“智慧政务”中的文本挖掘应用

摘要

近年来，随着网络问政平台逐渐成为政府了解民意、汇聚民智、凝聚民气的重要渠道，各类文本数据量急剧攀升。因此，运用文本挖掘技术对群众问政留言记录的研究具有重大意义。

对于问题 1，利用 jieba 中文分词工具将留言内容进行分词，利用 word2vec 方法将处理好的文本转化为词向量，然后通过 TF-IDF 算法计算得到权重，从词向量中选取出特征项的集合。最后利用 SVM 算法建立了关于留言内容的一级标签分类模型。

对于问题 2，热度评价采用留言内容相似数、点赞数、反对数这三个指标进行评价，得到热度指数。其中留言内容相似数算法如下：对留言内容进行分词、去停用词，利用 Tf-idf 对分词结果提取前 20 个关键词，由 corpora.Dictonary 建立词典，同时利用 doc2bow 将提取关键词结果转化为稀疏向量，形成语料，通过 TF-idf 模型训练，从而求得留言内容之间的相似度。确定中文命名实体识别对象为地点、机构和人名，对留言主题采用哈工大 LTP 工具。

对于问题 3，从结构化特征和文本特征两个角度构建答复意见指标评价体系，利用九个指标来刻画答复意见的质量，最后利用熵值法计算答复意见的综合得分。

关键词：jieba 分词，TF-IDF 算法，SVM，熵，相似度，命名实体识别

ABSTRACT

In recent years, as the online political platform has gradually become an important channel for the government to understand public opinion, gather people's wisdom and build up people's spirit, the amount of various text data has risen sharply. Therefore, the application of text mining technology to the study of the people's political message records is of great significance.

For problem 1, jieba Chinese word segmentation tool is used to segment the message content, word2vec method is used to convert the processed text into a word vector, and then tf-idf algorithm is used to calculate the weight, and the set of feature items is selected from the word vector. At last, the first level tag classification model of message content is established by SVM algorithm.

For question 2, the heat index was obtained by evaluating the similarity number, thumb up number and opposition number of message contents. The message content similarity number algorithm is as follows: participles, the content of the message to stop words, using Tf - idf 20 keywords before extraction, with the result of segmentation by corpora. Dictonary dictionary, at the same time using doc2bow extract keyword results into a sparse vector, form the corpus, through training, Tf - idf model to obtain the similarity between the message content. The Chinese named entity is identified as place, institution and name, and the LTP tool is used for the message subject.

For question 3, the evaluation system of response opinion index is constructed from the perspectives of structural feature and textual feature, and nine indicators are used to describe the quality of response opinion. Finally, the entropy method is used to calculate the comprehensive score of response opinion.

Key words: jieba word segmentation, tf-idf algorithm, SVM, entropy, similarity, named entity recognition

目录

1 挖掘目标

（一） 根据附件 1 给出的数据和附件 2 的分类标签，建立关于留言内容的一级标签分类模型，并使用 F-Score 对分类方法进行评价。

（二） 根据附件 3 所提供的数据，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，定义合理的热度评价指标，并给出评价结果。

（三） 针对附件 4 相关部门对留言的答复意见数据，对答复意见的质量给出一套评价方案。

2 群众留言分类

对于问题 1，首先根据所给的数据，利用 jieba 中文分词工具将留言内容进行分词，并使用哈工大停用词表去除留言内容中不需要的词汇及符号。其次，在数据预处理完后，利用 word2vec 方法将文本转化为词向量，然后通过 TF-IDF 算法计算得到权重，从词向量中选取出特征项的集合。最后利用 SVM 算法建立了关于留言内容的一级标签分类模型。

2.1 流程图

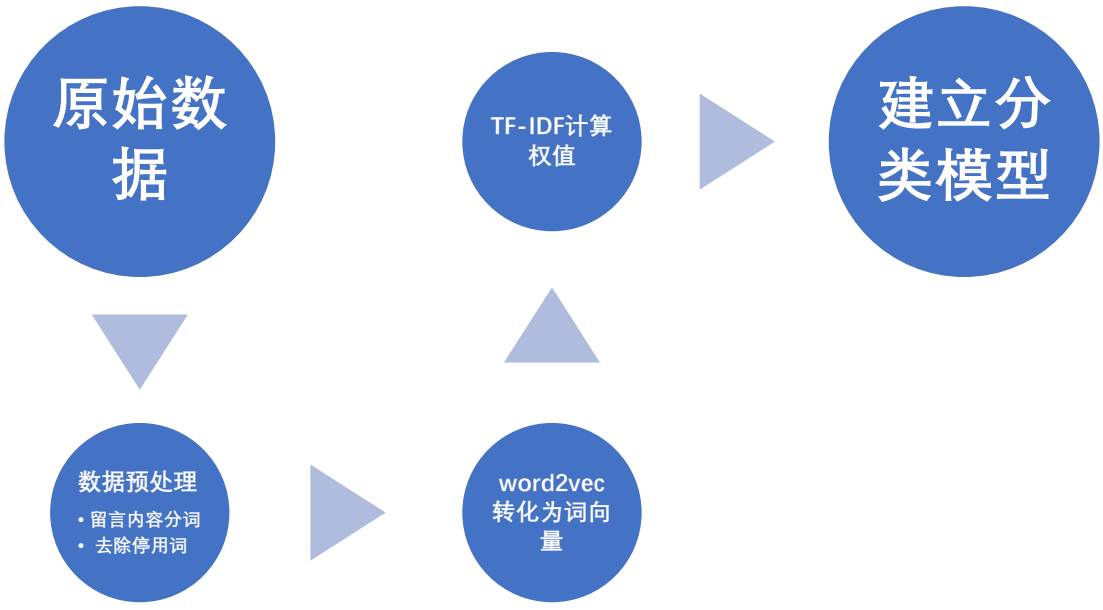

图 2-1 群众留言分类流程图

2.2 数据预处理

在对留言内容进行分类之前，第一步需要做的就是对文本进行预处理，通过这一步，可以有效的将文本内容结构化，减少不必要的维度以及噪声，提高分类的效果。此过程主要包括分词以及去除停用词。

2.2.1 留言内容分词

分词指的是通过一定的规则把连续的字符序列划分为间隔的基本单位，一般来说分词指的是对中文文本词语的划分。对于中文文本来说，单词之间都是连续的，因此必须进行分词处理，才能把词项与词项以空格分割开来。因此利用 python 中的中文分词包 jieba 对留言内容进行分词处理。

2.2.2 去除停用词

在对留言内容进行预处理的过程中，为了提高效率和节省存储空间，对于分类结果没有实际意义的符号及词语都应剔除掉。比如“的”，“而”，“是”等词，还有汉语标点符号、特殊符号以及数字串等。因此选用哈工大停用词表对这类符号及词语进行剔除。

2.3 Word2vec 训练词向量

Word2vec 是 Google 于 2013 年开源的一款训练词向量的高效工具[2]，其主要优点包括：

（1）在 word2vec 中，输入第一层时，无需将文本词汇对应的词向量进行排序，而是将这些向量相加，因此计算效率有所提高。

（2）模型中不含有隐藏层，在减少计算量的同时，结果也不错。

（3）Word2vec 含有两种训练模型，分别为 CBOW 和 Skip-gram，这两种模型都能使 Word2vec 模型具有充分考虑上下文的特点。

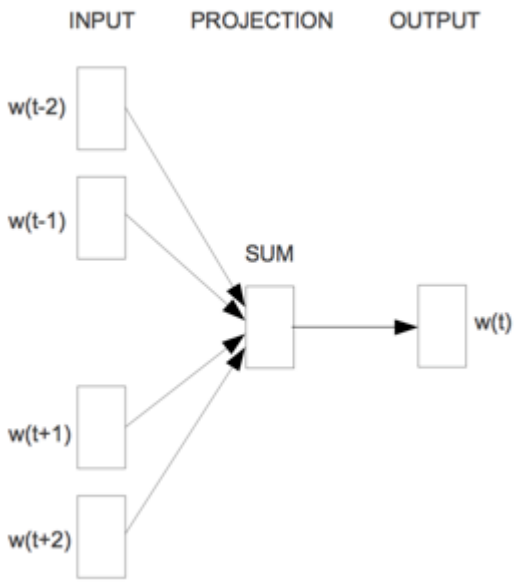

图 2-2 CBOW 模型

CBOW（Continuous Bag-of-Word Model）又称连续词袋模型，是一个三层神经网络。CBOW 模型的公式为：

$$\mathrm{P}(w_t|w_{t-k},\cdots w_{t-1},w_{t+1,\cdots}w_{t+k})$$

公式中，w_t 为一个留言内容中一个词汇，通过和w_t 相邻的窗口大小为 k 的词来预测w_t 出现的概率。

该模型的特点是输入为已知的上下文，输出对当前文本词汇的预测。而 skip-gram 和 CBOW 正好相反，使用单个的词汇作为输入，经过训练然后输出目标上下文。

（4）针对层次 softmax，word2vec 是根据单词的词频使用霍夫曼（huffman）编码，通过霍夫曼编码可以利用词汇的词频特征进行去层次化。

Word2vec 常用的网络结构图如下图所示：

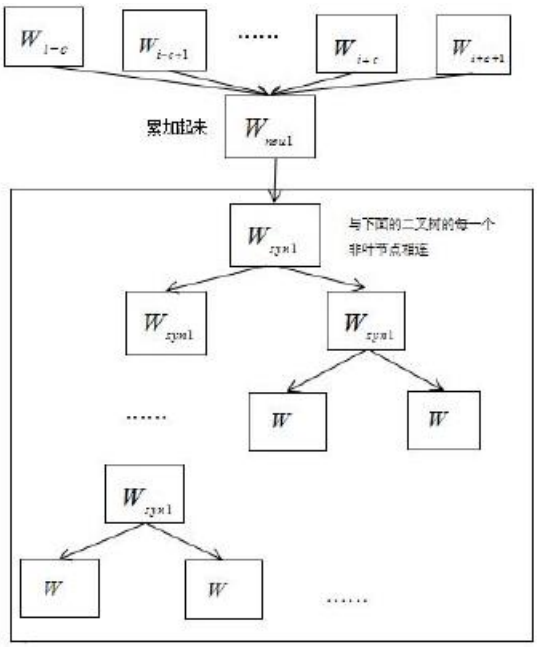

图 2-3 word2vec 的网络结构

2.4 TF-IDF 算法

在对留言内容预处理结束后，需要计算这些词语的权重，并转化为权重向量，此处采用 TF-IDF 算法，其具体原理主要为：

词频权重（Term Freqency，TF）表示某一个特征项在指定文档中的出现次数，计算方式如下：

$$TF = TF_{ij}$$

公式中，TF_{ij}表示文本集中第 j 个特征项在文本 i 中一共出现的次数。

逆文档频率权重（Inverse Document Frequency, IDF）是一种词语与文档之间权重的计算方法，其计算公式如下：

$$IDF = \log(\frac{N}{n_j + 0.01})$$

其中 N 代表中的文本总数量，n_j 代表包含第 j 个特征项的文本数量。该方法的思想是如果一个特征项在整个文本集中的出现频率越高，那么该特征项所包含的类别信息就越低。

TF-IDF 权重是将 TF 和 IDF 两种不同的权重计算方法相结合的权重方法。它的计算公式为：

$$TF - IDF = tf_{ij} \times idf_{ij}$$

实际分析得出 TF-IDF 的值与一个词汇在留言内容中出现的次数成正比，即该词的重要性越大，则 TF-IDF 值越高。

2.5 SVM 支持向量机算法

SVM（Suport Vector Machine）[25]是由 Boser 等人提出的一种基于统计学习理论的识别方法，它的主要特点有：

（1）将学习的问题归纳为凸二次的规划问题，得到了理论上的最优解，因此解决了在神经网络中经常出现的局部极值问题；

（2）引入了核函数从而可以使得解平面从线性到非线性，使用一种非线性的映射将训练数据映射到高维度的空间，通过在高维度空间上寻找最佳的超平面来完成分类；

（3）能够解决数据的维度问题，具有直观的几何解释和简明的数学形式，需要人为进行调试的参数比较少，也便于使用。

（4）由于 SVM 对于稀疏性和特征相关性不太敏感，并且在解决高维度的数据问题时优势明显，这也使得 SVM 被视作解决文本分类问题的最佳方案之一。SVM 可分为线性支持向量机和核支持向量机，前者针对线性分类问题，后者针对非线性分类问题。

2.5.1 线性支持向量机原理

（1）间隔最大化

SVM 数据线性可分的情况，即可以找到一条直线，把两类样本区分开来，并且可以找到更多的分离直线，当样本的数据达到 n 维时，这条直线就成为了一个超平面，如图所示：

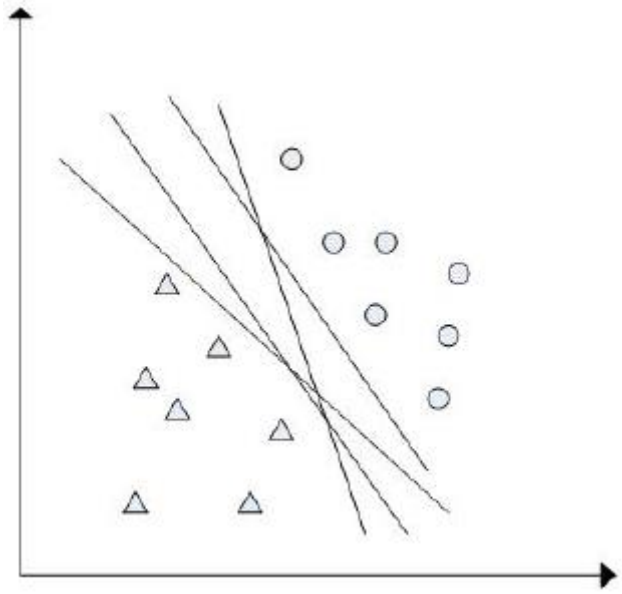

图 2-4 线性可分数据

其中 SVM 所要做的就是构建一个分类超平面来解决该问题，即 $W \cdot X + b = 0$。最大可能的间隔分离两类样本，从而使得分类的误差最小，构造的这个超平面能使两个不同类别之间的边缘（Margin）最大，两个间隔边界的距离 $d = \frac{2}{||\omega||}$ 被定义为边缘。如图所示：

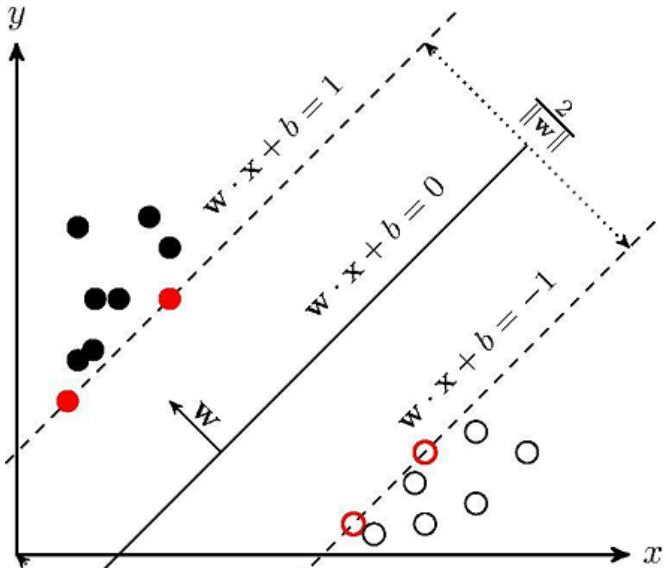

图 2-5 支持向量

（2）松弛变量

在大多数情况下，由于数据中包含噪音，所以很难用一条直线将两个样本完整的区分开，此时的目标函数约束条件则无法满足。因此引入松弛变量，对于每一个样本点赋予一个松弛变量的值：如果该点落在最大边缘超平面正确的一侧，

则松弛变量为 0，即$\xi = 0$。此时的目标函数为：

$$\min_{w,b}(\frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\xi_i$$

$$s.t. \quad y_i((w)^T x^i + b) - 1 \geq \xi_i, \quad i = 1,2,\cdots,n$$

$$\xi_i \geq 0$$

其中ξ_i为第 i 个样本的松弛变量，n 为样本个数。

（3）惩罚系数 C

引入松弛变量之后，目标函数中新加入$C\sum_{i=1}^{n}\xi_i$，即所有样本松弛变量的和乘以惩罚系数 C，表示模型对错误分类的容忍度。当 C 取较大值时，分类器对错误的容忍度较低，从而会产生较高的训练集正确率。当 C 取较小值时，则相反。

2.5.2 核支持向量机

（1）核函数

核支持向量机的思想是将非线性分类问题通过映射到高维空间，使之转化为线性分类问题。随后对于在高维空间下的数据，使用线性支持向量机进行分类，此时线性支持向量机对偶问题的目标函数为：

$$\boldsymbol{Max}_{\boldsymbol{\alpha}}[\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}}\boldsymbol{\alpha_i} - \frac{\boldsymbol{1}}{\boldsymbol{2}}\sum_{\boldsymbol{i,j=1}}^{\boldsymbol{n}}\boldsymbol{y_i y_j \alpha_i \alpha_j \phi(x_i)^T \phi(x_j)}$$

（2）常用的核函数

在实际的应用中，通常使用的核函数包括：线性核函数，多项式核函数，Sigmoid 核函数，高斯核函数等，本文采用的为高斯核函数：

$$k(x,y) = \exp(-\gamma||x-y||^2$$

这种核函数是一种局部性强的核函数，无论大样本还是小样本都有比较好的性能，而且其相对于多项式核函数参数要少。

（3）分布系数$\boldsymbol{\gamma}$值

γ值是支持向量机中重要的一个参数，本文使用到的高斯核函数中也包含γ值。γ值决定了原始数据映射到高维之后，在高维特征空间中的分布情况。

2.6 建立文本分类模型

2.6.1 分类性能评估

目前对于分类性能的评估测量指标主要有查全率（Recall）、查准率(Precision)以及 F-Score 测量（F-Measure）和精确率(Accuracy)。查全率和查准率主要用来评价算法特征提取项的好坏，而 F-Score 测量值是查全率和查准率结合起来的一种评价方法：

$$\mathrm{F-Score}=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$$

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

2.6.2 参数的确定

在对原始留言内容进行预处理之后，通过 word2vec 将文本转换为词向量，并使用 TF-IDF 进行加权处理之后，将其结果作为 SVM 的输入，从而进行样本的训练和预测。

其中 word2vec 对经过预处理之后的留言内容进行训练，word2vec 主要的参数设置如下表所示：

表 2-1 word2vec 的训练参数

参数	说明	取值
size	词向量的维数	200
window	窗口大小	5
hs	是否采用 softmax 体系	1
min-count	词语出现的最小阈值	5

之后确定 svm 的参数，如下表所示：

表 2-2 SVM 的训练参数

参数	取值
核函数	高斯核函数
惩罚系数 C	10
分布系数γ	0.1

2.7 分类结果

将全部数据按照 9：1 的比例分为训练集与测试集。使用 python 得到最

后结果，F-Score=0.812。

3 热点问题挖掘

3.1 流程图

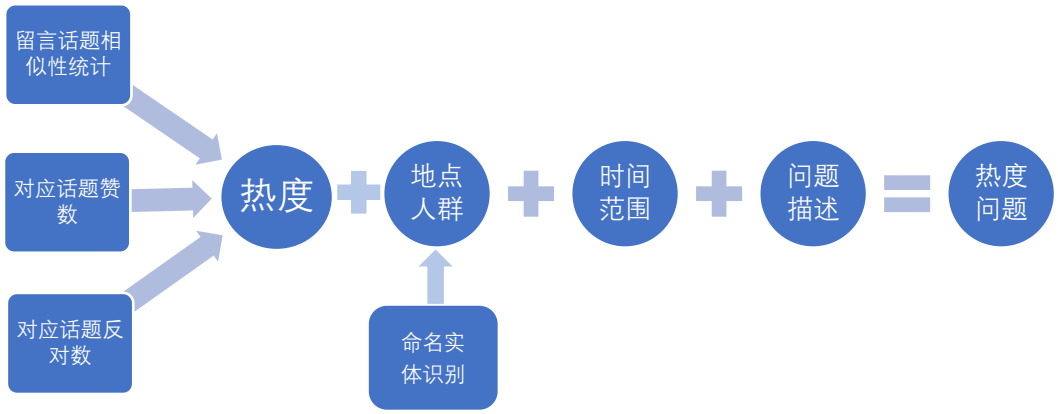

图 3-1 热点挖掘流程图

3.2 数据预处理

在对留言信息进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。

（1）数据清洗。查看是否有空值，附件 3‘留言详情’和‘留言主题’没有缺失值，有效数据量 4326 条。同时，移除数字、字母，汉字以外的所有符号，为接下来的分词做准备。

（2）分词、去停用词。这里采用 python 的中文分词包 jieba 进行分词，基于前缀词典实现词图扫描，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，较好实现中文分词。值得一提的是，本次的数据是脱敏之后的数据，为达到后面识别的准确性，此处我们创建一个自定义词典，收录一些数据中的特定词语（如：A 市、B1 区等），完整词典见 newdic1.txt 文件。在分词同时，利用‘哈工大停用词’对分词结果去停用词。

（3）关键词提取。由于‘留言详情’文本通常较长，分词后数量较多，通过关键词提取可以降低维度的同时保留文本主题和主要意思。采用无监督学习 TF-IDF 算法，抽取每个留言详情中的前 20 个关键词作为文本相似性计算的基础。

3.3 留言相似性统计

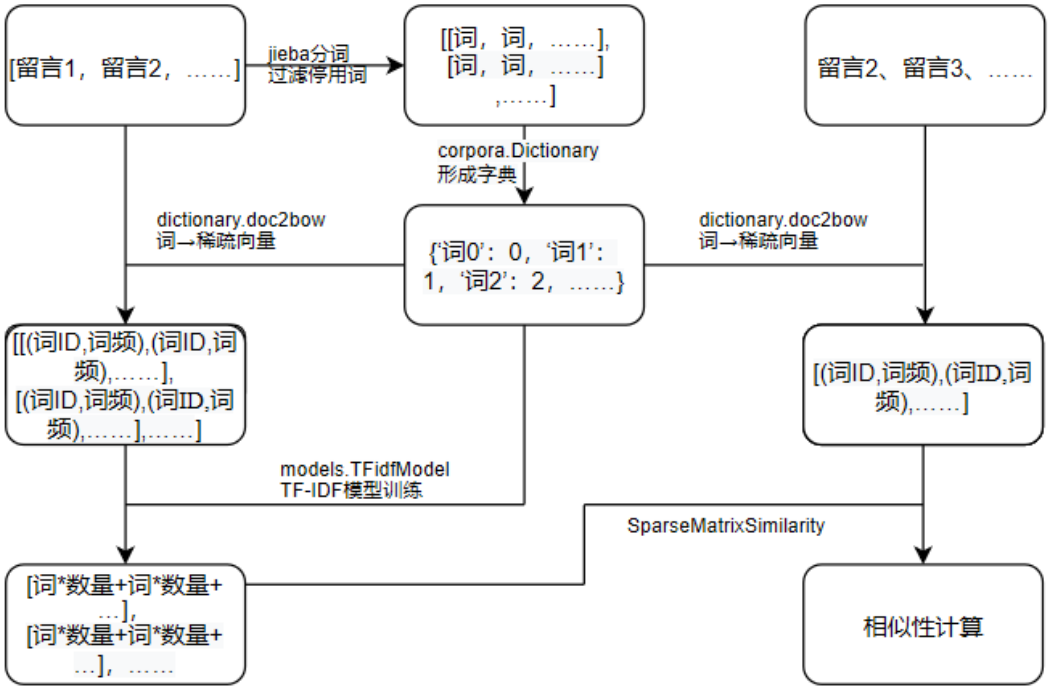

图 3-2 留言相似性统计流程

3.3.1 Tf-idf 算法

这里采用 TF-IDF 算法，提取分词过后的留言关键词，在相似计算中把留言详情信息转换为权重向量。TF-IDF 算法的具体原理如下：

（1）计算词频 TF

$$词频(TF) = 某个词在文章中的出现次数$$

（2）计算逆文档频率 IDF

在语料库（corpus）的基础上，模拟语言的使用环境。

$$逆文档频率(IDF) = \log(\frac{语料库的文档总数}{包含该词的文档数 + 1})$$

（3）计算 TF-IDF

$$\text{TF} - \text{IDF} = \text{词频（TF）} \times \text{逆文档频率（IDF）}$$

计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的职位描述表中文本的关键词，同时 TF-IDF 考虑了词的重要性，计算出的相似性效果更有价值。

3.3.2 doc2bow 原理

本文通过 doc2bow 将分词结果转化为稀疏向量，形成语料。Doc2Bow 是 Gensim 中封装的一个方法，主要用于实现 Bow 模型。BoW 模型原理：Bag-of-words model (BoW model) 最早出现在自然语言处理（Natural Language Processing）和信息检索（Information Retrieval）领域.。该模型忽略掉文本的语法和语序等要素，将其仅仅看作是若干个词汇的集合，文档中每个单词的出现都是独立的。在这种模型中，文本（段落或者文档）被看作是无序的词汇集合，忽略语法甚至是单词的顺序。函数 doc2bow 只是计算每个不同词汇的出现次数，将词汇转换为整数词汇 id，并将结果作为一个词袋（bag-of-words）——一个稀疏向量返回，形式为（ word_id1，word_count1），（ word_id2，word_count2），（ word_id3，word_count3）。

3.3.3 相似性算法

具体算法如下：

（1）这里使用 python 中 gensim 包的 corpora 模块，将提取的关键词形成的二维数组生成词典；

（2）将二维数组通过 doc2bow 稀疏向量，形成语料库；

（3）采用 LsiModel 模型算法，将语料库计算出 Tfidf 值；

（4）获取词典 token2id 的特征数；

（5）计算稀疏矩阵相似度，建立一个索引；

（6）逐行读取‘留言详情’数据，通过 doc2bow 计算测试数据的稀疏向量；

（7）求得测试数据与样本数据的相似度，得到留言文本的相似性统计。

3.4 热度评价算法

热点问题的及时发现，有助于相关部门进行有针对性地处理，从而提升服务的效率。就目前对话题热度计算的研究，主要都是从话题的媒体关注度和用户关注度这两方面考虑的，从附件 3 数据来看，问题 2 的热度研究属于用户关注度，用户关注度则是从舆情数据的接收方来考察话题的热度，其主要包括浏览数、评论数等一些用户的行为信息[3]。同时，同一话题下留言具有相似性的特征，结合留言的相似性，本文热度算法如下：

$$热度 = 留言话题相似数 + 对应话题赞成数 - 对应话题反对数$$

3.5 中文命名实体识别

某一时间段内群众集中反映特定地点或特定人群的问题称为热点问题。问题二则是要求对热点问题进行挖掘，故此处就需要对留言进行命名实体识别，确定实体的类别为地点、人名、机构名。通过对原始数据的探查，“留言主题”中绝大部分都涉及到命名实体，而“留言详情”中会涉及大量的命名实体（如：地铁线路的走向问题）。为更好的凸显识别效果，选择对“留言主题”进行识别。

3.5.1 实体识别

命名实体识别命名实体识别（Named Entity Recognition ，NER）：指识别文本中具有特定意义的实体，主要涉及人名、地名、机构名、专有名词等。现在做命名实体识别的方法较多，例如：NLTP、StanfordNLP、哈工大的 LTP 等。

本文根据方法所能识别的 NE 种类以及识别模型标注的形式，选择 LTP 对分词结果进行。LTP 的 NER 识别模型采用的是 O-S-B-I-E 标注形式。标记含义见下表。

表 3-1 LTP 的 NER 模块能识别的三种 NE

标记	含义
Ns	地点
Nh	人名
Ni	机构名

表 3-2 标注形式的含义

标记	含义
O	该词不是 NE
S	该词单独构成一个 NE

B	该词为一个 NE 的开始
I	该词为一个 NE 的中间
E	该词为一个 NE 的结尾

3.5.2 输出结果

调用哈工大的 pyltp 包，对 jieba 分词后的文本数据进行识别，输出结果如图。

```
['A4区,ns,B-Ni', '君悦,nz,I-Ni', '幼儿园,n,E-Ni', '直,d,O', '改为,v,O', '普惠,nh,S-Nh', '公办,v,O']
['A市,ns,S-Ns', '垃圾,n,O', '分类,v,O', '上海,ns,S-Ns', '罚款,v,O']
['希望,v,O', 'A市,ns,B-Ns', '金桥,ns,E-Ns', '国际,n,O', '设,v,O', '地铁站,n,O']
['A市,ns,B-Ni', '鹏,nz,I-Ni', '基诺,nz,I-Ni', '亚,j,I-Ni', '山林,n,I-Ni', '爱尔,nz,I-Ni', '眼科医院,n,E-Ni', '后花园,n,O']
['A市,ns,S-Ns', '2017,m,O', '年,q,O', '出租车,n,O', '油补,v,O', '发放,v,O']
['A市,ns,S-Ns', '尿毒症,n,O', '病人,n,O', '医保,n,O']
['A市,ns,B-Ni', '科金域,n,I-Ni', '蓝湾,ns,I-Ni', '新房,n,E-Ni', '质量,n,O', '堪忧,v,O']
```

图 3-3 实体识别输出结果（部分）

图 3-3 可以看到，“A4 区君悦幼儿园”识别结果为：Ni（机构名），“A 市金桥”识别结果为：Ns（地点）。最终识别完成率为 98.4%，剩下的 1.6%的“留言主题”不涉及命名实体。建议政府网站的留言页面在“留言主题”栏设置特殊的格式，这样会提高后期处理数据的准确性和效率。

3.6 结果分析

分别统计同一类相似留言下的赞成数、反对数，得到话题热度的统计：

表 3-3 热点问题部分结果表

热度排名	热度指数(%)	问题 ID	时间范围	地点	问题描述
1	98.22	214238	2019/01/08 至 2019/07/08	A4 区公安派出所	请问 A4 区公安派出所对 58 车货一案办案的进度如何了
2	86.91	208636	2019/08/19 至 2019/08/19	A 市 A5 区汇金路五矿 K9 县	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
3	72.78	223297	2019/04/11 至 2019/04/11	A 市金毛湾	反映 A 市金毛湾配套入学的问题
4	27.79	191951	2019/08/23 至 2019/09/05	A4 区绿地海外滩小区	A4 区绿地海外滩小区距渝长厦高铁太近了
5	10.02	193091	2019/06/19 至 2019/06/19	A 市	A 市富绿物业丽发新城强行断业主家水

表 3-3 给出排名前 5 的热点问题，热点话题全部排名在“热点问题排名.xls”可详见。相应热点问题对应的留言信息，在“热点问题留言明细表.xls”中可详见。

4 答复意见质量的评价方案

4.1 评价流程图

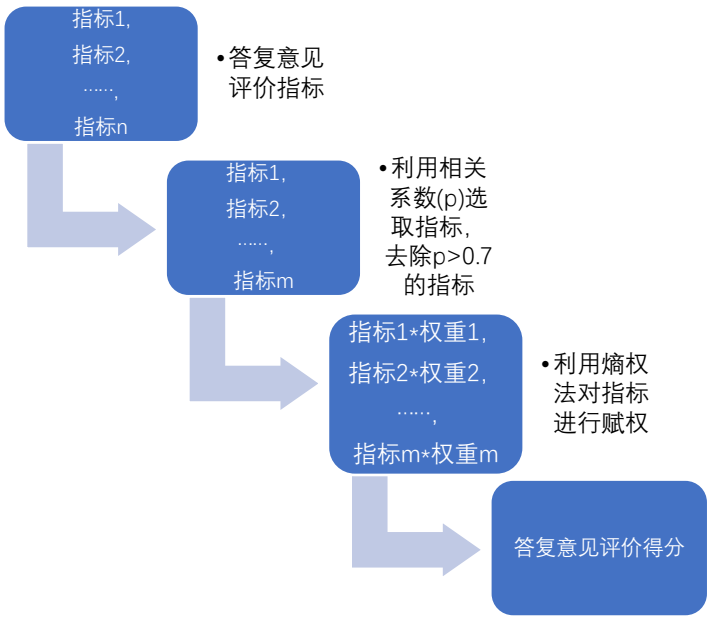

图 4-1 答复意见评价流程图

4.2 评价指标体系的建立

4.2.1 结构化特征挖掘

主要包括能够直观反映答复意见属性的特点，本文包括文本长度、标点符号占比、平均句长、是否含有参考文件、格式是否良好、是否有外部链接以及答复时间跨度。

表 4-1 特征变量说明

特征变量	特征类型	特征说明
文本长度	数值型	文本长度与信息量正相关[4]
标点符号占比	数值型	标点符号占比过大，其所包含的有价值信息就相对不足[5]
平均句长	数值型	长句和短句在表达力方面存在显著差异[6]
是否含有参考文件	二值变量	推断出答复意见中是否有相关政府规定
格式是否良好	二值变量	是否包含尊称，感谢词等

是否有外部链接	二值变量	是否提供电话，网站，微信公众号等
答复时间跨度	数值型	推断答复的及时性

结构化特征可以很好的解释说明答复意见的规范性、完整性、及时性、可解释性等特点。

4.2.2 文本特征

文本特征是指在文本中难以直观表达或统计出的特征，能在一定程度上反映文本质量。因此，文本分析是判断答案质量的一条有效途径。主要包括文本熵和留言内容与答复意见的主题相似度。

4.3 文本熵

4.3.1 信息熵

“熵”的概念由德国物理学家 CLAUSIUS K 于 1854 提出，用来阐明热力学第二定律。熵理论经过近百年的研究，逐渐拓展到其他科学领域，较为经典的有：通信领域中由香农提出的香农信息熵。在信息论中，信息熵被用来度量信息量，信息的不确定度越大，熵值越大。信息熵还可以作为系统混乱程度的度量，一个系统越是混乱，信息熵就越大；相反一个系统越是稳定，熵值越小。

定义：一个随机变量 X 的熵$H(x)$定义为：

$$H(x) = -\sum_{x} p(x)\log p(x)$$

式中对数一般取 2 为底，单位为比特。但是，也可以取其它对数底，采用其它相应的单位，它们间可用换底公式换算。一个随机变量 X 的熵$H(x)$是该列分布$p(x)$的函数，它衡量了包含在 X 中的平均信息量[7]，依据此公式计算中文文本的信息熵。

4.3.2 统计语言模型

假定 S 表示某一个有意义的句子，由一连串特定顺序排列的词 w1,w2,...wn 组成，n 为句子的长度。现在想知道 S 在文本中出现的可能性，即 P(S)。此时需要有个模型来估算，不妨把 P(S)展开表示为 P(S)=P(w1,w2,...,wn)。利用条件概率的公式，S 这个序列出现的概率等于每一个词出现的条件概率相乘，于是 P(w1,w2,...,wn)可展开为：

P(w1,w2,...wn)=P(w1)P(w2|w1)P(w3|w1,w2)...P(wn|w1,w2,...,wn-1)

其中 P(w1)表示第一个词 w1 出现的概率；P(w2|w1)是在已知第一个词的前提下，第二个词出现的概率[]；以此类推。当句子长度过长时，P(wn|w1,w2,...,wn-1)的可能性太多，无法估算，因此本文将其简化，取 N=1，每个词出现的概率与其他词无关的一元模型，对应 S 的概率变为：

P(S)=P(w1)P(w2)P(w3)...P(wi)...P(wn)

4.3.3 中文语料预处理

本文是基于词的统计语言模型，因此在进行挖掘分析之前，需要对答复意见句子进行分词，这里采用 jieba 中文分词工具对句子进行分词。由于一元模型不需要考虑上下文关系，直接移除文本中除数字、字母、汉字之外的所有符号，留下中文字符和 utf-8 编码下字节数为 1 的标点符号，去掉这些符号后再进行繁简转换和分词，得到所需要的语料库。

4.3.4 词频统计

对‘答复意见’列的数据逐行读取，每一个答复意见形成一个语料库，统计每个词在语料库中出现的频数，得到词频表。

4.3.5 计算文本熵

一元模型的信息熵计算公式为：

$$H(x) = -\sum_{x \in X} P(x) log P(x)$$

其中$P(x)$近似等于每个词在语料库中出现的频率，本文对数底数取 2。

4.4 问题与答复的相似性统计

4.4.1 流程图

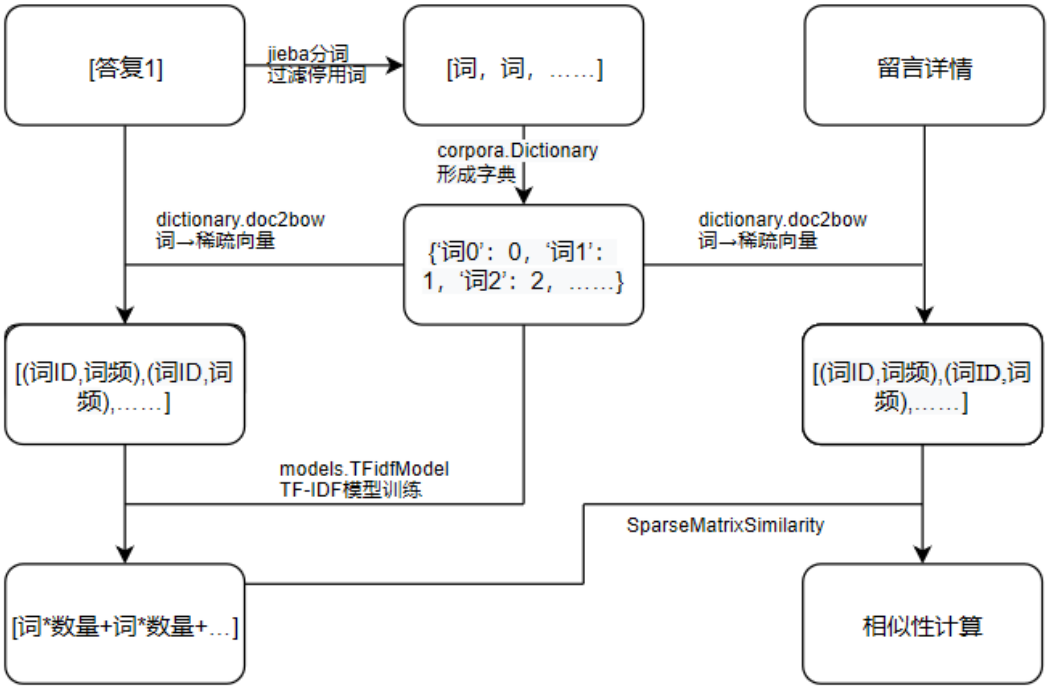

图 4-2 计算相似性流程图

4.4.2 相似性算法

（1）取 ID_n 下的答复意见，对其进行中文分词，返回分词列表；

（2）基于答复意见建立词典，获取特征数；

（3）基于词典建立语料库；

（4）取同一 ID_n 下的留言详情，将其转化成稀疏向量；

（5）用 ID_n 答复意见建立的语料库训练 TF-IDF 模型；

（6）计算同一 ID_n 下的留言详情与答复意见的相似度；

（7）循环遍历所有 ID，直至取完所有样本。

4.5 熵值法

在信息论中，熵是对不确定性的一种度量。信息量越大，不确定性就越小，熵也就越小。根据熵的特性，可通过计算熵值来判断一个事件的随机性及无序程

度，也可以用熵反应指标的离散程度，指标的离散程度越大，该指标对综合评价的影响（权重）越大。熵值法步骤如下：

（1）选取 n 个评价对象，m 个指标，则x_{ij}为第 i 个评价对象的第 j 个指标的数值（i=1,2，…，n；j=1,2，…，m）；

（2）指标标准化，对于正项指标具体方法如下：

$$x'_{ij} = \frac{x_{ij} - \min\{x_{1j,\cdots,}x_{nj}\}}{\max\{x_{1j,\cdots,}x_{nj}\} - \min\{x_{1j,\cdots,}x_{nj}\}}$$

为方便起见，归一化后的数据仍记为x_{ij}。

（3）计算第 j 项指标下第 i 个评价对象占该指标的比重：

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}，\ \text{i=1，…，n；j=1，…，m}$$

（4）计算第 j 项指标的熵值：

$$e_j = -k\sum_{i=1}^{n} p_{ij}\ln(p_{ij})，\ \text{其中}k = \frac{1}{\ln(n)} > 0$$

（5）计算信息熵冗余度：

$$d_j = 1 - e_j$$

（6）计算各项指标的权值：

$$\omega_j = \frac{d_j}{\sum_{j=1}^{m} d_j}$$

（7）计算各评价对象的综合得分：

$$s_i = \sum_{j=1}^{m} \omega_j . p_{ij}$$

4.6 结果分析

表 4-2 答复意见评价部分结果表

文本长度	标点符号占比	平均句长	是否有参考文件	格式良好	是否有外部链接	答复时间跨度	文本熵	问题与答复相似度	得分
454	0.158590308	27.28571	1	1	0	15.22550926	4.06	0.471878	0.314753

305	0.17704918	17.92857	0	1	0	14.73993056	4.1	0.143839	0.03468
357	0.168067227	29.7	1	1	0	14.75636574	3.608	0.577934	0.318048
310	0.251612903	16.57143	1	1	1	14.77930556	3.847	0.507093	0.679406
161	0.273291925	14.625	0	1	0	15.70012731	3.386	0.564933	0.049582
232	0.137931034	25	0	1	1	31.05888889	3.125	0.343401	0.410403
245	0.155102041	17.25	1	1	1	40.93443287	3.621	0.562617	0.682888
624	0.147435897	20.46154	1	1	0	28.52153935	4.001	0.446332	0.315715
505	0.142574257	24.05556	0	1	0	16.23244213	3.898	0.39552	0.047633

表 4-2 中给出了这次答复意见的部分评价结果，评价得分由文本长度、标点符号占比、平均巨长、是否有参考文件、格式良好、是否有外部链接、答复时间跨度、文本熵、问题答复相似度 9 个指标，通过熵值法赋权而计算得到，样本答复意见评价的全部结果可在“答复意见质量评价结果表.xls”详见。

5 结论

本文的主要目的是利用自然语言处理技术从三个角度对政务留言文本数据进行挖掘分析。一是群众留言分类模型。利用 word2vec、TF-IDF 和 SVM 建立留言内容的一级标签分类模型，最终 F-Score 测量值达到 0.812；二是热点问题挖掘，热度评价采用留言内容相似数、点赞数、反对数这三个指标进行评价，最后将值进行标准化处理，得到热度指数。通过对识别对象和标注形式的考量，最后采用哈工大 LTP 对本部分的实体进行识别；三是答复意见质量评价方案，从结构化特征和文本特征构建 9 个指标来评价答复意见的质量，采用熵值法进行综合评价。9 个指标权重依次是：0.090609、0.006913、0.028463、0.263743、0.041339、0.366255、0.163555、0.003890、0.035234。最终评价结果见“答复意见质量评价结果表.xls”

6 参考文献

[1]Boser B E, Guyon I M, Vapnik V N. A Training Algorithm for Optimal Margin Classifiers[J].Proceedings of Annual Acm Workshop on Computational Learning Theory, 1996, 5:144--152.

[2]冯贵川. 基于 Word2vec 的文本建模及分类研究[J].

[3]潘夏晖, 虞欣平, 邹军. 层次化在线话题热度算法[J]. 名城绘, 2019(4):595-595.

[4]CHUJO K,UTIYAMA M. Understanding the role of text length，sample size and vocabu lary size in determining text coverage［J］. Reading in a foreign language,2005,17(1):1-2 2.

[5] CHAFE W. Punctuation and the prosody of written language［J］. Written communication，1988,5(4) : 395-426.

[6] MC LAUGHLIN G H． SMOG grading-a new readability formula［J］. Journal of reading，1969，12(8) : 639 － 646.

[7] 靳 锐 , 张 宏 莉 , 张 玥 , 王 星 . 中 文 公 众 事 件 信 息 熵 计 算 方 法 [J]. 软 件 学 报,2016,27(11):2855-2869.

[8]吴军. 数学之美(第二版) 第 2 章 统计语言模型. 人民邮电出版社.