + 留言数 × 1.1 + 留言平均参与量 × 1.2

对每一类问题的热度指数进行计算，按照热度指数的大小，对每一类留言进行排序，根据热度排名，得到热度排名前五的五类热点问题。

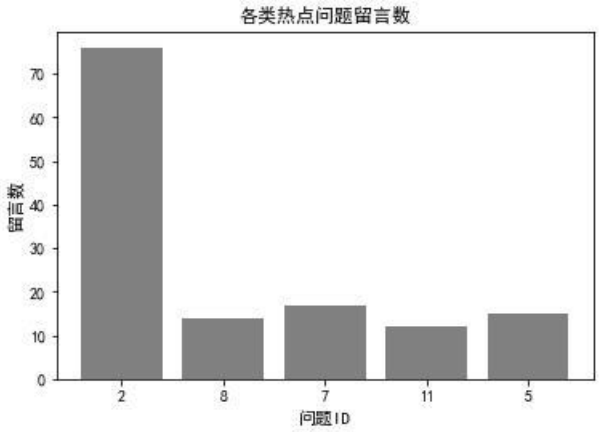

图17 排名前五的问题ID

2.2.2.3.4. 进行问题描述

根据排名前五的热点问题的问题ID，观察留言内容，人工地进行问题描述。

2.2.2.3.5.构建热点问题表

选择每一类留言中点赞数与反对数最高的一条留言中涉及的地点/人群作为代表，放入热点问题表中，构建最终的热点问题表

	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
0	1	2	378.594737	2019/01/08至2020/01/06	联合广铁集团	捆绑销售车位
1	2	8	327.571429	2018/11/15至2019/12/02	A市高级技师	购房补贴问题
2	3	7	305.405882	2019/01/02至2019/12/25	西地省直公积金管理中心	公积金贷款问题
3	4	11	301.600000	2019/01/08至2019/12/29	楚府中学学生	学校对学生进行违规补课
4	5	5	291.620000	2019/01/07至2019/12/16	西地省中交驾校	驾校收费退费违规

图18热点问题表

2.2.3.问题3的具体步骤

2.2.3.1.流程图

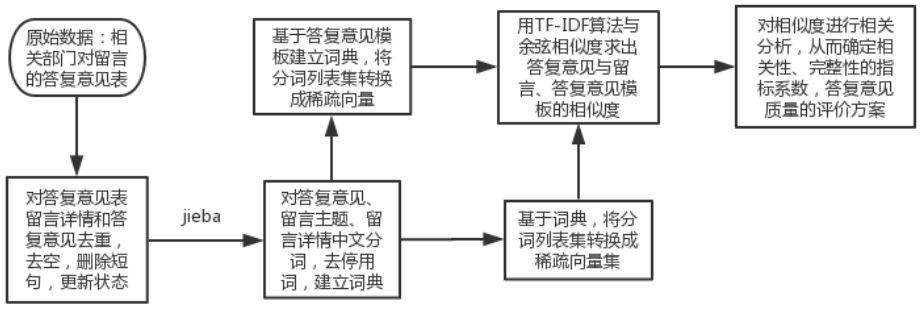

图19 问题3流程图

2.2.3.2.数据预处理

2.2.3.2.1.答复意见表的去缺失值、去重、去除短句

在题目给的附件答复意见表中有一些重复的留言详情，答复意见。考虑到一些用户的留言较早时候相关部门先回复正在调查，后面再出现留言的时候才回复调查之后的相应结果，所以保留较晚的留言答复意见，去掉较早的留言答复意见，更新数据状态。答复意见中一些只有短句，对后面数据挖掘无意义，删除掉短句数据，更新数据状态。

2.2.3.2.2.中文分词，去停用词，建立字典，稀疏向量

在进行文本数据挖掘之前，先将留言主题，留言详情，答复意见进行中文分词。这里采用 Python 的中文分词包 jieba 进行分词，jieba 采用了基于前缀词典实现的高效词图扫描，生成句子中汉字所有可能成词情况所构成的有向无环图（DAG），同时采用了动态规划查找最大概率路径，找出基于词频的最大切分组合，对于未登录词，采用基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法，以便更好的实现中文分词。为了在信息检索中，节省存储空间和提高搜索效率，对分词后的数据进行去停用词。然后将去停用词的分词数据建立字典，基于字典将分词列表集转换成稀疏向量（语料库），以供文本挖掘分析使用。

2.2.3.2.3.TF-IDF 算法、基于词向量的余弦相似度

TF-IDF 是一种统计方法，用以评估一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。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。基于词向量的余弦相似度算法，是将文本作为一个多维空间的向量，计算两个文本的相识度即计算判断两个向量在这个多维空间中的方向是否是一样的。而这个多维空间的构成是通过将文本进行分词，每个分词代表空间的一个维度。当文本一的词向量为 $A=\{A_1,A_2,\cdots,A_n\}$，文本二的词向量为 $B=\{B_1,B_2,\cdots,B_n\}$，其余弦相似度公式为：

$$S=\cos(\theta)=\frac{A\cdot B}{\|A\|\cdot\|B\|}=\frac{\sum_{i=1}^{n}A_i\cdot B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}}$$

S 的值越接近 1 时两个向量的夹角越小，也代表两个文本越相似。

用 TF-IDF 算法求出稀疏向量即语料库中的文本特征，根据文本特征得出待比较的文本词向量，再根据基于词向量的余弦相似度求出答复意见与留言主题及详情、答复意见模板的相似度。这里使用 python 中的 gensim 库来进行操作处理。

2.2.3.3.答复意见的质量评价指标

答复意见的相关性指的是答复意见的内容是否与问题相关，这里选取答复意见与留言的相似度，相似度越高，代表其与留言内容的相关的可能性越大；答复意见的完整性指的是是否满足某种规范即指答复模板，根据附件所给的答复意见，确定一个答复意见模板，并取答复意见与模板之间的相似度来表示答复意见的完整性。对答复意见与留言内容、答复意见模板的相似度统计分析，确定评价指标为这两种相似度。答复意见跟模板再相似，跟留言的问题不相关，其对于留言者来说，没有解决问题，体验感差；其对于答复的相关部门来说，也没有解决问题，可能会再次收到相应留言和“投诉”，留言者对其印象不好。答复意见的相关性比完整性对其质量的影响大，其权重应比完整性的大。

2.2.3.4.结果分析

2.2.3.4.1.答复意见与留言、答复意见模板的相似度结果

通过合并留言主题与留言详情，作为留言分词、去停用词，建立词典，并将答复意见分词、去停用词，将分词列表集转换成稀疏向量集即语料库，根据 TF-IDF 算法建立 TF-IDF 模型，用语料库来对 TF-IDF 模型训练，求出特征词的 tfidf 值，通过余弦相似度算法求出答复意见与留言的相似度，分组统计之后的结果如下图。从图中答复意见与留言的相似度的频数直方图分布可以判断其服从正态分布。

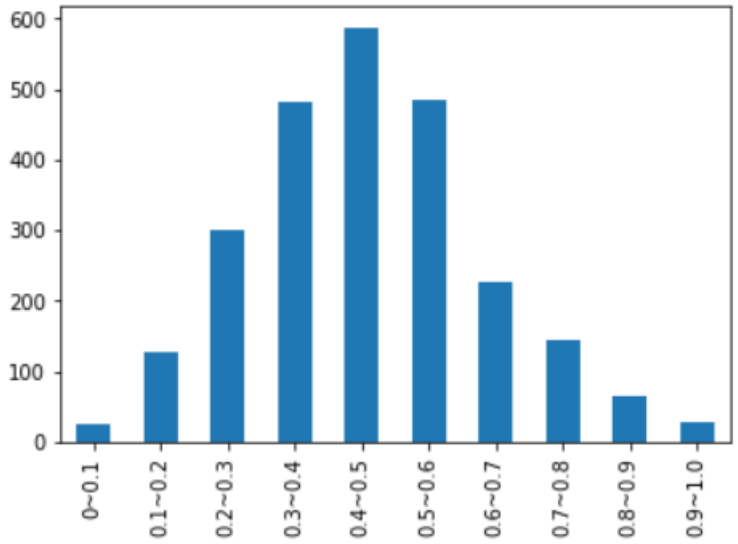

图 20 答复意见与留言的相似度的频数直方图

将答复意见模板分词，建立词典，重复上述操作，得到答复意见与答复意见模板的相似度结果如下图。

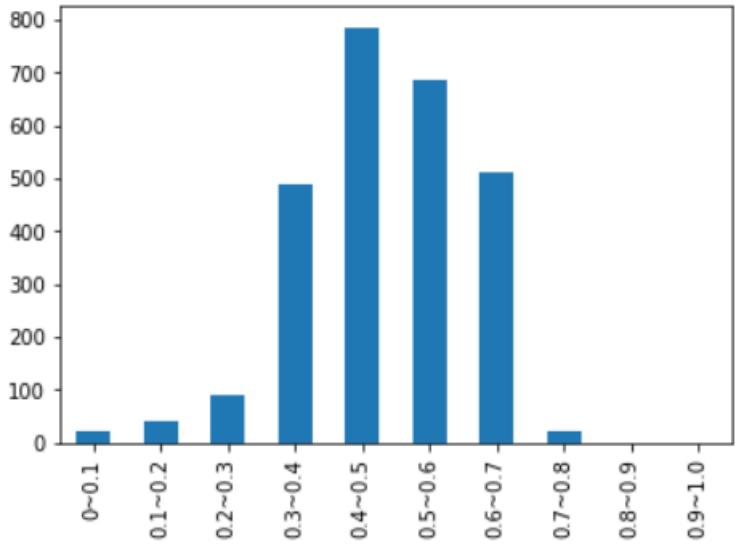

图 21 答复意见与答复意见模板的相似度的频数直方图

图 21 的数据分布说明对于答复模板，大部分答复意见是有相似的，在附件 4 中大部分答复意见是有模板的格式的，说明了相似度的准确性。

2.2.3.4.2. 答复意见评价指标得分结果

将答复意见与留言、模板的相似度作为相关性、完整性的指标，其中相关性的权重比完整性的权重大，则分别取 0.6，0.4。根据附件 4 中大部分答复意见与留言是相关的和正态分布的性质，其与留言的相似度大于 0.3 时，答复意见与

留言大概率相关，故将关于相关性的得分分段；而模板相似度则还是表示相似度越大，答复意见越完整。则答复意见评价得分公式为：

$$score=\begin{cases}(0.6\cdot sim1+0.4\cdot sim2)\times 100 & sim1\le 0.3\\ (0.6+0.4\cdot sim2)\times 100 & sim1>0.3\end{cases}$$

其中 $sim1$ 表示答复意见与留言的相似度， $sim2$ 表示答复意见与模板的相似度， $score$ 表示得分。计算得分之后结果如图 22。

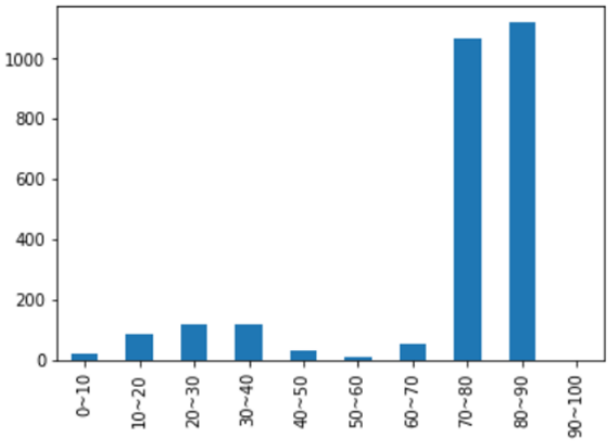

图 22 得分图

对比前面相似度的数值分布，其得分分布层次明显，留言相似度较好的基本评分较好。附件 4 中的答复意见大部分都是相关的，有模板格式的，其得分会比较高，评价效果不错，具体得分数值可以查看附件中的“zon1.csv”文件。

2.3. 结果分析

2.3.1. 问题1的结果分析

对LSTM神经网络模型训练的结果进行分析：在进行模型训练时，不断调整训练次数和学习速率，在最终选择的7次训练中，学习速率为0.001。

```
Train on 7368 samples, validate on 1842 samples
Epoch 1/7
7368/7368 - 121s - loss: 1.0571 - accuracy: 0.7036 - val_loss: 0.7044 - val_accuracy: 0.8252
Epoch 2/7
7368/7368 - 105s - loss: 0.6670 - accuracy: 0.8143 - val_loss: 0.6366 - val_accuracy: 0.8208
Epoch 3/7
7368/7368 - 106s - loss: 0.4789 - accuracy: 0.8629 - val_loss: 0.5019 - val_accuracy: 0.8561
Epoch 4/7
7368/7368 - 105s - loss: 0.4059 - accuracy: 0.8802 - val_loss: 0.4850 - val_accuracy: 0.8588
Epoch 5/7
7368/7368 - 105s - loss: 0.3521 - accuracy: 0.8966 - val_loss: 0.4069 - val_accuracy: 0.8833
Epoch 6/7
7368/7368 - 105s - loss: 0.2868 - accuracy: 0.9191 - val_loss: 0.4572 - val_accuracy: 0.8735
Epoch 7/7
7368/7368 - 103s - loss: 0.2606 - accuracy: 0.9245 - val_loss: 0.4038 - val_accuracy: 0.8865
```

图 23 训练过程图

并且在对模型进行评估时如图，f1分数为0.89。

	precision	recall	f1-score	support
0	0.89	0.73	0.80	123
1	0.90	0.94	0.92	320
2	0.93	0.93	0.93	398
3	0.83	0.83	0.83	241
4	0.86	0.94	0.90	159
5	0.89	0.82	0.85	407
6	0.85	0.97	0.91	194
accuracy			0.89	1842
macro avg	0.88	0.88	0.88	1842
weighted avg	0.89	0.89	0.89	1842

图24 模型评估图

再把其准确率和损失值可视化为以下两个图

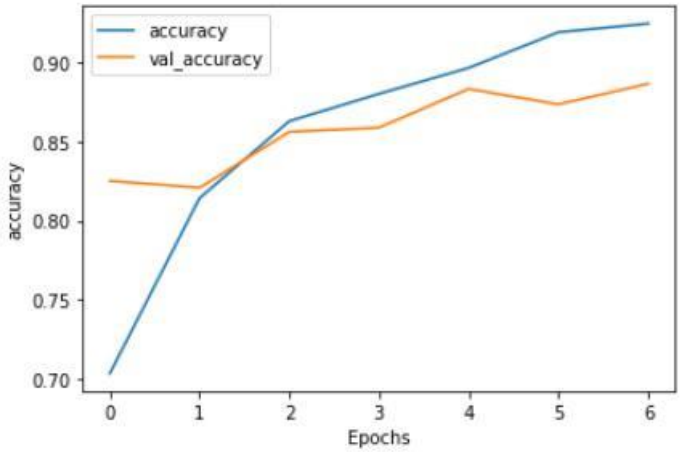

图25 可视化图1

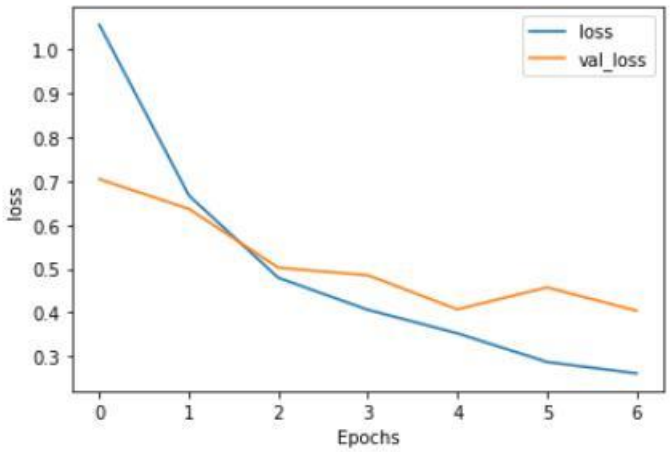

图26 可视化图2

从数据和可视化图像可以看出，在处理网络问政平台的群众留言时，运用LSTM神经网络建立的文本分类模型来实现留言内容的一级标签分类模型效率高，使用 F-Score 对分类方法进行评价，可以看出该模型的准确率高，损失值低。

解决现有依靠人工根据经验按照一定的划分体系、对留言进行分类这样导致的工作量大、差错率高等问题。也便于今后电子政务系统的运用及数据的智能化、数据化处理。

2.3.2.问题2的结果分析

2.3.2.1 .DBSCAN聚类分类结果

由于-1类是噪声项，将DBSCAN聚类的结果作为“问题ID”插入附件3后，将“问题ID”等于-1的样本剔除，进行排序后，得到热点问题留言明细表，部分数据如下，得到12类不同的热点问题对应的留言信息。及时发现热点问题，也有助于相关部门进行有针对性地处理，提升服务效率。

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
0	228423	A00051608	A3区西湖街道茶场村五组不属于拆迁部分的村民该何去何从	2019/1/9 14:19:17	\n\t\t\t\t\t\n\t\t\t\t\t各位领导，你们好！我是A3区西湖街道茶场村五...	0	0
0	214447	A00051608	A3区西湖街道茶场村五组是如何规划的?	2019/4/1 17:07:36	\n\t\t\t\t\t\n\t\t\t\t\t请问局长，（A3区山后山）的A3区西湖街道茶...	0	0
0	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	\n\t\t\t\t\t\n\t\t\t\t\t请问胡书记，A3区西湖街道茶场村六组已经因为...	0	0
0	189739	A00051608	请问A3区西湖街道茶场村五组是如何规划的	2019/9/12 8:30:47	\n\t\t\t\t\t\n\t\t\t\t\t请问领导，政府对于A3区西湖街道茶场村五组是...	0	0
0	240551	A00051608	A3区西湖街道茶场村五组什么时候能拆迁	2019/7/9 17:09:32	\n\t\t\t\t\t\n\t\t\t\t\t请问领导，A市A3区西湖街道茶场村五组何时能...	0	0
...	...	...	...	...	...	...	...
11	260007	A00036889	A6区二中要求学生提前到校补课	2019/8/22 7:25:15	\n\t\t\t\t\t\n\t\t\t\t\tA6区二中老师在家长群通知：8.24号可能提...	0	0
11	256328	A000100973	A市长郡楚府中学强制学生周六补课	2019/10/22 15:40:17	\n\t\t\t\t\t\n\t\t\t\t\t我是A市长郡楚府中学初三的学生，我们学校强制...	0	0
11	278791	A00024813	A市雅礼中学强制高二学生周六补课	2019/3/1 15:01:54	\n\t\t\t\t\t\n\t\t\t\t\tA市雅礼中学违反教育局规定，强制高二学生在周...	0	0
11	249309	A00015335	A市同升湖学校寒假对学生实施有偿补课	2019/1/8 9:59:37	\n\t\t\t\t\t\n\t\t\t\t\tA市同升湖学校在今年寒假对学生实施有偿补课：...	0	1
11	276340	A00030644	A7县第一中学违规补课收费	2019/11/30 22:12:32	\n\t\t\t\t\t\n\t\t\t\t\tA7县第一中学违规补课收费将于12月1日进行...	1	3

图27 12类热点问题对应的留言信息

2.3.2.2.BERT+LSTM+CRF网络模型训练结果

构建Bert+LSTM+CRF网络，用于训练人民日报语料库。模型结构如下：

```
Model: "model_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #   
=================================================================
input_1 (InputLayer)         (None, None)              0         
_________________________________________________________________
input_2 (InputLayer)         (None, None)              0         
_________________________________________________________________
model_2 (Model)              multiple                  101322240 
                                                                 
_________________________________________________________________
bidirectional_1 (Bidirectional) (None, None, 128)      426496    
_________________________________________________________________
crf_1 (CRF)                  (None, None, 7)           966       
=================================================================
Total params: 101,749,702
Trainable params: 101,749,702
Non-trainable params: 0
_________________________________________________________________
```

图28 Bert+LSTM+CRF网络模型图

训练好模型后，对留言主题进行命名实体识别，但是该模型无法提取到市级，只能提取到下一级地点，因此通过正则表达式，地点提取出来。得到部分数据如下：

0	中海国际社区三期与四期
1	A7县开元路深业睿
2	号线保利麓谷林语段
3	万家丽马王汽堆配城地铁5号线
4	保利麓谷林语小区
5	
6	市四医院滨水星城院区
7	才市场站
8	室岭站
9	五家岭天健城
10	

图29 地点提取数据

2.3.2.3.构建热度指数并计算排序结果

请根据附件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整合，将每一类热点问题中点赞数和反对数最高的留言整合归纳在一起。定义热度指数作为评价指标，按照热度指数大小进行排序，跟据排名前5的热点问题ID，再进行人工问题描述，相关数据如下所示，在联合广铁集团存在捆绑销售车位现象，A市高级技师存在购房补贴问题，西地省直公积金管理中心存在公积金贷款问题，楚府中学学生反映学校对学生进行违规补课问题，西地省中交驾校存在收费退费违规问题，这五个热点问题关注的，也便于引起政府相关部门的重视。

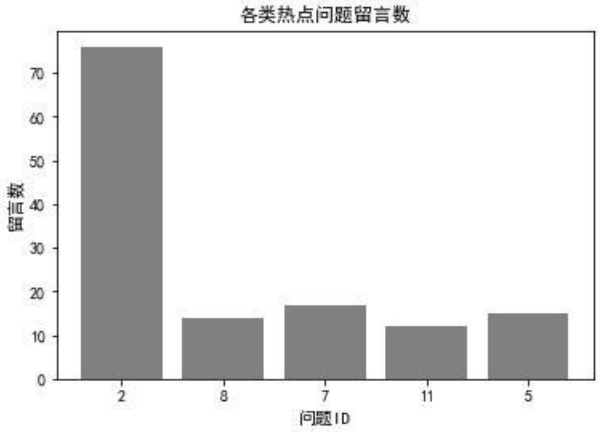

图30 五类热点问题

	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
0	1	2	378.594737	2019/01/08至2020/01/06	联合广铁集团	捆绑销售车位
1	2	8	327.571429	2018/11/15至2019/12/02	A市高级技师	购房补贴问题
2	3	7	305.405882	2019/01/02至2019/12/25	西地省直公积金管理中心	公积金贷款问题
3	4	11	301.600000	2019/01/08至2019/12/29	楚府中学学生	学校对学生进行违规补课
4	5	5	291.620000	2019/01/07至2019/12/16	西地省中交驾校	驾校收费退费违规

图31 热度问题对应的内容

2.3.3.问题3的结果分析

2.3.3.1.答复意见与留言、答复意见模板的相似度结果

通过余弦相似度算法求出答复意见与留言的相似度，分组统计之后的结果如图。从图中答复意见与留言的相似度的频数直方图分布可以判断其服从正态分布。

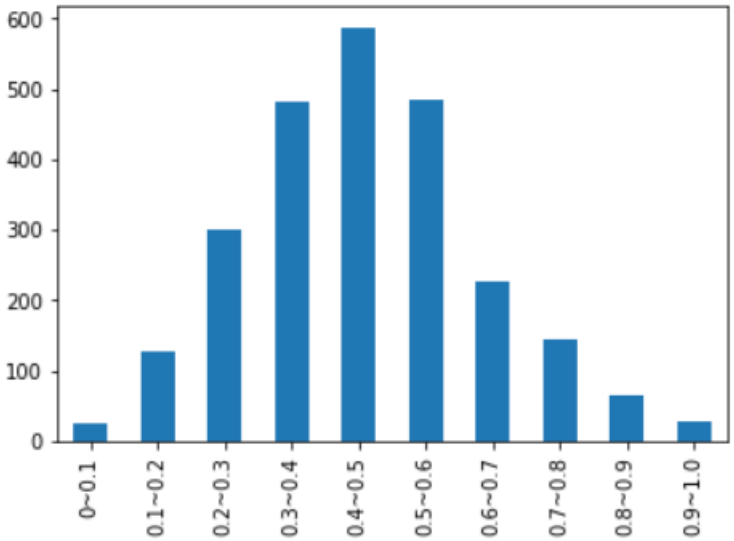

图32 答复意见与留言的相似度的频数直方图

按同样方法得到答复意见与答复意见模板的相似度结果如图 3。从数据分布说明对于答复模板，大部分答复意见是有相似的，从附件 4 中大部分答复意见的模板格式也看出相似度的准确性。

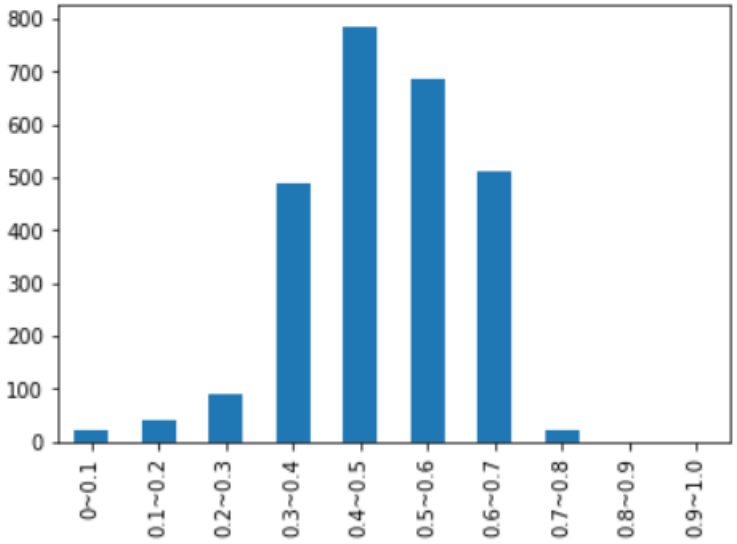

图 33 答复意见与答复模板的相似度的频数直方图

2.3.3.2.答复意见评价指标得分结果

将答复意见与留言、答复意见与答复模板的相似度作为相关性、完整性的指标，其中相关性的权重比完整性的权重大，并且根据相关性和正态分布的性质，将关于相关性的得分分段；而模板相似度越大，答复意见越完整。因此构建答复意见评价得分公式，计算得分之后结果如图 4。

对比前面相似度的数值分布，其得分分布层次明显，留言相似度较好的基本评分较好。答复意见大部分也都是相关的，有模板格式的得分会比较高，评价效果不错，总体从答复的相关性、完整性、可解释性等角度可以看出相关部门对留言的答复意见质量较高，基本分数段在 70 到 90 之间。

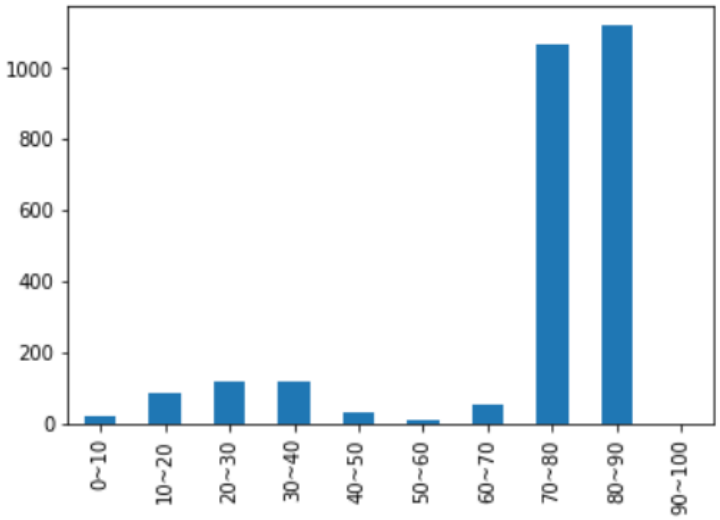

图 34 得分图

3. 结论

对网上群众留言信息进行分析研究，是政府了解社情民意的重要渠道，也是社会治理创新发展的新趋势，对政府相关部门都具有重大意义，同时也是文本数据分析的一个课题、一个难题。对于传统的人工文本分类和整理已经不能满足数据量庞大的网络问政，群众留言信息。本文采用根据DBSCAN聚类方法，统计群众目前最关注的热点问题类型，并给予相关问题负责任的答复， 深入分析社会热点问题现状。

由分析结果可以看出，网络群众留言中所需要的解决的热点问题主要有车位销售问题、购房补贴问题、公积金贷款问题、学校违规补课问题、驾校收费退费违规问题这五类。相关部门也及时给予相应的答复，从答复意见质量评价指标上看，例如从答复的相关性、完整性、可解释性等角度上看，基本是很高效，负责任的。

通过相关的得分分布规律，以及相关可视化图形结果上看，运用网络文本分析和数据挖掘技术对网络问政信息的处理能力强，速率高，因此网络问政平台对提升政府的管理水平和施政效率具有极大的推动作用，具有广泛的应用和发展前景。

4. 参考文献

[1]Yan Song, Shuming Shi, Jing Li, and Haisong Zhang. Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings. NAACL 2018 (Short Paper)

[2]史国荣,戴洪德,戴邵武,陈强强.基于长短期记忆网络的时间序列预测研究[J].仪表技术,2020(02):24-26+29.（LSTM算法）

[3]朱志远.基于数据挖掘的网络招聘系统是设计与实现.电子科技大学.硕士学位论文.2013

[4]DBSCAN聚类原理及详解_#define微光-CSDN博客 https://blog.csdn.net/weixin_43224539/article/details/85003631

[5]jieba 分词原理_Python_樱花落瓣-CSDN 博客 https://blog.csdn.net/Sakura55/article/details/88286293

[6]tf-idf_360 百科 https://baike.so.com/doc/433640-459181.html

[7]词向量余弦算法计算文本相似度_Java_相逢一笑-CSDN 博客
https://blog.csdn.net/qq_28851503/article/details/97616249
[8]BERT 项目地址：
https://github.com/google-research/bert#fine-tuning-with-bert
[9]知乎：https://www.zhihu.com/question/35866596/answer/236886066
[10]Hanlp 项目地址：https://github.com/hankcs/pyhanlp

5. 附件

附录清单：

附件一：data:存放程序使用到的数据

1. Parameter：问题二中 bert 模型的部分数据
2. 5000-small.txt：5000 个腾讯词向量（缩小版）
3. stop_words.txt：问题一和问题二所使用的停用词表
4. train1.txt 与 test.txt:问题二中命名实体识别模型的训练数据和测试数据
5. stoplist.txt“问题三使用的停用词表
6. 热点问题留言明细表
7. 热点问题表

附件二：Question_1：问题一的程序

1. datapreparel.py：数据预处理程序
2. Classification_model.py：文本分类模型程序

附件三：Question_2：问题二的程序与结果文件

1.Q_2_1：Message_schedule.py：热点问题留言明细表程序

2.Q_2_2：（1）BertLstmCrf.py：命名实体识别模型

（2）DealWithData.py：数据处理程序

（3）model_train.py：模型训练程序

（4）Hot_issue_schedule.py：热点问题表程序

附件四：Question_3：问题三的程序与结果文件

1.问题三计算相似度和评价得分的 python 程序

2.问题三预处理之后的相似度、得分数据 zon1