智慧文本中政务系统的挖掘应用

摘 要

近年来，随着网络问政平台逐步成为政府了解民意、凝聚民气的重要渠道，群众的留言数量攀升，以往主要依靠人工完成划分留言和整理热点问题的工作面临着极大挑战。运用大数据、人工智能等技术来利用自然语言处理和文本挖掘的方法处理工作任务，极大地提高了工作效率，从而也提高了政府的管理水平和施政效率。

针对问题一，首先对留言主题进行数据预处理，我们在机器学习中采用三个模型，分别是朴素贝叶斯模型，KNN模型，BP神经网络模型，在深度学习中也采用了三个模型，分别是LSTM模型，Bi-LSTM模型，GRU模型，我们对模型进行融合，运用根据权重构建词袋模型， 推导预测结果， 最后使用f1_score对分类预测的精准度进行评判。

针对问题二，首先对留言主题进行数据预处理，通过采用k-means聚类算法的模型，我们对留言进行了分类处理，并建立初步的热点问题留言明细表。其次，我们引入hanlp处理包，进行分词和词性分析，并将地点信息提取。然后，我们引入pyhanlp处理包，分离地点信息和问题内容以获取热点问题，并求出三个关键词。再使用jieba库中的analyse对三个关键词进行提取和优化，得到最终的关键词。最后，使用自定义函数根据地点和关键词，求出包含的问题。然后清空原有的热点问题留言明细表，将相对应的内容填充到表格中。接着，通过热点问题留言明细表中的时间跨度，求出问题的时间范围。最后，我们根据地点信息、留言问题和时间范围，求出热点问题表。此时，我们得到了问题二中所需要的两个表格。

针对问题三,首先对附件 4 的留言主题与答复意见的文本数据进行数据处理，然后从相关性，规范性，及时性这三个环节来评判两者之间的关系，并根据建立评分标准进行评分，最后将三个环节的分数逐次累加，我们得到了对答复意见的评分分值。

此次实验分析并评估了该系统处理留言分类，挖掘热点问题，留言与答复意见关系的能力，并简要介绍了未来的计划，以此来更好地完善此政务系统。

关键词：深度学习，机器学习，模型融合 ，k-means聚类，文本相似度

Abstract

In recent years, as the Internet platform for asking government questions has gradually become an important channel for the government to understand public opinions and gather people's spirit, the number of messages from the masses has increased. In the past, the work of dividing messages and sorting out hot issues, which was mainly done manually, faced great challenges. The use of big data, artificial intelligence and other technologies to deal with work tasks using natural language processing and text mining methods has greatly improved the work efficiency, thus improving the government's management level and efficiency.

In response to question one, we first preprocessed the message topic. We used three models in machine learning, namely Naive Bayesian model, KNN model and BP neural network. We also used three models in depth learning, namely LSTM model, Bi-LSTM model and GRU model. We fused the models and used the analysis method of proportion weight to predict the accuracy of classification.

In response to question 2, we preprocessed the message topic and classified the messages by using the model of k-means clustering algorithm. Secondly, we introduce hanlp processing package, carry out word segmentation and part of speech analysis, and extract location information. Then, we introduce pyhanlp processing package to obtain hot issues and extract and optimize keywords. We set up a hot topic table, look for the contents of its six tags and fill them in. Then we empty the original list of hot issues and fill the corresponding contents into the tables. At this time, we get the two tables needed in question 2.

In response to question 3, first of all, the text data of the message subject and the reply opinions in Annex 4 are processed, then the relationship between the two is judged from the three links of relevance, standardization and timeliness, and the score is made according to the established scoring standard. Finally, the scores of the three links are accumulated one by one, and we get the score of the reply opinions. This experiment analyzed and evaluated the system's ability to deal with message classification, dig hot issues, and the relationship between message and reply opinions. It also briefly introduced the future plans, in order to better improve the government affairs system.

Key words: Deep Learning， Machine learning, model fusion, k-means clustering, text similarity

目 录

第 1 章 引言

1.1 问题背景

近年来，随着微信、微博、市长信箱等网络问政平台逐步成为政府了解民意的重要平台，各类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问题整理的相关部门的工作带来了极大挑战。尤其是在疫情期间，政府督查局需要及时处理群众反映的问题和提出的意见，只有妥善解决群众关心的交通出行、看病就医等方面的焦点问题，才能得到广大群众的认可。然而需要靠人工来进行留言划分的工作是艰巨的，若不能及时完成分类任务，那么给后续的热点问题查找和留言回复工作也带来了诸多不方便。

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，它对提高政府的管理水平和施政效率具有极大的推动作用。

1.2 问题重述

对于问题一，群众留言分类问题，参赛者需要参考附件 1 所提供的内容分类三级标签体系，建立关于附件 2 中留言内容的一级标签分类模型。同时，参赛者需要使用 F1-score 模型对分类方法进行评价，以确保留言分类的准确性。

对于问题二，热点问题挖掘，参赛者要将某一段时间内反映特定地点或特定人群的留言问题进行归类与整理，同时要对热点问题定义合理的热度评价指标，并展示评价结果。参赛者要按“表 1”的格式选出排名前 5 的热点问题，并保存为“热点问题明细表”，按表 2 的格式给出相应热点问题的留言信息，并保存为“热点问题明细表”。

对于问题三，答复意见的评价，参赛者要根据附件 4 相关部门对留言的答复意见，从答复意见的相关性、完整性、可解释性等角度，制定出一套合理且高质量的评价方案，并尝试实现。

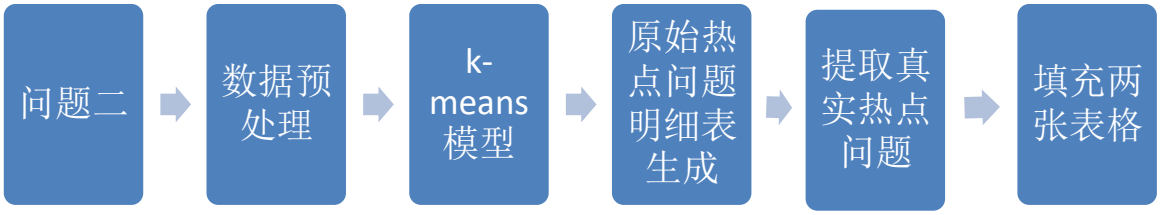
基于对政务系统的理解和认识，本文将立足于以上背景和问题，构建模型，并完成基于限定文本数据的留言处理操作。在完成对题目所给问题集的数据分析以及预处理工作后，该模型与其他主流方案相比，在准确率以及泛化能力.上都表现出优越的效果。本文包括引言、系统模型、问题处理、实验结果四个部分。

第 2 章 模型框架

2.1 问题一

此模型分为三个阶段，第一阶段是对留言主题进行预处理工作，第二阶段是构建机器学习与深度学习的模型，共有六个模型，第三阶段是多种模型进行融合，预测留言分类的精准度。

2.2 问题二



此模型分为五个阶段，第一阶段是对留言主题进行预处理工作，第二阶段是运用 k-means 模型处理数据，第三阶段是原始热点问题留言明细表生成，第四阶段是提取真实热点问题，第五阶段是填充两张表格，只选取热度排名前 5 的部分。

2.3 问题三

此模型分为四个阶段，第一阶段是对文本中留言主题和答复意见的数据分别处理，第二阶段是从相关性的角度评判答复意见的质量，第三个阶段是从答复意见规范性的角度评判，第四阶段是从回复留言的及时性角度评判，最后将三个角度的评判分值累加。

第 3 章 数据预处理

高质量的数据文本是我们进行模型匹配和优化的基础，对整个数据文本进行分析和处理可以使我们对整个数据文本有更加全面的了解，从而更好地对数据进行特征工程编码表示，进一步提高解决问题的精准度。因此，在解决问题前，我们选择对数据文本进行预处理工作，为后续留言分类处理和热点问题挖掘进行铺垫。

3.1 数据分词

在解决问题的过程中，我们需要提取文本的特征向量。句子形式的样本特征不够明显。因此，我们需要对句子进行词语的划分即数据分词。此过程我们选择使用 python 的“jieba”库，因为它带有一个自己的“dict.txt”词典，囊括两万多词语，且具有查找快速的优势。“jieba”分词的词典中基本可以囊括在句子中出现的词语，若没有的词语可以自定义添加，使分词结果的准确率更高。

例：A 市西湖建筑集团占道施工有安全隐患

分词后：A，市，西湖，建筑，集团，占，道，施工，有，安全隐患

3.2 去停用词

在分词结果中，它含有许多对表达主题没有实际意义的噪音词，这些噪音词应作为停用词被去除。去除停用词可以降低句子噪音对句子的影响，使句子的主题更加明确，特征更加鲜明，提高留言分类的准确性。在附件 2 留言主题中，代表地址或者序号的字母也将使用正则表达式去除，因为这些符号对理解留言主题并没有实际意义，而且出现频率较高，容易造成噪音。

例：A，市，西湖，建筑，集团，占，道，施工，有，安全隐患

去除停用词后：西湖，建筑，集团，占，施工，安全隐患

3.3 封装调用

去除停用词的处理后，文本中以“，”作为分隔符，它会对后续处理带来不必要的麻烦，所以我们将“，”改为“ ”来拼接文本词语。完成此步骤后，对经过处理后的文本数据进行封装，将以上步骤定义到一个单元格（Jypyter-Notebook 的基本执行单元）中，在调用数据时，直接将单元格复制并运行，避免了重复书写。

第 4 章 问题一

问题一是对留言主题进行分类处理，将其分成七类，在此过程中我们采用了多种模型，根据模型融合的思想，对模型预测的结果进行了进一步优化。

4.1 TF-IDF 模型

在使用 TF-IDF 模型前，我们将附件 1 中的二级标签和三级标签也作为留言主题分类，并将同种标签进行去重处理，在一定程度上，增加了留言主题的数量。

词频-逆向文件频率模型（TF-IDF）[1]的主要思想是指在一篇留言中，某个词语的重要性与该词语在这篇留言中出现的次数成正相关，同时与整个留言库中出现该词语的留言数成负相关。

TF(term frequency):词频，表示一个词语与一篇留言的相关性。计算时用该词在一篇留言中出现的次数除以留言的总词数。

IDF(inverse document frequency):逆向文件频率，表示-一个词语的出现的普遍程度。可以表示为 log（总留言数/出现该词语的留言数)。

一篇留言中某个词语的重要程度，可以标记为词频和逆向文件词频的乘积。基于词频-逆向文件频率模型的主要思想，我们构建了这样一个智能阅读模型:

$$TF = \frac{N}{M} \tag{4-1}$$

其中，N：词语在某留言中的频次　M：该留言的词语数

$$IDF = \log(\frac{D}{D_w}) \tag{4-2}$$

其中，D_w：出现了该词语的留言数　D：总留言数

计算得出：TF-IDF=TF*IDF

在基于留言的词语重要性与词语在留言中出现的位置不相关的假设下，根据.上述公式我们可以获取留言中每个词语的 TF-IDF 值。我们在使用 TF-IDF 模型前，先将预处理后的词频转换成向量函数，然后再将其转化成 TF-IDF 权重矩阵，即可得到留言主题中每个词语的 TF-IDF 值。

4.2 机器学习

4.2.1 KNN 聚类

KNN 聚类模型[2]的原理是，一个样本在特征空间中，总会有最相似的 K 个样本，即“近朱者赤，近墨者黑”。在群众留言中，我们将留言主题进行分类处理区分留言种类，此时会有相似的几个留言主题被划分为一类留言。此过程我们只需确定 K 临近法的三个基本要素“距离度量、k 值的选择及分类决策规则”，当三要素确定后，此时留言的分类情况也基本确定。KNN 聚类模型可以用来处理本次留言分类，它对数据没有假设精准度高，但留言主题不平衡的时候，对稀有主题的预测准确率低，会导致留言分类的精准度低。

KNN 做分类预测时，一般是选择多数表决法，即训练集里和预测的样本特点最近的 K 个样。一类是橙色圆点，一类是蓝色，而红色三角形是被分类的数据。如下图所示:

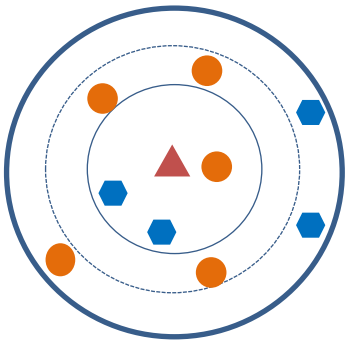


图 1 KNN 聚类训练示例图

如果 K=3，那么离红色点最近的有 2 个蓝色六边形和 1 个橙色的圆形，由于蓝色六边形所占比例为 2/3，所以红色的这个待分类点属于蓝色的六边形。

以此类推，如果 K=6，红色的这个待分类点属于橙色的圆形。

当 K=9 时，红色的这个待分类点也属于橙色的圆形。

我们在采用 KNN 模型时，把经过 TF-IDF 模型转化后的数值带入 KNN 模型，得到预测的结果集，为后续的处理做准备。

4.2.2 BP 神经网络

BP 神经网络算法具有泛化能力，对没有训练过的样本，有很好的预测能力和控制能力，对于多种多样的留言，BP 神经网络[3]具有更强的预测能力,用来预测也较为合适。

BP 神经网络的拓扑结构，一般包含三层前馈网，即输入层、中间层（也称隐含层）和输出层。它的特点是：各层神经元仅与相邻层神经元之间相互全连接，同层内神经元之间无连接，各层神经元之间无反馈连接，构成具有层次结构的前馈型神经网络系统。这样的神经网络结构来处理留言主题的分类，各主题之间不会相互影响，也不会造成分类之间的混乱。BP 神经网络的模型示例：

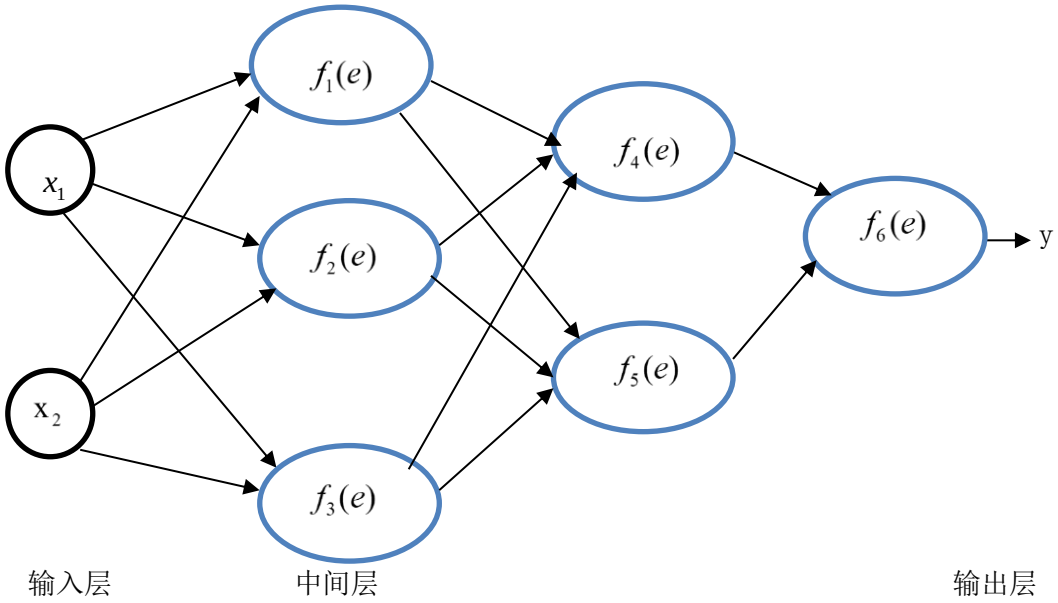


图 2 BP3 神经网络的拓扑结构

我们在使用 BP-神经网络算法时，将经过 TF-IDF 模型处理后的数值，分为训练集和测试集，处理后的样本数据作为 BP 神经网络的输入向量，对应的分类情况作为输出向量。样本数据训练网络，不同的输入向量得到不同输出值，输出值与期望值进行比较，当输出值与期望值之间的误差小于一个设定值时，神经网络模型训练完成，得到预测的结果集，为后续的处理做准备。

4.2.3 朴素贝叶斯模型

朴素贝叶斯算法[4]是一种直观的分类算法，也是最简单的贝叶斯分类器，具有很好的可解释性，且操作简便，对于同一分类主题的留言不会呈现出太大的差异性。留言之间的关系相对独立，采用此种模型的精确度可能会提高。

对于给出的待分类留言主题，在此项条件下各个分类的出现概率，哪个最大，就认为此待分类留言项属于哪个分类。其原理如下：

假设 A={ $X_1, X_2, X_3, \cdots X_n$ }为一个待分类项，每个 X 为 A 的一个特性，B={ $Y_1, Y_2, Y_3 \cdots Y_n$ }为类别的集合。

我们需要计算以下概率， P（$Y_1|A),P(Y_2|A)\cdots P(Y_n|A)$,并取得中间最大的作为我们的分类结果。过程可以分为，首先我们需要找一个已知分类的待分类项集合，作为训练集;统计得到在各个类别下各个特征属性的条件概率估计，即：

$$P(X_1|Y_1),P(X_2|Y_1)\cdots P(X_m|Y_1),\ \ P(X_1|Y_2),P(X_2|Y_2)\cdots P(X_m|Y_m)$$

朴素贝叶斯分类器是建立在一个条件独立性假设的基础之上，根据各个特征属性是条件独立的，则根据贝叶斯定理推导：

$$P(Y_i|A)=P(A|Y_i)P(Y_i)/P(A) \tag{4-3}$$

其中分母 P(A)表示 A 事件发生的概率，对于所有类别，都是同一个常数，所以，问题就转换成分子 $P(A|Y_i)P(Y_i)$ 最大化即可，所以有：

$$P(A|Y_i)P(Y_i)=P(X_1|Y_i),P(X_2|Y_i)\cdots P(X_m|Y_i)P(Y_i) \tag{4-4}$$

最终，取得 $P(A|Y_i)P(Y_i)$ 最大值，便得出分类结果。

我们在使用朴素贝叶斯分类算法时，将经过 TF-IDF 模型处理后的数值，作为样本输入，得到预测的结果集，为后续的处理做准备。

4.3 深度学习

构建模型时,我们选择使用 python 的 keras 库，因为它具有丰富的深度学习模型框架,易于构建深度学习的模型。在构建模型前,我们设置最频繁使用的 50000 个词，设置每条 cut_review 最大的词语数为 30，设置 Embeddingceng 层的维度为 50，同时设置 12 个训练周期，同时 batch_size 为 64，此时我们得到了相同的词语的个数。然后，填充 X_train,X_test，使每条数据的维数相同。

4.3.1 LSTM 模型建立

LSTM 模型（long-short term memory）[5]是一种特殊的 RNN 模型，在传统的 RNN 中，无法体现出长期记忆的效果，因此需要一个存储单元来存储记忆。在 LMST 模型中，有一个内存块（memory block ）包含 3 个控制门和一个记忆节点，它会对留言主题与分类标签相关的信息进行记忆，对不相关的信息进行遗忘，从而提高分类的准确度。

LSTM 模型示例：

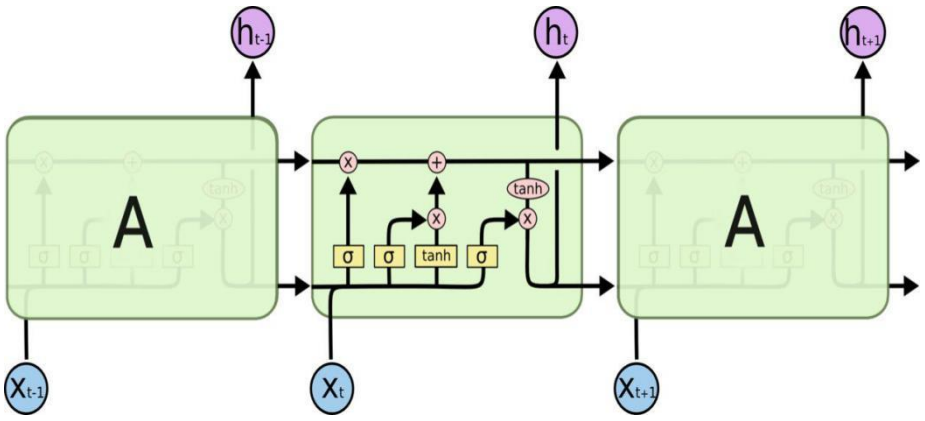


图 3 LSTM 模型示例图

（1）LSTM 的核心是“cell state”，被命名为细胞状态，也就是上述图中最顶的传送线。

（2）LSTM 中有 3 个控制门：输入控制门，输出控制门，遗忘控制门。

（3）LSTM 模型的工作原理：

①forget gate:遗忘控制门可以决定哪些信息被遗忘。随着时间的不断推进，一些没有用的历史信息被永久的遗忘。

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{4-5}$$

②input gate：输入控制门用于控制当前时刻的输入的留言主题是否对该分类标签有影响。因为某些特定情况下，一些隐层节点需要利用的是历史信息，另一些隐层节点可能利用当前信息，此时我们先将两部分的内容联合起来。

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4-6}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4-7}$$

③将过去与现在的记忆进行合并，即可以作为更新细胞状态:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{4-8}$$

④output gate：输出控制门用于控制此节点的输出在当前时刻是否起作用。

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{4-9}$$

$$h_t = o_t * \tanh(C_t) \tag{4-10}$$

接下来我们定义一个 LSTM 的序列模型: 模型的第一层是嵌入层(Embedding)，它使用长度为 50 的向量来表示，而且每一个词语 SpatialDropout1D 层在训练中每次更新时，都将输入单元的按比率随机设置为 0，这有助于防止过拟合的情况发生，最后在 LSTM 层包含 50 个记忆单元，输出层为包含 7 个分类的全连接层。LSTM 序列模型构建的信息图表，如下图所示：

Layer (type)	Output Shape	Param #
embedding_3 (Embedding)	(None, 30, 50)	715200
spatial_dropout1d_1 (Spatial	(None, 30, 50)	0
lstm_2 (LSTM)	(None, 50)	20200
dense_5 (Dense)	(None, 7)	357

Total params: 735,757
Trainable params: 735,757
Non-trainable params: 0

图 4 LMST 的序列模型构建信息图表

4.3.2 Bi-LSTM 模型

Bi-LSTM 模型[6]使用两个 LSTM 神经网络，已经成功应用在在字符分类标注、机器阅读理解研究中，知道上下文信息，对于提高字符序列标注准确率非常益。Bi-LSTM 模型包括前向计算和后向计算，水平方向表示时间序列的双向流动，竖直方向表示从输入层到隐藏层以及从隐藏层到输出层的单向流动，如下图所示：

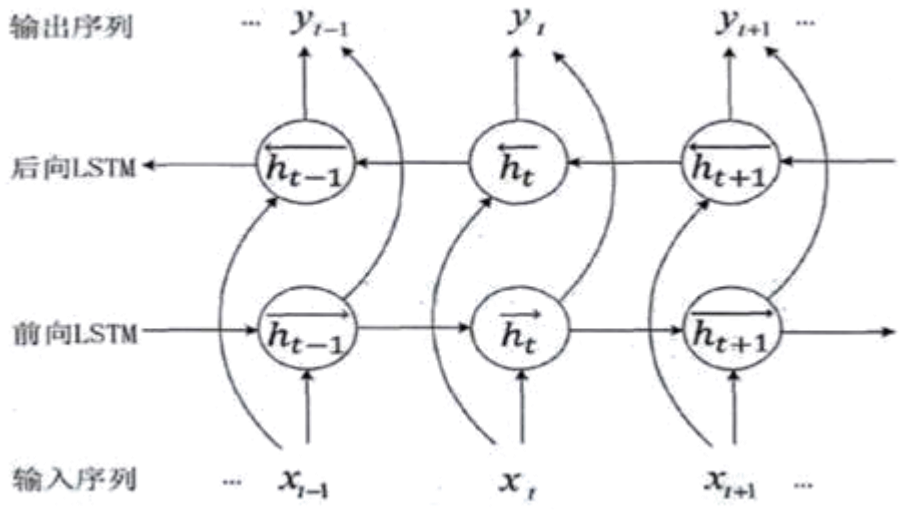


图 5 BI-LSTM 模型展开图

Bi-LSTM 神经网络[7]采用两个 LSTM 神经网络，分别是前向计算隐向量 $\vec{h}_t$，后向计算隐向量 $\overleftarrow{h}_t$，将正向输入序列和反向输入序列的输出结果结合，则 y_t 为最终输

出结果:

$$\vec{h}_t = LSTM(x_t, \vec{h}_{t-1}) \tag{4-11}$$

$$\overleftarrow{h}_t = LSTM(x_t, \overleftarrow{h}_{t-1}) \tag{4-12}$$

$$y_t = g(W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t + b_y) \tag{4-13}$$

接下来，我们构建了一个 Bi-LSTM 模型，其模型构建的信息图表如下图所示：

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_4 (Embedding)      (None, 30, 50)            2500000
_________________________________________________________________
dropout_2 (Dropout)          (None, 30, 50)            0
_________________________________________________________________
bidirectional_3 (Bidirection (None, 30, 400)           401600
_________________________________________________________________
dropout_3 (Dropout)          (None, 30, 400)           0
_________________________________________________________________
flatten_2 (Flatten)          (None, 12000)             0
_________________________________________________________________
dense_6 (Dense)              (None, 7)                 84007
=================================================================
Total params: 2,985,607
Trainable params: 2,985,607
Non-trainable params: 0
```

图 6 BI-LSTM 模型构建的信息图表

4.3.3 GRU 模型

GRU 即门控循环单元模型[8]。GRU 保持了 LSTM 的效果同时又使结构更加简单计算量更小，是 LSTM 模型的变种结构。GRU 把 LSTM 中的 forget gate 和 input gate 用 update gate 来替代。 把 cell state 和隐状态 htht 进行合并，在计算当前时刻新信息时，GRU 模型有了自己的特色方法。GRU 模型示例图，如下所示：

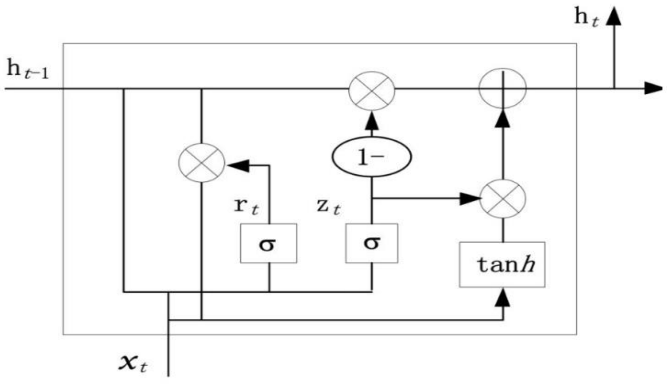


图 7 GRU 模型示例图

GRU 模型有两个门，分别为更新门 zt 和重置门 rt。更新门用于控制前一时刻的状态信息被带入到当前状态中的程度，更新门的值越大说明前一时刻的状态信息带入越多；重置门用于控制忽略前一时刻的状态信息的程度，重置门的值越小说明忽略得越多。GRU 模型向前传播公式：

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{4-14}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{4-15}$$

$$\tilde{h}_t = \tanh(W_{\tilde{h}} \cdot [r_t * h_{t-1}, x_t]) \tag{4-16}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{4-17}$$

$$y_t = \sigma(W_o \cdot h_t) \tag{4-18}$$

接下来，我们构建了一个 GRU 模型，其模型构建的信息图表如下图所示：

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding_3 (Embedding)      (None, 30, 50)            715200
_________________________________________________________________
spatial_dropout1d_1 (Spatial (None, 30, 50)            0
_________________________________________________________________
lstm_2 (LSTM)                (None, 50)                20200
_________________________________________________________________
dense_5 (Dense)              (None, 7)                 357
=================================================================
Total params: 735,757
Trainable params: 735,757
Non-trainable params: 0
```

图 8 GUR 模型构建的信息图表

4.4 模型融合

为了提高预测的精确度，我们采用模型融合的方法，进一步确定留言主题的分类。我们采取的模型融合时权重模型融合方法。该模型融合的主要思想是若对于一个分类问题，有多个不同的预测模型，采取权重的方法对不同模型的预测结果进行加和，占比最大的分类结果，就是最终的分类结果。

例如：

我们有 A、B 两个模型，留言有两个分类，分别为种类 X、种类 Y。在 A 模型中，种类 X 评判的精准度为 0.7，种类 Y 评判的精准度为 0.5。在 B 模型中，种类 X 评判的精准度为 0.6，种类 Y 评判的精准度为 0.8。

假设我们现有一留言主题需要进行分类，在模型 A 中被分为种类 X，在模型 B

中被分为种类 Y，由于模型 A 中种类 X 的精准度为 0.7，小于模型 B 中种类 Y 的精准度 0.8，所以此留言主题被划分为种类 Y.

我们在进行模型融合时，采用基于权重的分析方法，与上述的简单模型一样，只是增加了种类和模型数量。我们进行了多个模型融合，通过多次实验对比，最终得到了精准度不同的融合结果。

模型融合后得出的结果相对于融合前，有着不小的提升。但是，在深度学习过程中，随着训练周期的增加，会产生过拟合现象，该现象会对精确度造成影响，模型融合未能解决此问题。

第 5 章 问题二

问题二主要是根据附件三制作热点问题表（表 1）和热点问题留言明细表（表 2 ）两份表格。两张表中的数据需要通过下面模型和方案求出。

5.1 k-means 模型

在 k-means 聚类分类处理前，先使用 TF-IDF 模型将留言主题的文本数据转化为向量数据，以方便后续模型的使用。k-means 聚类算法[9]是典型的基于原型的目标函数聚类方法的代表，它是数据点到原型的某种距离作为优化的目标函数，对处理大数据集，该算法保持可伸缩性和高效性。k-means 聚类对热点问题大致分成五类。数据 x_i 可表示为 $x_{\mathrm{i}}=(x_{i1},\mathrm{x}_{i2},\cdots x_{im})$。

样本 x_i 和 x_j 的欧式距离：

$$d(x_i,x_j)=\sqrt{\sum_{d=1}^{m}(x_{id}-x_{jd})^2} \tag{5-1}$$

样本 x_i 到所有样本距离的平均值：

$$w_i=\frac{1}{n}\sum_{j=1}^{n}d(x_i-x_j) \tag{5-2}$$

样本 x_i 的方差：

$$v_i=\frac{1}{\mathrm{n}}\sum_{j=1}^{n}(d(x_i,x_j)-w_i)^2 \tag{5-3}$$

待聚类数据集的平均距离：

$$w_{AV}=\frac{2}{n(n+1)}\sum_{i=1}^{n}\sum_{j=1}^{i}d(x_i,x_j) \tag{5-4}$$

误差平方和:

$$E=\sum_{i=1}^{k}\sum|x-\bar{x}_i|^2 \tag{5-5}$$

式子中：b_i 表示第 i 个类簇所包含的样本集合。

5.2 使用 K-means 划分原始热点问题留言明细表

首先，我们读取附件 3 中的数据，并对留言主题进行筛选和处理，如果留言主题的长度小于 4，我们会把留言详情的内容赋值给留言主题，这样会使留言主题涉及更多的详细信息，以此来保证留言分类的准确性。然后我们将留言主题进行预处理，包括：结巴分词，去除停用词，将“，”转化为“”，等步骤，再将预处理完成的结果转换成 tf-idf 模型的数据。此时，我们已经完成基本的处理操作，可以用 k-means 模型进行分类。

其次，我们将数据导入 k-means 模型进行训练，运用算法将所有样本划分为 20 个不同的类别，并将其转化成列表模式。此时，我们计算各样本与类中心的距离，可以分别求出二十个类别的中心热点问题。然后，我们得到了原始热点问题留言明细表，和二十个中心热点问题的位置序号。再将二十个中心问题的信息填充到热点问题表中，此时，我们就得到了热点问题表。

5.3 提取真实热点问题

首先,我们引入 hanlp 处理包,hanlp 是面向生产环境的自然语言处理工具包，它可以对句子进行语法和句法分析，词法分析具体情况如下：

例句：

他在浙江金华出生，他的名字叫金华。

词法分析：

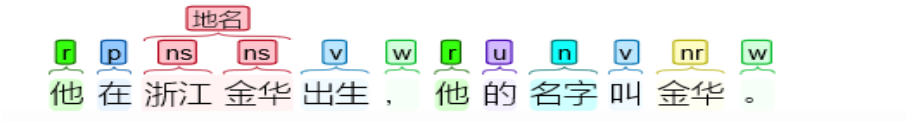


图 9 hanlp 例句词法分析图

词法分析 2.0：

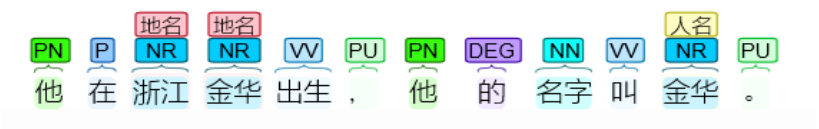


图 10 hanlp 例句词法分析 2.0 图

我们采用 hanlp 处理包的两个函数做分词，词性识别等处理，我们先将中心问题分词，再获取分词后的词性，其具体的处理过程如下：

导入函数

```
import hanlp
tokenizer = hanlp.load('PKU_NAME_MERGED_SIX_MONTHS_CONVSEG') #分词
tagger = hanlp.load(hanlp.pretrained.pos.CTB5_POS_RNN_FASTTEXT_ZH) #词性标注
recognizer = hanlp.load(hanlp.pretrained.ner.MSRA_NER_BERT_BASE_ZH)  #地点识别
```

例句:

请 A 市加快国家中心城市建设力度

分词结果:'请','A 市','加快','国家','中心','城市','建设','力度'

分词词性:'VV','NR','VV','NN','NN','NN','NN','NN'

然后我们采用 hanlp 处理包识别中心问题的地点信息，并将地点信息进行提取。但由于有些地点信息无法获取，所以我们使用自定义函数获取。自定义函数的原理时：无法获得的地名必定不是特殊地名，而是含有字母的字符串，如 A7，所以在此字母后面加一个名词，如“县”作为其地名即可，这样就可以保证地点信息的完全提取。

例如：

['A3', '区', '金茂梅', '溪湖', '别墅', '违建', '现象', '非常', '严重']
['CD', 'NN', 'NR', 'NR', 'NN', 'VV', 'NN', 'AD', 'VA']

在上述的示例中，“A3”是地名，但没有被识别出，我们在“A3”后面加“区”，即可作为地名信息被提取。

其次，我们采用 pyhanlp 来处理数据获取中心问题，先去除空项，再去除地点信息，获取到问题，部分问题展示如下：

[‘违建现象非常严重’，
‘部分路口信号灯建议’，
‘违规补课收费’，
‘捆绑销售车位’，
‘周边安全隐患问题’，
……]

然后，我们做提取关键词处理，先将中心问题中词性为“NN”的词语进行提取处理，若关键词中少于三个，则需要进行第二次的提取处理，处理规则：将不是“NN”“VE”“PU”“SP”“DEG”“ETC”“AD”等词性的所有词语进行提取。最后，我们对关键词进行优化，得到最具代表性的关键词，优化的三个步骤为：

1. 将词语长度小于 2 个字符的关键词去除；

2. 若关键词的个数大于 3，则只取后面三个关键词。

3. 使用“jieba”库中的 analyze 函数，将三个关键词中最重要的提取出来

完成上述步骤后，我们对原始热点问题留言明细表做优化处理，如果地点信息和关键词在留言详情中都有出现或者是在留言主题中都有出现，我们就认为该信

息时属于此类的一条信息，据此我们得到了真正的热点问题留言明细表。

5.4 填充两张表格

首先，我们填充热点问题表格，它有六个标签，分别为：热度排名,问题 ID,热度指数,时间范围,地点人群,问题描述。我们要将上述已经求出的二十类别信息填入到表格相应的标签下，此时仍有“问题 ID”“热度排名”“热度指数”和“时间范围”等标签内容空缺。

关于热度排名，我们使用自定义的热度值来作为评判标准。

热度值的标准：

$$热度值=\sum_{i=样本1}^{样本n}\left(\frac{点赞数-反对数}{2}\right)+此类热点问题的个数 \tag{5-6}$$

由以上公式即可求出每一类热点问题热度值。此时，我们将热度排名和热度指数填入到问题表中，并更新表格内容。

然后我们清空原有的热点问题留言明细表，只留下标签，根据热度排名的数值将真正的热点问题留言明细表赋值给原始热点问题留言明细表，按问题 ID 顺序排列。我们还需要求得热点问题表中的时间范围。我们处理留言时间采用 python 的 Datetime 模块。Datetime 可以帮助我们识别并处理与时间相关的元素，如日期，小时，分钟，秒，月份，年份等。

我们选取一个固定的时间，要求此时间在留言时间之前，求出此类热点问题中所有留言时间与此固定时间的时间间隔（单位天）。

例如：选取固定时间 2000-1-1，

若留言时间是 2019-1-2，时间间隔 t=6941,

若留言时间是 2019-6-13，时间间隔 t=7103

若留言时间是……，时间间隔是……

我们将所有时间间隔中，最小 t 值所对应的留言时间为最小时间，最大 t 值所对应的留言时间为最大时间。最小时间和最大时间即为我们所求的时间范围，并将格式改为规范格式，例：2019-1-2 至 2019-6-13。

至此，我们已经求出了热点问题表和热点问题明细表的所有内容，现将相关内容填入到表中，划分出热点排名前 5 的热点问题和热点问题对应的留言信息。

第 6 章 问题三

问题三是评判附件 4 中相关部门对留言的答复意见情况，从多种角度对答复意见的质量制定出一套评价方案。我们主要是从留言主题与答复意见的相关性，答复意见的规范性，答复留言的及时性等几个方面来制定评价方案，同时要将三个环节的分数累计相加并用百分制数值的形式来表现答复意见的质量。

6.1 数据处理

首先，我们对附件 4 的数据进行预处理工作。我们需要对留言主题文本数据和答复意见文本数据进行结巴分词、去停用词处理等。完成数据预处理，有利于提高接下来的相关性评判的准确度。

6.2 相关性评判

在做答复意见的相关性评判时，我们将使用余弦函数来计算留言主题与答复意见之间的文本相似度。因此，我们建立了余弦定理计算相似度的函数：

$$\cos\theta = \frac{\sum_{i=1}^{n} x_{1i} * x_{2i}}{\sqrt{\sum_{i=1}^{n} (x_{1i})^2} * \sqrt{\sum_{i=1}^{n} (x_{2i})^2}} \tag{6-1}$$

其中，通过计算 x 和 y 之间对应特征向量的余弦夹角来判断留言主题与之间的相似度，当余弦值越接近 1 时，夹角就越接近 0 度，两个向量则越相似[10]。

然后，我们使用余弦函数进行文本相似度计算的步骤如下所示。

表 1 基于余弦相似度的留言主题与答复意见相似度计算

输入：留言主题与答复意见两者的文本数据
1.分别建立留言主题与答复意见的词典。 2.将两个词典合并为两个同样的词典，并对新建立的两个词典进行清零处理。 3.把留言主题与答复意见分别对应到新词典中，并计算每篇的词频。 4.生成两篇专利各自的词频向量。 5.计算两个向量的余弦相似度。
输出：两者之间的余弦相似度的数值

通过上述的步骤计算，我们可以得到余弦值的具体数值，并求出其四分位数，其计算结果如下图所示:

	余弦值
0	0.172917
1	0.154303
2	0.409514
3	0.346844
4	0.214423
...	...
2811	0.000000
2812	0.000000
2813	0.106199
2814	0.064134
2815	0.157135

图 11 文本相似度的余弦值

	余弦值
count	2816.000000
mean	0.189338
std	0.134798
min	0.000000
25%	0.087039
50%	0.174236
75%	0.279202
max	0.766032

图 12 余弦值的四分位数值

最后，我们根据余弦值的四分位数值建立一个自定义的分段函数，此分段函数可将余弦值转化成百分制，其数学公式表示为：

$$p=\begin{cases} cos\theta * 400, \cos\theta < 0.174236 \\ (cos\theta - 0.174236) * 150 + 0.174236 * 400, 0.174236 < cos\theta < 0.279202 \\ (cos\theta - 0.279202) * 20 + 0.279202 * 400 + (0.279202 - 0.174236) * 150, \cos\theta < 1 \end{cases} \quad (6\text{-}2)$$

将其余弦值转化为百分制分数的规则是

在余弦值的 50%分段处，乘 400，约为 69.7 分。

余弦值的 50%-75%分段间，乘 150，在 75%分段处约为 85.4 分。

余弦值的 75%-max 分段间，乘 25，在最大值处约为 95.1 分。

其函数图像为：

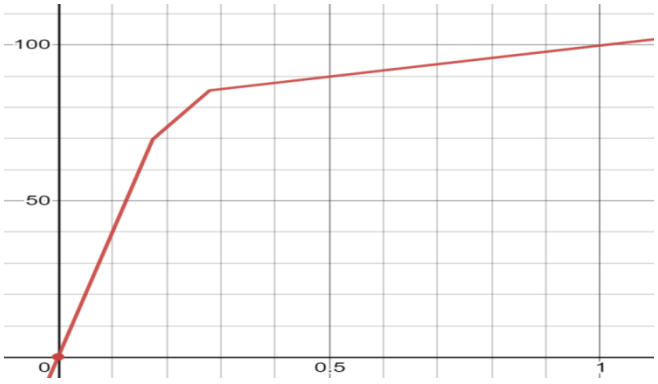


图 13 余弦值转化为百分制数值函数图像

当输入数据时得到了百分制数值，同时得到了相关性的评分。其具体的分值和四分位数值如下所示：

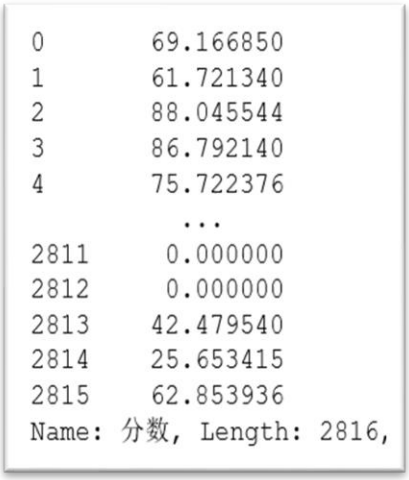


图 14 相关性的百分制评分

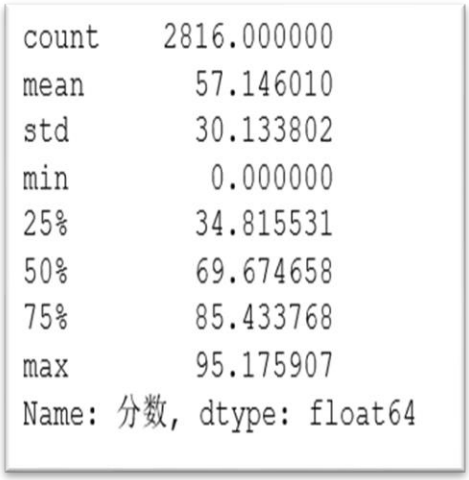


图 15 相关性分数的四分位数

此时，我们得到了第一环节相关性评判的分数。

6.3 规范性评判

在此环节中，我们要评判答复意见的规范性和礼貌性，因此我们添加了自定义的敬语词表对留言添加进一步的评分。敬语加分词表如下表所示：

表 2 敬语词汇及其附加分值

敬语词汇	附加分值
您好	2
感谢	2
答复	1
…	…

在敬语词表的参考下，根据文本数据中出现的敬语词语的频数和其对应的分数，在相关性的百分制分数的基础上对分数进行了第二环节的累加，评判了答复意见的礼貌性和规范性。我们可以得到其分数累加的情况，其具体数值和四分位数值结果如下图所示：

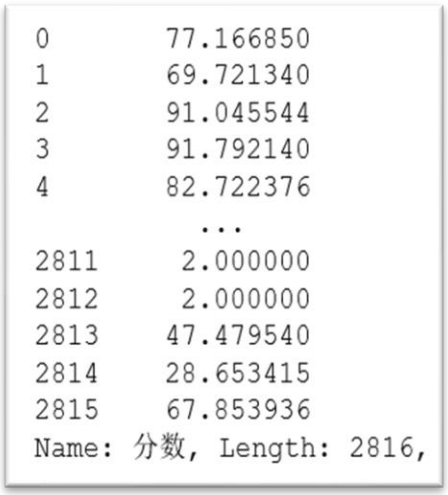


图 16 规范性评判后分值累加情况

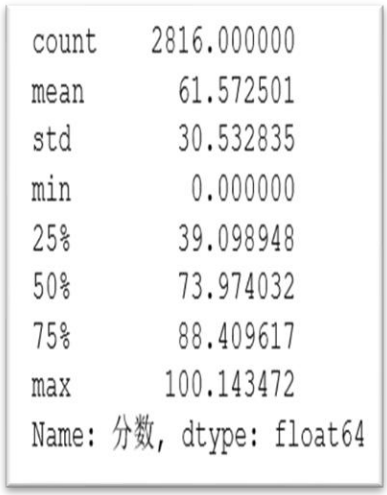


图 17 规范性评判后四分位数情况

6.4 及时性评判

在上述完成相关性和规范性的基础上，对回复留言的及时性也做出了评判标准，使答复意见的分数更具备说服力。此过程中，选用留言时间与答复时间两个时间数据，并对两个数据做差值，求出时间间隔，其时间间隔的四分位数值图表如下表格所示：

表 3 时间间隔的四分位数值分布情况

count	2816.000000
mean	20.821733
std	40.627094
min	1.000000
25%	5.000000
50%	12.000000
75%	23.000000
max	1161.000000

然后，我们根据时间间隔的情况做了第三环节及时性的评分标准，具体内容如下表格所示：

表 4 时间间隔的评分标准

时间间隔（t代表天数）	分数变动
0≤t≤5	+(5-t)
6≤t≤20	0
20<t<60	-(t-20)
60≤t	分数为0

注：若为负分直接归零。

通过对及时性的评判，分数进行进一步变动，但仍为百分制分数。

第 7 章 实验结果

7.1 问题一

7.1.1 评价指标

任务一中对留言主题的分类，主要采用的评价指标是 F1-Score。F1-Score 是精确率(precision)和召回率(recall rate)的调和函数。通常使用 F-Score 对分类方法进行评价:

$$F_1 = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \frac{2P_i R_i}{P_i + R_i} \tag{7-1}$$

其中 P_i 为第 i 类的查准率,R_i 为第 i 类的查全率。

7.1.2 实验结果

我们进行了多次融合实验，每次融合的模型的种类和数量不同，我们可以构建一个表格来对比不同模型的融合结果。

表 5 模型融合的结果对比

类别	融合模型	精准度
三个机器学习和三个深度学习	KNN，朴素贝叶斯，BP 神经网络，LSTM 模型，Bi-LSTM 模型，GRU 模型	0.8662983425414365
两个机器学习和三个深度学习	KNN，朴素贝叶斯，LSTM模型，Bi-LSTM模型，GRU模型	0.9337016574585635
一个机器学习和三个深度学习	朴素贝叶斯，LSTM模型，Bi-LSTM模型，GRU模型	0.9303867403314917
一个机器学习和两个深度学习	朴素贝叶斯，LSTM 模型，Bi-LSTM 模型	0.9237569060773481

注：当一个机器学习与三个深度学习融合时，虽然以朴素贝叶斯作为机器学习的模型，但通过结果可知，无论使用哪个机器学习模型，结果都相同。

经过上述多种融合情况的对比，我们可知，当两个机器学习和三个深度学习融合时，精准度最高，此时精准度为 0.9337016574585635。

7.2 问题二

在运用 k-means 模型的基础上，我们划分不同数量的类别，进行了多次试验，最终得到两个比较合适的试验结果。

7.2.1 实验结果一

运用 k-means 模型处理附件 3 数据，将数据分成二十类，最终可得到以下热点问题表和热点问题留言明细表，其表格如下所示：

表 6 二十聚类下的热点问题表

	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	1	203.0	2019-1-9至2020-1-1	A市	A 市公交线路建议
1	2	2	55.0	2019-1-2至2020-1-6	A市	A 市加快国家中心城市建设力度
2	3	3	45.0	2019-7-18至2019-9-1	伊景园滨河苑	捆绑销售车位
3	4	4	44.5	2018-11-15至2019-12-15	A市	询 A 市人才购房购房补贴实施办法相关问题
4	5	5	43.5	2019-11-2至2020-1-25	丽法新城小区	搅拌站噪音扰民

	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	1.0	189247	A000107866	建议在"我的A市"app中尽快接入法律服务的意见	2019-12-21 09:45:46	\n\t\t\t\t\t\t\n\t\t\t\t\t\t互联网时代，"我的A市"app对于推介A市，...	0	0
1	1.0	191111	A00072923	对A市公交线路的建议	2019-02-10 18:02:44	\n\t\t\t\t\t\t\n\t\t\t\t\t\t对A市公交线路的建议。1、必须时时牢记公交...	0	0
2	1.0	191381	A0008159	给A9市城区南西片区城铁站设立的建议	2019-07-04 18:52:39	\n\t\t\t\t\t\t\n\t\t\t\t\t\t肯定是选择在A9市西南角，支持A9市西南角设...	0	0
3	1.0	195915	A00018309	建议A市规划局东延地铁7号至泉塘片区，便民出行	2019-02-26 08:22:42	\n\t\t\t\t\t\t\n\t\t\t\t\t\t您好！泉塘片区位于星沙，占了A市经开区70%...	1	21
4	1.0	203208	A00043904	建议调整A5区新华都学校西南角红绿灯设置	2019-02-25 23:00:53	\n\t\t\t\t\t\t\n\t\t\t\t\t\tA市一中A5区新华都学校位于A市大道北辅道北...	0	0

图 18 二十聚类下的部分热点问题明细表

7.2.2 实验结果二

运用 k-means 模型处理附件 3 数据，将数据分成三十五类，得到了以下热点问题表和热点问题留言明细表：

表 7 三十五聚类下的热点问题表

	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	1	55.5	2019-1-2至2019-6-13	A市	A 市加快国家中心城市建设力度
1	2	2	53.0	2019-7-18至2019-9-1	伊景园滨苑	捆绑销售车位
2	3	3	41.0	2019-1-6至2019-12-30	A市	询 A 市办理户口迁出可以申请办理
3	4	4	40.0	2018-11-15至2019-12-2	A市	询 A 市人才购房购房补贴实施办法相关问题
4	5	5	38.5	2019-11-13至2020-1-25	丽法新城小区	搅拌站噪音扰民

	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	1.0	190213	A00031618	请A市加快自来水深度净化改造力度	2019-01-16 12:53:26	\n\t\t\t\t\t\n\t\t\t\t\t地处时代倾城小区的自来水烧过开水后电水壶就有...	0	0
1	1.0	191872	A00031618	请A市加快轨道交通建设力度	2019-03-01 15:19:28	\n\t\t\t\t\t\n\t\t\t\t\t地处中部中心城市的A市在高铁和地铁建设极为落...	2	9
2	1.0	193514	A00031618	请加快A市月亮岛片区公共服务力度	2019-03-20 16:39:22	\n\t\t\t\t\t\n\t\t\t\t\t地处月亮岛片区近年人口迅猛增长，成了全区人口...	0	4
3	1.0	196908	A00031618	请加快A市汉王陵考古遗址公园建设力度	2019-01-04 15:00:57	\n\t\t\t\t\t\n\t\t\t\t\t地处银杉路旁边的国家级汉王陵考古遗址公园已规...	0	9
4	1.0	200013	A00031618	请加快A市师大附中星城实验学校周边交通安全设施建设力度	2019-02-02 18:01:47	\n\t\t\t\t\t\n\t\t\t\t\t地处师大附中星城实验学校一小重点学校周边马路...	0	1

图 19 三十五聚类下的部分热点问题留言明细表

7.2.3 结果分析

上面两种聚类情况，一是二十个聚类，二是三十五个聚类，进行对比分析，我们认为，在二十个聚类下的热点问题表热度排名第一的问题描述为：A 市公交线路，这一问题的描述范围比较大，没有具体化，所以我们优先选择三十五个聚类的模型作为结果并导出其对应的 Excel 表格。

7.3 问题三

我们在对答复意见评判的过程中，从相关性、规范性、及时性这三个方面做出了评判标准，我们现在可以得到答复意见的具体分数和其四分位数值，当答复意见的具体分数小于 0 时，记为 0 分，当具体分数大于 100 时，记为 100 分。具体情况如下图所示：

	分数
0	77.166850
1	69.721340
2	91.045544
3	91.792140
4	82.722376
...	...
2811	0.000000
2812	0.000000
2813	47.479540
2814	0.000000
2815	0.000000

图 20 答复意见的质量的分值

	分数
count	2816.000000
mean	43.240082
std	37.686650
min	0.000000
25%	0.000000
50%	45.133474
75%	82.096021
max	99.171316

图 21 答复意见的质量的分数分布情况

参 考 文 献：

[1] 石凤贵.基于 TF-IDF 中文文本分类实现[J].现代计算机，2020,06:51-54++75
[2] 田琳.KNN 文本分类算法的研究[D]:[硕士学位论文].陕西省：西安理工大学，2019.
[3] 石莉,陈诚,邵艺.基于 BP 神经网络的大学生实践教学效果评价研究[J].扬州大学学报,2020,02:1-7
[4] 何伟.基于朴素贝叶斯的文本分类算法研究[D]:[硕士学位论文].江苏省：南京邮电大学，2018.
[5] 朱肖颖,赖绍辉,陆科达.基于 LSTM 算法在新闻分类中的应用[J].梧州学院学报，2018,06:10-20
[6] 闫捷.基于深度学习的唇语识别方法研究[D]:[硕士学位论文].北京市:北方工业大学，2019.
[7] 艾山·吾买尔,魏文琳,早克热·卡德尔. 基于 BiLSTM+Attention 的体育领域情感分析研究[J].新疆大学学报，2020,37(2):8
[8] 方炯焜，陈平华，廖文雄. 结合 GloVe 和 GRU 的文本分类模型[J].计算机工程与应用，2020.04：0-9
[9] 汪晶,邹学玉,喻维明,孙咏.分布式 MVC-Kmeans 算法设计与实现[J].长江大学学报(自然科学版)，2019,16(06):113-119
[10] 艾楚涵，姜迪，吴建德.基于主题模型和文本相似度计算的专利推荐研究[J].信息技术，2020,04：65-70