基于自然语言处理和文本挖掘的“智慧政务”

摘要

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、 汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、 云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

针对问题 1，首先利用内容分类的三级标签体系，获取分类所需的语料，观察每个标签下的样本数的总体特征。然后对文本进行预处理，其中包括剔除无意义的符号信息，特定地点识别，去除停用词，分词处理等。并针对类不平衡问题，利用 EDA 数据增强的生成少数类的样本。在文本的数值化中，我们利用 Word2vec 和 Skip-gram 模型，得到文本的词向量。最后，我们对比了传统机器学习朴素贝叶斯算法、逻辑回归、随机森林等算法和递归卷积神经网络 textRCNN 在文本分类下的表现，我们选择 textRCNN 模型作为分类模型的分类器，该分类模型在测试集上准确率达到了 93%。

针对问题 2，通过数据预处理，分词，中文实体解析，触发词解析，时间解析等步骤提取问题的关键词，然后基于改进的 Single-Pass 话题发现算法从众多留言中获取相似留言，对于每个话题通过特定的地点和人群等关键词特征将相似的问题归并为一类，最后定义热度评价指标，并基于层次分析法得到每个指标的权重，对每个热点问题进行分析，深度挖掘出每个热点问题的留言数目、留言人数、点赞数和时间跨度等特征，通过定义的热度分数公式得到每个问题的热度分数。

针对问题 3，我们首先通过查询相关部分针对留言回复意见的指导性文件，结合所提供的数据，定义答复意见评价的四个指标，相关性、完整性、答复时间间隔、可解释性，并量化。相关性是通过文本间相似度来量化，完整性和可解释性通过提取答复文本中是否存在一些符合留言答复规范的语句内容来量化得分，答复时间间隔通过留言答复指导文件的一些要求，将其分为三类。最后，通过分析数据给不同指标分配不同的权重来建立答复质量评价模型。

关键词：Word2Vec，逻辑回归，递归卷积神经网络，Single-Pass 算法

目录

1 挖掘目标

1.1问题背景

随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、 汇聚民智、凝聚民气的重要渠道，各类社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的相关部门的工作带来了极大挑战。同时，随着大数据、 云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治 理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用。

因此，此题需要根据互联网公开来源所采集的群众问政留言记录，及相关部门对部分留言的答复意见，运用自然语言处理技术和文本挖掘的方法，建立有效的数学模型，对留言进行划分，整理热点问题以及对相关部门的答复意见进行科学的评价。

1.2 挖掘目标

根据附件 1 和附件 2 的数据，按照内容分级标签体系，提取留言数据。通过对留言数据的分析和挖掘建立留言内容的以及标签分类模型。

根据附件 3 的数据识别留言的相似数据，并根据某一时间段特定地点和特定人群问题对留言进行归类。建立热度评价指标模型，对热点问题进行排名。

根据附件 4 的数据，从答复的相关性、完整性、可解释性、答复的效率等角度建立答复意见的评价模型。

2　问题分析

2.1问题一的分析

对附件 2 的数据进行分析，首先提取每个类别对应的留言信息，对留言数据进行处理。利用中文文本处理技术，去除数据中无意义的冗余信息和停用词。利用中文实体识别方法提取留言数据中特定的地点，建立自定义的词典。通过自定义的词典和中文分词方法对数据进行分词处理。在对进行文本数值化处理后，建立相关的分类模型对留言进行分类。

2.2 问题二的分析

问题二要求我们挖掘留言中的热点问题，并定义合理的热度评价指标对热点问题进行排名。首先需要做的是从众多的留言数据中识别相似的留言，可以通过提取留言数据中的特定地点，特定人群，事件、以及留言的主题等，利用

话题发现算法将相似主题的留言合并到一起，然后通过识别特定地点和人群，对每个话题下的文本进行聚类，将特定地点、特点人群或事件将问题归类。热度评价指标可以从点赞数，留言的数目，留言的人数，时间跨度等因素来考虑。其中原始数据中有冗余数据，即同一用户在相同时间的同一留言不止一条，需要在数据预处理阶段去除。各个热度评价指标可以利用层次分析法来确定。

2.3 问题三的分析

问题三需要我们对留言的答复意见的质量评价，评价指标可以从答复内容的相关性、完整性、可解释性及答复的时间间隔等角度考虑。量化这些指标，通过分析有关部门针对留言答复工作的指导性文件来分析获得量化指标的标准和依据。最后应该结合所提取的数据以及有关部门的指导性文件来建立留言答复质量评价模型。

3　留言内容的一级标签分类模型

3.1文本预处理

3.1.1数据清洗

在中文文本分类问题中，我们面对的原始数据经常存在许多影响最终分类效果的部分，这部分数据或文本都需要在文本分类最开始的时候进行数据清洗，否则很容易导致导致“trash in,trash out 问题。通常中文分类需要进行数据清洗和处理的数据包括

(1)非文本数据：文本中数据附带有 HTML 标签、URL 地址等非文本内容，这部分内容对分类一般情况下没有什么帮助。

(2)长串数字或字母：通常情况下中文文本中长串数字代表手机号、车牌号、用户名 ID 等文本内容，在此题进行留言内容的以及标签分类的背景，这部分内容对分类的帮助并不大。

(3)无意义文本：过滤掉剩余文本中诸如广告内容、版权信息和个性签名的部分，毫无疑问这些也都不应该作为特征被模型学习。

(4)停用词：停用词指的是诸如代词、介词、连接词等不包含或极少包含语义的刺，另外标点符号也可以被认为是一种停用词。通常情况下，在文本中去掉这些停用刺能够使模型更好地拟合实际的语义特征，从而提高模型的泛化能力。

3.1.2文本分词

中文分词是中文文本处理的一个基础步骤，也是中文人机自然语言交互的基础模块。不同于英文的是，中文句子中没有此的界限，因此在进行中文自然语言处理时，通常需要先进行分层次，分词效果将直接影响词性、句法树等模块的工具。中文分词根据实现原理和特点，主要分为以下两个类别：

1、 基于词典分词算法

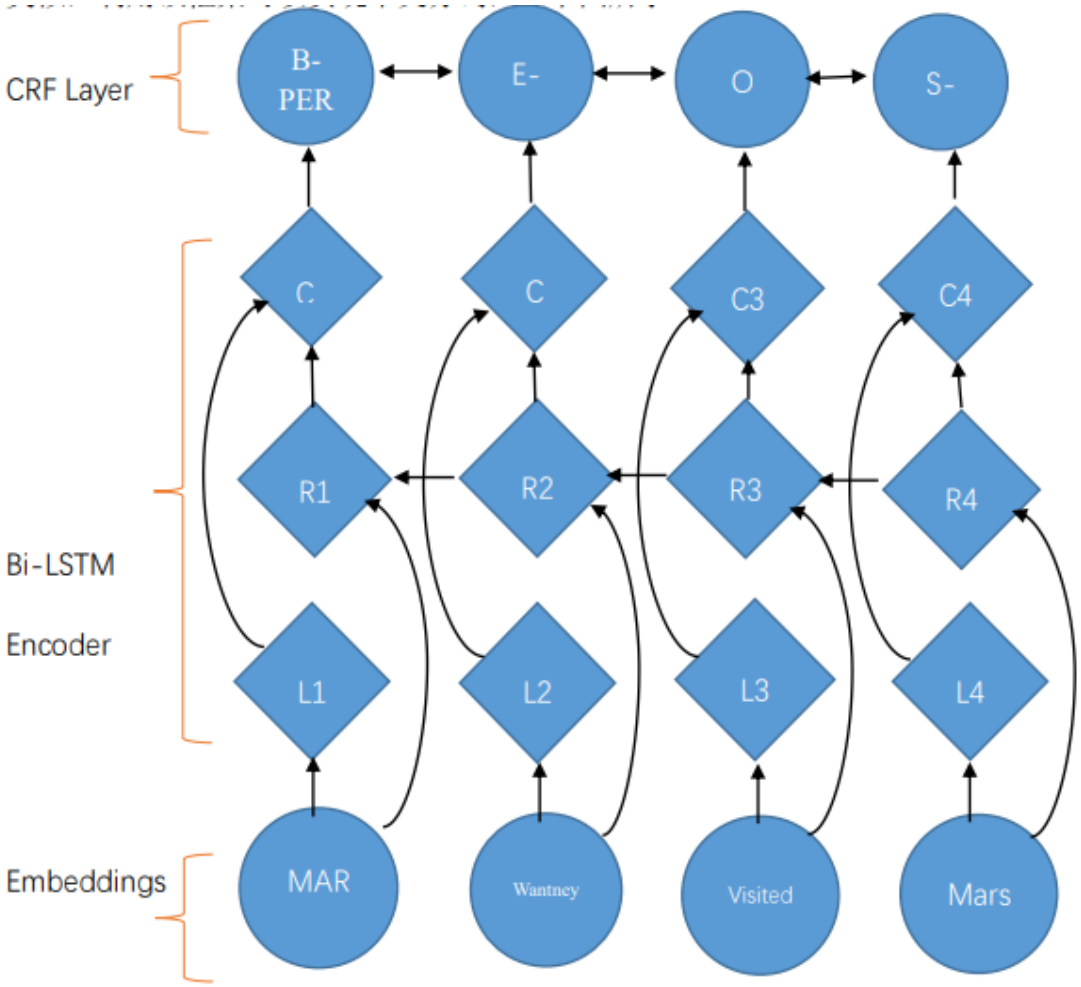
基于词典分词算法也称字符串匹配分词算法。该算法是按照一定的策略将待匹配的字符串和一个已建立好的“充分大的”词典中的词进行匹配，若找到某个词条，则说明匹配成功，识别了该词。常见的基于词典的分词算法分为以下几种：正向最大匹配法、逆向最大匹配法和双向匹配分词法等。

基于词典的分词算法是应用最广泛、分词速度最快的。很长一段时间内研究者都在对基于字符串匹配方法进行优化，比如最大长度设定、字符串存储和查找方式以及对于词表的组织结构，比如采用 TRIE 索引树、哈希索引等

2、 基于统计的机器学习算法

这类目前常用的是算法是 HMM、CRF、SVM、深度学习等算法，比如 stanford、Hanlp 分词工具是基于 CRF 算法。以 CRF 为例，基本思路是对汉字进行标注训练，不仅考虑了词语出现的频率，还考虑上下文，具备较好的学习能力，因此其对歧义词和未登录词的识别都具有良好的效果。Nianwen Xue 在其论文《Combining Classifiers for Chinese Word Segmentation》中首次提出对每个字符进行标注，通过机器学习算法训练分类器进行分词，在论文《Chinese word segmentation as character tagging》中较为详细地阐述了基于字标注的分词法。常见的分词器都是使用机器学习算法和词典相结合，一方面能够提高分词准确率，另一方面能够改善领域适应性。

随着深度学习的兴起，也出现了基于神经网络的分词器，例如有人员尝试使用双向 LSTM+CRF 实现分词器，其本质上是序列标注，所以有通用性，命名实体识别等都可以使用该模型，据报道其分词器字符准确率可高达 97.5%。算法框架的思路与论文《Neural Architectures for Named Entity Recognition》类似，利用该框架可以实现中文分词，如下图所示:



本文采用了 jieba 分词工具，jieba 分词是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，采用了基于汉字成词能力的 HMM 模型，使用了 Viterbi 算法。

由于文本的部分地名是以字母开头的虚拟地名，因此在分词过程中我们建立自定义词典，使这些地名在分词过程中不被分开。

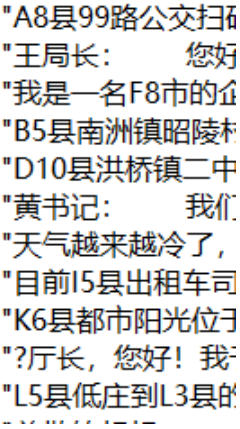


图 3.1 部分文本

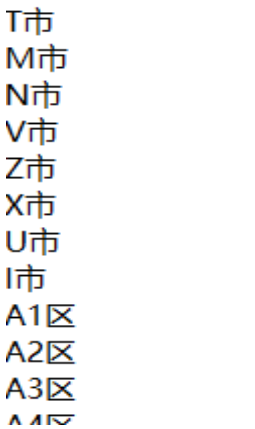


图 3.2 部分自定义词典

下面是数据清洗之后分词前后的语句对比：

表格 3.1 数据清洗前后语句

分词前	″近日由于西岭路段施工，一些施工材料及隔离牌施工牌的放置，导致西岭路段比较拥堵。今日一辆车牌尾号为“1952”的门票稽查车副驾驶上的工作人员态度极其嚣张，无视堵车情况强行加塞，甚至一度下车踢翻路边的施工警示牌和隔离施工区域的水马，辱骂其他单位工作人员，态度嚣张到令人发指。事后还没有恢复掉落在沟里的施工警示牌，如果驾车人员未及时发现施工区域而发生交通事故，请问谁负责？还有，身为景区工作人员，态度如此恶劣，这种行为是否合适？
分词后	近日 西岭 路段 施工 施工 材料 隔离牌 施工牌 放置 导致 西岭 路段 拥堵 今日 一辆 车牌 尾号 1952 门票 稽查 车 副驾驶 工作人员 态度 嚣张 无视 堵车 情况 强行 加塞 下车 踢翻 路边 施工 警示牌 隔离 施工 区域 水马 辱骂 单位 工作人员 态度 嚣张 令人发指 事后 恢复 掉落在 沟里 施工 警示牌 驾车 人员 未 及时 发现 施工 区域 发生 交通事故 请问 负责 身为 景区 工作人员 态度 恶劣 合适

下图是分词过后绘制的每个类别留言中的前 100 高频词的词云图：

图 3.3 每个类别的高频词

3.1.3类不平衡问题

机器学习中尝尝会遇到数据的类别不平衡，也叫做数据倾斜。类别不平衡指的是分类任务重不同类别的训练样本数目差别较大的情况，这种情况会对学习过程造成困扰。目前解决类别不平衡问题主要有三类做法：

1、 欠采样:直接去除训练集中样本数较多类别中部分样例，代表性算法

EasyEnsemble 利用集成学习机制，将多数类划分为若干个集合供不同学习器使用，这样对每个学习器来看都进行了欠采样，但在全局来看又不会丢失重要信息。

2、 过采样：增加少数类中样本数目，代表性算法是 SMOTE，通过对训练集中少数类中的样本进行插值来产生额外的样本。

3、 代价敏感学习：直接基于原始训练集进行学习，但在用训练好的分类器进行与预测时，调整阈值。

下面是本题中不同类别样本的数目，从图中来看，不同类别间样本数目相差较大，例如城乡建设类别下的样本数是交通运输三倍多，因此我们需要处理类别不平衡的情况。

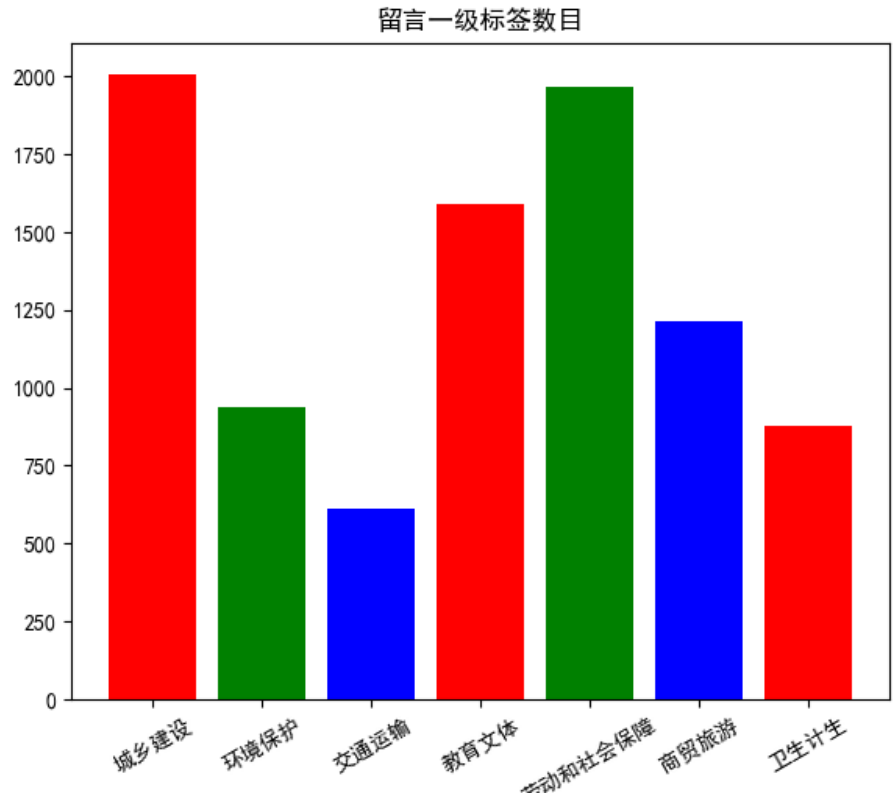


图 3.4 每个类别样本数

本文采用的方法是基于《EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks》EDA 由四个简单但功能强大的操作组成：同义词替换，随机插入，随即交换和随机删除。

对训练集中的每个句子，执行下列操作：

1、 同义词替换（Synonym Repalcement，SR）：从句子中随机选取 n 个不属于停用词集的单词，并随机选择其同义词替换它们；

2、 随机插入（Random Insertion，RI）：随机的找出句子某个不属于停用词集的词，并求出其随机的同义词，将该同义词插入句子的一个随机位置。重复 n 次；

3、 随即交换（Random Swap，RS）：随机的选择句子中两个单词并交换它们的位置，重复 n 次；

4、 随机删除（Random Deletion，RD）：以概率 P，随机的移除句子中的每个单词。

长句子相对于短句子，存在一个特性：长句比短句有更多的单词，因此长句在保持原
的类别标签的情况下，能够吸收更多的噪声。为了充分利用这个特性，EDA 提出基于句子长度变化改变的单词数。也就是说，不同的句长，因增强而改变的单词数可能不同。

EDA 提高文本分类效果的原理是，生成类似原始数据的增强数据会引入一定程度的噪声，有助于防止过拟合，使用 EDA 可以通过同义词替换和随机插入引入新的词汇，允许模型泛化到那些在测试集中但不在训练集中的词。

我们在针对英文语料 EDA 实现基础上进行修改，实现适合中文语料的数据增强 EDA 实现。下面是利用 EDA 生成的数据：

表格 3.2 **数据增强语句**

原数据	作为 B 市民两车，近期发现出租车乱象。每天出门坐公交上班，在公交站等车的时候，占到展台前面总是五号线有出租车占着，等公交坐公交都很不方便
EDA 生成数据	身为 B 市民，出租车乱象。每天乘坐公交上班，在公交站等待的时候，展台前面总是有出租车占着，等公交坐公交都很不方便

3.1.4文本的数值化

文本的数值化，即使用数字代表特定的词汇，因为计算机无法直接处理人类创造的词汇。为了让计算机能够理解词汇，我们需要将词汇信息映射到一个数值化的语义空间中，这个语义空间我们可以称之为词向量空间（词向量模型）。

文本的数值化方式有很多种，例如：TF-IDF、BOW、One-Hot、word2vec 和 Glove 等。

本文使用的是经典的 word2vec 工具。Word2vec 是一种无监督的学习模型，可以在一个语料集上（不需要标记，主要思想是“具有相似邻近词分布的中心词之间具有一定的语义相似度”），实现词汇学信息到语义空间的映射，最终得到一个词向量模型（每个词汇对应一个指定维度的数组）。Word2vec 实际上一种浅层的神经网络结构，它有两种网络结构，分别是 CBOW（Continues Bag of words）和 Skip-gram。

CBOW 的目标是根据上下文出现的词语来预测当前词的生成概率；而 Skip-gram 是根据当前词来预测上下文中各词的生成概率，如图 5 所示。

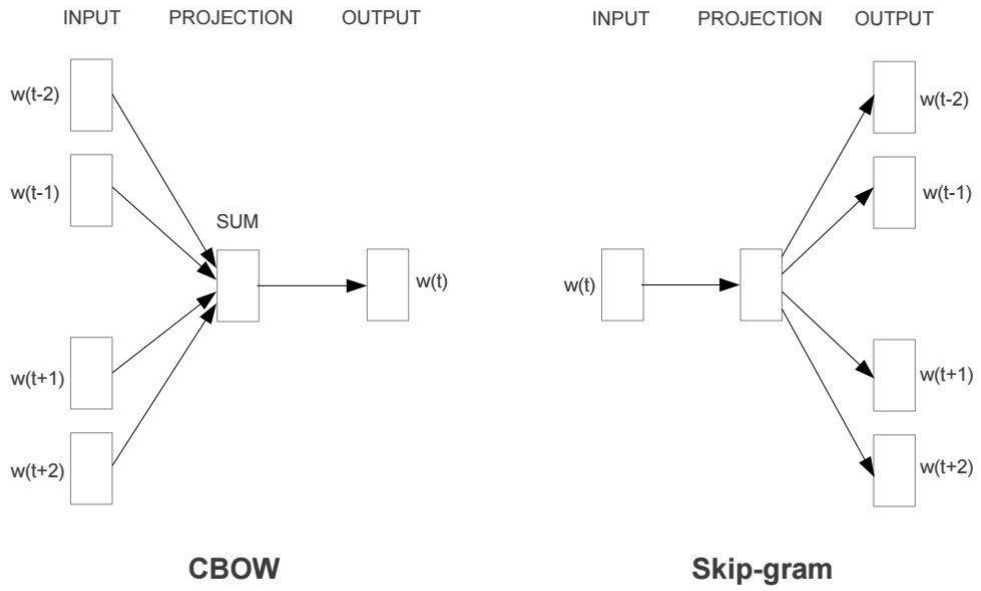


图 3.5　word2vec **的两种网络结构**

其中$\mathrm{w(t)}$是当前所关注的词，$\mathrm{w(t-2)}$、$\mathrm{w(t-1)}$、$\mathrm{w(t+1)}$、$\mathrm{w(t+2)}$是上下文出现的词。这里前后滑动窗口大小设为 2。

CBOW 和 Skip-gram 都可以表示成由输入层（input）、映射层（Projection）和输出层（output）组成的神经网络。

输入层中的每个词都由独热编码方式表示，即所有词均表示成一个 N 维向量，其中 N 为词汇表中单词的总数。在向量中，每个词都将与之对应的维度置为 1，其余维度的值均设为 0.

在映射层中，K 个隐含单元的取值可以由 N 维输入向量以及连接输入和隐含单元之间的$\mathrm{K \times N}$维权重矩阵计算得到。在 CBOW 中，还需要将各个输入词所计算的隐含单元求和。

同理。输出层向量的值可以通过隐含层向量（K 维），以及连接隐含层和输出层之间的$\mathrm{K \times N}$维权重矩阵计算得到。输出层也是一个 N 维向量，每维与词汇表中的一个单词相对应。最后，对输出层向量应用 Softmax 激活函数，可以计算得到的每个词的生成概率。Softmax 激活函数的定义为

$$p(y = w_n|x) = \frac{e^{x_n}}{\sum_{k=1}^{N} e^{x_k}} \tag{3.1}$$

训练方式中主要有 Hierachical Softmax 和 Negative Sampling 两种。

3.2基于传统机器学习的分类模型

我们选择三个个常用于文本分类传统机器学习算法，贝叶斯算法（NB）、Logistic 回归算法（LR）和随机森林算法（RF）。

1、 NB

NB 算法是基于贝叶斯定理与特征条件独立性假设的分类方法。对于给定的训练数据集，首先基于特征条件独立性假设学习输入输出的联合概率分布；然后基于此吗，，哦行，对给定的输入x，利用贝叶斯定理求出后验概率最大的输出y。NB 算法实现简单，学习与预测的效率都很高，是一种常用的方法。

算法 4.1：native Bayes algorithm

输入：训练数据$T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$,其中

$x_i = \left(x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(n)}\right)^T$, $x_i^{(j)}$是第 i 个样本的第 j 个特征。$x_i^{(j)} \in \{a_{j1}, a_{j2}, \dots, a_{jS_j}\}$, a_{jl}

是第j个特征可能取的第l个值，$j = 1,2, \dots, n, l = 1,2, \dots, S_j, y_i \in \{c_1, c_2, \dots, c_K\}$;实例$x$；

输出：实例x的分类。

（1） 计算先验概率及条件概率

$$\mathrm{p}(\mathrm{Y} = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k)}{N}, k = 1,2, \dots, K \tag{3.2}$$

$$P\left(X^{(j)} = a_{ij} \middle| Y = c_k\right) = \frac{\sum_{i=1}^{N} I\left(x_i^{(j)} = a_{jl}, y_i = c_k\right)}{\sum_{i=1}^{N} I(y_i = c_k)} j = 1,2,..,n; \tag{3.3}$$

$$l = 1,2, \dots, S_j; k = 1,2, \dots, K$$

（2） 对于给定的实例$\mathrm{x} = \left(x^{(1)}, x^{(2)}, \dots, x^{(n)}\right)^T$,计算

$$p(Y = c_k) \prod_{j=1}^{n} p(X^{(j)} = x^{(j)} | Y = c_k), k = 1,2, \dots, K \tag{3.4}$$

（3） 确定实例x的类

$$y = arg \max_{c_k} p(Y = c_k) \prod_{j=1}^{n} p(X^{(j)} = x^{(j)} | Y = c_k) \tag{3.5}$$

2、 LR

多项逻辑斯谛回归模型，用于多分类。假定离散型随机变量Y的取值集合$\{1,2,\dots,K\}$,那么多项逻辑斯谛回归模型是

$$p(Y=k|x)=\frac{exp(w_k\cdot x)}{1+\sum_{k=1}^{K-1}exp(w_k\cdot x)},k=1,2,\dots,K-1 \tag{3.6}$$

$$p(Y=K|x)=\frac{1}{1+\sum_{k=1}^{K-1}exp(w_k\cdot x)} \tag{3.7}$$

LR 算法的参数估计通常采用的是梯度下降法及拟牛顿法。

3、 RF

随机森林（RF）是集成学习中 Bagging 方法的代表算法。RF 在以决策树为基学习器构建 Bagging 集成的基础上，进一步在决策树的训练过程中引入随机属性选择。具体来说，传统决策树在选择划分属性时是在当前节点的属性集合（假设为 d）选择一个最优属性；而在 RF 中，对基决策树中的每个节点，先从该节点的属性集合中随机选择一个包含 k 个属性的子集，然后再从这个子集中选择一个最优属性用于划分。这里的参数 k 控制了随机性的引入程度；若令 k=d，则基决策树的都见与传统决策树相同；一般情况下，推荐$k=log_2 d$。

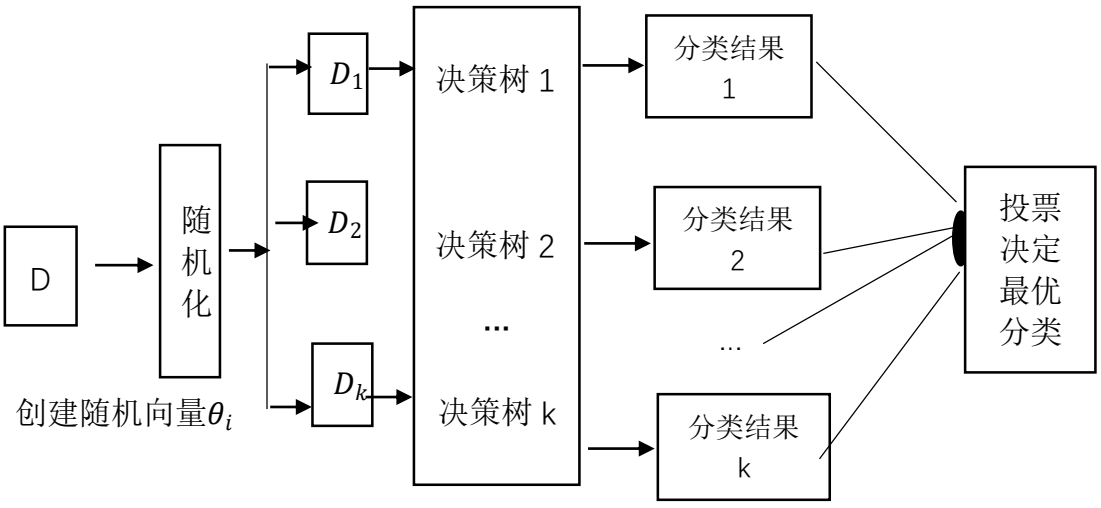


图 3.6 RF 示意图

RF 简单、容易实现、计算开销小，而且它再很多现实任务中展现强大的性能，被誉为“代表集成学习技术水平的方法”。可以看出，RF 对 Bagging 只做了小改动，但是与 Bagging 中基学习器的“多样性”仅通过样本扰动（通过对初始训练集采样）而来不同，RF 中基学习器的多样性不仅来自样本扰动，还来自属性扰动，这就使得最终集成的泛化性能可通过个体学习器之间差异度增加而进一步提升。

下表是个每个模型在测试集上准确率、召回率和 F1 分数：

表格 4.3 模型分类效果评价表

模型	精确率	召回率	F1-score
NB	0.871	0.863	0.861
LR	0.905	0.905	0.905
RF	0.862	0.861	0.860

针对平均准确率最高的 LR 模型，我们查看混淆矩阵，并显示预测标签和实际标签之间的差异；

混淆矩阵的主对角线表示预测正确的数量，除主对角线其余都是预测错误的数量，从上面的混淆矩阵可以看出“交通运输”类预测最准确，只有一例预测错误。“劳动和社会保障”预测错误数量最多。其中实际类别是“教育文体”，而模型预测为“劳动和社会保障”的数量。

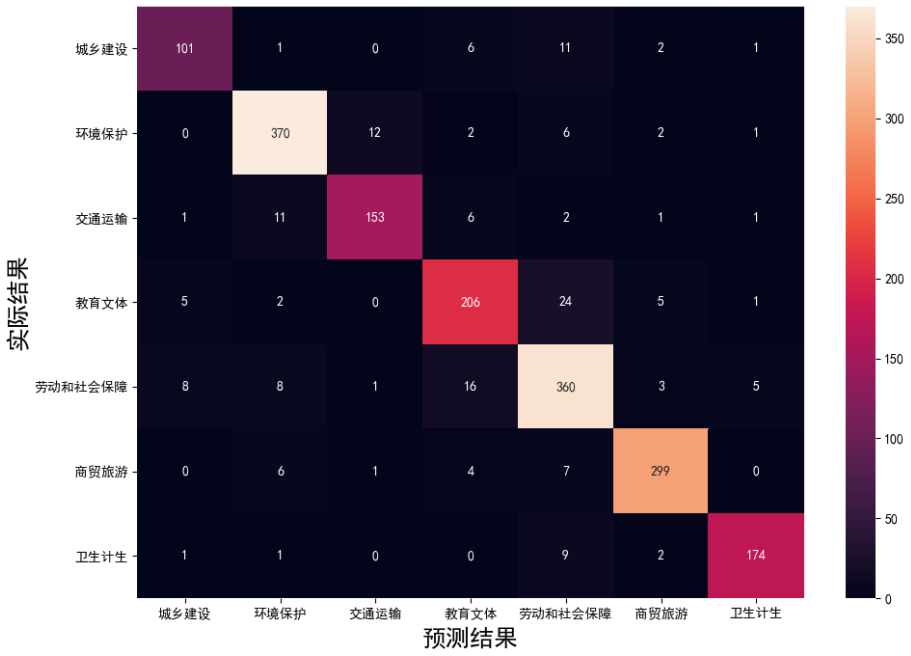


图 3.7 LR 模型混淆矩阵

3.3 基于 textRCNN 的留言一级标签分类模型

4.3.1 textRCNN 模型原理

textRCNN 模型即递归卷积神经网络的文本分类模型，在 textRCNN 模型中，学习单词表示时，应用递归结构来尽可能地捕获上下文信息，于传统的基于窗口的神经网络相比，这可能引入更少的噪声。RCNN 还采用一个更大的缓冲层，该层可以自动判断哪些单词在文本分类中起关键作用，以捕获文本中的关键成分。该模型在文档级的数据集上效果更好。

网络的输入是一个文档 D，它是一个词序列$w_1, w_2, \dots, w_n$,网络的输出包含类元素，我们使用$\mathrm{p(k|D,\theta)}$表示文档 D 属于类别 k 的概率，这里θ是网络的参数。定义$c_l(w_i)$为词w_i的左上下文，$c_r(w_i)$为词w_i的右上下文，$c_l(w_i)$和$c_r(w_i)$都是包含|c|个实值的密集向量。$c_l(w_i)$和$c_r(w_i)$通过以下方式计算：

$$c_l(w_i) = f\left(W^{(l)}c_l(w_{i-1}) + W^{(sl)}e(w_{i-1})\right) \tag{3.8}$$

$$c_r(w_i) = f\left(W^{(r)}c_r(w_{i-1}) + W^{(sr)}e(w_{i-1})\right) \tag{3.9}$$

其中，$e(w_{i-1})$是单词w_{i-1}的词嵌入向量，它是包含|e|个实值的密集向量。$W^{(l)}$是一个矩阵，将一个隐含层转换为下一个隐含层。$W^{(sl)}$是一个矩阵，用于将当前词的语义与下一个词的左上下文结合起来。f是非线性激活函数。

如公式 4.8 和公式 4.9 所示，上下文向量捕获所有的左侧和右侧的上下文语义，以这种方式，与仅使用固定窗口的常规神经网络模型相比，使用上下文信息，能够更好地消除w_i一词的含义。

$$x_i = [c_l(w_i); e(w_i);\ c_r(w_i)] \tag{3.10}$$

循环结构可以在文本的正向扫描中获得所有c_l，在文本的反向扫描中可以获得所有c_r。在获得单词w_i的表示形式x_i之后，我们将线性变化和 tanh 激活函数一起应用于x_i，并将结果发送到下一层。

$$y_i^{(2)} = \tanh\left(W^{(2)}x_i + b^{(2)}\right) \tag{3.11}$$

$y_i^{(2)}$是一个潜在语义向量，其中每个语义将分析特征以确定最有用的特征用于表示文本。

RCNN 模型的卷积神经网络旨在表示文本，从卷积神经网络的角度来看，我们前面提到的递归结构是卷积层。

当获得所有词表示向量被计算之后，我们使用了一个最大池化层。

$$y^{(3)} = \max_i y_i^{(2)} \tag{3.11}$$

RCNN 模型的最后一部分是输出层，类似于传统的神经网络，它定义为：

$$y^{(4)} = W^{(4)} y^{(3)} + b^{(4)} \tag{3.12}$$

最后将 softmax 函数应用于$y^{(4)}$，可以讲数字转化为概率。

$$p_i = \frac{\exp\left(y_i^{(4)}\right)}{\sum_{k=1}^{n} \exp\left(y_k^{(4)}\right)} \tag{3.13}$$

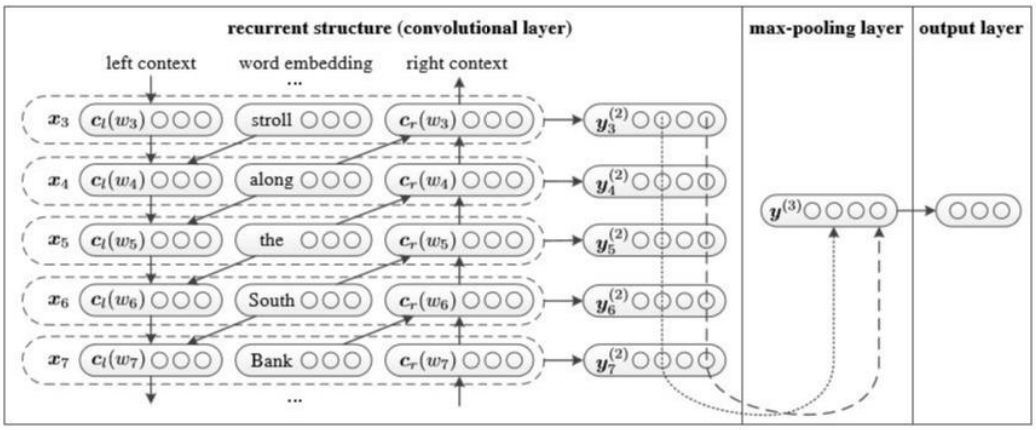


图 4.6 RCNN 的网络结构

4.3.2 模型分类结果

我们利用预处理和数据增强过的文本数据，应用 RCNN 模型，其中模型的超参数设置如下：

表格 4.4 textRCNN **模型超参数设置**

超参数	设置
类别数	7
批处理大小	256
衰减学习率之前步数	6000

衰减学习率	0.9
序列最大长度	200
嵌入尺寸	64
迭代次数	30
Dropout 比率	0.5

模型在测试集上的准确率：

表格 4.5 textRCNN **模型分类效果**

模型	精确率	召回率	F1-score
TextRCNN	0.933	0.924	0.928

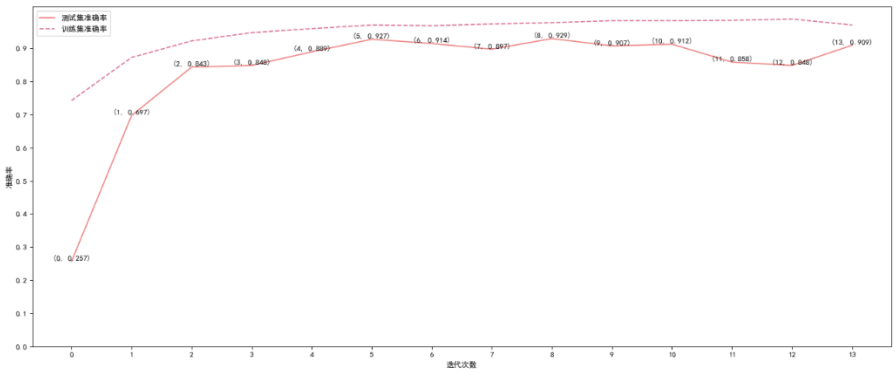


图 4.7 textRCNN 模型在测试集上的准确率

由图 6 可以看出 textRCNN 模型在迭代 9 次之后就能达到 0.927 的准确率，高于传统机器学习算法中表现最好的 LR 模型。

综合传统机器学习模型和深度学习的方法的表现，我们选择 textRCNN 模型作为留言一级标签分类模型的分类器。

4　热点问题挖掘

针对附件 3 的数据，我们通过建立以下模型实现热点问题的挖掘：

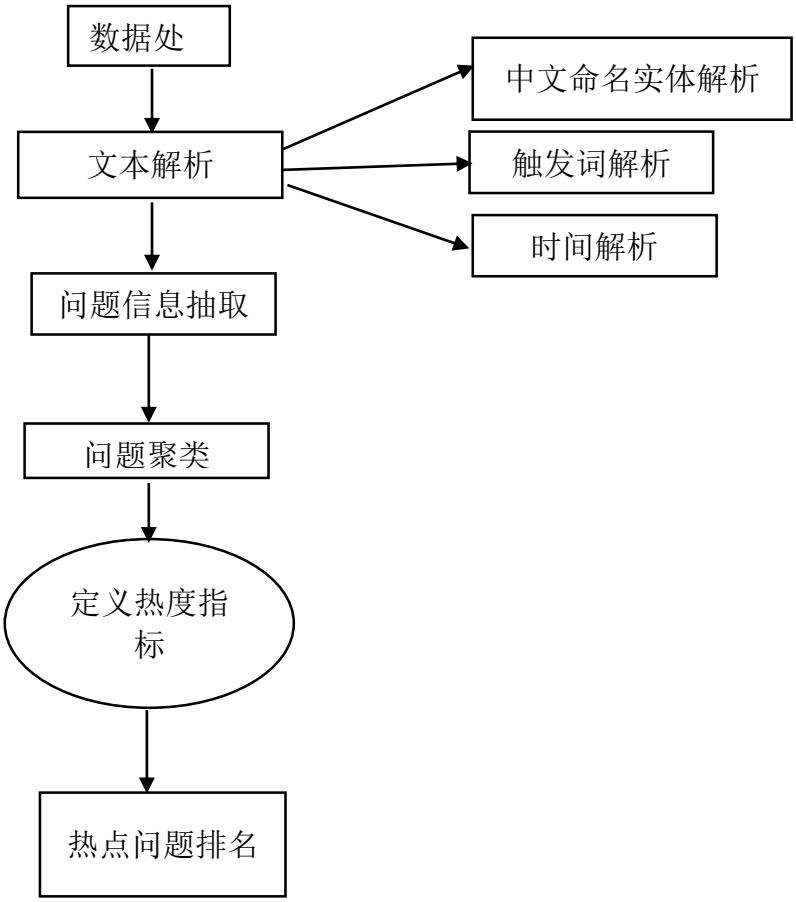


图 4.1 热点挖掘流程图

4.1数据处理

热点问题的挖掘，首先需要识别留言中相似的留言，我们基于 4.1 节文本预处理的方法，提取附件 3 中的留言数据。

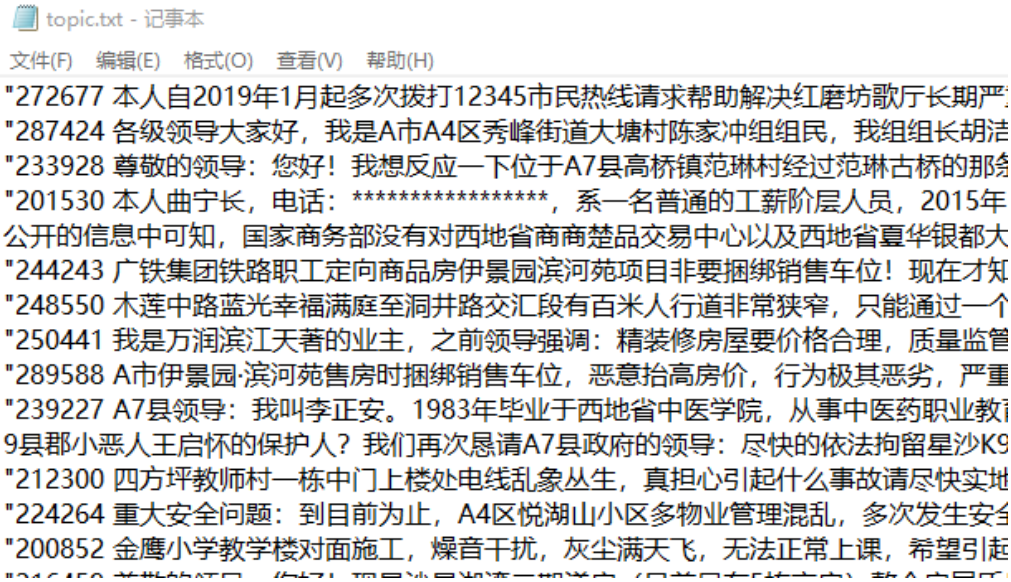


图 5.2 部分留言数据

5.1.1 关键句提取

由于留言数据较长，无意义的表达太多，而如果只利用留言主题，又无法很好地捕捉文本之间的相似性为了提高聚类的效果，我们利用文本摘要的方法提取留言详情中的关键信息结合留言主题从而形成具有概括性含义的短文本。这里我们使用的是 textRANK 算法。

TextRANK 的思想借鉴于网页排序算法—pageRanK，是一种用于文本的基于图的排序算法。通过把文本分割成若干组成单元（句子），构建节点连接图，用句子之间的相似度作为边的权重，通过循环迭代计算句子的 TextRank 值，最后抽取排名高的句子组合成文本摘要。具体迭代公式为：

$$\mathrm{WS}(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in out(V_j)} w_{jk}} WS(V_j) \tag{5.1}$$

其中$\mathrm{WS}(V_i)$表示节点V_i的 rank 值，$In(V_i)$表示前驱节点集合，$out(V_i)$表示后驱节点结合，d 表示阻尼银子，用于做平滑处理。w_{ji}表示两个节点之间的边的连接权重。

TextRank 关键词抽取算法如下：

（1） 把给定的文本 T 按照完整句子做分割，即$\mathrm{T} = [S_1, S_2, \ldots, S_n]$

（2） 对每个句子S_i进行分词和词性标注处理，并过滤掉停用词，只保留指定词性的单词，如名词、动词、形容词，即

$$S_i = [t_{i,1}, t_{i,2}, \ldots, t_{i,n}]$$

其中$t_{i,j}$是保留后的候选关键词。

（3） 构建候选关键词图 G = (V,E)，其中 V 为节点集，由（2）生成的候选关键词组成，然后采用共现关系（co-occurrence）构造任两点之间的边，两个节点之间存在边仅当它们对应的词汇在长度为 K 的窗口中共现，K 表示窗口大小即最多共现 K 个单词。

（4） 根据上面公式，迭代传播各节点的权重，直至收敛。

（5） 对节点权重进行倒序排序，从而得到最重要的 T 个单词，作为候选关键词。

（6） 由 5 得到最重要的 T 个单词，在原始文本中进行标记，若形成相邻词组，则组合成多词关键词。

TextRank 提取关键词短语：提取关键词短语的方法基于关键词提取，可以简单认为：如果提取出的若干关键词在文本中相邻，那么构成一个被提取的关键短语。

TextRank 生成摘要：

将文本中的每个句子分别看做一个节点，如果两个句子有相似性，那么认为这两个句子对应的节点之间存在一条无向有权边。考察句子相似度的方法是下面这个公式：

$$\text{SImilartity}(S_i, S_j) = \frac{|\{w_k | w_k \in S_i \& w_k \in S_j\}|}{\log(|S_i|) + \log(|S_j|)} \tag{5.2}$$

公式中，S_i, S_j分别表示两个句子词的个数总数，w_k表示句子中的词，那么分子部分的意思是同时出现在两个句子中的同一个词的个数，分母是对句子中词的个数求对数之和。分母这样设计可以遏制较长的句子在相似度计算上的优势。

我们将摘要的留言详情结合留言主题作为我们后面模型的语料集，对其进行数据预处理，然后对预处理后的分别进行命名实体识别、触发器解析、时间解析，得到每条留言的基本事件信息。

文本摘要前后的语句如下：

表格 4.1 **文本摘要**

摘要前	本人自 2019 年 1 月起多次拨打 12345 市民热线请求帮助解决红磨坊歌厅长期严重扰民及无证非法经营的情况，但 A 市 A2 区文化市场综合执法局（以下简称“文化局”）一直不正面处理问题，长期推诿扯皮，鼓励非法经营，偷税漏税，存在严重渎职的行为。具体情况如下：红磨坊歌厅位于西地省 A 市 A2 区新开铺路 732 号附近，有音响设备，搭了大棚运营，有商业行为但无娱乐场所经营许可证，长期深夜扰民。该歌厅位于黑石安置小区、香芙嘉 …… 2.请求严肃处理 A 市文化局推诿扯皮，不正面处理问题，多次鼓励非法

	经营的恶劣行径。请政府机构依法执法，杜绝包庇长期违法机构、违法个人的恶劣情况，还依法纳税、遵纪守法的居民以安宁的生活环境，感谢！〞
摘要后	本人自 2019 年 1 月起多次拨打 12345 市民热线请求帮助解决红磨坊歌厅长期严重扰。鼓励非法经营。红磨坊歌厅位于西地省 A 市 A2 区新开铺路 732 号。否则该歌厅未来也没有取得许可证的可能。文化局无人出面处理问题。且多次鼓励无证非法经营方日常经营。

5.1.2 命名实体识别技术

命名实体识别主要包含实体边界的识别和实体类型的识别两个方面，由于狭义的命名实体是指的是人名、地名、组织名等名词，所以一般情况下，本文主要关注命名实体边界的识别。

本文是在基于中文分词的基础上，采用基于中文维基百科数据库的方法来识别留言内容中的命名实体。维基百科。定位是包含人类所有领域的百科全书，是目前全球最大的百科全书，其包罗万象，具有海量的知识条目，内容比较全面。维基百科另外一大特点是允许大众的广泛参与，每个网名可以通过网页对其进行编辑或者创建新的条目。

5.1.3 触发词抽取规则

所谓的触发词就是指文本中最能代表是件本质和时间所属性质大的词语，是该事件的核心。每个句子中的触发词信息不仅是事件抽取的重要组成部分，更是计算两条留言内容相似度的一个重要依据。因此句子中时间触发词抽取效果对下面将要进行的问题聚类和问题归并的精确率都有直接的影响。

5.1.4 时间抽取规则

在热点问题中，问题的时间信息是信息抽取的重要载体。随着中文分词系统功能的不断强大，现在已经可以是别处大部分的时间表达式。本文主要针对相对的时间表达，通过规格和词典相结合的方式对其进行处理解析，处理解析后，获得一个相对时间的偏移量，把这个偏移量和参考时间进行计算，得到时间的绝对时间表达。

最后就是要把时间和问题对应起来，规范问题的世间怨俗，以确定某个问题或状态发生的时间。

4.2 基于 Single-Pass 话题发现算法的问题挖掘模型

我们利用 5.1 得到文本特征向量，首先基于 Single-Pass 话题发现算法，从留言数据中获取相应的主题并整合，然后基于特定的地点和特定的人群事件，对每个话题下的文本提取地点、人群等特征进行聚类，将相似的留言归并为同一问题。

4.2.1Single-Pass 算法

Single-Pass 话题算法不同于 LDA 等话题模型，不需要事先指定话题的数量。在本题目中，我们无法事先知道话题的数量或者相似问题的数量，因此使用 Single-Pass 算法比较适合磁此场景下识别相似的问题。Single-Pass 算法是一个代表性的增量式聚类算法，处理效率高且简单容易理解。

首先文本的相似度计算方式为—余弦相似度。

在向量空间模型（VSM）下，文档D_m和D_n间的相似度用$\mathrm{Sim}(D_m, D_n)$;

$$\mathrm{Sim}(D_m, D_n) = cos\theta = \frac{\sum_{i=1}^{N} w_{mi} \times w_{ni}}{\sqrt{(\sum_{i=1}^{N} w_{mi}^2)(\sum_{i=1}^{N} w_{ni}^2)}} \tag{5.3}$$

其中，w_{mi}和w_{ni}分别是文档D_m和D_n第 i 个特征项的权值，N 表示两个向量特征项的总数，并且 1<=i<=N。

Single-Pass 算法的基本思想是，首先按照顺序对每个文档进行处理，设定第一个文档为一个类别，然后一次计算将剩下的文档依次和类簇集合中的各个类的相似度，如果要处理的文本与类簇中的相似度大于某个阈值，则把该文本归入到有最大相似度的类簇中，否则重新建立一个类别，更新聚类中心。本文在原来的基础上，针对原始 Single-Pass 算法选取聚类中心比较随便的缺点做了改进，首先设置领域半径 eps 和最小密度阈值 MinPts，计算每一篇文档 d 在领域半径 eps 内的所有文档数目 T，如果 T>MinPts，则文档 d 就被设置为初始聚类质心。

改进的 Single-pass 算法：

输入：D—留言文档集合

Eps—半径

MinPts—在半径 eps 的范围内，使给定点称为初始聚类中心的最小领域点的个数。

θ—相似度阈值

输出：目标类簇集合

1） 对于输入的所有文档，计算每一篇文档D_j在领域半径 eps 内的所有文档数目T_j。如果T_j>MinPts,则将文档D_j设置为初始聚类中心，否则将D_j加入到未被处理的集合 F 中；
2） 判断所输入的所有文档是否处理完毕，如果是的话则执行步骤 3）,否则继续执行步骤 1）；
3） 计算文档集合 F 中文档F_i与初始聚类 C 中每个话题质心的相似度，利用递归算法找到最大相似度；
4） 比较步骤 3）计算出最大相似度与相似度阈值θ的大小，若最大相似度值比θ小，则新建一个话题簇，否则将D_j与F_i聚为一类；
5） 判断文档集合 F 的所有文档是否处理完毕，如果没有处理完毕，则返回步骤 3）继续循环执行，否则输出聚类结果。

通过 Single-Pass 话题发现算法提取了 212 个话题，部分话题的关键词如下：

```
Topic #0: 书记 解决 华颖 农民 加班 村民 部门 村委 侵占 村上
Topic #1: 90 ktv 歌厅 中央 男子 大厦 督察组 70 平方米 超过
Topic #2: 青竹 学校 小区 业主 外国语 湖楚 政府 义务教育 保利 加州
Topic #3: 村民 鄱阳 社区 鉴定 a市 人员 a5区 筹委会 方案 住房
Topic #4: 规划 火车 高架 西站 a6区 a市 高升 出口 高压线 请问
Topic #5: 违法 执法 酒店 部门 建筑 拆除 a市 施工 a1区 违规
Topic #6: 广胜村 村民 毛骏 涉案人员 政府 a市 苏纳 集体 资产 组织
Topic #7: 医院 手术 公司 工程 证书 希望 医生 道路 受理 信访
Topic #8: a9 大道 黄星 发展 居民 开元 和平 西南 小区 广场
Topic #9: 车辆 道路 路口 车道 路段 停车 交通 通行 交警 设置
Topic #10: 医院 a市 当事人 生殖 执行 技术 患者 试管 楚雅 2019
Topic #11: 违建 小区 城管 业主 公共 占用 搭建 投诉 私自 绿化
Topic #12: a市 停水 政府 农机 a8县 企业 频繁 亮之星 公寓 开发
```

4.2.2问题归并

针对 5.2.1 获取的话题，对每个话题中的通过抽取的地点、组织名和时间等信息，对相似的留言归并为同一问题。以下是留言数前十的问题数量

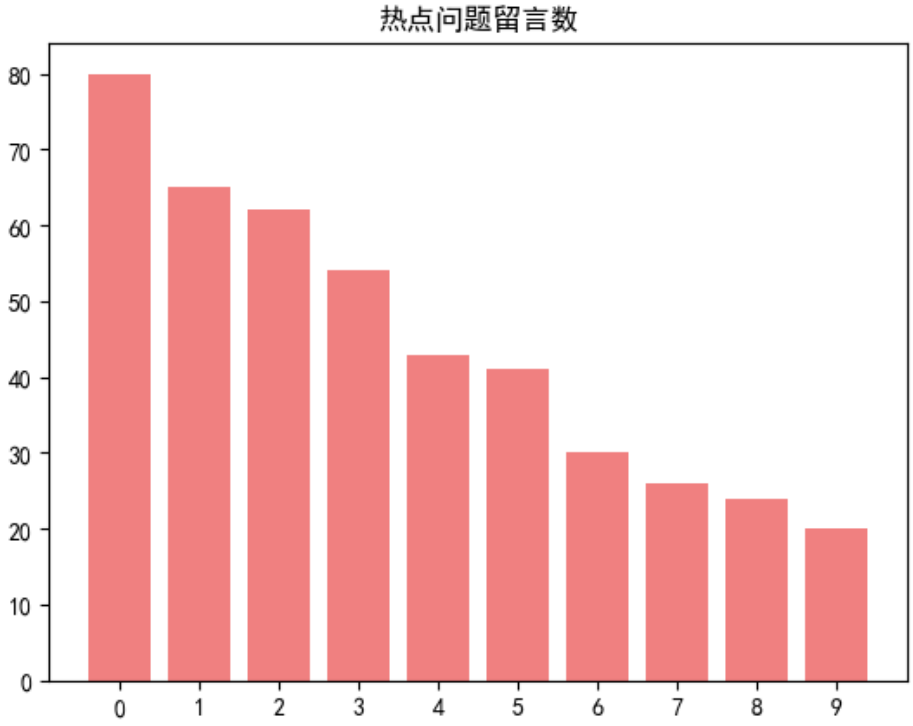


图 4.2 热点问题留言数

4.3问题热度评价模型

结合所提供的留言数据和文献，本文主要研究和计算提取的以下热度评价指标：

热度评价指标	留言的人数
	点赞数
	留言的数目
	留言的时间跨度

综合信息评价方法和本题目背景，各指标的得分标准定义如下：

评价指标	评分标准
留言的人数	$y_1 = k_1 n$
点赞数	$y_2 = k_2 m$
留言的数目	$y_3 = k_3 T$
留言的时间跨度	$y_4 = -k_4 t$

其中n，m,T,t分别是留言人数、点赞数、留言的数目和留言的时间跨度归一化后的数据。k_1，k_2，k_3，k_4分别是系数，其中热度随着时间跨度的增加而衰减，k_4是其衰减系数。因为通过分析数据发现，有部分留言用户在不同时间发表了相同的留言内容，所以我们考虑留言人数这个指标。

评价指标的权重，通过层次分析法得到。

层次分析法：

本文采用层次分析法解决问题的具体步骤为：

1， 热度评价指标建立。

2， 构造比较判断矩阵并求权重。

现采用标度法构造重要程度层次：

数字	重要程度含义
1	两元素相比，其同等重要
3	两元素相比，前者比后者稍重要
5	两元素相比，前者比后者明显重要
7	两元素相比，前者比后者强烈重要
9	两元素相比，前者比后者极端重要
2,4,6,8	上述判断的重要值
	若元素 i 与元素 j 的重要性之比为a_{ij},则
	元素 j 与元素 i 的重要性之比为$a_{ji} = 1/a_{ij}$

根据标度法最终得到的比较矩阵如下所示

$$M = \begin{bmatrix} 1 & 1/2 & 1/3 & 1 \\ 2 & 1 & 1 & 2 \\ 3 & 1 & 1 & 3 \\ 1 & 1/2 & 1/3 & 1 \end{bmatrix}$$

根据比较矩阵，对列向量归一化得到对应矩阵 Q 和其特征向量ω：

$$Q = \begin{bmatrix} 0.0625 & 0.0625 & 0.05 & 0.0625 \\ 0.125 & 0.125 & 0.15 & 0.125 \\ 0.1875 & 0.125 & 0.15 & 0.1875 \\ 0.0625 & 0.0625 & 0.05 & 0.0625 \end{bmatrix}$$

$$\omega = [0.063 \;\; 0.13 \;\; 0.16 \;\; 0.063$$

3，对构造的判断矩阵进行一致性判断，并以此来确定权重是否合理，公式如下：

$$CR = \frac{CI}{RI}$$

$$CI = \frac{\lambda - n}{n - 1}$$

其中，CR为一致性比例，CI为一致性指标，λ为最大特征根，n 为判断矩阵的阶数。当CR<0.1 时，即通过一致性检验，当CR>0.1 时，则判断矩阵不符合一致性要求，需要对该矩阵进行重新修正。经计算，一致性比例均小于 0.1，权重合理

3， 最终得到的热度评价指标权重如下：

评价指标	权重/w
留言的人数	0.635
点赞数	0.635
留言的数目	0.16
留言的时间跨度	0.16

最后根据综合评价公式算出每个热点问题的热度，其热度分数公式为：

$$\mathrm{P} = \sum_{k=1}^{n} yw$$

其中 p 热度得分，y 为其单项指标得分，w为其单项指标权重。

排名前 5 的热度问题如下表所示：

表格 5.3 **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**
1	132	72.5	2019/12/15 至 2020/1/20	A 市 A2 区丽发新城小区业主	小区搅拌厂噪音扰民，污染环境，经营不合法。
2	64	69.3	2019/08/18 至 2019/09/04	A 市 A5 区魅力之城小区	小区附近餐饮店油烟噪声扰民

3	19	67.9	2019/1/27 至 2019/4/23	A 市地铁七号线	A 市地铁七号线何时开工建设，增加地铁站
4	33	64.3	2019/4/8 至 2019/12/5	A7 县小塘路	A7 县小塘路路况太差，路灯太暗，建议改造
5	21	62.0	2017/06/08 至 2019/11/22	A 市经济学院学生	学校强制安排学生去企业实习

5 答复意见评价模型

对于相关部门对留言意见的答复意见，可以从答复意见与留言的相关性，答复间隔的时间长短，答复内容的完整性，答复内容的可解释性等角度考虑。

通过查阅政府有关部门针对网络留言回复工作的指导意见，我们总结出以下基本要求：

1、 专人办理：对于一些带倾向性、普遍性问题的留言，应有相关部门安排专人办理。

2、 办理时限：对于一般咨询及其他类留言办理一般不超过 7 天，适当时可放宽至 20 天。信访类留言办理一般不超过 30 天。

3、 回复质量：回复内容要严格依据现行的方针政策和工作实际，回复要做到语言朴实、简洁、亲切，既要表现人性化，更要展示人性化。

4、 实事求是：对需要调查核实的留言，办理部门需要坚持到现场了解情况，不得采用利益相关单位的答复。

5.1 评价指标提取

结合附件 4 所提供的数据以及相关部门针对网络留言办理的相关要求。我们定义以下指标来对留言的答复意见进行评价：

评价指标	相关性
	内容完整性
	答复时间间隔
	答复的可解释性

1、 相关性（Correlation）

答复意见和留言内容的相关性，是指答复意见回复的内容是否是留言中提到的内容。如果答复和留言的相关性很高，那么留言中关键词或者所表达的主题应该与答复意见中出现的关键词和主题接近。

在经过数据预处理、关键句提取、分词、之后，我们首先基于 LDA 主题模型提取文本主题作为文本的表达，然后基于 KL 距离来计算留言和答复之间的相关性。

基于 LDA 的文本相关性

LDA 模型通过类似于词聚类的办法将相似词聚类为一个个主题，使得主题与主题之间具有语义上的意思，对于在同一个主题中的词项一般具有近义词特性，而在不同主题中的同一个词项具有多义词特性，从而在文本相似性的计算过程中免去了计算词项之间的相似度。而利用文本的主题分布可以计算文本之间的相似性，而且其计算在不需要外部词典的情况下，其计算结果也具有语义效果。

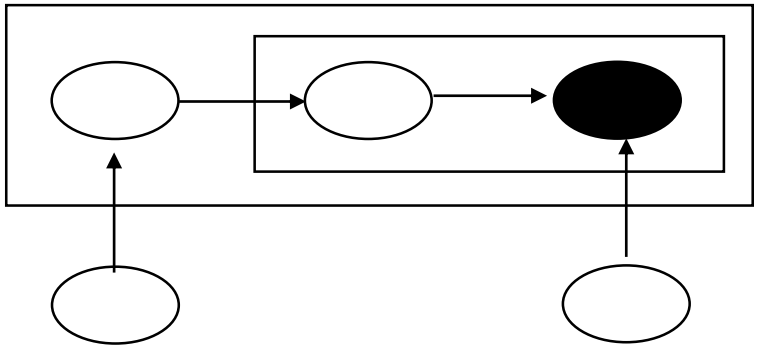


图 5.1 LDA 的模型图表示

（1） 基于 Gibbs 抽样算法，将文本的词向量空间映射为文本的主题向量空间。

（2） 通过计算两个文本对应的主题概率分布，得到两个文本的相似度。

由于主题是词向量的混合分布，因而本文使用的是 JS(Jensen-Shannon)距离作为相似度的度量标准，JS 距离如下所示

$$D_{JS}(p,q)=\frac{1}{2}\left[D_{KL}\left(p,\frac{p+q}{2}\right)+D_{KL}\left(q,\frac{p+q}{2}\right)\right] \tag{6.1}$$

$$D_{KL}(p,q) = \sum_{j=1}^{T} p_j ln \frac{p_j}{q_j} \quad (6.2)$$

其中p,q分别是留言文本和答复文本的主题分布。

以下部分留言与其答复意见之间的相关性

表格 5.1 留言与答复之间的相关性

留言编号	留言用户	留言内容	答复意见	相关性
3756	UU0081500	A9 市北盛镇对泉水村塘下组土地征收存在违规行为，主要问题如下……依法依规妥善解决征地事宜。此致 A 市人民政府、A9 市人民政府	网友“UU0081500”您好！您的留言已收悉。现将有关情况回复如下：据查，中南袜业园项目系 A9 市重点招商项目…….	0.86
3994	UU0082156	A 市 A5 区劳动路尊邸华庭路段由西往东方向跨二环线全程没有人行横道、天桥和地下通道，行人通行全部都在机动车道上非常危险。……	网友“UU0082156”您好！您的留言已收悉。现将有关情况回复如下：关于 A5 区劳动路尊邸华庭路段，近年来市城管局收到过有关修建天桥的建议，并就有…….	0.92

2、 完整性（completion）

答复意见的完整性是通过答复意见的回复格式来判断，通过人工观察质量较好的答复意见来看，一般情况下，答复意见包括以下三个部分：

收到留言-----相关问题答复-----感谢

我们通过文本匹配的方式，识别答复意见的结构，以上结构的我们定义为 ST。 不满足 ST，但文本长度大于 50 的答复意见完整度 completion 设为 0.5，

既不满足以上结构且文本长度小于 50 的答复意见 completion 设为 0.2。满足 ST 结构的答复意见完整度 completion 设为 1.0。

3、 答复时间间隔（interval）

根据政府有关部门针对留言办理时限的指导意见，我们将答复时间间隔归为三类，第一类，答复时间间隔小于七天；第二类，答复时间间隔在七天和二十天之间（包括二十天）；第三类，答复时间间隔在二十天以上的；

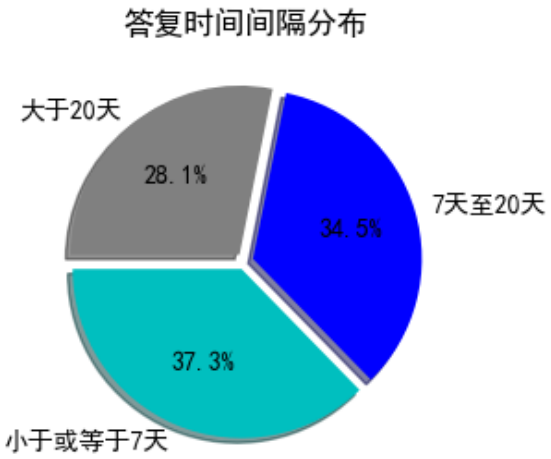


图 5.1 答复时间间隔分布图

4、 可解释性（explanation）

根据相关部分对留言答复工作的指导意见，回复内容要根据相关的政策和法规，回复要做到语言朴实、简洁、亲切，既要表现人性化，更要展示人性化。

因此，本文中通过识别答复内容是否有相关法律法规或者政策内容出现，以及回复内容的逻辑性来表示答复意见的可解释性。

其中，出现“《》”表示的内容表示内容中引用了相关的政策和法律法规。回复内容的逻辑性，通过检测文本数据是否存在逻辑性表达来呈现。最后结合这两部分的内容来判断答复意见的可解释性。

如果答复内容出现政策或法律法规的内容，具有逻辑性表达

explanation=1，否则 expalnation=0.5

5.2 答复意见评价模型

各评价指标单项得分定义如下：

相关性得分y_1： $y_1 = Corralation * 100,\ Corralation \in (0,1)$

完整性得分y_2：$y_2 = \ Completion * 100,\ Completion \in [0.2,0.5,1.0]$

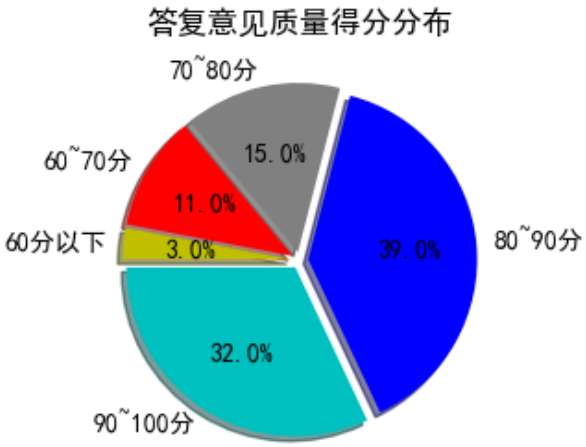
答复时间间隔得分y_3：$y_3 = p * 100, \mathrm{p} = \begin{cases} 1 \ if \ interval \leq 7 \\ 0.5 \ if \ 7 < interval < 20 \\ 0.1 \ if \ interval \geq 20 \end{cases}$

可解释性得分y_4：$y_4 = expalnation * 100$

根据有关部门的对于留言答复工作的指导意见以及我们所提取的数据，我们认为答复时间间隔在答复意见质量评价所占的权重应该大于其他几个指标。
最终答复意见的得分通过如下的公式计算：

$$\mathrm{y} = 0.2y_1 + 0.2y_2 + 0.5y_3 + 0.1y_4$$

答复意见质量得分如下：



总结

本文的主要目的是通过自然语言处理和文本挖掘的方法，对网络问政平台的留言进行智能留言划分、热点问题整理、答复意见评价的工作，从而解决由于社情民意相关的文本数据量不断攀升，给以往主要依靠人工来进行留言划分和热点整理的工作带来的挑战，对提升政府的管理水平和施政效率具有重要的意义。

首先，利用数据清洗、词性标注、中文分词等文本数据的处理方法对于留言数据进行预处理，通过 word2vec 将文本数据数值化。针对留言数据类别不平衡问题，利用 EDA 生成少数类的样本。在留言分类器的选择上，通过对比了朴素贝叶斯分类算法、逻辑回归、随机森林以及递归神经网络模型的分类效果。最终选择递归神经网络分类模型作为留言分类的分类器，该分类模型在测试集上准确率达到了 93%。然后通过特定地点抽取、事件抽取、文本摘要，利用话题发现算法和聚类算法将相似留言进行归并。为热点问题定义评价指标，计算每个问题的热度分数，对问题进行排序。最后结合有关部门对于留言答复问题的指导要求，定义并量化答复意见评价指标，通过设计评价模型对答复意见进行分析、评价。

参考文献

[1] https://github.com/Edward1Chou/Textclassification

[2] https://github.com/FesonX/cn-text-classifier

[3] Nianwen Xue, Susan P. Converse, Combining Classifiers for Chinese Word Segmentation, Computer Science,2002;

[4] Nianwen Xue, Chinese Word Segmentation as Character Tagging, international Journal of Computational Linguistics & Chinese Language Processing,2003;

[5] Guilaume Lample, Miguel Ballesteros et al, Neural Architectures for Named Entity Recognition, Computer Science,2016;

[6] Jason Wei, Kai Zou, EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks

[7] 《机器学习》周志华;

[8] 孙昌年，郑诚，夏青松，基于 LDA 的中文文本相似度计算，计算机技术与发展，2013;

[9] 齐向明，孙煦骄，基于语义簇的中文文本聚类算法;

[10] 李劲，张华，吴浩雄，向军，基于特定领域的中文微博热点话题挖掘系统，BTopicMiner，计算机应用，2012;

[11] https://wenku.baidu.com/view/43117d20a5e9856a56126030.html

[12]http://www.yinchuan.gov.cn/xxgk/bmxxgkml/scgj_2566/xxgkml_2569/rdhy_8418/201912/t20191226_1904505.html