“智慧政务”中的文本挖掘应用

摘要

本文针对“智慧政务”中的文本挖掘问题，运用*Jieba*分词、Word2*V*ec词向量训练、一维卷积神经网络、独热编码、*K – means*聚类等方法，构建了“关于留言内容的一级标签分类模型”，找到了群众关心的排名前五的热点问题，并且就有关部门对留言的回答意见进行了综合性的评价。运用*python*软件建模求解。

本文的特色是：卷积神经网络进行词分类、定义了用来评价的热度得分等，具体体现在方法、研究思路、建模过程中。

针对问题一，要求建立关于留言内容的一级标签分类模型。首先，本文运用*Jieba*分词，将句子分割成词语，再找出有研究价值的词语进行Word2*V*ec词向量训练，将每一个单词训练成150维的向量；然后将每一个句子表示成一个500*150的矩阵，运用一维卷积神经网络分类，输出一个7*1的矩阵；最后*Softmax*函数将文本分为了七类，并且得到了*F – S*core得分。

针对问题二，要求将留言进行归类后合理定义热度评价指标。首先运用*TF – IDF*提取文章的高频词，然后用独热编码将留言向量化；再用*K – means*聚类算法对文本进行分类，在这些分类里识别出留言里特定的地点或人群；最后综合留言数、点赞数对每一条留言本文都给出一个热度得分，得到热度评价指标。

针对问题三，要求对有关部门的答复意见给出一套评价方案。我们从相关性、及时性、完整性、可解释性的角度设置四个指标，来发现问题解决情况和相关部门的服务水平。

最后本文还提出了对模型的进一步思考和改进，对模型的优缺点进行了客观的评价。

关键词： *Jieba*分词　一维卷积神经网络　*TF – IDF*算法　热度得分

目录

1. 符号说明

表 1 符号说明表

w_i	高频词排序	N_w	词条出现的次数
A_i	第 i 条句子	num_i	留言数量指标
d_i	i 类问题的留言总量	$influence_i$	用户活跃度指标
$corn_i$	关注度指标	$Heat_i$	热度评价指标
$correlation$	相关性指标	$Itergrity$	完整性判断

2. 问题一模型的建立与求解

2.1. 问题一的分析

问题要求处理网络问政平台的群众留言，建立留言内容的一级标签分类模型，并用 $F-S$core 对分类方法进行评价。要想对文本进行分类，就要将文本的主题提取出来，现有的机器学习无法对文本直接识别，我们需要将文本先进行向量化。在向量化之前还需要对文本进行分词、去除无关词等操作。由于留言中出现的词语很多，若直接使用会造成维数灾难，对计算机运行要求高，而且运行速度慢。所以我们采用词向量嵌入的方式，并且得到每一条留言对应的向量。最后，用一维卷积神经网络将所有留言分成七类，建立了“关于留言内容的一级标签分类模型”。最后再运用 $F-S$core 对分类方法进行评估。

分析流程图如下：

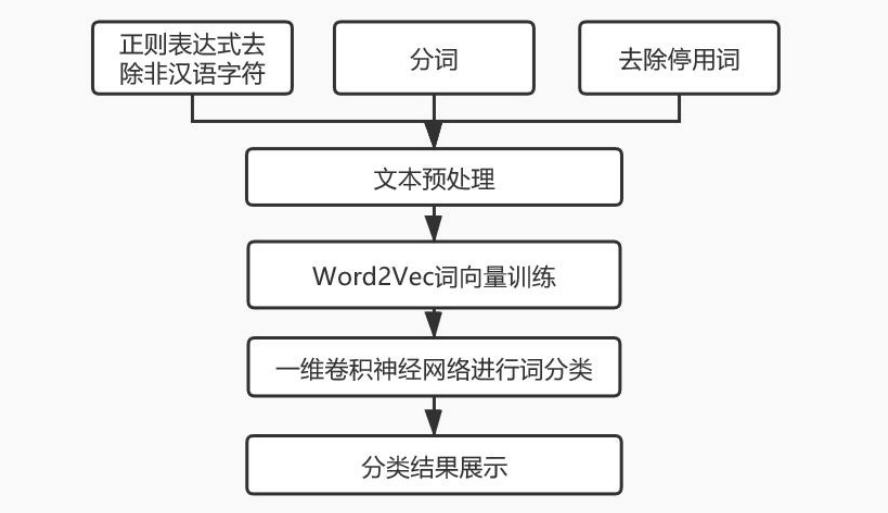

图 1 问题一分析流程图

2.2. 问题一模型的建立

2.2.1. 文本预处理

Step1：*Jieba* 分词

英语文本中，每个单词都由空格分开，非常方便进行词语的划分[1]以及进一步的文本处理问题。在汉语的文本处理问题中，我们首先需要进行汉语的分词，目前 *python* 中汉语分词工具很多，比如盘古分词、*Yaha* 分词、*Jieba* 分词、清华 *THULAC* 等。我们选择使用功能强大的 *Jieba* 分词。

Jieba 分词涉及到的算法包括：

(1) 基于 *Trie* 树结构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* ）；

(2) 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；

(3) 对于未登录词，采用了基于汉字成词能力的 *HMM* 模型，使用了 *Viterbi* 算法。

Jieba 中文分词目前支持三种格式：

1.精确模式：将句子精确地分开，适合文本分析。

2.全模式：把句子中所有的词扫描出来，但是不能解决歧义问题。

3.搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

结合我们题目的特点，我们选取精准模式并且采用 *HMM* 新词发现作为分词方法。

下面给出一个例子来查看分词结果：

- **原句：**

  地铁 5 号线施工导致 A 市锦楚国际星城小区三期一个月停电 10 来次

- **全模式：**

  地铁/5/号/线/施工/导致/A/市/锦/楚国/国际/星/城/小区/三期/一个/一个月 /停电/10/来/次

- **精确模式（使用 *HMM* ）：**

  地铁/5/号线/施工/导致/A/市锦楚/国际/星城/小区/三期/一个月/停电/10/来次

- **精确模式（不使用 *HMM* ）：**

  地铁/5/号/线/施工/导致/A/市/锦/楚/国际/星/城/小区/三期/一个月/停电/10/来/次

整个流程如下图：

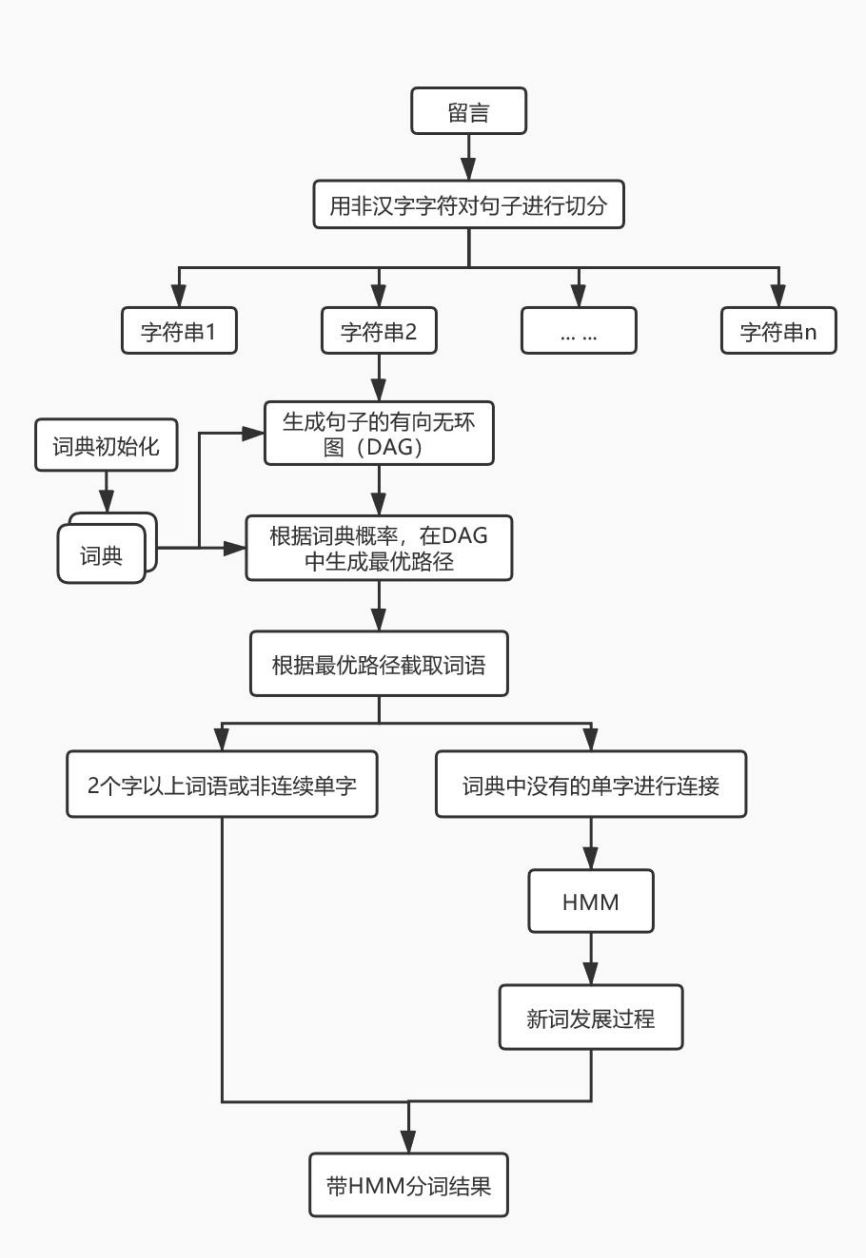

图 2 jieba 分词有向无环图

Step2：去除文本中的停用词

停用词在文本中，没有什么实际含义，且出现的频率较高，如：“的”、“了”等。停用词过多会影响我们在使用深度学习进行文本分类时的特征提取，进而影响分类效果，并且过多停用词的出现会导致神经网络训练中过多的无意义参数，导致训练时间过长甚至有可能出现梯度消失梯度爆炸的问题。去除停用词可以提高训练速度和精度。

下面我们展示了城县建设类留言中针对词频做的词云图，可以看出去除停用词后的词云图包含更多的实用信息。

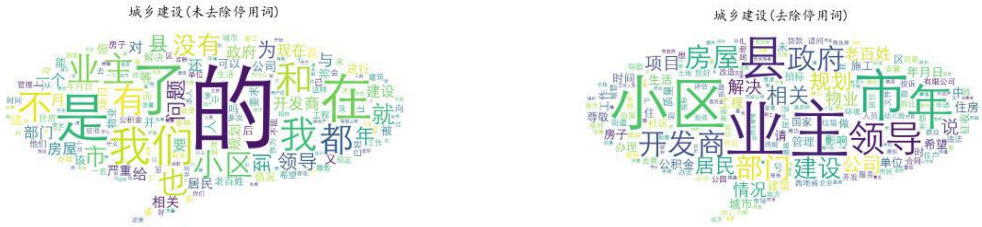

图 3 去除停用词前后词云对比图

2.2.2. Word2*Vec* 词向量训练

要想使用神经网络模型对文本进行分类，需要把已经处理好的文本数据进行词向量化。

目前常用的词向量化的方法有两种，第一种是独热编码形式，数据中词汇表的大小就是向量的维数，每个单词对用的位置为 1，其余为 0。独热编码解决了分类器不好处理离散数据的问题，在一定程度上也起到了扩充特征的作用，但是它在文本分类方面的缺点非常突出。首先，它是一个词袋模型，忽略了词与词之间的顺序；其次，它假设词与词之间是相互独立的，没有考虑到现实中词与词之间的联系；而且，它得到的特征是稀疏离散的，即词的数量便是生成向量的维数（一条 1000 个词的评论，将会生成 1000 维的向量，这在计算机运行中会消耗大量的内存也不利于运算效率）。

第二种方法是：词向量嵌入。它的思路是，通过训练将每个词都映射到一个较短的词向量上，生成的所有词向量构成一个向量空间，从而便于研究词与词之间的联系。并且向量之间的欧氏距离大小体现了现实中词语实际意义的相似度。比如训练后的‘城镇’对应的向量跟‘城市’对应向量之间的距离比‘城镇’对应向量与‘卫生’对应向量之间的距离更近，并且在训练后向量计算中，‘男人’对应向量减去‘女人’对应向量是约等于‘男孩’对应向量减去‘女孩’对应的向量。

下边我们以‘卫生’为例，展示‘卫生’的词向量训练结果，和训练出的与‘卫生’予以最相近的词以及相似度：

('市景树', 0.5112833380699158)

('公共卫生', 0.4472726881504059)

('健康', 0.4259778559207916)

('服务中心', 0.4102417230606079)

('消毒', 0.40944361686706543)

本文使用 *python* 自带的 Word2*Vec* 包进行词向量嵌入， Word2*Vec* 进行词向量训练的本质使用是简单神经网络，输入语料库单词的独热编码，隐藏层没有激活函数。输出层维度跟输入层的维度一样，用的是 *Softmax* 损失函数。当这个模

型训练好以后，我们并不会用这个训练好的模型处理新的任务，我们真正需要的是这个模型通过训练数据所学得的参数，例如隐层的权重矩重。这个模型输入输出方式有两种，一般分为*CBOW* 与 *Skip – Gram* 两种模型，*CBOW* 模型的训练输入是某一个特征词的上下文相关的词对应的词向量，而输出是这特定的一个词的词向量。*Skip – Gram* 是输入是特定的一个词的词向量，而输出的是特定词对应的上下文词向量。*CBOW* 对小型数据库比较合适，而 *Skip – Gram* 在大型语料中表现更好。

本文采用附件二中的留言内容作为语料库进行训练，选择使用*CBOW* 算法，训练轮数为 50，剔除词频小于 5 的单词，根据经验，词语向量的维度一般在 100-200 之间，我们暂且把向量维度设置为 150。词语向量的维度对使用神经网络进行分类有举足轻重的影响，维度过低，无法把特征提取充分，维度过高，导致可能出现过拟和或者梯度问题，因此在分类完成后我们会设置不同的向量维度，尝试计算不同的维度下最后分词的精度，进而选取效果最好的维度。至此我们得到了留言内容中 1021062 个词语的向量。现摘出一例进行展示：

```
>>> model['卫生']
array([-7.56740093e-01, -5.36050439e-01, -3.98038059e-01,  9.08117712e-01,
        2.51031846e-01, -1.43774736e+00, -2.65668064e-01, -8.02722871e-02,

                                ... ...          ... ...

        1.58034050e+00,  9.86083984e-01, -1.90329778e+00, -1.72852647e+00,
        2.78445625e+00,  1.53881893e-01,  4.26619723e-02, -1.82484174e+00],
      dtype=float32)
```

图 4‘卫生’向量示意图

2.2.3. 一维卷积神经网络进行词分类

a. 原理介绍

卷积神经网络[2][3]是目前深度学习技术领域中非常具有代表性的神经网络之一，它可以很好地识别出数据中的简单模式，然后使用这些简单模式在更高级的层中生成更复杂的模式。当我们希望从整体数据集中较短的（固定长度）片段中获取感兴趣特征、该特性在该数据片段中的位置不具有高度相关性时，一维卷积神经网络是非常有效的。并且一维卷积神经网络同样具有特征识别的可移不变形，也能够控制参数，抑制过拟和现象。

无论是一维、二维还是三维，卷积神经网络都具有相同的特点和相同的处理方法。可以根据输入数据的维数以及特征检测器（或滤波器）如何在数据之间滑动的不同来区分。二维卷积神经网络中的二维是指滤波器的移动，是从上向下、从左往右的，这样可以获得图像中不同位置的信息，提取不同位置的特征。而一维卷积神经网络是滤波器从上向下滑动，这是因为对于自然语言处理问题，矩阵的每一行就是一个单词，从上向下滑动可以提取句子中单词的整体之间的信息。

（图片来自网络）

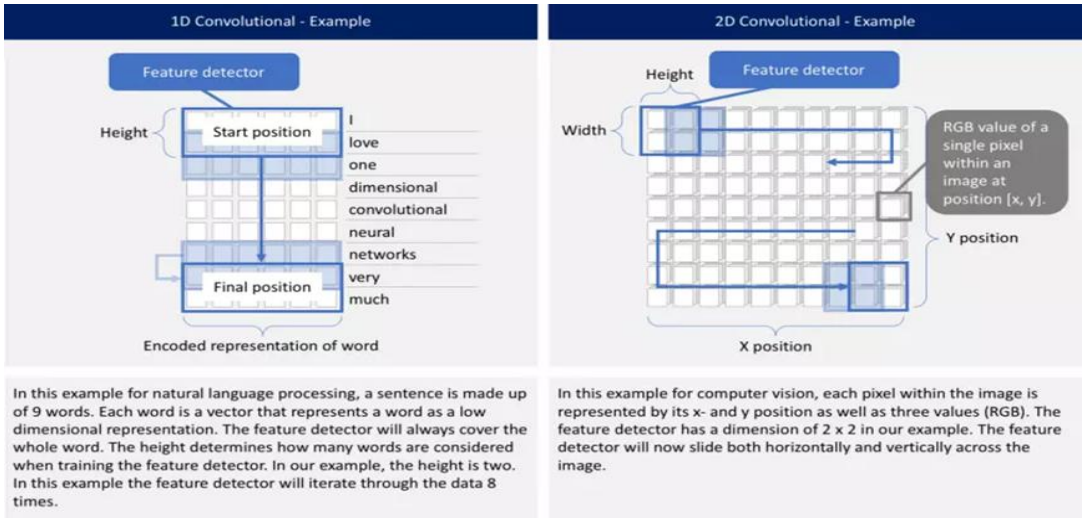

图 5 一维、二维卷积神经网络对比图

b. 具体步骤

在上文中我们已经对文本进行了处理，对于每一条留言详情，通过 jieba 分词使其变成词语的集合，并且在经过语料库训练后，每个词语都对应一个长度为 150 的向量。这样每个句子有唯一一个矩阵与其对应，这个矩阵的行数是 jieba 分词后的词语数量，列数是 150 维。经过探索后发现，留言分词后每个句子中单词个数最大为 3399，最短为 3，平均值为 110。在 9201 条留言中，有 9002 条是单词个数少于 500 的，故我们暂且将输入矩阵的维度设置为 500*150（后文会对矩阵维度寻求最优解）。留言内容中短于 500 个词的部分用 0 补齐，多于 500 的截断处理，经过这样的处理所有的句子都可以表示成一个 500*150 的矩阵。

下边是本文设计的卷积神经网络结构：

- 第一个 1D CNN 层，设置高度为 1 的滤波器，一个滤波器可以提取一个单一特征，这可能还不够，因此我们定义了 50 个滤波器，这样就能在第一层中提取 50 个特征，第一层输出一个 500*50 的矩阵。
- 第二个 1D CNN 层，来自第一 CNN 层的结果将被传送到第二 CNN 层。在第二个卷积层再次定义 50 个滤波器，与第一层原理相同，输出矩阵维数变为 500*50。
- 为了减少输出的复杂性并防止数据过拟合，通常在 CNN 层之后使用池化层，我们使用池化层的大小为 3，这样第三层的输出矩阵维数变为 166*50。
- 第三个和第四个 1D CNN 层，学习更高级别的特征。在这两层之后输出 166*100 的矩阵。
- 平均池化层，进一步避免过拟和，这次不是取最大值，而是取神经网络中两个权重的平均值。输出矩阵的大小为 1 x100 个神经元。每个特征检测器在这一层的神经网络中仅剩一个权重。
- 丢弃层，采取比率为 0.5，随机丢弃百分之五十的神经元，不仅可以提高运

算速度，增加准确性，还能提高模型的泛化能力，避免过拟和。

- 连接层：通过 *Softmax* 激活函数进行六大类的预测，*Softmax* 函数输出预测结果数每一类的概率，取概率最大的为最终预测结果。

$$Softmax(f)_y = p(y \mid x) = \frac{\exp(f_y)}{\sum_{c=1}^{C} \exp(f_c)} \quad (1)$$

此步的工作流程图如下：

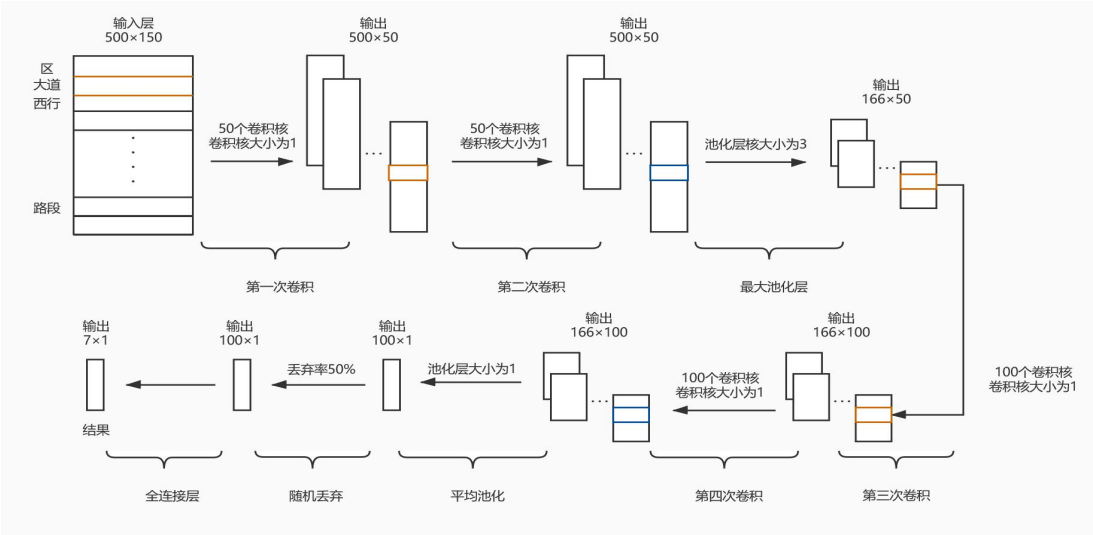

图 6 一维卷积神经网络工作步骤

由上图可以看出，最后输出一个 7*1 的矩阵，故通过这一步我们将留言归为了七类，下面给出分类结果表：

表 2 留言分类表

	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
真实个数	2009	938	623	1589	1969	1215	877
预测个数	2294	956	479	1568	1973	1057	883

2.3. 问题一模型的求解

在训练卷积神经网络时，我们随机在所有留言中选择百分之二十作为验证集，我们采用尝试使用不同的优化器，最后发现使用自适应矩估计优化器训练速度最快，训练精度高。可以看出随着训练次数的增加，训练集损失不断下降，但是验证集损失不再变化，并且训练集和验证集精度上升极其缓慢，训练集精度在94%左右，验证集进度维持在 89%左右。为了避免出现过拟和现象，我们将训练次数设置为 50，既能得到尽可能精确的训练结果，同时增加了模型的泛化能力。

下图给出卷积神经网络训练过程图：

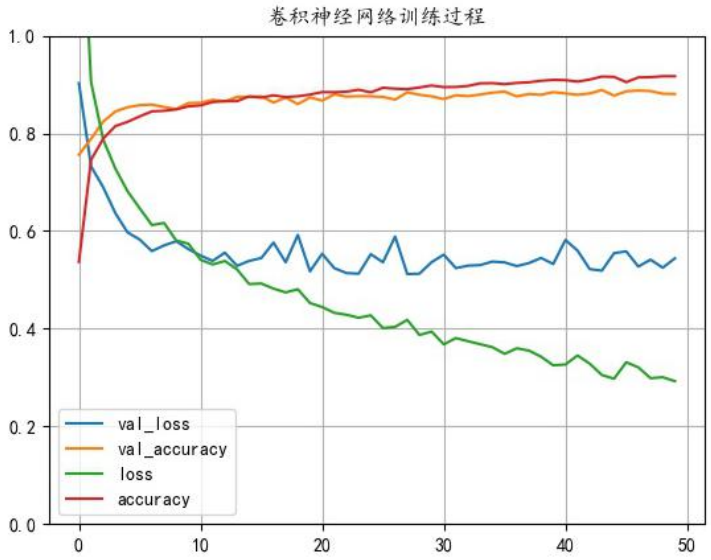

图 7 卷积神经网络训练过程图

最终我们通过一维卷积神经网络分类的 $F-Score$ 值是 0.930364，在模型中，词向量的维度和句子中词语的长度是影响训练结果好坏的重要参数，因此我们在词向量维度和句子长度分别选取不同的参数组合，进行超参数搜索，得到以下结果：

表 3 超参数检索表

	词向量维度					
句子长度	80	100	120	140	160	180
80	0.907534	0.907982	0.908763	0.909836	0.909932	0.90882
100	0.907723	0.908654	0.909376	0.912473	0.916437	0.912122
120	0.907893	0.909372	0.909984	0.912434	0.913432	0.918463
140	0.908373	0.909983	0.915333	0.917843	0.923745	0.920001
160	0.918463	0.918987	0.921355	0.926453	0.938974	0.933212
200	0.919983	0.924753	0.926784	0.928947	0.938495	0.930002
250	0.923432	0.926743	0.928943	0.934754	0.942732	0.940213
300	0.926473	0.928943	0.938495	0.947863	0.948895	0.932342
350	0.917468	0.918479	0.919984	0.927484	0.936475	0.920374
400	0.917839	0.918937	0.928394	0.912747	0.927454	0.920493
450	0.9128476	0.918475	0.919983	0.924543	0.928375	0.913462
500	0.9065443	0.910374	0.918476	0.923744	0.917364	0.902124

可以看出，在词向量维度为 160，句子长度是 300 时，得到 F-Score 最高，达到 0.9488，因此我们最终将该组合作为最终模型的参数。

3. 问题二模型的建立与求解

3.1. 问题二的分析

为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，并发现热点问题，我们首先对文本使用问题一的方法进行初步处理，然后计算 $TF-IDF$ 提

取高频词，使用独热编码实现留言向量化，然后使用 $K-means$ 聚类对文本留言问题进行分类，对每一类问题通过命名实体识别对地区/人群进行识别，并匹配每类问题的反映时间，为发现所有问题类中的热点问题，我们选择了三个指标留言数量、关注程度、用户活跃度来定义热度指数，最终挑选出五个热门问题进行分析展示。

3.2. 问题二模型的建立

3.2.1. $TF-IDF$ 提取高频词

首先对于附件 3 中的 4326 条留言做初步处理，运用问题一中的方法去除停用词、进行 jieba 分词。

然后计算留言中词语的 $TF-IDF$ 值，找出前 30 位高频词。但是我们发现有很多词诸如“问题”、“投诉”等，在实际分析中无特征指向的词语。经过对比、筛选后，去除“咨询”、“投诉”、“问题”、“反应”、“请问”、“解决”、“建议”、“涉嫌”、“希望”、“举报”、“请求”等词。

重新计算 $TF-IDF$ 值最高的前 30 个词，并记为 w1,…,w30。TF 的计算式如下：

$$TF_w = \frac{N_w}{N} \tag{2}$$

其中 N_w 是在某一文本中词条 w 出现的次数，N 是该文本总词条数。IDF 的计算式如下：

$$IDF_w = log(\frac{Y}{Y_w + 1}) \tag{3}$$

其中 Y 是语料库的文档总数，Y_w 是包含词条 w 的文档数，在分母上加一可以避免 w 未出现在任何文档中从而导致分母为 0 的情况。$TF-IDF_w$ 的计算式如下：

$$TF-IDF_w = TF_w * IDF_w \tag{4}$$

高频词提取以及得分结果如下表：

表 4 高频词提取表

小区	扰民	街道	噪音	县星沙	社区	业主	地铁	车位	新城
0.2181	0.1553	0.0781	0.0713	0.0708	0.0554	0.0549	0.0499	0.0413	0.0367
溪湖	居民	泉塘	物业	景园	国际	拖欠	违建	公交车	规划
0.0364	0.0358	0.0334	0.0332	0.0316	0.0311	0.0310	0.0304	0.0304	0.0297
安置	油烟	购房	房屋	大道	滨河	加快	拆迁	开发商	公园
0.0279	0.0272	0.0269	0.0268	0.0254	0.0252	0.0252	0.0251	0.0243	0.0236

3.2.2. 独热编码留言向量化

独热编码又称一位有效编码，主要采用 N 位状态寄存器来对 N 个状态进行

编码。在对特征进行分类时，特征值通常是离散的，特征之间距离的计算或者相似度计算都是非常重要的。这些计算大都是基于欧氏空间来完成，使用独热编码，将离散的特征取值扩散到后欧氏空间，离散特征值的某个取值就对应欧氏空间的某个点。将离散特征使用独热编码，可以使得特征值之间的距离计算更加合理。

我们将每一条留言都对应一个 30 维的向量，该条留言出现某个词时，这个向量的某个位置就对应为 1，否则为 0。对应公式如下：

$$A_i = (a_{ij})_{1*50} \ ，\ i = 1,2,...,4326, j = 1,...,50 \tag{5}$$

$$a_{ij} = \begin{cases} 0, \text{句子} A_i \text{中没有出现} w_j \\ 1, \text{句子} A_i \text{出现} w_j \end{cases} \tag{6}$$

3.2.3. *K – means* 聚类算法总结热点问题

K 均值聚类算法是一种迭代求解的聚类分析算法，是一个将数据集中在某个方面、相似的数据进行分类的过程。算法流程如下：

Step1：对 4326 个文本向量做聚类，选择 30 个点作为初始中心点；

Step2：按照距离初始中心点最小的原则，把所有样本分到各个中心点所在的类中；

Step3：每类中有若干个样本，计算 k 个类中所有样本的均值，作为第二次迭代的 k 个中心点；

Step4：根据这个中心重复第 2、3 步，直到中心点不再改变或达到指定的迭代次数，过程结束。

进行以上步骤，将文本分为 30 类。

3.2.4. 实体识别

实体识别是指，识别文本中具有特定意义的实体，包括：人名、地名、专有名词等。在本文中，每类问题都归属于特定地点或特定的人群，我们通过命名实体识别进行地点或人群提取。

首先先用 jieba 分词提取地名、人名，如“小区”、“花园”、“学院”等。再通过编程实现将地名之前的专有名字与地名/人群联系起来，得到以下数据包：'特立东路'、 '长雅中学',、'向家坡小学'、 '星沙锦璨家园'、 '涉外经济学院'、'碧楚社区'...通过这些对原有留言数据进行分类，就可以将高频词-高频留言-特定地方人群建立联系。

3.3. 问题二模型的求解

在以上的处理后，我们已经得到一个用特定地方和人群分类过的文本数据，我们还要对评价的“热度”进行定义，即什么是热度评价？热度得分又该怎么构成？于是我们通过综合留言数量、关注程度、用户活跃度来给出最后的“热度得分”。

3.3.1. 热度评价

热点问题[4]，是一定时间、一定范围内，公众最为关注的问题。通过公众留言，正确挖掘热点问题并及时解决热点问题有助于相关部门有针对性地处理问题、提升服务效率。那么制定评价某问题是否为热点问题的标准就显得尤为重要。

问题热度需从：该问题的留言数量（反映人数）、该留言的点赞数（反应关注程度）、留言用户的活跃度三方面考虑，所以我们定义了三个指标来定义问题热度。

- **留言数量**

在上述步骤中，我们得到了 30 个问题分类，附件 3 中总留言数为 4326 条，平均分配到每类问题为 144 条。定义留言数量指标：

$$num_i = d_i / 144 \tag{7}$$

d_i 是对一类问题的留言总量。

- **关注程度**

一个问题受到人们广泛关注，不仅表现在留言数量上，还体现在这类问题的留言反馈得到的浏览用户的赞同或反对数。

我们认为，无论是赞同票的数量多还是反对票的数量多，都能反映问题受到公众关注的程度。如果赞同票占优势，说明用户反映的此类问题真实性较高相关部门需要尽快解决这个问题；如果反对票占优势，说明某些用户反映的问题没有得到更多人的认同，可能此问题的个人主观性比较强，只是少数人遇到过此类问题或此类问题个人层面可以解决；如果赞同与反对势均力敌，则说明这个问题存在很强的争议性，相关部门应重视并且实地调查来发现问题的根源。所以定义关注程度指标：

$$corn_i = \sqrt{\frac{n_1 + n_2}{d_i}} \tag{8}$$

其中 n_1 为赞同数， n_2 为反对数

- **用户活跃度**

热心公众对和谐社会建设充满热情与责任心，他们总能敏锐地发现日常生活的各种问题并且及时反映，但是也存在同一个人恶意举报的行为。于是我们认为，同一个用户在不同的话题下都比较活跃的成为活跃用户，而同一个用户在同一个话题下连续评论，只选取字数较多、点赞数较高的那一篇进行分析。定义用户活跃度指标：

$$influence_i = log(1 + \frac{comment - count_i}{user - count_i}) \ , i = 1,2,...,30 \tag{9}$$

其中。 $comment$ 这一类问题的总留言数， $count_i$ 为第 i 类问题的留言数量， $user$ 为用户 ID 个数。

- **热度指标**

利用层次分析法，确定三指标在热度计算中的权重，得出以下热度得分指标：

$$Heat_i = \frac{1}{\frac{\alpha}{\sqrt{num_i}} + \frac{\beta}{\sqrt{corn_i}}} \cdot \gamma \cdot influnence_i \tag{10}$$

其中 $\alpha = 0.4$ 、 $\beta = 0.25$ 、 $\gamma = 0.16$ 。

3.3.2. 热度得分

计算出前二十的热度得分如下：

[2.1082006754603286, 1.3354076819593572, 1.1191594561794602,
0.7589031139643797, 0.7095248206144906, 0.6808039236645196,
0.6584263156717132, 0.6471535326895371, 0.6184877213144495,
0.5713436616713958, 0.5145201747356636, 0.4695893897009422,
0.4671220442686689, 0.46478873672338483, 0.432532562708727,
0.4270225094298833, 0.40332033944258633, 0.40322955173707686,
0.3957546793766537, 0.3860408494101679]

其中热度前五的话题见下表：

表 5 热点问题表

热度排名	热度指数	时间范围	地点/人群	问题描述
1	2.1082	2019/07/18 - 2019/09/01	伊景园滨河苑	捆绑销售车位
2	1.3354	2019/03/28 - 2020/01/26	丽发新城小区	搅拌站噪音扰民和污染环境
3	1.1191	2019/01/06 - 2019/12/03	辉煌国际城	交通乱象
4	0.7589	2019/02/20 - 2020/01/03	安置小区	噪音扰民
5	0.7095	2019/03/05 - 2020/01/07	泉塘街道	全天候施工扰民

4. 问题三模型的求解

4.1. 问题三的分析

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。公众对生活中的不便和问题在网络平台上留言反馈，相关部门及时发现问题并解决问题，才是留言机制设置的目的。而相关部门的回复可以反映问题的解决情况，所以我们关于留言回复，从相关性、及时性、完整性、可解释性的角度设置四个指标，来发现问题解决情况和相关部门的服务水平。

4.2. 问题三的求解

4.2.1. 相关性

相关部门及时回复的同时，要确保回复是与反映的问题相关[5]的，而不是答非所问、文不对题的。故我们设置相关性指标*correlation*，从短文本的语义相似度出发去衡量回复内容与留言主题的相似性。公式为：

$$sim(T_1,T_2)=\frac{1}{2}\left[\frac{\sum_{w\in\{T_i\}}(\max(sim(w,T_2)\times idf(w))}{\sum_{w\in\{T_i\}}idf(w)}+\frac{\sum_{w\in\{T_2\}}(\max(sim(w,T_1)\times idf(w))}{\sum_{w\in\{T_2\}}idf(w)}\right] \quad (11)$$

4.2.2. 及时性

公众进行问题反馈，相关部门应当及时核实、解决问题，并且及时给予群众回复。一个问题一般都具有时效性、季节性，如果不及时解决，问题虽然可能暂时消失但还会卷土重来。故我们设置及时性指标：

回复时间间隔：*Intime*

4.2.3. 完整性

相关部门给出的回复应当是有一定格式和要求的，不能太过随意、没有逻辑，所以我们针对回复内容的格式问题，设置完整性指标*Integrity*。通过查找政府相关部门回复文件的格式，对比本文中的政府回应，我们制定了以下模板，并且通过匹配回复内容中有无此类结构来评判回复的完整性。

- 现将网友在平台...栏目...留言反应...的调查...核实情况...答复如下:...感谢您对我区工作的理解和关心。
- 网友..您好!针对您反应...的问题，现回复如下:...请您看到回复后，致电...反映相关问题。感谢您对我们工作的关心、监督与支持。
- 网友...您好...您的留言已收到...现将有关情况回复如下:...目前，该项工作正在积极推进中...感谢您对我们工作的支持、理解与监督!

$$Itergrity=\begin{cases}1,\text{回复存在完整结构模式}\\0,\text{回复无完整结构}\end{cases} \quad (12)$$

4.2.4. 可解释性

广义上的可解释性指在我们需要了解或解决一件事情的时候，可以获得的我们所需要的足够的可以理解的信息。针对问题答复，可解释性体现在”问题是否解决”、“有没有明确答复”。如已解决，解决方案是否进行解释与告知；如未解决，原因是否给出合理解释。故我们给出可解释性指标*Interpretability*。

4.3. 评价方案综述

最后本文给出评价方案如下：

关于相关部门的回复，我们采用五星制，即五分制。基础分数为一星，上述四个指标，每达标一项，加一星，可以合理考虑半星的情况。如此-来，便得到相关部门留言回复的得分，可以进一步问题分析。下面给出两条留言的得分情况：

表 6　留言得分表

留言编号		12193	16631
留言内容		……保障小朋友出行安全的人行天桥或地下通道依然遥遥无期，恳切的希望您能够在基于确保小学生出行安全的大方向下，尽快能够解决我们现在的出行难题，真正保证孩子们的安全！	……请求在路口增设天桥，消除安全隐患。 同时，特立路松雅湖二小入口处与对面马路也可以考虑增设人行天桥，感谢领导!
相关部门回复		留言 12193 的回复：网友：您好！留言已收悉	……针对城区修建人行天桥及地下通道，目前规建局、交警部门、城管部门正在拿方案报政府办……
相关性	相关度	0	1.1
	得分	0	1
及时性	留言时间	2014/7/1 12:29:31	2018/4/27 8:39:26
	回复时间	2014/7/21 9:55:55	2018/5/4 13:59:34
	时间间隔	20 天	7 天
	得分	1	1
完整性		0	1
可解释性		0	1
最终星级			

5. 模型改进

- 使用更加精确的分词库，提高分词效果和实体识别精度，进而提高问题发掘效果。

- 针对有用户多次留言反映同一问题，设置不同的权重。
- 第一问可以采取卷积神经网络和长短记忆循环神经网组合的复合网络，设置记忆门等，进而提高分类效果和训练速度。

6. 模型的优缺点

6.1. 优点

- 本文在评价留言热度与相关部门回复质量时从多角度给出评价指标，体现了评价的全面性与严谨性，有利于提升评价的真实性和准确性。
- 本文明确给出关于文字预处理与后续分类等操作的方法与操作，并经过多次试验设置参数，使模型得到不断优化，实用性好。

6.2. 缺点

- 问题三做实体识别时，有部分地区不能被识别到，导致分类时可能有缺失留言，本文模型在使用时，可以事先将被应用区市的地点名称导入命名实体识别库，提高识别精确率。

7. 参考文献

[1] 祝永志，荆静.基于 Python 语言的中文分词技术的研究[J].通信技术，2019，52（07）：1612-1619.
[2] 杨锐， 陈伟,何涛,等.融合主题信息的卷积神经网络文本分类方法研究[J].现代情报,2020,40(4).
[3] 陶文静.基于卷积神经网络的新闻文本分类研究[D].北京交通大学.2019
[4] 吴靓弹媛.基于社区发现的网络舆情热点主题识别研究[D].江苏省:南京理工大学,2017.
[5] 张敏.短文本语义相似度计算研究[J].微型电脑应用,2019,35(10).