政务留言智能管理初探

摘要

随着社交平台的不断发展，在各个领域能够更加方便且广泛的收集到反馈、评论以及建议等，因此对于大量的留言数据进行有效的智能管理是实现问题发现以及反馈信息提炼的有效途径。基于留言平台的兴起，民众问政也得一实现，便于政务机构有效熟知民众需求建议等。因此对于政务系统用留言平台的智能管理也具有相应的应用价值。

本文则以群众问政留言记录及相关部门对部分群众留言的答复意见作为研究对象，对于政务留言数据的智能管理进行一个探索。文中利用 Jieba 分词与 word2vec 词向量对原始文本数据进行处理并将其向量化，并提出了基于留言时间、点赞数、反对数这三个属性来构造的政务留言热度算法。此外还针对相关部门对留言的答复意见设计基于相关性、完整性与可解释性三个层次的留言答复评价系统。在实验环节中，本文针对样本不均衡的问题，通过重采样与过采样调整训练样本，将文本转换为向量后，使用 Softmax 回归模型对留言文本进行多分类，并以准确率，召回率和 F1 值作为评价分类器质量的指标，实验证明了算法的有效性。针对热点挖掘问题，本文设计了新的政务留言热度算法来得到热点问题，定义了合理的热度评价指标，同时利用层次聚类对群众留言内容进行归类以及热点问题发现。最后，针对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。

关键词：词向量、文本分类、Softmax 回归、层次聚类、热度值、答复评分

Abstract

With the development of social platforms, feedback，comments and suggestions can be more convenient and widely collected in various fields. Therefore, effective intelligent management of bulk text data is an effective way to realize problem discovery and feedback information extraction. Based on the rise of the message platform, people's political inquiry has to be realized, so that government agencies can effectively know the public's needs and suggestions. Therefore, the intelligent management of the message platform for the government system also has the corresponding application value.

This paper takes the record of people's comments on government affairs and the comments made by relevant departments on some people's comments as the research object to explore the intelligent management of government affairs' comments. In this paper, Jieba word segmentation and word2vec word vector are used to process the original text data and vectoquantify it. In addition, a message response evaluation system based on relevance, integrity and interpretability is designed for the comments of relevant departments. In the experiment, aiming at the problem of unbalanced samples, this paper adjusted the training samples through resampling and oversampling. Converted the text into a vector, and used the Softmax regression model to classify the message text, and used the accuracy rate, recall rate and F1 value as the indicators to evaluate the quality of the classifier. Experiments prove the effectiveness of the algorithm. Aiming at the hot spot mining problem, this paper designs a new algorithm of government affairs message heat to get the hot spot problem, defines the reasonable heat evaluation index, at the same time USES the hierarchical clustering to classify the mass message content and discover the hot spot problem. Finally, according to the comments of the relevant departments on the comments, from the relevance, integrity, interpretability and other aspects of the quality of the comments to give a set of evaluation program.

Keywords: Word vector, Text Categorization, Softmax regression, Hierarchical clustering, Hot value, Reply Rating

目录

1 绪论

在我们的工作学习生活中无时无刻不在接受来自各个方面的信息，而信息的形式又是多样化的。众多信息形式中，文本类信息是主要形式之一，例如电子邮件、办公文档、网页信息、留言评论等。在这些海量信息中也存在很多潜在的、有价值的信息，因此处理海量资源并从中获得有效信息是目前值得关注的话题。

随着社交平台的不断发展，使得各个领域的信息收集变得更加便捷。例如在近年来，随着市长信箱、阳光热线等问政平台的逐步发展与完善，其成为了政府部门了解民意、汇聚民智、凝聚民气的重要渠道之一，而在其中，依靠人工来进行数据整理和处理的工作方式面临着“大数据时代”带来的巨大挑战。

在实时获取民意的同时又能够快速准确进行有效信息的提取分类以及实时热点问题的挖掘成为了一个关注点，随着大数据发展带来的信息化时代以及人工智能在各个领域的不断渗入，想要解决上述问题，就需要借助基于自然语言处理以及文本挖掘的相关技术来实现。

在“智慧政务”问题中要解决的问题有三个。问题一的任务是实现对众多群众留言信息的自动分类，将群众留言信息中所涉及到的诉求、意见等信息分配到与之相关的主要政务部门，提升信息分类处理的效率，也能及时实现民众诉求的传递。问题二的任务则是热点问题挖掘，“政务热点”也即是在一段时间内民众反应得较为高频的事件或者建议。在这一问题中则是期望通过文本挖掘，从众多的留言信息中搜寻出热点问题。问题三较为开放，希望构建一个评价体系，实现对留言答复的评分。

问题一的核心任务是实现文本分类器的构建，需要解决的问题主要集中在如何有效处理非结构化的文本类数据，如何利用标签数据来进行多类别分类器的构建以及分类效果的评定情况。

问题二的主要目的是热点问题挖掘。本文将一段时间内民众集中反映的问题称为等价问题群，热点问题挖掘可以分解为等价问题群发现与热度定义排序这两个问题的复合。等价问题群发现可以理解为一个文本匹配问题，这是自然语言处理领域的核心问题之一。热度定义则是如何定义热点问题，通过构建相应的热度衡量指标来对等价问题群进行热度值计算，实现热点问题挖掘。

问题三的任务则是对群众留言的答复构建评价系统。这需要指定相应的衡量指标来构成评价系统，例如可以从答复的完整性，可解释性等方面来构建评价指标。

2 文本预处理

在文本数据处理的流程中，要进一步对文本数据进行处理分析，首先就是对所有的文本数据进行预处理。通常的数据的预处理主要是排除数据中的异常，对数据进行变换以降低处理难度。在文本数据预处理中，通常有数据清洗，文本分词，去停用词，文本表示四个步骤[1]。

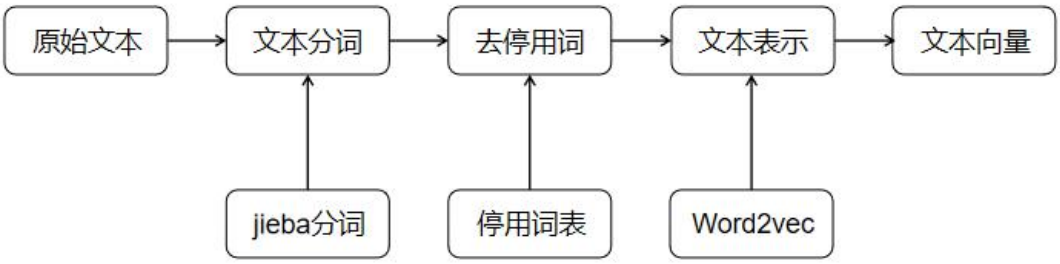


图 1：数据预处理流程

2.1 数据清洗

通过对赛题数据的整理，可以发现文本中存在大量的空格符、换行符等特殊符号。这些特殊符号对我们的后续处理并没有帮助，甚至会产生负面效果，因此对文本中的特殊符号进行删除。

2.2 文本分词

分词，即是将一个连续完整的文本，按照一定的规范，切分为一个个单独词的过程。分词作为文本处理的基础工作，其好坏直接影响后续的分析效果，因此良好的分词结果至关重要[2]。

现有的分词方法主要有三大类：基于词典的分词方法、基于理解的分词方法和基于统计的分词方法。而在实际的应用中，通常将基于词典的分词方法与基于统计的分词方法结合使用，这样即可以发挥词典分词速度快、效率高的特点，又可以识别利用未录入词典的生词，消除歧异分词的优点[2]。

基于对语料库的统计，可以构造包含了各个单词的词频、词性信息的前缀词典，并根据前缀词典对输入文本进行切分，而所有可能的切分方式可以构造为一个有向无环图 G。

从统计的思想出发，我们以出现的概率衡量一个切分的准确性，即正确的切分出现的概率是最大的。图 G 中的每条路径都有对应的权值 w，对于在前缀词典里面的词语，其权重就是它的词频，在前缀词典已经构建的情况下，所有词的词频之和是确定的，单个词的词频自然反映了它出现的概率，而一个切分。因此，我们需要找到一个切分，该切分所经过路径的权值之和最大，问题进而转化为动

态规划问题，众多的动态规划算法都可用于解决这一问题。

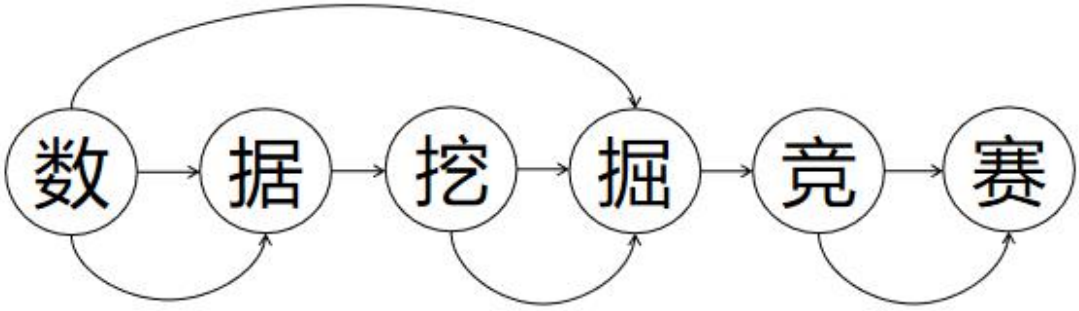


图 2：一个有向无环图

Jieba 分词是当前非常流行的中文分词工具，基于内置的前缀词典，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，并采用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，而对于句子中存在未登录词，即内置的前缀词典中未包含的词的情况，使用了 Viterbi 算法进行求解。利用 Jieba 分词工具，我们可以便利的对留言文本进行分词。

2.3 去停用词

在对文本数据的处理中，为了提升处理效率，节省处理时间，将对文本进行一个无用词组的过滤，即去停用词。停用词通常指代的是对于文本的语义贡献度较低的特征词，因此去停用词能够达到一个初步数据提取的目的，提升后续分析效率，减少特征空间维度，去除噪声特征干扰，有利于后续的分析。统计停用词是一项复杂的工作，幸运的是，很多自然语言处理研究组织公开了他们的停用词表，本文则使用了来自百度公司的停用词表。

2.4 文本表示

语言作为人类在进化了几百万年所产生的一种高层的抽象的思维信息表达的工具，具有高度抽象的特征。文本是符号数据，两个词只要字面不同，就难以刻画它们之间的联系，即使是“麦克风”和“话筒”这样的同义词，从字面上也难以看出这两者意思相同。因此，一个合适的文本表示方法是进行自然语言处理任务的关键[3]。

NLP 中最简单，最直观的词表示方法是 One-hot 编码。这种方法把每个词表示为 1*N 的向量，N 是字典 D 中所有元素的个数。在这个 1*N 的向量中，绝大多数元素为 0，只有一个维度的值为 1，这个维度就代表了当前的词。显然，这种词向量是稀疏的，浪费了大量的存储空间，更关键的是，它忽略了词之间的相关性。

词的分布式表示正是为了克服以上缺点。它通过训练神经网络，将词映射成一个固定长度（该长度由用户自定义）的向量，所有的这些向量构成一个词向量空间，进而可以定义每个词之间的距离，并通过距离判断他们之间的相似性。

Word2vec 算法近年来使用非常广泛，且具有多种形态[4]。本文使用的是 Skip- Gram 模型，该模型通过输入的一个单词，来预测上下文。

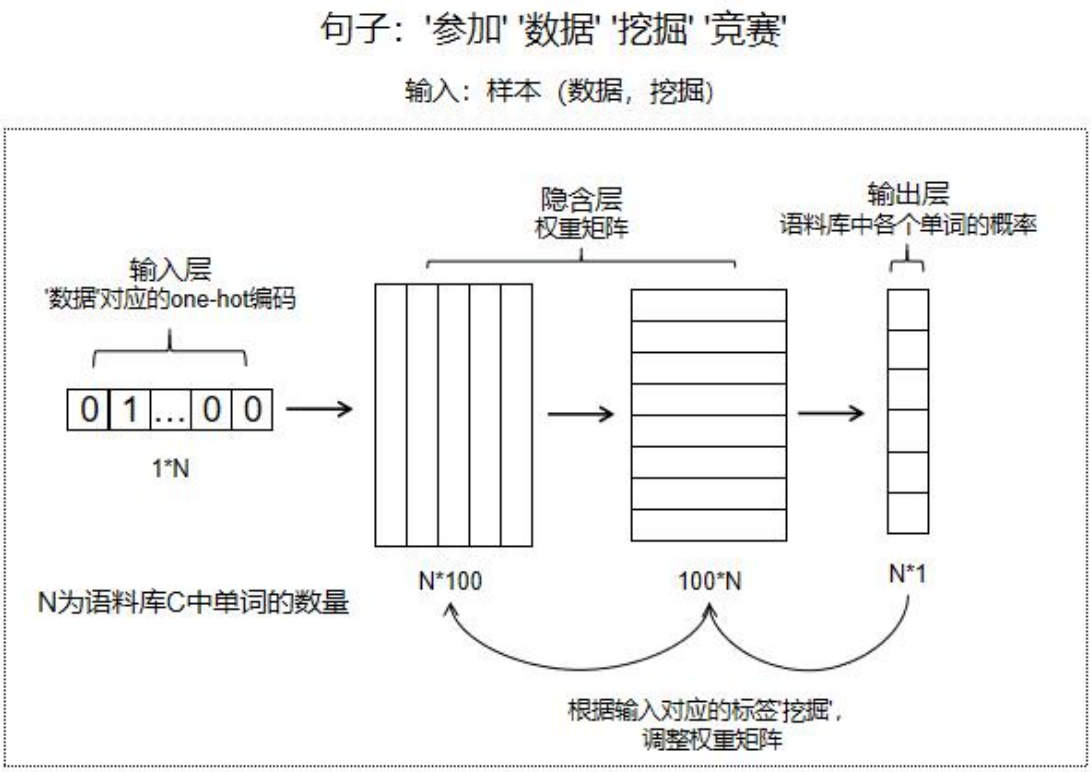


图 3：Skip-Gram 模型工作原理

使用语料对 Skip-Gram 模型进行训练，该模型中的第一个权重矩阵即是所谓的词向量。在得到词向量后，我们可以将句子、段落也转换为向量的形式。一个通常的做法是，对于由单词 $\{x_i\}, i=1,2,\ldots,n$ 组成的文本 sen，sen 对应的文本向量 sen_vec 为各个单词 x_i 对应的词向量 vec_i 的和的平均[5]。

$$sen\_vec = \frac{1}{m}\sum_{i=1}^{m} vec_i \tag{1}$$

3 留言文本分类

文本分类是自然语言处理领域最经典的使用场景之一，根据是否使用深度学习方法将文本分类主要分为一下两个大类[3]：

1）基于传统机器学习的文本分类，如 TF-IDF 文本分类。

2）基于深度学习的文本分类，如 FastText 文本分类，Text-CNN 文本分类

目前，基于深度学习的文本分类是非常流行的技术，但由于附件 2 提供的留言数量有限，各个分类的留言样本数量在 1500 条左右，并不适合使用深度学习进行处理。基于该考虑，文本针对样本不均衡的问题，通过重采样与过采样调整训练样本，将文本转换为向量后，使用 Softmax 回归模型进行多分类，并以准确率，召回率和 F1 值作为评价分类器质量的指标。

3.1 欠采样与过采样

附件 2 提供了 9212 条已完成分类的用户留言数据，对其中各类样本的数量进行统计。

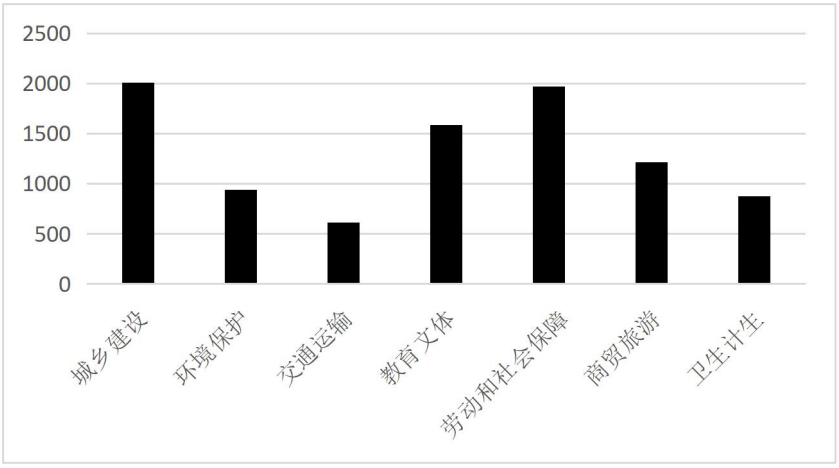


表 1：附件 2 样本数量统计

可见，样本的分布并不均匀，城乡建设、劳动和社会保障、教育文体三类样本数量较多，交通运输、卫生计生、环境保护三类样本明显较少。因此，以同比例从留言中抽取训练样本会导致不同标签的训练样本数量相差较大，在训练时，导致模型产生偏移。

过采样、欠采样是解决样本分布不均匀的方案之一，过采样通过消除占多数的类的样本来平衡类分布，而欠采样相反，通过复制少数类来增加其中的实例数量。当然过采样、欠采样存在对应的问题。欠采样会丢失一部分样本信息，而过采样会增加产生过拟合的可能性[6]。

基于以上分析，对交通运输、卫生计生、环境保护这三类样本进行随机的同

义词替换、改变语序，以实现过采样，由于不是直接对样本进行复制，可以减少产生过拟合的可能性；对于城乡建设、劳动和社会保障、教育文体这三类样本、进行有限度的欠采样。最终使得各个标签的训练样本数量变为 1400 条左右。

3.2 Softmax 回归模型

Softmax 回归模型是 Logistic 回归模型在多分类问题上的推广，对于训练集 $\{(x_1,y_1),\ldots(x_m,y_m)\}$，$y_i\in\{1,2,\ldots,k\}$，样本 x 对应的标签 y 可以取 k 个不同的值[7]。

对于一个输入 x，Softmax 回归模型计算 x 属于每一类的概率 $p(y=j\mid x)$，并选取概率最大的类作为 x 的分类。与逻辑回归类似，定义

$$h(x_i)=\begin{bmatrix} p(y_i=1\mid x_i;\theta)\\ p(y_i=2\mid x_i;\theta)\\ \vdots\\ p(y_i=k\mid x_i;\theta)\end{bmatrix}=\begin{bmatrix} \dfrac{e^{\theta_1^T x_i}}{\sum_{j=1}^{k}e^{\theta_j^T x_i}}\\ \dfrac{e^{\theta_2^T x_i}}{\sum_{j=1}^{k}e^{\theta_j^T x_i}}\\ \vdots\\ \dfrac{e^{\theta_k^T x_i}}{\sum_{j=1}^{k}e^{\theta_j^T x_i}}\end{bmatrix}=\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x_i}}\begin{bmatrix} e^{\theta_1^T x_i}\\ e^{\theta_2^T x_i}\\ \vdots\\ e^{\theta_k^T x_i}\end{bmatrix} \tag{2}$$

其中 $\theta=[\theta_1,\theta_2,\ldots,\theta_k]$ 是参数。定义代价函数

$$J(\theta)=-\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k}\sigma_j(y_i)\log\frac{e^{\theta_j^T x_i}}{\sum_{l=1}^{k}e^{\theta_l^T x_i}}\right] \tag{3}$$

$$\sigma_j(y_i)=\begin{cases}0, & y_i\neq j\\ 1, & y_i=j\end{cases}$$

在 $J(\theta)$ 中，$\sigma_j(y_i)\log\dfrac{e^{\theta_j^T x_i}}{\sum_{l=1}^{k}e^{\theta_l^T x_i}}$ 表示 Softmax 回归模型中样本 x_i 的标签为 y_i 的对数概率，由于我们选择概率最大的类作为对 x 的分类，我们自然希望该项取得最大值。当对每一个样本，该项都取得最大值时，$J(\theta)$ 达到最小值。因此，通过梯度下降法等算法，我们可以优化 Softmax 回归模型中的参数 θ，使代价函数达到最小，以得到最优的分类器。

3.3 模型实现及结果分析

在 64 位 win10 系统下，使用 python3.6 对算法进行实现。和一般的分类不同，群众留言分类由留言主题和留言详情组成，在算法实现中，我们使用了留言

主题和留言详情的拼接。将样本空间中 70%的样本作为训练样本，30%的样本作为测试样本，对 Softmax 模型进行训练，训练结果可见表 2。

	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
准确率	0.75	0.88	0.74	0.86	0.79	0.76	0.79
召回率	0.79	0.88	0.72	0.84	0.86	0.66	0.72
F1 值	0.77	0.88	0.73	0.85	0.83	0.71	0.75

表 2：Softmax 回归分类结果

可见，该模型对大多数类别都可以实现较为准确的分类，但是由于交通运输、商贸旅游、卫生计生三类，样本较少，过采样只能略微提高效果，无法达到和其他类别同等的水平。

4 热点事件挖掘

热点问题挖掘可以细分为等价问题群发现与热度定义排序这两个步骤。等价问题群发现可以作为一个文本匹配问题来考虑，更准确的说，是文本匹配中的文本语义匹配问题。常见的文本语义匹配做法是计算两个文本的相似度，以相似度作为两个文本语义是否一致的依据。而等价问题群发现更为复杂，相似度的计算不仅仅是一对一的，也存在一对多的可能。通过将余弦相似度转换为欧几里得距离，我们可以从聚类的角度解决这一问题。在实现等价问题群发现的基础上，本文充分利用了附件 3 中的信息，将等价问题群大小、时间间隔、点赞数、反对数作为计算问题群热度值的依据，实现热点问题的排序[8]。

4.1 文本距离度量

在文本匹配中，通常以文本向量 A,B 间的余弦相似度作为相似性的度量[9]

$$\cos(A,B)=\frac{A\cdot B}{\|A\|\times\|B\|} \tag{4}$$

显然 $|\cos(A,B)|\le 1$，而 $1-\cos(A,B)$ 则被称为余弦距离。但是在距离的定义中，余弦相似度只满足非负性和对称性，并不满足直递性，因此无法直接用余弦相似度作为距离进行聚类。

A，B 间的欧几里得距离为

$$E(A,B)=\|A-B\|^2=(A-B)^T(A-B)=\|A\|^2+\|B\|^2-2A^TB \tag{5}$$

当对文本向量向量 A,B 进行模的归一化后，A,B 的模等于 1，可以发现

$$\|A-B\|^2=2(1-A^TB)=2\left(1-\frac{A^TB}{\|A\|\times\|B\|}\right)=2(1-\cos(A,B)) \tag{6}$$

可见此时欧几里得距离与余弦距离等价。因此，在将留言文本表示为向量的基础上，对其进行归一化，进而使用欧几里得距离作为度量，进行层次聚类。

4.2 层次聚类实现等价问题群发现

实现等价问题群发现有以下两个主要问题。

1）等价问题群个数未知。

2）群间距离如何定义。

对应这两个问题，本文选择层次聚类算法实现等价问题群的发现。层次聚类算法可根据初始类的个数分为凝聚的层次聚类算法和分裂的层次聚类算法两类，本文要使用的是凝聚的层次聚类算法[10]。

凝聚的层次聚类并没有全局目标函数，也没有局部极小问题或是选择初始点的问题[11]，该特点使得我们可以回避等价问题群个数未知的问题。我们定义两个热点问题群 O，T 之间的距离为两者中元素 o_i，t_j 的最短欧几里得距离，即

$$dis(O,T)=\min(E(o_i,t_j)),o_i\in O,t_j\in T \tag{7}$$

在聚类的开始，每个个体都是一个类，计算类与类之间的距离，将距离最近的类合并，并重新计算类间距离，多次重复，最后所有个体属于一个类。通过设置类间距离的阈值，我们可以提前终止层次聚类，得到需要的问题群。

4.3 热度值定义

群众留言主要具有留言时间 T、点赞数 a、反对数 n 这三个属性。其中，点赞/点踩数反映了浏览者对该留言的看法，而留言时间反映了留言在时间维度上的特征。

我们将某一时段内群众集中反映的某一问题称为热点问题群 W，将 W 中的留言按时间先后顺序记为 $\{W_i\}$。本文假设，当某条留言 W_i 出现后，在一定时间内，群众对问题 W 的关注度会得到提高，并随着时间推移，热度逐渐冷却。新留言 W_{i+1} 的出现会重新提高群众对问题 W 的关注，并与留言 W_i 所产生的热度叠加。该模型可以解释为群众长期、多次反映的某一问题为热点问题，即受影响群

众多，且长时间未解决，群众对此有很大的意见。

基于以上分析，我们设计了政务留言热度模型

$$H_P = \sum_{W_i \in W} hot(W_i)$$
$$hot(W_i) = \left(H_0 + a - n\right) * e^{-k\left(T_{P_{i+1}} - T_P\right)} \tag{8}$$

其中，H_0为初始热度值，k 为的衰减系数，在本文的实现中，我们令 $H_0 = 10$，$k = 0.1$。从图 4 的对比可以看出，该模型较为符合本节中对热度变化的假设。

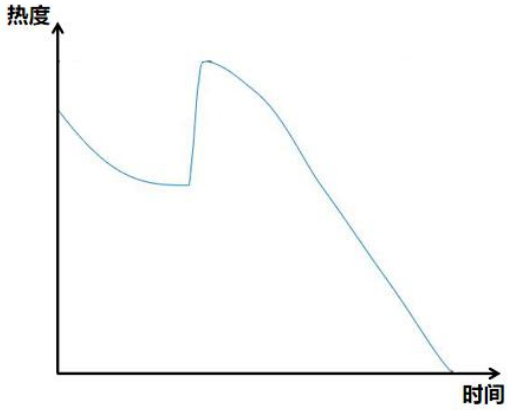


图 4(a):基于本文假设的热度曲线

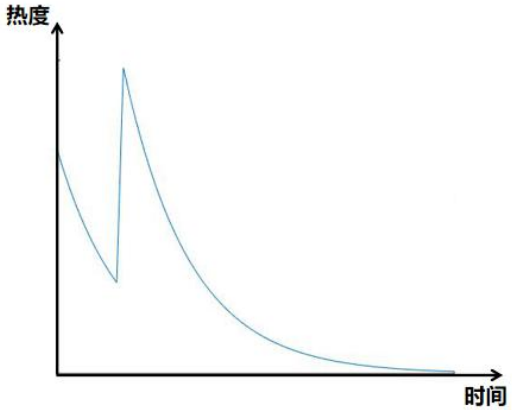


图 4(b):热度模型曲线

4.4 算法实现

在 64 位 win10 系统下，使用 python3.6 对算法进行实现。图 5(a)展示了通过层次聚类得到的前 10 大等价问题群，其中，我们以 0.3 为阈值，用最短距离法计算类间距离，当类间距离大于 0.3 时，停止聚类。在得到等价问题群的基础上，使用 4.3 中的模型，计算得到了对应的热度值，图 5(b)展示了热度值最高的 5 个问题。更为具体的信息请见附件中的《热点问题表》和《热点问题明细表》。

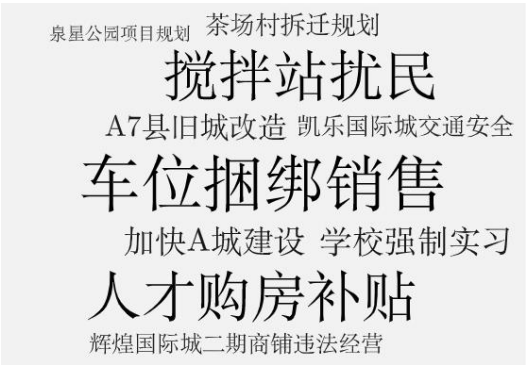


图 5(a):前 10 大等价问题群

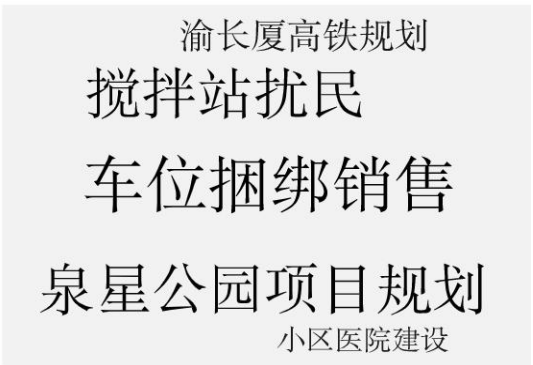


图 5(b):前 5 大热点问题

5 答复意见评价

在自然语言处理任务中，答复意见评价属于自动评分问题。但与常见的作文评分、主题题评分等应用场景不同，自动答复意见评价更为复杂和困难。留言反映了用户在生活中遇到的各种各样的问题，随机性相对其他场景而言更强，很难预先确定某个问题的准确答案所具有的特征。这也就导致从传统角度，即内容上评价答复意见是困难的，基于这一考虑，本文提出从留言内容与答复内容的相关性，答复内容的规范性，答复是否具有可解释性三个角度构建答复意见的评价体系，并尝试实现。同时，考虑到赛题的需求，本文在假设存在专家经验的前提下，对完整性进行了定义，并对答复的完整性进行评分。

5.1 相关性、规范性、可解释性的定义

相关性用于反映答复内容是否对应留言内容，即判断是否存在答非所问的现象。本文使用同公式(4)的余弦相似度来体系这一性质，对于留言 A 和对应的答复 B，相关性函数为

$$cor(A,B)=\begin{cases}0, & \cos(A,B)<\delta \\ 1, & \cos(A,B)\geq\delta\end{cases} \tag{9}$$

其中 1 代表相关，0 代表不相关，δ 为阈值，通过对附件 4 提供的样本的分析，本文认为该值取 0.6 较为合理。

规范性用于评价答复内容是否规范，即是否遵循规范的格式。通过对样本的分析，本文构造了答复内容模板

网友“xxxxxxxx”您好！您的留言已收悉。现将有关情况回复如下：

……

感谢您对我们工作的支持、理解与监督！xxxx 年 x 月 x 日。

通过对该模板进行匹配，我们可以计算答复文本 B 对关键点的匹配数量 L_B。

具有可解释性的答复往往会通过引用相关条文（如政府文件，法律法规），证明自己的正确性，即这个答复的依据是什么，以此增强说服力。通过对答复内容的识别，我们可以判断答复文本 B 中是否引用了相关条文。

$$Q_B=\begin{cases}0, & \text{未引用条文} \\ 1, & \text{引用了条文}\end{cases} \tag{10}$$

5.2 评价体系

在 5.1 中提到的三项指标中，相关性是答复的基本要求，而完整性反映了答复人员的工作态度，少量缺失可以允许，相关条文的引用增加了答复的说服力，

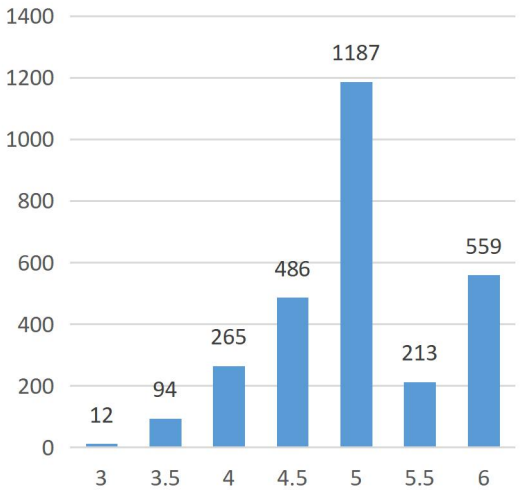
综合评分公式如下

$$score(A,B)=S_0\cdot cor(A,B)+S_1\cdot L_B+S_2\cdot Q_B \tag{11}$$

其中，S_0代表基础得分，本文中取 3；S_1代表规范得分，本文中取 0.5；S_2代表奖励分，本文中取 1。

5.3 算法实现

在 64 位 win10 系统下，使用 python3.6 对算法进行实现。可见约 42%的答复意见得到了 5 分，得到 6 分的答复意见约占总体的 20%，96%以上的答复意见都达到 4 分以上。同时，该分布也与正态分布接近，说明提出模型较为科学。



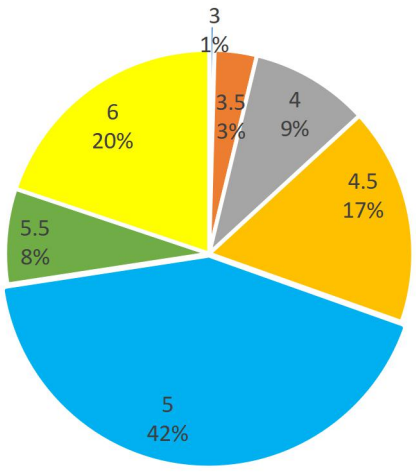


表 3：评分结果可视化

5.4 基于专家经验的完整性评分

一条群众留言中可能包含多个问题，对应的有效答复应对每一个问题都做出了相应的回答。因此，本文将这种特征定义为答复的完整性。基于大量的专家经验，即一个留言问题的最佳答复，我们可以将一对<留言，答复>定义为一个主题，进而构造主题空间 L，在此基础上，训练 Labeled-LDA 主题模型[12]。通过 Labeled-LDA 主题模型，可以得到一个句子对应的主题。将留言文本和答复文本进行分句，输入 Labeled-LDA 主题模型，可以得到两者的主题分布，通过比较二者的主题是否一致，来对答复的完整性进行一个评分，最终得到答复的完整性方面的评定。

6 总结

本文本文首先对原始数据进行预处理，并通过 Jieba 分词和 word2vec 将原始文本向量化。针对样本不均衡的问题，通过重采样与过采样调整训练样本，之后，使用附件 2 提供的样本 Softmax 回归模型进行训练，并以准确率，召回率和 F1 值作为评价 Softmax 回归模型的质量的指标。从结果中可以看出：该模型对大多数类别都可以实现较为准确的分类，其 F1 值在 0.7-0.9 之间，但是由于交通运输、商贸旅游、卫生计生三类，样本较少，过采样也只能略微提高效果，召回率较低，总体而言，分类器效果良好。

本文进一步地在文本向量化的基础上，将余弦相似度转换为欧几里得距离，通过计算向量之间的欧几里得距离，利用层次聚类方法对群众留言内容进行归类。同时基于留言时间、点赞数、反对数这三个属性建立了新的政务留言热度模型，最终得到排名前五的热点问题，他们分别是：伊景园车位捆绑销售问、搅拌站扰民问题、泉星公园项目规划问题、渝长厦高铁规划路线问题、诺亚山林小区医院设置问题。

最后，本文提出从留言内容与答复内容的相关性，答复内容的完整性，答复是否权威三个角度构建答复意见的评价体系，从结果可以得到：约 42%的答复意见得到了 5 分，得到 6 分的答复意见约占总体的 20%，96%以上的答复意见都达到 4 分以上。同时，该分布也与正态分布接近，说明提出模型较为科学。

7 参考文献

[1] 程显毅, 朱倩. 文本挖掘原理[M]. 科学出版社, 2010.

[2] 汪文妃, 徐豪杰, 杨文珍, 等. 中文分词算法研究综述[J]. 成组技术与生产现代化, 2018, 35(03):5-12.

[3] 闫琰. 基于深度学习的文本表示与分类方法研究[D]. 2016.

[4] 罗钰敏, 刘丹, 尹凯,等. 加权平均 Word2Vec 实体对齐方法[J]. 计算机工程与设计, 2019, 040(007):1927-1933.

[5] Mikolov T , Chen K , Corrado G , et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer ence, 2013.

[6] StevenBird, EwanKlein, EdwardLoper. Python 自然语言处理[M]. 人民邮电出版社, 2014.

[7] 刘亚冲，唐智灵. 基于 Softmax 回归的通信辐射源特征分类识别方法[J]. 计算机工程，2018,2（44）：98-102.

[8] 贾威. 基于武汉城市留言板的舆情热点监控研究[D]. 华中师范大学, 2019.

[9] 陈二静, 姜恩波. 文本相似度计算方法研究综述[J] . 数据分析与知识发现, 2017 , 1(6): 2-3.

[10] 叶枫, 江永省. 基于聚类融合欠采样的不平衡分类方法[J].计算机应用与软件，2020,1（37）：292-297.

[11] Chrupała, Grzegorz. "Hierarchical clustering of word class distributions." Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure. Association for Computational Linguistics, 2012.

[12] Ramage, Daniel, et al. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009.