一种基于 WMD 和卷积神经网络的智慧政务模型

摘要

近年来“智慧”这一概念逐渐渗透到我们生活的方方面面。对于政府来说，了解民意、汇聚民智、凝聚民气不可或缺，但是由于互联网的普及，各式各样的网络问政平台（如微博、政府网站、市长信箱等）已经取代了原先的传统平台，而各类民生民意相关的文本数据也因此急剧攀升，给原先的传统人工体系带来了巨大的挑战。除此之外，智慧政务平台取代传统政务平台也是时代所趋，随着大数据、人工智能等技术的蓬勃发展，“智慧”政务平台能够高效、全面且准确地处理各类民生事件、政务问题，这是传统人工体系无法企及的高度。因此，本文通过比较不同的方法，加入适当的改进以适应具体情形，来构建基于自然语言处理技术的智慧政务模型。

首先我们对留言进行了预处理。清晰了解了训练集中数据的特征后，我们对训练集进行去噪、去重、去空等处理，以提高训练集质量，防止干扰模型训练。

针对**问题一群众留言分类**，我们使用朴素贝叶斯和卷积神经网络两种方法完成留言分类任务。使用朴素贝叶斯模型之前我们使用 TF-IDF 方法对留言的特征进行提取。使用卷积神经网络之前我们使用 One-hot 对文本进行编码。通过实验发现：通过逐步增加训练量，朴素贝叶斯模型的准确率比较稳定，而卷积神经网络的准确率与训练量呈正相关。因此我们设计了一个更深更复杂的改进版卷积神经网络模型。对各条留言进行了 conv1d 卷积操作后，经过 Max-pooling（最大池化），再将池化后的向量拼接成一体。随后，我们加入了 ReLU 激活函数，这个函数能够赋予模型复杂的非线性解题能力，并且能够防止反向传播过程中的梯度消失、梯度爆炸等问题。我们采用准确率、召回率和 F1 值对留言分类结果进行评价，结果显示这两种方法的训练效果类似，朴素贝叶斯模型和卷积神经网络模型的这三项指标分别为 81.48%、80.15%、79.59%和 81.18%、80.71%、80.82%。

针对**问题二热点问题挖掘**，我们首先利用正则表达式提取出地区，然后将留言按照地区进行预分类以提高聚类效果。接着我们使用 TF-IDF 提取特征值。我们采用两种线路挖掘热点问题。在线路一中我们先使用 PCA 对预分类的数据进行降维，然后再使用 DBSCN 算法进行聚类。在线路二中我们直接对预分类的数据进行 K-means 聚类。通过实验发现，K-means 更适合进行热点挖掘的任务。因为使用线路一对文本进行分类之后会使留言的分类非常分散，不利于挖掘热点问题，而线路二中有部分合适的事件适合进行热点挖掘。通过筛选，我们发现了适合进行热点挖掘的事件的特征，这些事件的留言条数基本分布在 5—25 条之间，通过这一特征我们将各个地区的事件纳入备选热点问题当中。接着，我们基于紧急度、反应度、关注度三个指标，利用线性回归计算得到了备选留言的热度。最终我们总结出了排名前 10 的热点问题。

针对**问题三答复意见的评价**。我们定义了相关性、及时性、完整性和可解释性四项评价指标。首先，我们基于 WMD 方法算出了留言与留言答复意见的文本相似度，得到了回复的相关性。然后基于留言时间、回复时间得到回复的及时性，接着定义了完整回复的模板，通过正则匹配方法得到回复的完整性，最后，根据回复的权威文件引用率得到回复的可解释性。基于上述四个指标，我们对每一条留言回复进行打分，并抽取了部分得分最高、最低的回复进行了分析。

关键词：朴素贝叶斯，卷积神经网络，K-means，WMD，回复品质综合评价体系， 正则表达式

目录

1 引言 4

1.1 主要工作 4
1.2 全文脉络 4

2 预处理 5

2.1 留言去重、去空 5
2.2 中文分词 5
2.3 去停用词 5

3 问题一：群众留言分类 6

3.1 问题分析 6
3.2 流程图 6
3.3 算法选择 6
3.3.1 特征提取 6
3.3.1.1 TF-IDF 算法 7
3.3.1.2 TextRank 算法 7
3.3.2 向量表示 8
3.3.3 分类模型 8
3.3.3.1 朴素贝叶斯 8
3.3.3.2 卷积神经网络 9
3.4 结果展示与评价 10
3.4.1 分类结果：混淆矩阵 10
3.4.2 分类结果评价：准确率，召回率 & F1 值 11

4 问题二：热点问题挖掘 13

问题分析 13
4.1.1 流程图 14
4.1.2 子问题——事件分类 14
4.1.3 子问题——热度计算 15
4.2 算法选择 15
4.2.1 降维算法 15
4.2.2 聚类算法 16
4.2.2.1 K-means 16
4.2.2.2 DBSCN 16
4.3 实现过程 17
4.3.1 预分类 17
4.3.2 线路一：PCA 降维+DBSCN 聚类 17
4.3.3 线路二：K-means 聚类 18
4.4 结果展示和分析 18
4.4.1 预分类 18
4.4.2 线路一：PCA 降维+DBSCN 聚类 19
4.4.3 线路二：K-means 聚类 20

5 问题三：答复意见的评价 21

5.1 问题分析 ——如何评价 21

5.2 算法介绍 21

5.2.1 相似度的计算 21

5.3 结果展示 22

5.4 结果评价 24

6 总结与展望 24

6.1 全文总结 24

6.2 展望 25

1 引言

1.1 主要工作

本文首先对留言数据进行去重、去空、分词、去停用词等预处理，然后将分别使用两种算法对留言进行分类。接着，我们根据本文对于热点事件的定义，将留言进行聚类之后再按照三个热度指标即紧急度、反映度、关注度对入选事件进行热度打分，最终整理得到排名前10的热点事件。最后，我们根据本文对留言回复的四个要求——及时性、相关性、完整性、可解释性，对留言的答复意见进行评价。

1.2 全文脉络

以下是本文的工作脉络图：

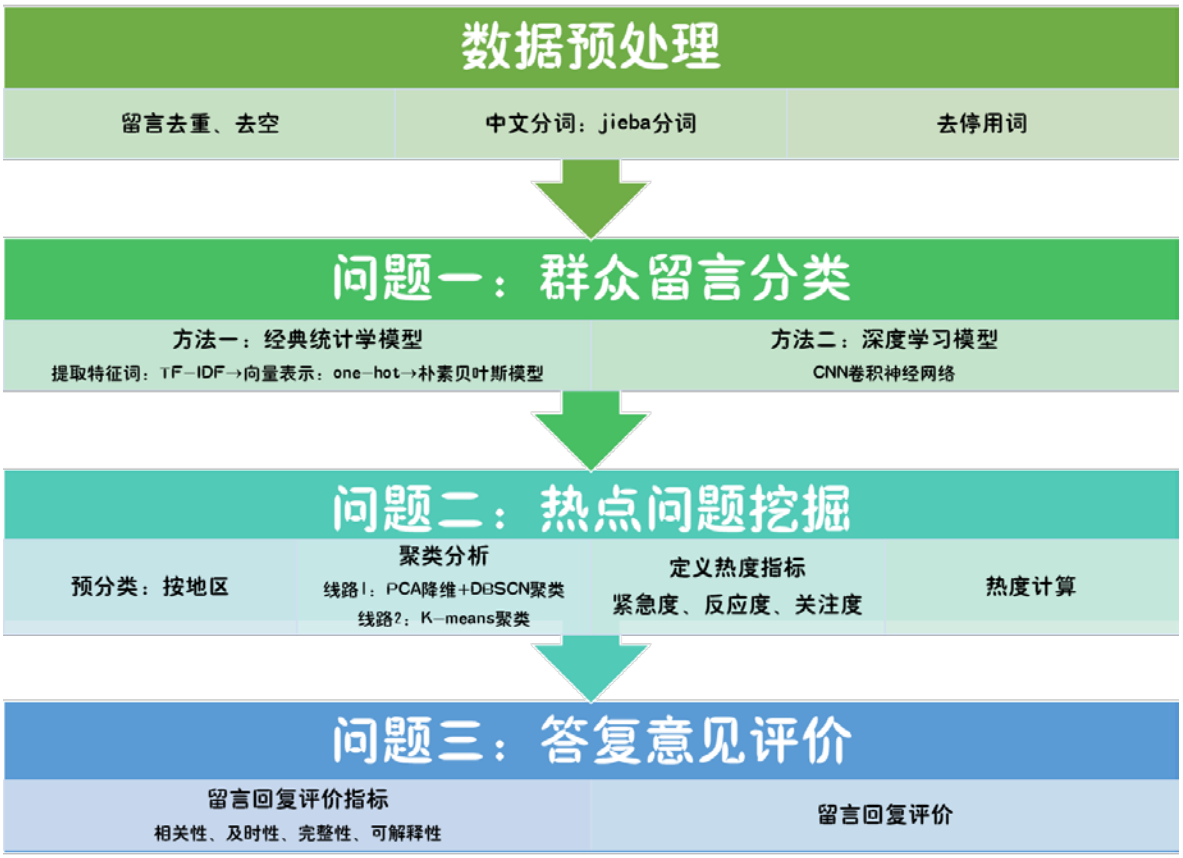

图 1 本文脉络图

2 预处理

2.1 留言去重、去空

去重分为两部分：留言之间去重，留言内部连续词去重。市民的留言中不乏有重复的留言出现，这可能是由于网络卡顿，或者个人情绪等主客观因素造成，题目所给的数据中也发现了一些重复的留言，这对后面文本的分类、聚类过程造成一定影响，降低了模型的准确率；而留言中重复的连续词本身并没有太大的意义，而且得到的文本还会影响之后 NLP 模型（特征提取等）的判断结果。故需要对文本进行去重、去空处理，以提高模型的准确率。

2.2 中文分词

汉语是以字为基本书写单位，单字很多时候表达不了语义，词往往能表达。因此进行中文自然语言处理通常是先将汉语文本中的字符串切分成合理的词语序列，然后再在此基础上进行其它分析处理。一条留言就相当于一个乐高模型，要想让计算机能够完美的“拼”出留言模型，我们就必须给计算机相应的“零件”，而分词的目的，便是制造这些“零件”。

这里我们采用 python 的中文分词库 jieba 进行分词，jieba 的分词原理大致如下：

- 基于前缀词典实现高效的词图扫描，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）。
- 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（即没有被收录在分词词表中但必须切分出来的词），采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

但是题目所给的数据中有大量的地名，并且这些地名大多是以“A 市”、“A6 区”、“A7 县”等形式表现，而 jieba 对这些词的处理不是很好，所以我们利用正则表达式，对 jieba 无法分词的词语进行提取，生成词典，再导入到 jieba 中进行分词，使分词结果更加符合常理。

2.3 去停用词

经过上述处理之后的文本中有很多无效的词，比如“并且”，“和”，“的”等，还有一些标点符号，这些功能词极其普遍，与其他词相比，功能词没有什么实际含义，所以我们并不想在文本分析的时候引入这些词，因此需要去掉，这些词就是停用词。将这些词去除之后可以减少索引量，增加检索效率，并且提高检索的效果。停用词主要包括英文字符、数字、数学字符、标点符号及使用频率特高的单汉字等。

3 问题一：群众留言分类

3.1 问题分析

为了正确归类网络问政平台的群众留言，我们首先要对文本进行分析，找出数据噪点，再进行相应的数据预处理。文本分类属于自然语言处理的一个主要分支，相关资料较为丰富且完善，主要的两种模型分别是统计学模型与深度学习模型，因此我们希望选择其中较为稳定高效的方法完成这项任务。

3.2 流程图

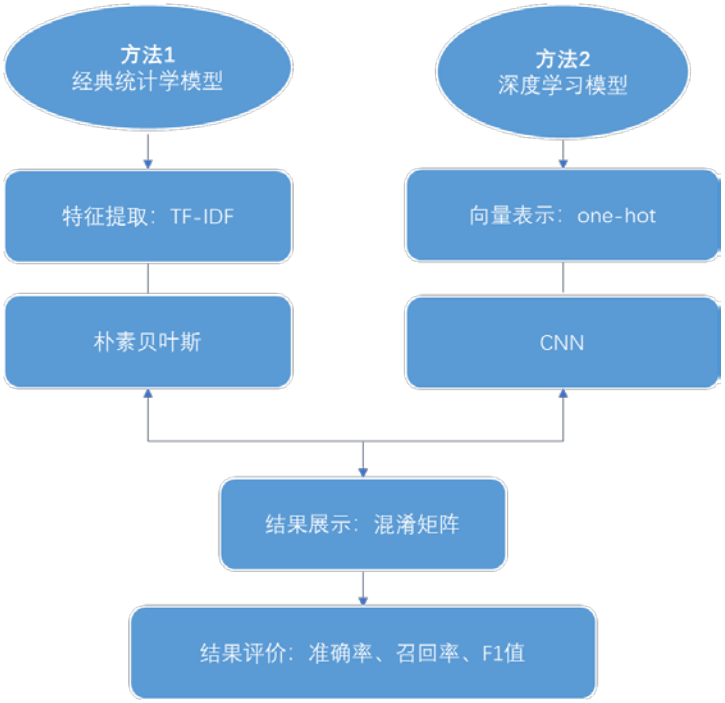

图 2 问题 1 流程图

3.3 算法选择

3.3.1特征提取

为了从众多词语中提取出那些对分类、聚类最有效的特征，从而实现特征空间维数的压缩，即提取一组“少而精”且分类错误概率小的分类待征，来尽可能的表示留言。常见的特征提取方法有：词频-逆向文档频率（TF-IDF）、TextRank 等。

3.3.1.1 TF-IDF 算法

TF-IDF 的主要思想是: 如果某个词或短语在一篇文章中出现的频率 (Term Frequency，TF）高，并且在其他文章中很少出现，即反文档频率（Inverse Document Frequency，IDF）低，则认为此词或者短语具有很好的类别区分能力，适合用来分类。那么该词可以看成该文章的关键性词语。具体计算公式如下:

$$TFIDF_{i,j} = TF_{i,j} \times IDF_i \tag{1}$$

其中$TFIDF_{i,j}$是指词i相对于文档j的重要性值。

$TF_{i,j}$指的是某一个给定的词语在指定文档中出现的次数占比。即给定的词语在该文档中出现的频率。这个数字是对词数（term count）的归一化，以防止它偏向长的文档。计算公式如下:

$$TF_{ij} = \frac{n_{i,j}}{\Sigma_k n_{k,j}} \tag{2}$$

其中

$n_{i,j}$是该词在文件d_j中出现的次数。

$\Sigma_k n_{k,j}$是在文件d_j中所有字词的出现次数之和。

IDF_i：指的是词i的逆向文档频率，是用总文档数目除以包含指定词语的文档的数目，再将得到的商取对数得到。这是一种度量词语重要性的指标。计算公式如下:

$$IDF_i = log\frac{|D|}{\left|\{j = t_i \in d_j\}\right|} \tag{3}$$

其中，

$|D|$为语料库中的文档总数，$\left|\{j = t_i \in d_j\}\right|$为包含词语$t_i$的文档数目。

3.3.1.2 TextRank 算法

TextRank 算法是一种用于文本的基于图的排序算法。其基本思想来源于谷歌的 PageRank 算法，通过把文本分割成若干组成单元（单词、句子）并建立图模型，利用投票机制对文本中的重要成分进行排序, 仅利用单篇文档本身的信息即可实现关键词提取、文摘。

TextRank 一般模型可以表示为一个有向有权图$G = (V, E)$，由点集合V和边集合E组成，E是$V \times V$的子集。图中任两点V_i，V_i之间边的权重为w_{ji}，对于一个给定的点V_i，$In(V_i)$为指向该点的点集合，$Out(V_i)$为点V_i指向的点集合。基本公式如下:

$$WS(V_i) = (1-d) + d * \sum_{v_j \in In(V_i)} \frac{w_{ij}}{\sum_{v_k \in Out(V_i)} w_{jk}} \mathrm{WS} \tag{4}$$

其中，d 为阻尼系数，取值范围为 0 到 1，代表从图中某一特定点指向其他任意点的概率，一般取值为 0.85。使用 TextRank 算法计算图中各点的得分时，需要给图中的点指定任意的初值，并递归计算直到收敛，即图中任意一点的误差率小于给定的极限值时就可以达到收敛，一般该极限值取 0.0001。

其中，

d表示阻尼系数，一般设置为 0．85（为经验值）；

V_i表示图中的任一节点；

$In(V_i)$表示指向顶点V_i所有顶白集合；

$Out(V_i)$表示由顶点V_i连接出去的所有顶点集合；

w_{ij}表示V_i和V_j的连接仅重；

$WS(V_i)$表示顶点的最终排序权重。

3.3.2向量表示

计算机无法直接处理文本信息，我们需要对文本进行处理，将文本表示成为计算机能够直接处理的形式，即文本向量表示。使用深度学习模型 CNN 之前我们采用 One-hot 来对文本进行编码。One-hot 是比较常用的文本特征提取的方法。One-hot 编码，又称“独热编码”。该方法用 N 位状态寄存器编码 N 种状态，每种状态都有独立的寄存器位，且这些寄存器位中只有一位有效，即只能有一种状态。

3.3.3分类模型

3.3.3.1　朴素贝叶斯

贝叶斯算法历史悠久，沉淀了丰厚的理论基础，很多高级自然语言处理模型都是以它为基础而发展起来的。因此，将朴素贝叶斯的思想运用在文本分类任务上非常必要。

$$P(\text{属于某类}|\text{具有某特征}) = \frac{P(\text{具有某特征}|\text{属于某类})P(\text{属于某类})}{P(\text{具有某特征})} \tag{5}$$

上述公式就是朴素贝叶斯算法的核心公式：

P(“属于某类”|“具有某特征”)：在已知某样本“具有某特征”的条件下，该样本“属于某类”的概率。叫做[后验概率]。

P(“具有某特征”|“属于某类”)：在已知某样本“属于某类”的条件下，该样本“具有某特征”的概率。

P(“属于某类”)：在未知某样本具有该“具有某特征”的条件下，该样本“属于某类”的概率。叫做[先验概率]。

P(“具有某特征”)：在未知某样本“属于某类”的条件下，该样本“具有某特征”的概率。

因此，我们的任务就是要判断 P(“属于某类”|“具有某特征”)是否大于 1/7(留言类型一共有 7 类)。朴素贝叶斯方法把计算“具有某特征的条件下属于某类”的概率转化成计算“属于某类的条件下具有某特征”的概率，很显然，后者的获取比较容易获得。我们只需要一些包含已知特征标签的样本，即可进行训练。而样本的类别标签都是明确的，所以朴素贝叶斯算法在机器学习里属于有监督学习方法。

3.3.3.2 卷积神经网络

CNN 在图像处理方面的成就有目共睹，在其基础上衍生的网络模型可以直接将图像数据作为输入，无需人工进行预处理等操作，真真实现了机器处理图像。但是 CNN 不仅仅是图像处理方面的宠儿，在自然语言处理领域，CNN 也逐渐崭露头角：情感分析、文本分类等。

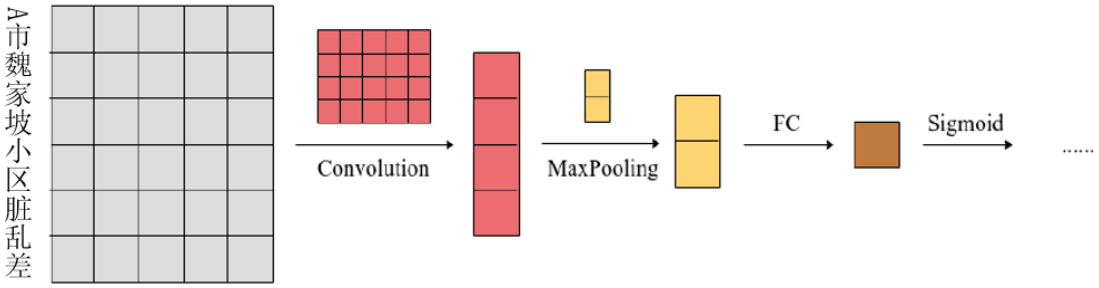

图 3 通用 CNN-NLP 模型示意图

对于自然语言处理，有一个通用的基于 CNN 的模型，模型主要由输入层、卷积层、池化层、全连接层以及激活函数组成。

输入层（Word Embedding Layer）：

模型的训练需要一段定长的文本序列，根据语料集的文本长度，我们需要综合确定一个平均文本长度 L，比 L 短的文本可以用 0 填充，比 L 长的文本需要截断，最终就可以形成完整的词向量矩阵输入模型，其中词向量有不同的表现形式：

1.static（静态词向量）

利用 Word2Vec、FastText、Glove 等词向量工具对公开语料库进行无监督训练，得到词向量后可以直接作为输入层，并且 CNN-NLP 模型在训练过程中不需要再调整词向量。

2.non-static（非静态词向量）

基于静态词向量，在 CNN-NLP 模型的训练过程中对词向量进行微调。

3.multiple channel（多通道）

模型 CNN 进行图像处理的 RGB 三通道思想，比如，可以用静态词向量和非静态词向量两种词向量搭建两个通道。

4.CNN-rand（随机初始化）

CNN-NLP 模型在训练过程中对不同单词的向量作随机初始化，在后续的有监督学习过程中，通过 BP 的方向更新输入层的各个词汇对应的词向量。

卷积层（Convolution Layer）：

在图像处理中，卷积核通常是对图像的一小块像素区域进行计算，而在自然语言处理中，卷积核通常是对文本所构成的词向量进行计算。一般卷积核的宽度与词向量的维度等宽，卷积核进行一维的滑动。

池化层（Pooling Layer）：

卷积层与池化层在分类模型的核心作用就是特征提取，即基于输入层，利用局部词序信息，提取初级特征，并组合初级特征为高级特征。相比较于传统机器学习，CNN-NLP 省去了中间的特征工程步骤。但 CNN-NLP 也有一个明显不足，文本信息在经过卷积层和池化层的操作后，丢失了词汇顺序、位置等信息，因而很难获取到文本中的否定、反义等语义信息。

全连接层(Fully Connected Layer)：

全连接层的作用相当于分类器，原始的 CNN-NLP 模型使用了只有一层隐藏层的全连接网络，相当于把卷积层、池化层提取的特征输入到一个 LR 分类器中进行分类。

激活函数（Activation Function）：

神经网络模型往往是为了逼近一个复杂的非线性函数，因此需要引入非线性激活函数帮助模型达到目标，否则，多层神经网络和单层也没有了区别，甚至无法实现简单函数。一般来说，常用的激活函数有 ReLU、Sigmoid 等。

3.4 结果展示与评价

附件二共有 9210 条留言，为了避免样本不均衡导致泛化能力低和过拟合等问题，我们选取了最少的留言数作为每一类训练的数据集。附件二中的留言涵盖城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生七个类别，分别用$x_1, x_2, x_3, x_4, x_5, x_6, x_7$表示。以下分别展示经典统计学模型和深度学习模型的混淆矩阵和分类结果评价。

3.4.1分类结果：混淆矩阵

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	130	0	0	0	3	0	8
x_2	1	159	2	6	2	0	2
x_3	3	2	133	5	11	14	0
x_4	1	2	0	139	1	0	0
x_5	7	1	4	8	113	2	8
x_6	5	2	3	6	17	119	1
x_7	37	1	1	10	33	3	67

表 1 方法一经典统计学模型分类结果（混淆矩阵）

	x_1	x_2	x_3	x_4	x_5	x_6	x_7
x_1	105	1	2	0	6	1	37
x_2	0	144	1	0	0	1	2
x_3	3	1	135	2	4	4	4
x_4	0	2	3	136	4	2	2
x_5	3	1	8	4	121	5	15
x_6	4	0	10	0	9	140	9
x_7	18	1	7	2	24	5	85

表 2 方法二深度学习模型分类结果（混淆矩阵）

由表 1 可得经典统计学模型对城乡建设类留言的分类结果准确率相对较高，由表 2 可得深度学习模型对环境保护类留言的分类结果的准确率较高。结合两个表来看，两类算法均容易将城乡建设类留言错分为卫生计生类，并且两类算法均容易将卫生计生类留言错分为城乡建设类和劳动和社会保障类。

3.4.2分类结果评价：准确率，召回率 & F1 值

我们采用以下三个指标对留言分类结果进行评价:

1. 准确率（Precision），查准率。即正确预测为正的占全部预测为正的比例。

$$P_i = \frac{TP_i}{TP_i + FP_i} \tag{6}$$

2. 召回率（Recall），查全率。即正确预测为正的占全部实际为正的比例。

$$R_i = \frac{TP_i}{TP_i + FN_i} \tag{7}$$

3. F1 值（H-mean 值）。F1 值为算数平均数除以几何平均数，且越大越好，将 Precision 和 Recall 的上述公式带入会发现，当 F1 值小时，TP 相对增加，即 Precision 和 Recall 都相对增加，即 F1 对 Precision 和 Recall 都进行了加权。

$$\frac{2}{F_i} = \frac{1}{P_i} + \frac{1}{R_i} \tag{8}$$

公式转化之后为:

$$F_1 = \sum_{i=1}^{7} \frac{2P_iR_i}{P_i + R_i} = \frac{2TP_i}{2TP_i + FP_i + FN_i} \tag{9}$$

以下是分类结果评价图：

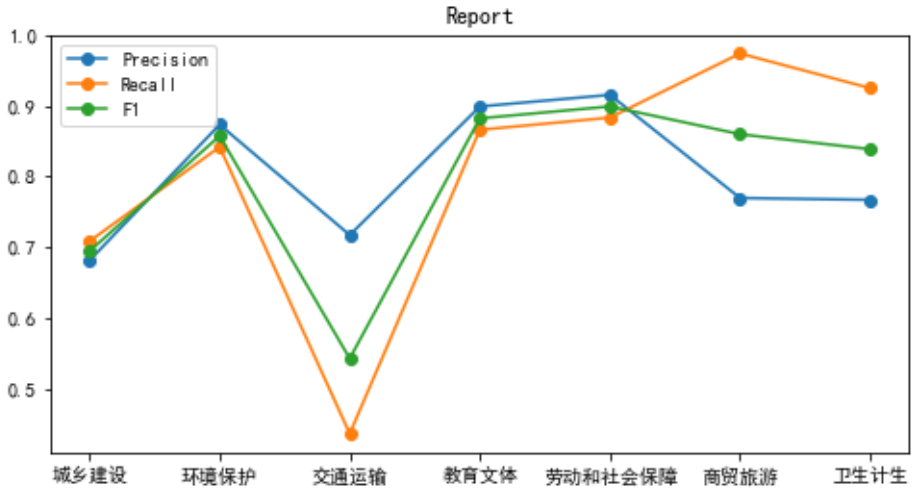

图 4 方法一经典统计学模型分类结果评价

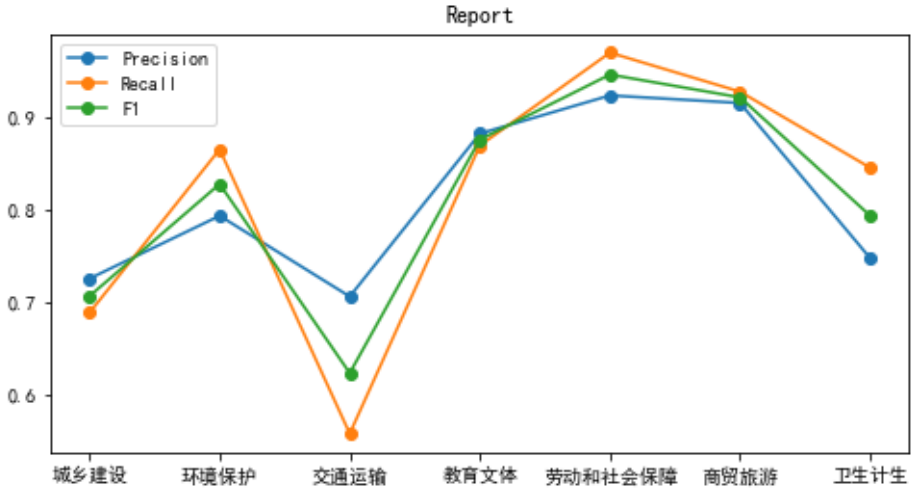

图 5 方法二深度学习模型分类结果评价

由图 4 可以看出，经典统计学模型对留言分类的准确率、召回率和 F1 在交通运输、商贸旅游和卫计卫生类中差异较明显。

由图 5 可以看出，深度学习模型对留言分类的准确率、召回率和 F1 在交通运输和卫计卫生类中差异较明显。

类别	准确率	召回率	F1	均值
经典统计学模型	81.48%	80.15%	79.59%	80.41%
深度学习模型	81.18%	80.71%	80.82%	80.91%

表 3 模型比较报告表

综合两个分类结果评价图来看，再结合表 3 来看，两种方法对该问题的处理效果均差不多，三个评价指标的均值分别为 80.41%和 80.91%均在 80%左右。因此两种方法对于该留言

分类问题的效果差不多。

不过，深度学习模型最明显的优势就是不需要预先提取文本的特征，它可以自动获取基础特征并组合为高级的特征，找到留言特征与目标分类之间的关系，省去了使用 TF-IDF 等构建特征的过程。但是这一模型的准确率也依赖于样本的数量。

从上述分析不难看出，两种方法的准确率有待提升，造成误差的原因可能为以下几个方面：

1) 部分数据集不适合我们选定的两类算法。拿深度学习模型来说，虽然应用深度学习的优势之一是不再需要构建特征，但是若不考虑语义，难免会产生误差。
2) 在算法的选择方面过于追求运行速度。迭代速度是决定算法项目成败的关键，而算法项目重要的不只是迭代速度，一定要关注迭代质量。在算法的质量上还有待优化。
3) 参数设置。由于时间的原因，也许是我们调试出最优的参数，使准确率下降。
4) 文本预处理欠佳。合适的文本预处理过程有注意提高算法的效果，在分词方法等的选方面仍有待提升。
5) 为了避免模型泛化能力受影响，防止过拟合，我们将所给数据集按照数量最少的类为标准进行平均切分，并将切分后的数据集以 7：3 的比例划分为训练集：测试集。

4 问题二：热点问题挖掘

4.1 问题分析

问题二要求我们根据附件 3 中的 4326 条群众留言对热点问题进行挖掘，在处理这一问题时，我们将该问题处理为两个子问题，一是对事件进行分类，二是计算事件的热度。在这一节中我们先在 1.1.1.中展示我们分析的流程图，然后在 1.1.2.和 1.1.3.中分别介绍两个子问题的分析过程。

4.2 流程图

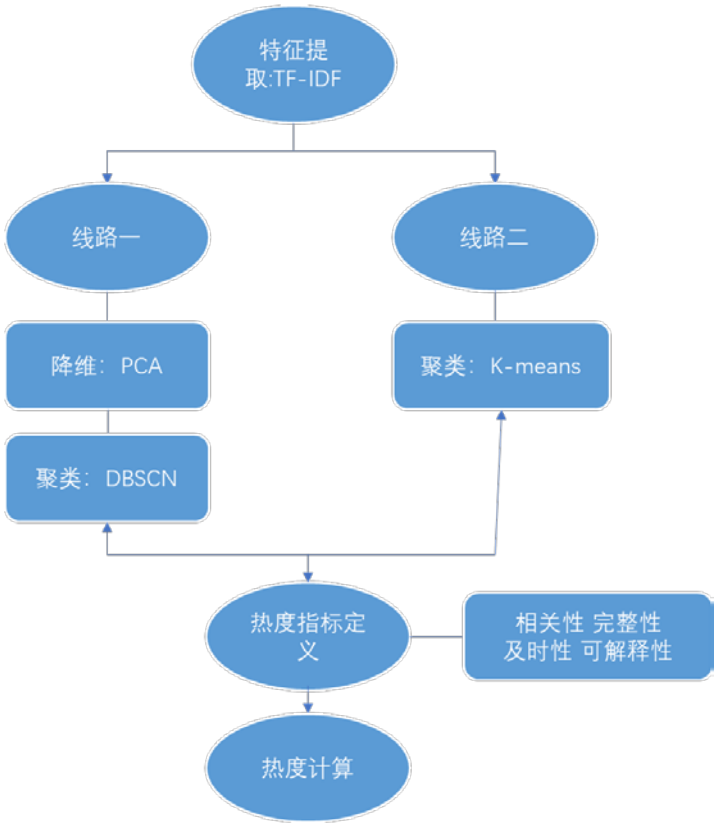

图 6 问题二流程图

4.3 子问题——事件分类

在对事件进行分类之前我们首先需要明确同一事件的定义是什么，即什么样的事件应该被归为同一个类型。我们的事件分类应该为政府的工作提供便利，因此，我们认为同一事件应该有以下特征：由同一地区的人反映、事件类型基本相同。这样分类之后，政府工作人员处理问题的时候就能够针对特定区域的特定问题进行集中处理。

处理子问题的第一步是对问题进行预分类。为了能够更加精准地对同一类型的事件进行聚类，按照前文提及的对同一事件的定义，我们首先需要把同一地区、同一类型的留言分到一起。预**分类**的详细过程将在 1.3 节介绍。

然后对留言进行聚类。聚类算法有两种，一种是规定聚类个数的聚类，另一类则没有规定聚类个数。选择不同的聚类算法意味着有不同的处理线路。通过我们的分析主要梳理出来了两条线路。

线路一适合无法事先确定聚类个数的聚类算法。通过查阅资料发现，大部分不事先确定聚类个数的聚类算法对于数据的维度有要求，因此需要先对留言数据进行降维，我们希望降维的过程中能够最大限度地保留信息，因此选择合适的降维算法显得尤为重要。降维算法和聚类算法将在 1.2 节（算法选择）介绍。

线路二省去了降维的步骤，直接对留言进行聚类。该方法的不足就是需要指定聚类的个

数，会增加工作量。

由于我们事先无法确定事件的个数，并且为了直观地得到事件的分类，我们偏好选择线路一。在 1.3 节（实现过程）中我们将进行两条线路并在 1.4（结果展示和分析）中对两条线路的效果进行分析。

4.4 子问题——热度计算

计算事件的热度之前首先应该确定热度指标。热点是指受广大群众关注、欢迎的新闻或者信息，或指某时期引人注目的地方或问题。本文主要从三个方面来研究事件的热度——紧急度、反映度、关注度。

- ✓ 紧急度：各个留言对该问题的反映时间的集中程度。如果一定时间内该事件的反映次数很多，那么这个事件的热度就越高。
- ✓ 反映度：该事件总共反映的次数。如果该事件反映的总数越多，那么该事件的热度就越高。
- ✓ 关注度：其他民众对于该留言的关注度。该关注度主要通过留言的点赞数来体现。如果该事件的点赞数越高，那么其他民众对于该事件的关注度也就越高，相应地，该事件的热度也就越高。

这三个指标我们赋予的权重分别为 3：5：2

热点是指受广大群众关注、欢迎的新闻或者信息，或指某时期引人注目的地方或问题。从热点的定义我们可以看出，一个话题若想成为的热点话题，首先要拥有广大的关注群众，这里体现在留言数量以及点赞、反对数量。其次，介于热点具有时效性，随着时间的流逝，人们对话题的关注度会明显下，故热点应该是在一定的时间段内引起人们广泛瞩目的事件。通过观察数据，我们发现，通过统计统一类别的留言的时间可以得出该类留言是否集中在一个时间段，而留言的数量以及留言的点赞、反对数可以反应该类的受关注程度，但是，留言数与点赞、反对数又有些不同，如果一个事件需要人们通过对政府留言来得到解决，那么他的重要性与紧急性可想而知，故我们定义紧急度、反映度、关注度三个指标，由于人们广泛关注是首要条件，事件的集中程度是次要条件，故我们选取 7:3 的权值来确定得分，前者又分为反映度和关注度，由于反映度更能体现事件的重要性，故我们对紧急度、反映度、关注度采取 3:5:2 的权值来对一类事件进行赋分计算热度。

4.5 算法选择

在 1.1.1 节我们提到了我们需要对留言的表示进行降维，然后在已经降维数据的基础上对留言进行聚类。

4.5.1降维算法

根据降维过程中评估准则的不同，降维方法主要分为线性和非线性 2 类。传统的线性降维方法，例如 PCA 法，根据原始矩阵奇异分解后的特征向量对原始矩阵做线性变换，抽取变换后矩阵中对矩阵方差保留度大的主成分特征组成低维特征集，从而将原始数据集从高维

空间投影到低维空间。线性降维的优势为处理速度快，但维度较低时信息丢失严重。非线性降维方法在保持数据原有分布的准则下使用非线性映射方式。T-SNE 算法由图灵奖得主 Geoffrey Hinton 提出。该方法和 PCA 方法的区别在于能够更好地可视化。用 t-SNE 能够很清晰的看到数据全景（感觉说这个没啥用）。该方法只能用来做投影，将高维度向量进行映射，但是不反映聚类关系，也不能用于预测。而本文使用降维算法的最终目的是用来聚类，因此不能选择该算法，我们最终决定用 PCA 算法来进行降维。

4.5.2聚类算法

文本聚类的方法很多，主要分为基于层次的方法、基于划分的方法、基于密度的方法、基于模型的方法、基于网格的方法。

4.5.2.1 K-means

K-means 是最常用的基于划分的方法。K-means 算法是利用质心与质心之间的距离迭代，选择最优质心距离来进行聚类划分的算法，当类间质心点的距离相互收敛时，算法质心的距离迭代结束，聚类计算分辨出具有共同特性的类别。优点是算法实现简单快捷，数据大小可伸缩，计算数据类高效。缺点是算法对初始距离质心点依赖性大，初始质心点的偏差可能导致产生的结果偏离实际的分类，算法在运行时需要不断地更改类质心距离，数据计算量比较大，算法开销的时间也比较大，聚类的种类是预先设定的，很难估计。

- 最佳聚类数的确定：手肘法

该方法的核心思想是：随聚类数 k 的增大、样本划分会更加精细、每个簇的聚合程度会逐渐提高，误差平方和 SSE 逐渐变小 。并且，当 k 小于最佳聚类数时，由于 k 的增大会大幅增加每个簇的聚合程度，SSE 的下降幅度会很大，而且，当 k 到达最佳聚类数时，再增加 k 所得到的聚合成都回报会迅速变小，所以 SSE 的下降幅度会骤减，然后随着 k 值的继续增大而趋于平缓，此时 SSE 和 k 的关系图是一个手肘的形状，肘部对应的 k 值就是数据的最佳聚类数。

4.5.2.2 DBSCN

DBSCAN 聚类算法是一种经典的基于密度的聚类算法[14]，其将一个高密度的区域分成多个簇，簇即是密度相连的点的最大集合，以该集合内的单位数目为标准，根据事先设置的阈值，将区域内的点划分为噪声点和核心点，同时将核心点区域内的单位划分为边界点从而实现聚类。

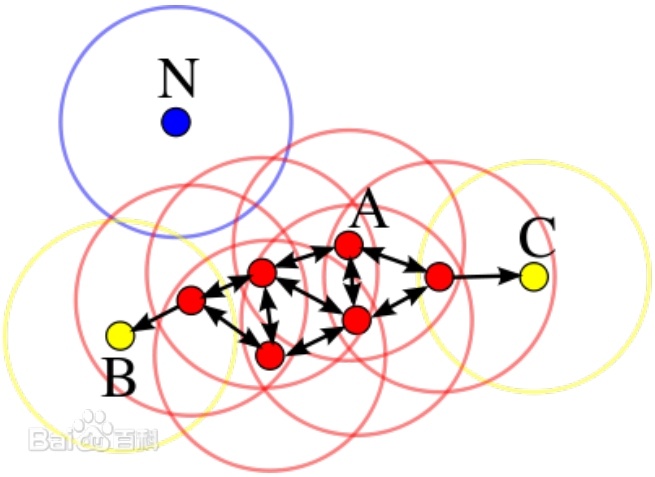

图 7 DBSCN 原理图

4.6 实现过程

4.6.1预分类

1. 用正则表达式提取留言中的地名
2. 依据地名对留言进行分类

4.6.2线路一：PCA 降维+DBSCN 聚类

1. 将按照地区分类好的留言进行使用 jieba 库进行中文分词
2. 将分词后的数据进行去除停用词处理
3. 对相同地区的留言进行特征提取(用的什么方法？选取的前多少个特征向量？)
4. 用 word2vec 进行词嵌入
5. 使用 PCA 对向量进行降维

PCA 的基本步骤如下:

① 对原始指标进行标谁化，以消除变量在数量极或量纲上的影响（

② 根据标准化后的数据矩阵求出相关系数矩阵 R

③ 求出 R 矩阵的特征根和特征向量

④ 取前两个主成分进行合成，得到降维后的留言

6. DBDCAN 聚类

4.6.3线路二：K-means 聚类

K-mean 聚类的算法步骤如下：（假设要把样本集分为 k 个类别）

（1）适当选择 k 个类的初始中心；

（2）在第 k 次迭代中，对任意一个样本，求其到 k 个中心的距离，将该样本归到距离最短的中心所在的类；

（3）利用均值等方法更新该类的中心值；

（4）对于所有的 k 个聚类中心，如果利用(2)(3)的迭代法更新后，值保持不变，则迭代结束，否则继续迭代。

4.7 结果展示和分析

4.7.1预分类

下表是用正则表达式提取的各个区域对应的留言数的统计结果

地名	A3区	A6区	A7县	D1区	K3县	A4区	G1区	A2区	A1区	C5市
数量	398	188	409	1	7	249	1	277	268	37
地名	E4区	A8县	A9市	A5区	L6县	M9县	L5县	M1区	K4县	E8县
数量	3	97	82	235	13	15	4	2	6	9
地名	K9县	J3县	E5区	D8县	M3县	I5县	C3县	G7县	B4区	D7县
数量	12	1	2	2	4	2	3	1	5	2
地名	M5市	J5县	B7县	G2区	M4市	E5县	E7县	K1区	F6县	L7县
数量	4	4	1	4	1	3	2	1	6	1
地名	G8县	F7县	I3县	C3区	C2区	E3区	M2县	H3县	C4市	I4县
数量	3	3	3	1	3	1	1	1	2	1
地名	F5县	M7市	G3县	J4县	E6县	K2区	F8市	K6县	A市	G市
数量	2	2	1	2	1	3	1	1	1985	8
地名	C市	B市	E区	F市	A区	K市	M市	D市	B区	L市
数量	28	21	4	17	5	9	11	10	4	12
地名	F区	I市	R区	D区	E市	M区	H市	J市	K区	C区
数量	2	15	2	4	8	1	4	2	1	2

表 4 预分类结果

由统计结果我们可以看出：

该统计表存在对同一地区的某条留言进行重复统计的情况。这是由于我们并不清楚每个区域对应的从属关系，加之赛题数据已被处理过，所以我们无法保证以下区域对应的留言条数之和与总留言数相等。在实际生活中，这种情况不会发生，因为某个区或县所属的市区是唯一的。

由表 4 可以看出 A3 区、A6 区、A7 县、A4 区、A2 区、A1 区的留言数比较多，可知留言的群众主要分布在这些区域，同时可以看到，A 市的留言条数几乎是其他区域的总和，由此可以大致推断出其他区域大部分是 A 市所管理的区域。

4.7.2线路一：PCA 降维+DBSCN 聚类

采用 PCA 降维进行降维之后，进行聚类工作。

接下来随机选取 A5 区进行分析。A5 区共被分为了 39 个类，其中有一个类是异常类。每个类对应的留言数如下表所示。

类别	-1	0	1	2	3	4	5	6	7	8
留言数	63	76	2	2	13	2	2	2	3	3
类别	9	10	11	12	13	14	15	16	17	18
留言数	2	2	4	2	2	2	2	3	2	2
类别	19	20	21	22	23	24	25	26	27	28
留言数	2	4	2	2	2	2	2	2	2	2
类别	29	30	31	32	33	34	35	36	37	
留言数	2	4	2	2	2	2	2	4	2	

表 5 线路一分类

由表可知，使用该线路得到的聚类结果比较分散，有？%的类当中只包含两三条留言。部分结果如下所示：

留言主题	类别
A5区时代阳光大道中建嘉和城附近防护绿地被侵占	1
A市山水庭院小区物业是否偷税漏税?	1
A5区万家丽中路泰禹家园住改商改成麻将馆，深夜麻将扰民无人管	2
A5区地质中学附近拥堵问题很严重	2
A5区明昇壹城小区业主未取得养犬许可证，违规养狗	3
A市万科金域华府三期出行难，断头路何时能通	3
反映A1区星城世家全体业主急盼关注解决六号路交通安全隐患问题	3
A5区井湾子机械化小区会拆迁或者改造吗?	3
A5区万科金域华府三期的业主出行不便	3
能不能在A5区环保科技园规划一个公园	3
反映A5区洞井接到鸿运佳苑出行问题	3
A5区赏识培训学校用刚装修的教室给孩子上课	3
咨询A市高升路往东向延伸等相关问题	3
A5区黎托村60号令老安置小区和粟塘小区小孩读书不方便	3
A3区江山帝景哈佛四期有安全隐患	3
咨询A市高升路往东向延伸等相关问题	3
A5区合丰安置小区现在完全变成了物流园区!	3
A市A5区树木岭世贸璀璨天城通宵施工噪音扰民	4
A市A5区梓园路紫金苑当街门面餐饮废气扰民!	4
A市云开一品开发商欺诈业主，产权缩水30年	5
A5区泰禹云开一品小区产权缩水30年	5

表 6 线路一结果

类 0 包含结果较多，留言内容比较杂乱，没有在表中进行展示。由表？完整展示了类 1-5 的留言主题情况，留言数有 13 条的第 3 类，涵盖了多类问题，而剩余的几个类别中，分类基本准确。另外，通过分析所有结果我们发现，该线路容易将同一类型的留言分到不同的类型中，

通过以上分析我们可以看出，该线路的分类结果中，留言较多的类别分类效果不佳，尽管分类较少的留言中表现不错，但是容易让留言的类别过于分散。

4.7.3线路二：K-means 聚类

为了方便分析，仍然选取 A5 区进行分析。

类别	0	1	2	3	4	5	6
留言数	49	86	14	31	9	29	17

表 7 线路二分类

由表 7 可知，直接用 K-means 对 A5 区进行聚类得到 7 个类别，聚类数明显小于通过线路一得到的聚类个数，留言个数相比线路一而言更加集中。

留言主题	留言
A5区学而思培优华盛花园培训点噪音扰民，影响居民生活	2
长A5区梓园路紫金苑西面临街门面油烟污染严重	2
A5区劳动东路魅力之城小区油烟扰民	2
A5区劳动东路魅力之城小区临街门面烧烤夜宵摊	2
A5区白田佳苑居民饱受高速噪音折磨	2
A5区劳动东路魅力之城小区底层餐馆油烟扰民	2
A5区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气，急需处理!	2
A5区粟塘小区曲塘路F市烧烤粥吧晚上店外经营，噪音扰民	2
A5区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气	2
A5区劳动东路魅力之城小区油烟扰民	2
A5区劳动东路魅力之城小区底层餐馆油烟扰民	2
A5区劳动东路魅力之城小区临街门面烧烤夜宵摊	2
A5区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气	2
A5区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气，急需处理!	2
西地省聚利人普惠投资有限公司涉嫌诈骗巨额资金	4
A市A5区利聚网平台不让提现	4
对西地省聚利网的联名控诉书	4
A市聚利网诈骗，派出所立案已超3个月依然毫无进展	4
西地省聚利人普惠投资有限公司涉嫌诈骗巨额资金	4
A市A5区利聚网平台不让提现	4
西地省利聚人普惠投资有限公司涉嫌诈骗巨额资金	4
西地省惠普利聚人投资有限公司涉嫌诈骗巨额资金	4
请A市A5区公安分局和候家塘派出所认真对待聚利网诈骗一案	4

表 8 线路二结果

留言数较多的类别不在这里进行展示，这些类别中的事件比较多样。我们抽取了留言数在 10 左右的类别进行分析，由上表可知，类 2 中的留言大部分反映劳动东路魅力之城小区烧烤摊油烟扰民问题，夹杂的不属于该区域的留言也与油烟扰民问题相关。类 4 中的留言均反映利聚人平台网络诈骗问题。

由以上分析我们可以看出，使用 K-means 算法进行直接聚类仍然会出现很多类型的事件聚集在一起的情况，但是比较好的是，该算法能够将集中发生的事件聚集到一起。聚集起来的这些事件符合本文对热点事件的定义，因此我们认为线路二更适合采集热点问题，及时发现热点问题。线路一中将 PCA 算法和 DSCAN 算法结合的思路则更适合在每条留言的处理上进行使用。（感觉我表达不清楚）比如在处理某一条留言的时候，可以使用线路一，找到与该条留言相近的个别问题，这在一定程度上也能提高政府的工作效率。

5 问题三：答复意见的评价

5.1 问题分析 ——如何评价

问题三要求我们根据附件 4 的?条留言对相关部门的答复意见进行评价。我们主要围绕及时性、相关性、完整性和可解释性对留言进行评价。

✓ 及时性

由相关部门的回复时间与群众留言时间的时间差计算得出。通过观察附件 4 可知相关部门回复的时间距离群众留言的时间并未集中分布在 24 小时以内，并且基本都在一天以上，因此我们采用天作为时间间隔的单位。时间间隔越短，得分越高。

✓ 相关性

通过计算相关部门留言与该留言的相似度得出。相似度越高，得分越高。

✓ 完整性

主要衡量留言回复的格式的规范性和完整性。我们主要度量了五个方面：①开头是否进行称呼；②是否使用了礼貌用语；③是否用合适的语句引出相关内容；④是否用合适的语句结束回复；⑤结尾是否写明留言时间。由于格式的规范性应该要求比较高，每个部分缺一不可，因此只要有一项不符合都不能得分。我们使用的正则表达式如下所示：

```
r'(.*)(尊敬的|你好|您好)(.*)((答复|回复)如下)(.*)(感谢)(.*)[0-9]+(年)[0-9]+(月)[0-9]+(日)'
```

✓ 可解释性

主要考查回复中是否引用了相关法律条文以及相关政策文件等具有权威依据的文件。如果涉及到相关内容，在我们的评价体系中讲得到更高的分数。

以下是我们计算的公式：

$$res_i = \frac{相关性_i + 及时性_i + 可解释性_i + 完整性_i}{4} \tag{10}$$

$$SCORE_i = \frac{res_i}{\sum res_i} \times 100 \tag{11}$$

5.2 算法介绍

5.2.1相似度的计算

计算相似度的算法有很多

基于词向量：余弦相似度、曼哈顿距离、欧几里得距离、明式距离（是前两种距离测度的推广），在极限情况下的距离是切比雪夫距离。

基于字符：编辑距离、simhash、共有字符数（有点类似 onehot 编码，直接统计两个文本的共有字符数）。

基于概率统计：杰卡德相似系数。

基于词嵌入模型：word2vec/doc2vec、WMD。这里不一一列举。词嵌入（word embeddings）已经在自然语言处理领域广泛使用，它可以让我们轻易地计算两个词语之间的语义相似性，或者找出与目标词语最相似的词语。

word2vec 其基本思想是通过求句子中所有单词词嵌入的平均值，然后计算两句子词嵌入之间的余弦相似性来估计两句子间语义相似度

本文主要介绍词嵌入模型 WMD。它具有以下几个优点：

效果出色：充分利用了 word2vec 的领域迁移能力

无监督：不依赖标注数据，没有冷启动问题

模型简单：仅需要词向量的结果作为输入，没有任何超参数

可解释性：将问题转化成线性规划，有全局最优解

灵活性：可以人为干预词的重要性

当然它也有一些缺点：

处理否定词能力偏差

处理领域同义词互斥词的能力偏差

时间复杂度较高：O(p3logp)O(p3log　p)（其中，p 代表两篇文本分词去重后词表的大小）

在利用 WMD 计算两条文本的相似度时，会进行以下步骤：

1、利用 word2vec 将词编码成词向量

2、计算出每个词在文本中所占权重，一般用词频来表示

3、对于每个词，找到另一条文本中的词，确定移动多少到这个词上。如果两个词语义比较相近，可以全部移动或移动多一些。如果语义差异较大，可以少移动或者不移动。用词向量距离与移动的多少相乘就是两个词的转移代价

4、保证全局的转移代价加和是最小的

5、文本 1 的词需要全部移出，文本 2 的词需要全部移入

5.3 结果展示

优质评价：

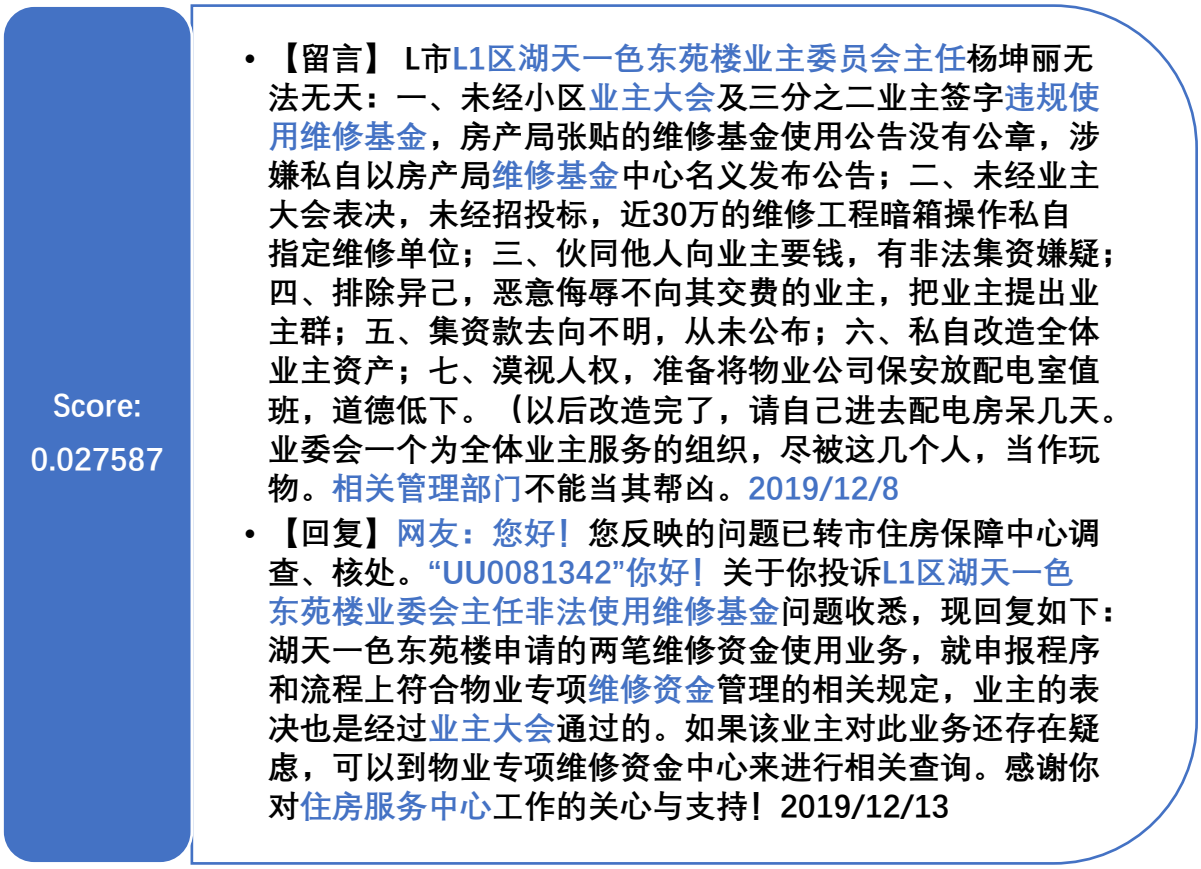

图 8 优质评价

1. 相关性：留言详情中提到的关键点在留言详情中几乎完全出现，而且匹配度较高。
2. 及时性:留言时间与回复时间相差只有 5 天，回复很及时。
3. 完整性：回复详情开头称谓——“网友，您好！”，礼貌又准确，且文末备注了回复时间，回复格式较好。
4. 可解释性：引用了“相关规定”等具体政策规范，具有一定说服力。

劣质评价：

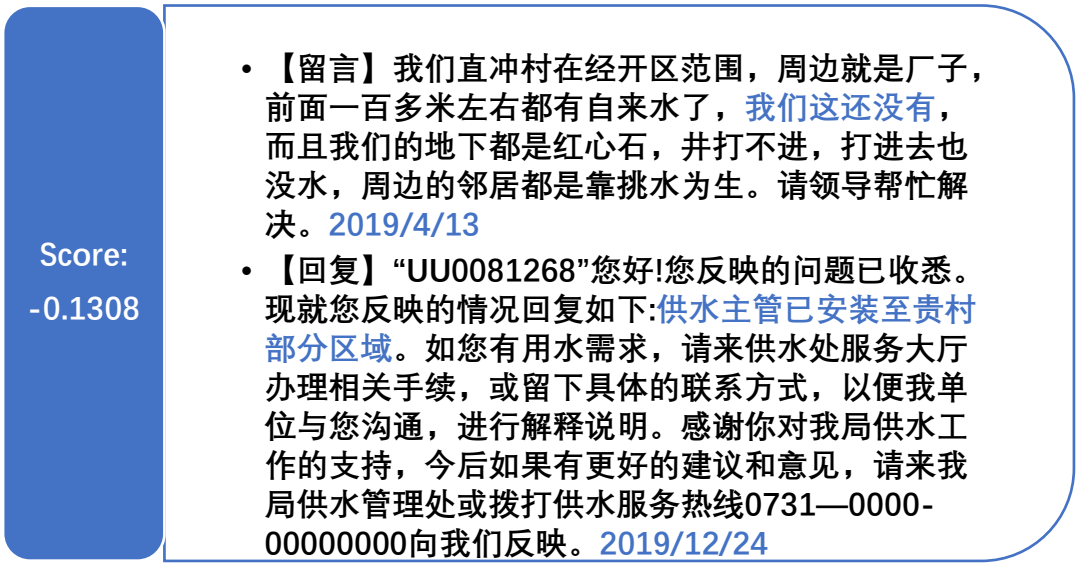

图 9 劣质评价

1. 相关性：留言详情中提到的关键点在留言详情中很少出现，而且匹配度不高。

2. 及时性：留言时间与回复时间相差近大半年，回复很不及时。
3. 完整性：回复缺少时间，格式随意。
4. 可解释性：未引用相关法律条文或政策规范，缺乏说服力。

5.4 结果评价

我们对于答复意见的评价是基于四个指标的，分别是留言与回复的相似度，及时性，可解释性，完整性，最终得出一个回复品质得分。

根据最终的得分，我们发现优质评价与劣质评价的主要影响因素是留言与回复的相似度、及时性两个指标，具体表现为：留言详情的关键词与回复详情的关键词匹配度越高，则回复品质得分越高；留言时间与回复时间的时间间隔越大，则回复品质得分越低。

我们的回复品质得分体系能够基本从留言与回复的相似度、及时性、可解释性和完整性四个方面，判断每一条回复的质量，再配合人工进行少量删选，能够有效提高准确率。

6 总结与展望

6.1 全文总结

通过一系列的操作，我们通过两种方法将群众留言由原始的无标签状态进行了初步的分类。然后我们把留言按照所属地区进行初步分类，接着通过聚类算法，将留言按照本文定义归纳成类，并按照本文对热点事件的定义对事件的热度进行评分。另外，我们还按照相应的标准对留言的回复进行评价。

总的来说，我们的成果如下：

1) 分别使用经典统计模型和深度学习模型对留言进行分类，分类效果类似，准确率分别为 81.48%、81.18%，但是我们发现前者的准确率比较稳定，而后者与数据集的大小呈正相关关系，我们认为政府部门实际会获得更大的数据量，因此我们的深度学习模型能够给政府提供有效的帮助。
2) 通过初步将留言按照所属地区进行分类后，创新地引入类间相似度概念，对每个类之间的留言的相似度进行计算，然后筛选出相似度符合要求的类别，进入最终热点类的评比，最终对符合要求的问题进行热度打分。该方法能够帮助政府快速发现热点问题，对急需处理的事件进行针对性处理。
3) 围绕四个角度，即相关性、及时性、可解释性、完整性，对留言回复进行了客观地评价，并且评分能够较好地客观反映留言的回复质量。该成果能够初步为政府的工作评价提供依据。

当然我们也还有应该改进的地方：

1) 在留言分类的准确率方面我们还有一定的提升空间。未来我们将用更优化的方式来对文本进行编码。
2) 在对留言的事件类型进行归类之后仍会出现不属于该事件的留言，当然事件的划分

具有一定的主观性，指定严格的标准将留言准确地划分具有一定的难度。

6.2 展望

文本挖掘是从文本数据中获取新发现的过程，也是一个结构化的过程。它是多方技术的一个综合。现阶段文本挖掘技术尚不够成熟，我们的算法也有一些缺陷，我们还有很长的一段路要走。

在接下来的工作中，我们将围绕以下方面进行：寻找更适合该问题的词嵌入模型、寻找更合适的算法以实现性能上的提升、寻找能够实现自动确定类数并且聚类效果良好的聚类算法。另外，为了让政府工作能够向更加智能的方向迈进，我们将把改进后的算法打包，制作成全自动的留言处理系统，该系统能快速分析出热点事件、热点事件类型，还能对问题处理情况进行实时反馈，不仅有利于政府部门对问题进行高效地处理，还能利于群众对政府部门的工作进行监督。

文献引用

[1] 邵峰晶，于忠清．数据挖掘原理与算法 [M]. 北京：中 国水利水电出版社 ,2003
[2] 张跃，李葆青，胡玲,等．基于 K-Mean 文本聚类的研究[J]. 中国教育技术装备,2014(18):50-52
[3] 张义军，刘泉凤。DKTC：一种中文文本聚类方法[J]。图书情报工作，2009（01）：35＋111-114。
[4] 傅德胜，周辰．基于密度的改进 K 均值算法及实现[J]. 计算机应用, 2011(02):146-148.
[5] 袁方，苑俊英．基于类别核心词的朴素贝叶斯中文文本分类[J]. 山东大学学报（理学版), 2006, 41(3):111-114.
[6] 张绮琦，张树群，雷兆宜．基于改进的卷积神经网络的中文情感分类[J]. 计算机工程与应用, 2017, 053(022):111-115.
[7] 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].
[8] 吴萍萍．基于信息熵加权的 Word2vec 中文文本分类研究[J]. 长春师范大学学报，2020(2):28-33.
[9] 黄昌，殷杰，侯芳。关键词：词义信息，TF-IDF，文本相似度测量 计算机学报，2011，34（5）：856-864。
[10] 苏金树，张博锋，徐昕．基于机器学习的文本分类技术研究进展[J]. 软件学报，2006(09):16-27.
[11] 田增山,王向勇,周牧,等．基于 DBSCAN 子空间匹配的蜂窝网室内指纹定位算法[J]. 电子与信息学报, 2017,39(5): 1157-1163.
[12] https://blog.csdn.net/u013063099/article/details/80964865
[13] https://www.jianshu.com/p/883df8e81450
[14] https://blog.csdn.net/Ding_xiaofei/article/details/81034058