基于循环匹配和改进的 TextRank 算法的热点问题挖掘研究

摘要

随着人们物质文化需求的提高，提升政府办公效率，促进政府人员了解民意、汇聚民智、凝聚民气，减小人工分类成本，用自然语言处理技术的智慧政务系统提升政府的管理水平和施政效率具有重要意义。至此，为降低大量的人工成本，通过机器学习获得便利，本文针对群众的留言详情，通过自然语言处理和文本挖掘，将群众留言进行分类处理，挖掘热点问题，对答复意见进行评分，本文解决思路如下：

针对问题一，首先根据附件一和附件二的数据进行分析，统一留言时间格式以方便后期数据运用，找出留言主题与一级分类标签之间的相互关系，通过对留言主题进行中文分词、随机抽样、去除停用词等数据处理，其次用预处理后的数据构建 $TF-IDF$ 权值矩阵，计算词频与词频逆反文档频率，将文本转化为词频矩阵，定义 $TF-IDF$ 权值矩阵与分类标签矩阵的相关分类算法，最后利用提供的数据进行训练逻辑回归模型，并将预测出的数据用 $F\text{-}Score$ 评价方法对分类算法进行评价。

针对问题二，首先对留言进行粗略分类，采用改进的聚类循环匹配算法进行三次不同的循环匹配，第一次匹配采取文本遍历的形式进行，通过设定阈值筛选提取粗略的一个集合 $S=\{S_1,S_2,S_3...\}$ 并分词备用，第二次匹配提取第一次匹配所分出的集合 S，将文本重新与集合 S 所提出的词进行匹配，重新设定一个新的阈值进行匹配，将其分为一个新集合 $G=\{G_1,G_2,G_3...\}$，采取先紧后松的原则提高准确度和匹配度，通过排序初略找出排名前五的分类，第三次匹配对这些分类的文本进行词频统计，通过词性分析找出排名前五的地点及人群，将此作为索引查找对应的原文本，并具体分出热点最高的五类问题。提取出每一类的留言主题并用改进的 TextRank 算法进行多文档中文留言主题提取，将每句留言主题转换为 $TF-IDF$ 权值矩阵并计算句子与句子之间的语义相似度，将特殊句子进行相应处理，最后提出出每个分类下权重最高的句子，进行修饰润色。

针对问题三，首先从答复的相关性、完整性、重复性、答复时间四个角度入手，第一，先进行相关性判断，匹配程度列公式表示为：$match(T_1,T_2)=F\left(f\left(T_1\right),f\left(T_2\right)\right)$，其中$T_1$为留言详情文本，$T_2$为答复意见文本，保留相关性分数$X$；第二，再进行完整性判断，通常情况下，一句话所提及的核心重点的权重都会很高，提取权重最高的三个词作为索引对答复意见做匹配，保留完整性分数W；第三进行答复时间计算，通过计算答复时间和留言时间的时间差，计算出回复的时间，保留答复时间分数T，最终拟定评分模型，中和答复的相关性、完整性、可解释性、答复时间四个不同方面的分数，计算出答复效率D，将其作为最终评价得分。

综上所述，此项目利用机器学习通过大数据人工智能降低人工成本，更大程度的减轻相关部门的工作量，同时又能够高效准确的处理群众反映的问题，提高处理率。而项目优势在于：一是创新解决思路，二是利用机器学习算法调试多种算法模型以及思路构建模型达到较高准确率；三是与实际情况相结合，合理分类文本及热点问题。

关键词：jieba 中文分词；TF-IDF；TextRank；逻辑回归

Abstract

Along with people's increasing material and cultural needs, improve the government's office efficiency, gather people's wisdom and spirit, reduce labor costs, smart government system of great significance is to improve government's management level and governance efficiency, which is adopted Natural language processing technology. Therefore, to reduce a large number of labor costs, though machine learning get convenience, for comments on the masses, this thesis use Natural language processing technology and Text mining, classify the message of the masses, explore the hot issues and rate responses. The ideas of this paper are as follows:

For the first issues, according to the data in annexesIandII， analysis data, unified message time format for data utilization at later period, find out the relationship between message subject and first class label, data processing of message subject which is Chinese word segmentation, random sampling, elimination of stop words, etc. Then use the data which has been processed to building weight matrix, calculation word frequency and word frequency anti document frequency, to turn text into word frequency matrix, define $TF-IDF$ weight matrix and classification label matrix. At last, training logistic regression model for data, and evaluate the final result by using F-score.

For the second issues, first, classified the message roughly, using the improved clustering cycle matching algorithm to do three different cycle matching. The first match takes the form of text convenience, extract a rough set by setting a threshold filter $S=\{S_1,S_2,S_3...\}$ and participle reserve, the second match extracts the set of the first match S, and match the text again with the words proposed by the set, reset a new threshold to match, divide into a new set $G=\{G_1,G_2,G_3...\}$, improve accuracy and matching by tighten first, loosen after, find top 5 though sort, the third match is to make word frequency statistics for these texts, find out the top 5 palces and groups through

part of speech analysis, and take this as index to find the original text, then find the top tangibly. Extract each category of message subject, use TextRank which is improved to extract Chinese message subject of multiple documents, turn every one of it into TF-IDF weight matrix and calculation semantic similarity between sentences, dealing with special sentences, at last, the sentences with the highest weight under each classification are extracted, to embellish the sentences.

For the third issue, first of all, the relevance, integrity and repeatability of the response time are taken as the entry points. First, judge relevance, formula of matching degree is: T_1 is text of message, T_2 is text of reply comments, retention relevance score X , second, judge integrity, as usual, the weight of the core points in a sentence will be very high, extract the three words with the highest weight as the index to match the reply, retention integrity score W . Third, calculate the response time , calculate the reply time by calculating the time difference between the reply time and the message time, Retention response time score T , then rating model, Scores of relevance, completeness, interpretability and response time, calculate the response efficiency D , take it as the final score.

In summary, this project uses machine learning to reduce labor cost through big data artificial intelligence, Reduce the workload of relevant departments to a greater extent. At the same time, it can effectively and accurately deal with the problems reflected by the masses. Improve processing rate. The advantage of this project: First, innovative solutions. Second, try various methods of machine learning models to make the data as accurate as possible. Last, combine with real situation to deal with the issue.

Keywords : jieba Chinese word segmentation; TF-IDF; TextRank; Logistic regression

目录

I. 挖掘背景及目标.....1
1.1 挖掘背景.....1
1.2 挖掘目标.....1
II. 问题分析.....1
2.1 问题一的分析.....1
2.2 问题二的分析.....2
2.3 问题三的分析.....2
III. 符号说明.....2
3.1 符号说明.....2
IV. 数据预处理.....3
4.1 文本数据的基本处理.....4
4.2 文本分类数字化及随机抽样.....4
4.3 jieba 中文分词模型.....4
4.3.1 jieba 中文分词模型的基本原理.....4
4.3.2 jieba 中文分词模型的应用.....5
4.4 停用词处理.....5
4.5 TF-IDF 构建词向量矩阵.....6
V. 问题求解.....7
5.1 群众留言一级标签分类.....7
5.1.1 提取并分析留言主题分布.....7
5.1.2 构建逻辑回归模型.....7
5.1.3 基于 F-score 算法的模型评价.....10
5.2 基于改进的聚类循环匹配算法与改进的 TextRank 算法的热点问题挖掘模型.....11
5.2.1 改进的聚类循环匹配算法.....11

5.2.2 改进的 TextRank 算法.....13
5.3 基于改进的循环神经网络文本匹配综合评价系统.....18
5.3.1 评分标准.....18
5.3.2 相关性判断.....18
5.3.3 完整性判断.....19
5.3.4 重复性判断.....19
5.3.5 答复时间.....19
5.3.6 评分模型.....20
VI. 参考文献.....20

I. 挖掘背景及目标

1.1 挖掘背景

随着生活质量的提高和大数据技术的发展，自然语言处理技术（NLP）被广泛运用在生活当中，例如文本分类和聚类 、信息检索和过滤、信息抽取问答系统、拼音汉字转换系统 、机器翻译 、新信息检测等。而现在大数据时代发展，越来越多的数据被储存在网络数据库当中，对于这些数据的处理就显得越来越重要，人工处理这些数据会花费大量的时间与精力，计算机的计算能力强大，通过不同的算法与模型对数据进行处理，利用计算机对数据进行处理节省很多人工资源。

1.2 挖掘目标

随着社区民意的大幅度的增加，社区问题难以避免，通过各种网络政务平台进行留言逐渐成为主要的群众留言方式，其中也会存在大量的重复信息，也就加大了人工回复处理的难度，也就形成了工作量大，但解决效率低等情况，故本文的主要目标：一是为解决社区群众民意，减轻政务工作难度节省人工资源，二是在大数据趋势下，研究自然语言处理技术[1]。

II. 问题分析

2.1 问题一的分析

针对问题一，首先根据附件一和附件二的数据进行分析，找出留言主题与一级分类标签之间的相互关系，通过对留言主题进行中文分词、随机抽样、去除停用词等数据处理，其次用预处理后的数据构建$TF-IDF$权值矩阵，计算词频与词频逆反文档频率，将文本转化为词频矩阵，定义$TF-IDF$权值矩阵与分类标签矩阵的相关分类算法，最后利用提供的数据进行训练模型，并将预测出的数据用$F\text{-}Score$评价方法对分类算法进行评价。

2.2 问题二的分析

针对问题二，首先对某一时段内反映特定地点或特定人群问题的留言进行分类，采用基于聚类算法的改良循环算法，初略找出排名前五的分类，再对这些分类的文本进行词频统计，通过词性分析找出排名前五的地点及人群，将此作为索引查找对应的原文本，并具体分出五类。

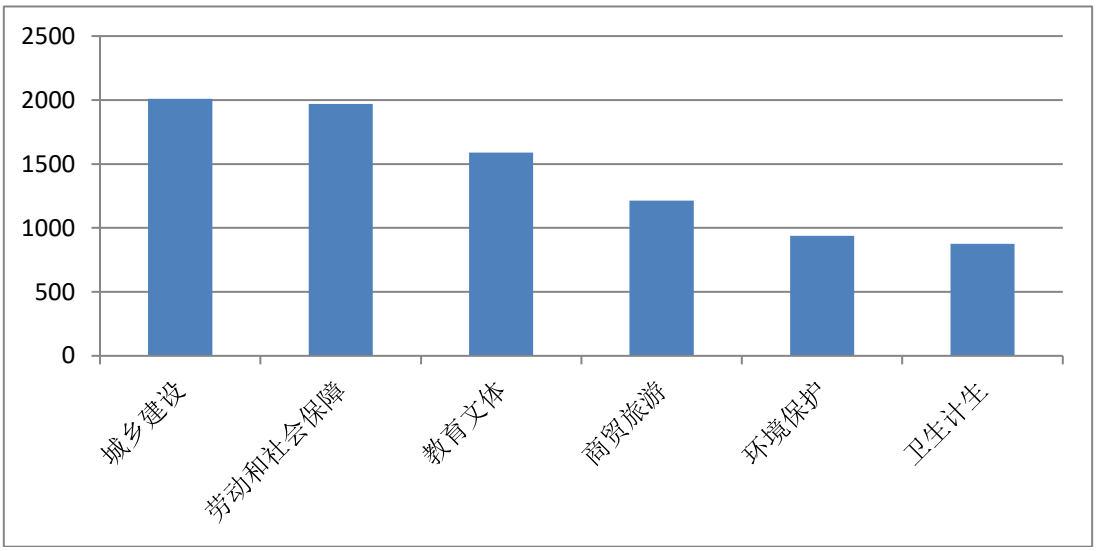


图 1 原数据每个分类个数

2.3 问题三的分析

针对问题三，首先分出评分标准的几个关键因素，从相关性、完整性、重复性、答复时间四个方面分别进行分析判断，再通过这四个关键因素不同的特点和关系拟定各自的评分标准，最终结合四个关键因素进行最终评分公式拟定。

III.符号说明

3.1 符号说明

符号	意义
d	文档
$\|D\|$	代表语料库中所有文档的数量
w	某词

$\lvert\{d_j \mid d_j \in D, w_i \in d_j\}\rvert$	表示语料库中含有 w_i 词的文档数量
$\beta_1, \beta_2, \cdots \beta_k$	多元线性回归模型的待定参数
ε_a	Logistic 函数的随机变量
$P(x)$	Logistic 函数的因变量
$y_1, y_2, \cdots, y_m$	最大似然估计的观测值
$L(\beta)$	对数似然函数
$\beta_0, \beta_1, \beta_2, \cdots \beta_n$	$L(\beta)$ 取得最大值时的参数
$Sim(s_i, s_j)$	句子相似度
T	句子中所包含词项的权重
$L(s)$	文档中句子的长度特征
$\lvert l(s_i)\rvert$	句子中所含词的个数
$lavg(d_i)$	代表文档 中每个句子长度的平均长度

IV.数据预处理

本文数据预处理主要分为 5 个部分，由于给的数据包含了时间、分类标签、数个语句，且有重复和空数据，故需要将时间转换为 *DataFrame* 格式，并进行去重去空处理。因为分类标签是中文不便于机器识别，于是进行数字化转换，根据统计各个分类标签总数发现总数不一，不利于算法的训练，于是进行随机抽样提取相同数据量的数据并对数据进行分词和停用词处理，利用 $TF-IDF$ 算法构建词向量矩阵，方便后续对热点问题挖掘的研究。

图 2 预处理流程图

4.1 文本数据的基本处理

读取 Excel 文件并转为 *DataFrame* 格式处理，经过多次观察发现留言时间的时间格式不统一，这不利于后期对数据的运用，所以将留言时间这一列进行了数据处理，将时间全部变成类似 *%Y-%m-%d* 这样的格式，并对整个 *DataFrame* 进行了去重和空值查询，保证基础数据的完整和可用性。

4.2 文本分类数字化及随机抽样

通过附件一所给的十五个一级分类，给予十五个分类分别赋值一个数值以便于机器可以更好的识别它。经过多次试验，发现如果一个分类的数据量过少，预测结果就将会出现过拟合的现象，为了避免预测结果的过拟合，我们将十五个分类随机取出相同的数据量的数据作为训练集进行训练。

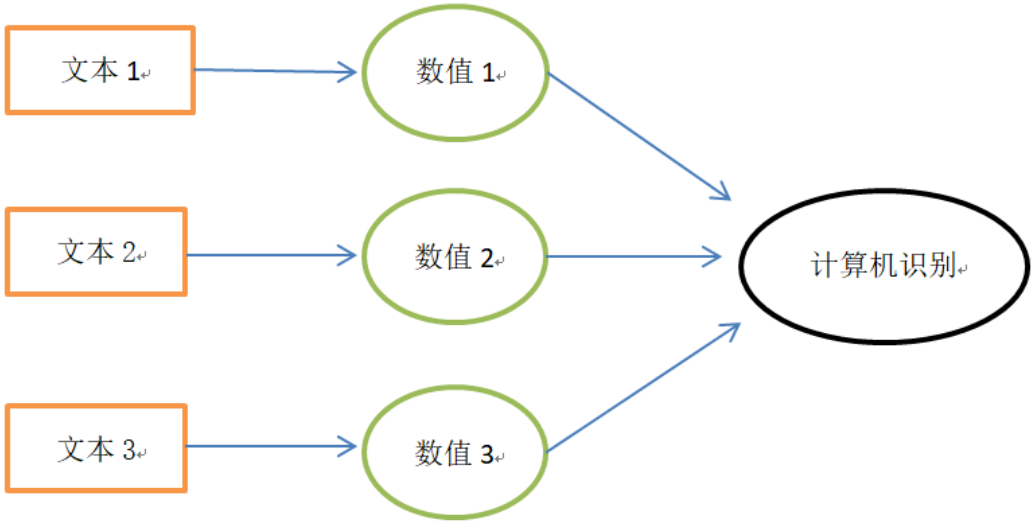


图 3 文本分类数字化

4.3 jieba 中文分词模型

4.3.1 jieba 中文分词模型的基本原理

1） *Jieba* 分词[2]采了基于 *Trie* 树结构的算法， *Jieba* 分词利用该算法高效实现了词图扫描并且利用词图扫描将得到句了中汉字所有的成词可能，并且将这些所有成词可能的情况构成有向无坏图(DAG)，为下一步打下基础。

通过分词 *Jieba* 分词的源码可以发现， *Jieba* 分词本身就包含了一个有 2 万多词条的词典。基于 *Trie* 结构的扫描就是将这些词条放到 *Trie* 的树结构之中，一旦扫描的词条中和这些放在 *Trie* 结构中的词条有着相同的前缀，那么就实现了快速查找。

2） *Jieba* 分词中最大概率路径的实现采用了动态规划查找的万法，动态规划查找可以找出根据词频的最大切分组合。分析 *Jieba* 分词的源码可以发现 *Jieba* 分词不仅仅将字典生成 *Trie* 树，同时，将把每个词的出现次数转换为了频率。

3） 对于在 *Jieba* 分词自带的词典中未出现的词， *Jieba* 分词采用了 Viterbi 算法，并且将用于汉字成词的 HMM 型应用其中。

4.3.2 jieba 中文分词模型的应用

由于是对中文文本的处理，所以用到了 *Jieba* 中文分词模型，分词模式使用精简模式，对文本数据进行拆分组词，对每个留言数据都进行分词，并以列表的形式保存。

4.4 停用词处理

停用词[3]是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为停用词。因为数据中停用词较多，影响了算法的训练，所以加入停用词字典，将分词后的数据进行停用词处理，把没有语义没有作用的词去除，其中还自定义加入了针对本文本数据的停用词，例如下表，更有利于数据后期的运用。

停用词例子			
请	建议	请求	质询
对	举报	投诉	希望
求	？	！	（、）

表 1 停用词表

4.5 TF-IDF 构建词向量矩阵

词频—逆向文档频率[4]（Term Frequency—Inverse Document Frequency, TF-IDF）是衡量一个词对于一个文档集中的其中一份文档的重要程度的重要参数，是关键词抽取领域影响深远的指标之一。使用 $TF-IDF$ 算法[5]抽取关键词的基本思想是：如果某个词 w 在一篇文档 d 中出现的频率很高，并且在其他文档中出现的频率很低，则认为词 w 具有很好的区分文档的能力，适合用来把文档 d 和其他文档区分开来，即认为词 w 是文档 d 中具有代表性的关键词。$TF-IDF$ 值通常使用以下公式来计算：

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{1}$$

其中，词频（TF）由以下公式定义：

$$tf_{i,j} = \frac{n_{i,j}}{N_j} \tag{2}$$

其中，$n_{i,j}$ 代表词 i 在文档 j 中出现的次数，N_j 代表文档 j 的总词数。

逆向文档频率（IDF）由以下公式定义：

$$idf_i = \log \frac{|D|}{1+|\{d_j \mid d_j \in D, w_i \in d_j\}|} \tag{3}$$

其中，$|D|$ 代表语料库中所有文档的数量，$|\{d_j \mid d_j \in D, w_i \in d_j\}|$ 表示语料库中含有 w_i 词的文档数量，分母中的+1 的目的在于对 IDF 值进行平滑，以防止语料库中没有任何文档包含词 w_i 的情况导致分母为 0，无法进行计算。

$$tf-idf=\begin{bmatrix}0&0&0&\cdots&\cdots&\cdots&0&1&0\\0&0&0&\cdots&\cdots&\cdots&57&0&0\\1&0&0&\cdots&\cdots&\cdots&0&0&0\\0&0&0&\cdots&\cdots&\cdots&0&0&0\\0&0&0&\cdots&\cdots&\cdots&0&0&3\\0&0&0&\cdots&\cdots&\cdots&0&0&0\\0&7&0&\cdots&\cdots&\cdots&0&0&0\\0&0&0&\cdots&\cdots&\cdots&0&5&0\\0&0&0&\cdots&\cdots&\cdots&0&0&0\end{bmatrix}$$

$TF-IDF$ 值越高，说明该词语区分这篇文档和语料库中的其它文档的能力越强，也就越有可能是这篇文档的关键词。

V. 问题求解

5.1 群众留言一级标签分类

5.1.1 提取并分析留言主题分布

通过 python 的 pandas 库读取训练数据，将 Excel 文件读取 DataFrame 格式以便对数据进行处理，由于附件二只有七个一级标签，所以将 0~6 分别赋值给标签，目的是为将标签数字化，便于计算机读取。通过统计附件二中一级标签的个数，发现数量最少的标签是交通运输，其数据个数为 613，数量最多的标签是城乡建设，其数据个数为 2009，由于两者的数据量差距较大，为了避免训练数据过少而出现预测结果过拟合现象，所以将每个类别随机选取数据的个数定为 600 个，再用 *jieba* 进行中文分词，并进行停用词去除，将分词后的数据进行格式处理，便于后面词向量矩阵的运用。

5.1.2 构建逻辑回归模型

逻辑回归(Logistic Regression)[6]是分析因变量为定性变量的常用统计分析方法，是当前比较常用的机器学习方法之一，可以用于对事件发生的可能性进行预测，也可以用于分类，属于一种概率型非线性回归方法。由于逻辑回归模型对数据的正态性、方差齐性及自变量类型不做要求，并且具有系数的可解释性等优

点，使得其在数量心理学、生物统计学、社会学、计量经济学、临床医学等领域得到了广泛的应用。

5.1.2.1 线性回归

线性回归是利用数理统计中的回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛。

回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。可以简单的表示为：

$$y = a + bx + c \tag{4}$$

式中 y 为因变量，x 为自变量，a 为常数项（截距项），b 为回归系数，c 为随机误差项。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。假设某一因变量 y 受 k 个自变量 $x_1, x_2, ..., x_k$ 的影响，其 n 组观测值为 $(y_a, x_{1a}, x_{2a}, ..., x_{ka}), a = 1,2,...,n$。那么，多元线性回归模型可表示为以下的形式：

$$y_a = \beta_0 + \beta_1 x_{1a} + \beta_2 x_{2a} + \cdots + \beta_k x_{ka} + \varepsilon_a \tag{5}$$

式中 β_0 为常数项，$\beta_1, \beta_2, \cdots \beta_k$ 为待定参数，即回归系数，ε_a 为随机变量。

5.1.2.2 Logistic 函数

Logistic 函数或 Logistic 曲线[7]是一种常见的 S 形函数，它是皮埃尔·弗朗索瓦·韦吕勒在 1844 或 1845 年在研究它与人口增长的关系时命名的。函数值在起初阶段大致是指数增长；然后随着开始变得饱和，增加变慢；最后，达到成熟时增加停止。其函数可以表示为如下公式：

$$P(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

其中 x 为自变量，$P(x)$ 为因变量，e 为自然对数函数的底数，函数图形如图所示。

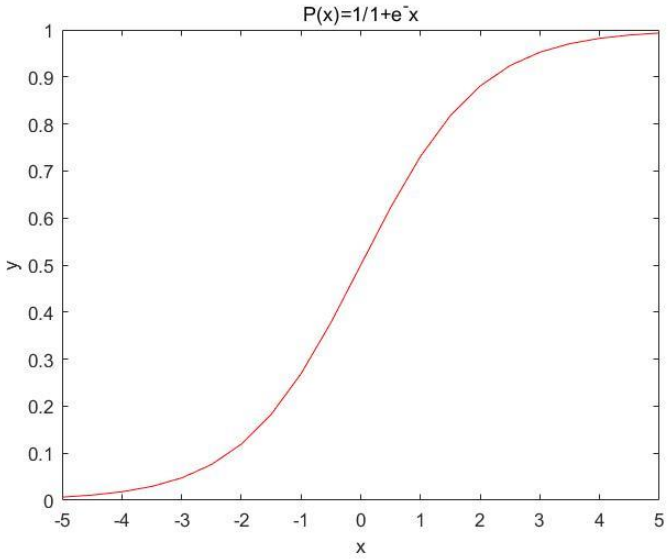


图 4 Sigmoid 函数

从图中可以看出，Logistic 函数具有明显的 S 型分布，当 $x=-\infty$ 时，$P(x)=0$，当 $x=+\infty$ 时，$P(x)=1$，即无论 x 的取值如何，$P(x)$ 的取值范围都在 $0-1$ 之间。

5.1.2.3 逻辑回归模型

1. 逻辑回归模型的公式

考虑具有 n 个变量的向量 $x=(x_1,x_2,\cdots,x_n)$，设条件概率 $P(Y=1|x)=p$ 为根据观测量相对于某事件发生的概率，则逻辑回归模型可表示为：

$$P(Y=1|x)=\pi(x)=\frac{1}{1+e^{-g(x)}}=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n)}} \tag{7}$$

式中，β_0 为截距项，$\beta=(\beta_1,\beta_2,\cdots,\beta_n)$ 为自变量的回归系数，x_n 为不同的自变量。如果 $g(x)$ 中还含有名义变量（分类变量）的话应将其变为虚拟变量，一个具有 k 取值的名义变量，将变为 $k-1$ 个虚拟变量，因此 $g(x)$ 变为：

$$g(x)=\beta_0+\beta_1x_1+\cdots+\sum_{i-1}^{k-1}\beta_{ji}D_{ji}+\beta_nx_n \tag{8}$$

可以看出，逻辑回归模型其实就是一个线性回归被 Logistic 方程归一化后的概率型非线性回归模型，$g(x)$ 通常被称为一系列影响事件发生概率的因素的线性函数。因为 $\pi(x)$ 的值域为 $[0,1]$，因此我们可以根据其取值来估计因变量 $Y=1$ 时发生的概率。

2. 最大似然估计

最大似然估计，也叫极大似然估计，此方法的基本思想是：当从模型总体随机抽取 组样本观测值后，最合理的参数估计量应该使得从模型中抽取该m组样本观测值的概率最大。这是一种迭代算法，它以一个预测估计值作为参数的初始值，根据算法确定能增大对数似然值的参数的方向和变动，估计了该初始函数后，对残差进行检验并用更新改进后的函数进行重新估计，直到对数似然值不再显著变化为止。

假设有m个观测样本，其观测值分别为$y_1, y_2, \cdots, y_m$，设$p_i = P(y_i = 1 | x_i)$为在给定条件下得到的$y_i = 1$的概率。在同样条件下得到$y_i = 0$的条件概率为$P(y_i = 0 | x_i) = 1 - p_i$。于是，可以得到一个观测值的概率为：

$$P(y_i) = p_i^{y_i}(1 - p_i)^{(1-y_i)} \tag{9}$$

因为式(9)中各项观测独立，所以它们的联合分布可以表示为各边际分布的乘积。

$$l(\beta) = \prod_{i=1}^{m} P(y_i) = \prod_{i=1}^{m} P(x)^{y_i}[1 - P(x)]^{1-y_i} \tag{10}$$

上式即通常所说的m个观测值的似然函数，结合式(9)，只要求得使得式(10)的值最大时的参数估计，就可以得到逻辑回归模型中的回归系数。对式(10)两边求自然对数，可得对数似然函数：

$$L(\beta) = \ln[l(\beta)] = \sum_{i=1}^{m}[y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}) - \ln(1 + e^{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_n x_{in}})] \tag{11}$$

要求解使得$L(\beta)$取得最大值时的参数$\beta_0, \beta_1, \beta_2, \cdots \beta_n$的值，通常对上式进行求导，而后应用牛顿－拉斐森(Newton-Raphson)[8]迭代法对求导得到的非线性方程进行求解。

5.1.3 基于 F-score 算法的模型评价

$F\text{-}Score$[9]是度量特征在不同类别间的区分度的一种指标，$F\text{-}Score$值越大，代表该特征在不同类别之间的区分度越强。假设x_k代表数据集中的样本

$(k=1,2,\cdots,N)$。n_+ 为正类样本的数量，n_- 为负类样本的数量，则数据集中第 i 个特征的 $F-Score$ 可由

$$F_i=\frac{(\overline{x_i}^{(+)}-\overline{x_i})^2+(\overline{x_i}^{(-)}-\overline{x_i})^2}{\frac{1}{n_+-1}\sum_{k=1}^{n_+}(x_{k,i}^{(+)}+\overline{x_i}^{(+)})^2+\frac{1}{n_--1}\sum_{k=1}^{n_-}(x_{k,i}^{(-)}+\overline{x_i}^{(-)})^2} \tag{12}$$

计算得到。式中：$\overline{x_i}$ 表示该特征在整个样本集上的平均值，$\overline{x_i}^{(+)}$ 和表示该特征在正类样本上的平均值，$\overline{x_i}^{(-)}$ 表示该特征在负类样本上的平均值，$x_{k,i}^{(+)}$ 表示第 k 个正类样本在第 i 个特征上的值，$x_{k,i}^{(-)}$ 表示第 k 个负类样本在第 i 个特征上的值。

利用 $F-Score$ 算法将预测出来的结果与真实值带入式中计算得到 0.88 的准确度，由此可见，逻辑回归模型对于此分类具有良好的分类效果。

5.2 基于改进的聚类循环匹配算法与改进的 TextRank 算法的热点问题挖掘模型

5.2.1 改进的聚类循环匹配算法

改进的聚类循环匹配算法[10]共进行了三次循环匹配

匹配算法步骤如下：

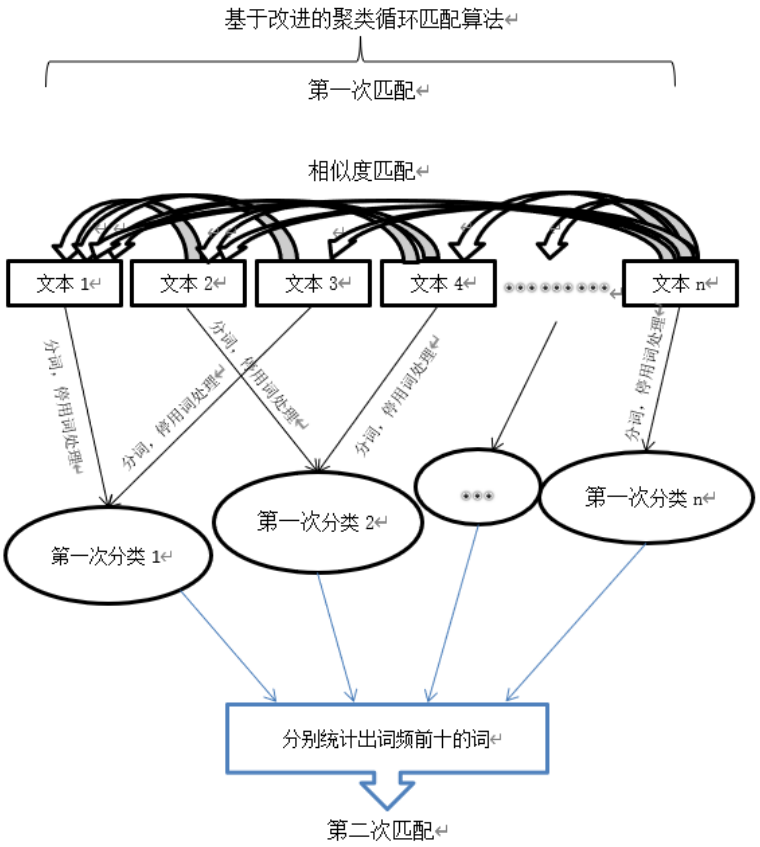
Step1：第一次匹配采取文本遍历的形式进行，以一个文本作为对比文本，将后面的文本与之进行匹配，设定一个较大的值 n 作为阈值，当匹配度大于阈值时，将文本粗略的分为一个集合 并分词备用

Step2：第二次匹配提取第一次匹配所分出的集合 ，进行词频统计，分别提炼出词频最高的十个词作为索引

Step3：将文本重新与集合 所提出的词进行匹配，设定一个合适的值 作为阈值，当匹配度大于阈值时，将其分为一个新集合 ，采取先紧后松的原则提高准确度和匹配度，最后排序得出热度前五的分类

Step4：第三次匹配对第二次匹配分出的新集合 进行词性分析，分别提出地名或人群词频最高的词作为索引

Step5：重新对整个文本进行索引匹配，提炼出五个热度最高的反映特定地点或特定人群问题的留言



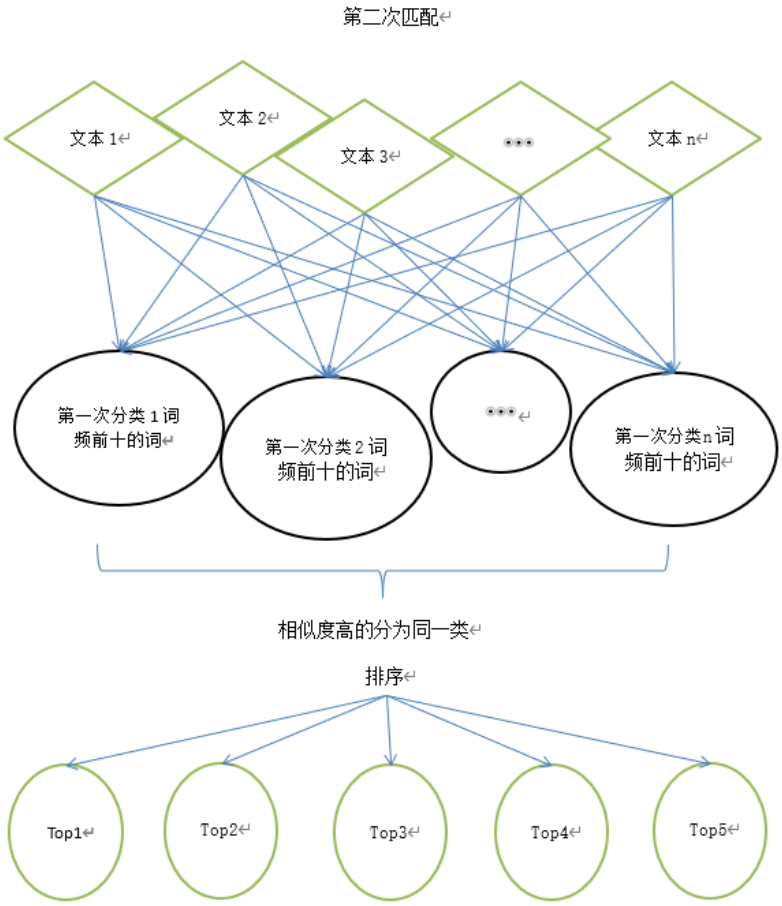


图 5 改进的聚类循环匹配算法流程图

5.2.2 改进的 TextRank 算法

5.2.2.1 传统的 TextRank 算法原理介绍

传统的 TextRank 算法[11]利用了 PageRank 网页排序的迭代思想对文档中的句子打分：

$$W_s(v_i)=(1-d)+d*\sum_{v_j\subseteq In(v_j)}\frac{w_{jk}}{\sum_{v_k\subseteq Out(v_j)}w_{jk}}Ws(v_j) \tag{13}$$

式中 $d\in[0,1]$ 称为阻尼系数，每个节点转移到其他节点的概率为 $(1-d)$，本实验中设置 d 为 0.9，公式从左到右依次为 $W_s(v_i)$ 表示节点 i 的权重，第一个求和

号表示每一个句子对所在文档的贡献度，分母的 $\sum_{v_k \subseteq Out(v_j)} w_{jk}$ 表示文本中相对应的部分句子的权重之和，$W_s(v_j)$ 表示上次迭代得到的节点 j 的权重。

接着将基于改进的聚类循环匹配算法分类出来的句子作为 TextRank 的输入构建图模型，若两个句子的相似度大于给定的阈值，那么就将两个句子通过一条边连接起来。每个节点的最终权重利用排序算法的迭代公式计算得到。

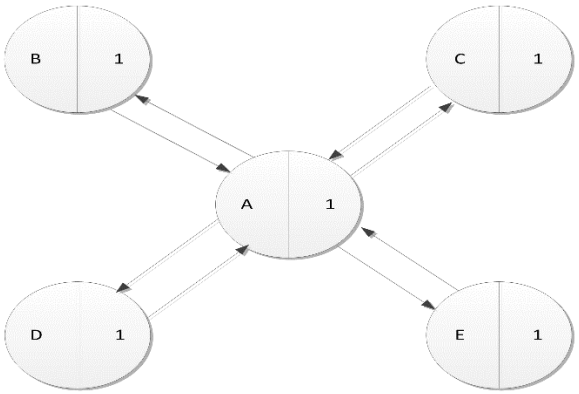


图 6 节点 A 的图模型

图中给出了经典 TextRank 算法句子打分机制的双向加权图。其中 A，B，C，D，E 五个节点分别为文档中的五个句子，边的长度为相邻两节点之间的相似度值，每个点的初始权值默认为 1，且每个点的得分大小和周围节点息息相关，最终的权重由公式(13)得到。

5.2.2.2 改进的 TextRank 算法介绍

经典的 TextRank 一般认为全部句子都是相邻的，边的构建往往通过计算两两句子之间相似度的策略，设定阈值 1，即当两个句子间包含相同词项个数大于阈值时，两个节点之间建立一条边。常用方法有欧式距离、余弦函数等函数计算相似度之后对边赋予权值。传统的计算相似度的方法，只是从统计的层面进行了比较，未从语义信息和句法结构的相似度等层面进行考虑。本文在基于统计的基础上提出结合 $TF-IDF$ 以及语义信息的计算方法[12]，具体过程如下：

1. 基于 $TF-IDF$ 算法计算句子相似度

$TF-IDF$ 的基本原理是“某个句子中，其中所包含的每一个词语的重要性与该词语所在句子中出现的次数成正比，同时与出现该词语的其他句子数成反比例关系。”具体表达式为：

$$TF-IDF(w_i)=TF(w_i)\times IDF(w_i)=TF(w_i)\times\log\frac{N}{DF(w_i)}\sum_{i=1}^{n}X_i^2 \qquad (14)$$

其中 TF 表示词 i 在句子中出现的概率（词频），IDF 表示该词 i 在其他句子中的出现情况并且取总句子数除以包含该词的句子数的商的对数，利用 TF 、IDF 的乘积得到词项 i 的权重。可以看出，$TF-IDF$ 句子中的重要性以及在句子集中的普遍性综合考虑来得到，$TF-IDF$ 值较高的词项与句子的相关度越大，重要性也就越高。因此，如果两个句子 s_i 和 s_j 中对应的词项 w_i 和 w_j 的 $TF-IDF$ 值分别在两个句子中最高，那么 s_i 和 s_j 的向量形式就可以分别表示为 $s_i=(w_i)$，$s_j=(w_j)$，然后利用余弦相似度公式计算 s_i 和 s_j 相似度，相似度值越大，两个句子越相似。

1. 基于句子间的语义相似度

基于语义信息的相似度计算方法[13]，考虑了词义、词语间的同义词、近义词等深层次的语义关系，相对来说，它使得相似度的准确率得到了一定的改善和提高。目前，它不足的地方在于，因为受到语义词典是否全面、语料库数据稀疏和噪声大小的问题等方面的影响，使得该方法获取的相关语义信息准确度不够高。这样就不能满足现代互联网或其他领域对文本语义相似度任务的需要，因此，本文使用卷积神经网络（Convolutional Neural Network，CNN）和词向量相结合的方式 来计算文本中句子间的语义相似度。具体计算过程如图：

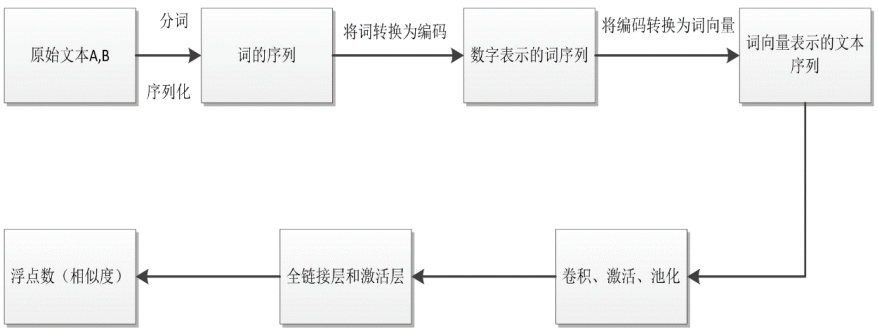


图 7 使用 CNN 和词向量相结合的方式计算相似度流程图

5.2.2.3 结合句子的其他特征

经过前面几个步骤已经对文档中的句子进行了比较合理的打分，但是，仍未将一些特殊句子进行单独考虑。以下将作具体分析：

1) 句子与标题之间的关系：一个句子与其所属文档的标题越相似，那么这个句子包含的关键信息就越多，越能表达文档的中心意思，即越有可能成为摘要句。因为，文档的标题就是文档的主题，整篇文档的具体内容都是围绕着这个主题展开叙述的。计算句子与标题相似度时，可以将文本信息向量化，用特征向量来表示句子，然后利用余弦相似度公式计算具体数值。具体过程为：设给定两个句子 s_i 和 s_j，其中 $s_i=(T_{i1},T_{i2},T_{i3},\ldots T_{in})$，$s_j=(T_{j1},T_{j2},T_{j3},\ldots T_{jn})$；那么，两个句子的相似度 $Sim(s_i,s_j)$ 可以使用如下公式得到：

$$Sim(s_i,s_j)=\frac{\sum_{i=1}^{n}T_{ik}\times T_{jk}}{\sqrt{(\sum_{i=1}^{n}T_{ik}^2)\times(\sum_{i=1}^{n}T_{jk})}} \tag{15}$$

公式(15)中 T 为句子中所包含词项的权重，夹角 θ 越小，相似度越高。

2) 句子的位置：通常，处于文档特殊位置（段首与段尾）的句子往往含有比较重要的信息。比如，我们做英语快速阅读时，往往将重点放在每个段落的段首或段尾。有研究结果显示：人工选取的摘要当中位于段首和段尾的句子成为摘要句的概率 要远远高于其他位置的句子且分别达到了 85%和 7%。这些句子的权重定义如公式(16)：

$$P(s_i)=\begin{cases}0.8 & s_i\text{为段首句时}\\ 0.2 & s_i\text{为断尾句时}\end{cases} \tag{16}$$

3) 特殊句子：在中文文档中如果存在某一个段落且仅仅只由一个单独的 句子组成，那么该段落在文档中所起的作用就比较突出，可能起承上启下或者过渡的作用。这些句子一般比较符合摘要的标准——简短、精炼、信息全面。因此，对于这一类句子应该赋予特殊的权重：

$$W(s)=F(s_i) \tag{17}$$

4） 句子的长度：热点问题描述是对热点事件精简的表达，要求摘要的内容简明扼要，同时要考虑句子中含有的信息量。通常较短句子（句子中词的个数少于六个）包含的信息量会很少，这些句子就没有保留的必要，因此选取的句子在预处理过程中，利用正则表达式去除掉短句子对热点问题详情的干扰。此时，定义文档 d_i 中的句子 s 的长度特征 $L(\mathrm{s})$ 为：

$$L(s_i)=1-\frac{|l(s_i)-l_{avg}(d_i)|}{l_{avg}(d_i)} \tag{18}$$

公式(18)中代表句子 s 中所含词的个数，$lavg(d_i)$ 代表文档 d_i 中每个句子长度的平均长度。

对以上句子各个特征值经过线性加权求和，可得到句子的最终权重为：

$$Score=\lambda_1\times Sim(s_i,s_j)+\lambda_2 P(s_i)+\lambda_3 \mathrm{F}(\mathrm{s}_i)+\lambda_4 L(s_i) \tag{19}$$

将附件 3 中经过改进的聚类循环匹配算法分类好的数据再次经过改进的 TextRank 算法提取出主题对应的关键句并计算出权重，如下表所示：

类别	权重	句子
1	0.02434489278416814	A3 区梅溪湖看云路一师润芳园小区临街门面油烟扰民
	0.024204274828557738	A3 区梅溪湖看云路一师润芳园小区临街门面油烟直排扰民
	0.022932009149663615	A3 区梅溪湖看云路一师润芳园小区临街门面烧烤夜宵摊
	0.022626019479411393	A3 区梅溪湖看云路润芳园小区油烟扰民
2	0.03784113279663813	A 市伊景园滨河苑定向限价商品房项目违规捆绑销售车位
	0.03768491036617089	惊 A 市伊景园滨河苑商品房捆绑销售车位
	0.03693877835226386	A 市伊景园滨河苑捆绑车位销售
	0.03693877835226386	A 市伊景园滨河苑捆绑销售车位
3	0.043095018350217236	A5 区劳动东路魅力之城小区底层餐馆油烟扰民

	0.04309501835021723	A5 区劳动东路魅力之城小区底层餐馆油烟扰民
	0.041700268963504074	A5 区劳动东路魅力之城小区油烟扰民
	0.04146766446667408	A5 区劳动东路魅力之城小区临街门面烧烤夜宵摊
4	0.05104226974723059	A2 区丽发新城小区非法搅拌站
	0.049914289424234634	A 市 A2 区丽发新城小区搅拌站明目张胆污染环境
	0.049798054744110046	A 市 A2 区丽发新城小区遭搅拌站污染
	0.04485642638715216	A 市丽发新城小区搅拌站噪音扰民污染环境
5	0.16594605561066778	A 市 A3 区中海国际社区三期四期空地夜间施工噪音扰民
	0.1470756137280706	A3 区中海国际社区三期四期空地夜间施工噪音扰民
	0.1317055795646122	A3 区中海国际社区空地夜间施工噪音太大
	0.12333715304966741	A3 区中海国际社区期北门那片空地通宵施工扰民

由此表可以看出，经过基于改进的聚类循环匹配算法和改进的 TextRank 算法相结合的热点问题挖掘结果，更加科学、客观，同时能够生成更加贴近中国汉语逻辑的语句。

5.3 基于改进的循环神经网络文本匹配综合评价系统

5.3.1 评分标准

中和答复的相关性、完整性、重复性、答复时间四个角度对答复意见的质量给出一套评分标准。

5.3.2 相关性判断

将留言详情文本和答复意见文本分别设为T_1和T_2，则其匹配程度可用下列公式表示：$match(T1,T2)=F\left(f\left(T1\right),f\left(T2\right)\right)$。这类问题目前有两种主流的做法：一种称

为 representation-focused，即倾向于首先将文本整体按照某种规则表示出来，即更注重上式中的函数 f，文章中使用过 CNN 和 RNN[14]来实现对 f 的拟合，获得了文本表示之后即可用余弦相似度等简单的函数计算文本的相似度；另一种方法称为 interaction-focused，即将两个文本以某种方式联系，更注重上式中的 F，对两者的联系（如矩阵）进行拟合。对于两个文本 $T_1=\{w_1,\ldots,w_m\}$ 和 $T_2=\{v_1,\ldots,v_m\}$，其中 w_i 和 v_j 分别为 T_1 和 T_2 中的第 i 和 j 个单词，我们可以通过 word embedding 获得其对应的词向量，此时单词之间的关系是可以通过向量的运算（如余弦、点乘等）体现的；此时构建一个关系矩阵 M，其元素 M_{ij} 表示单词 w_i 的向量和 v_j 的向量的相似程度；将矩阵 M 通过训练好的 CNN，获得两个文本之间的相似程度 S，打分方式为：$X=S\times 0.4$

5.3.3 完整性判断

通常情况下，一句话所提及的核心重点的权重都会很高，用 $TF-IDF$ 将文本向量化，用 LDA 模型找出关键词，并取出关键词中权重最大的三个词，与其对应的答复意见进行匹配，如果全部匹配成功，那么答复意见的完整性就很高，反之完整性就很差，综合匹配度进行打分，打分方式为：$W=n\times 0.4$

5.3.4 重复性判断

如果说对于同一类留言详情的答复意见相似度特别高或相同的话，那么就可以认为这一次的留言答复效果很差。通过改进的聚类循环匹配算法对留言详情进行匹配分类，计算出属于同一类的答复意见文本的相似度 S，相似度大于一定阈值说明答复意见过于敷衍，将会影响答复意见的质量，打分方式为：$C=S\times 0.2$

5.3.5 答复时间

去除留言时间和答复时间的时分秒，保留年月日，计算留言时间与答复时间的时间差，如果时间差越短，说明回复积极，反之越消极，其中答复时间分数 T=（答复时间-留言时间）（单位：天）/50，例如答复时间差为 10 天，那么 T=0.2

5.3.6 评分模型

结合答复的相关性、完整性、可解释性、答复时间四个不同方面的分数，计算出答复效率 D，如果时间差距越短，质量越好，那么其答复效率就越高，反之效率越低，将答复效率 D 做为最终的评价得分。

最终评分公式为： $D=\left(X+W-C\right)/T$

VI.参考文献

[1]刘小冬.自然语言理解综述[J].统计与信息论坛,2007(02):5-12.

[2]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.

[3]李楚贞,余育文.中文微博数据预处理常用方法研究[J].科技经济导刊,2019,27(33):23.

[4]覃世安,李法运.文本分类中 TF-IDF 方法的改进研究[J].现代图书情报技术,2013(10):27-30.

[5]石凤贵.基于 TF-IDF 中文文本分类实现[J].现代计算机,2020(06):51-54+75.

[6]孙广路,齐浩亮.基于在线排序逻辑回归的垃圾邮件过滤[J].清华大学学报(自然科学版),2013,53(05):734-741.

[7]谢忠红,张颖,张琳.基于逻辑回归算法的微博水军识别[J].微型机与应用,2017,36(16):67-69+72.

[8]张瑞生.板式反应精馏塔的模拟计算——修正的 Newton-Raphson 法[J].华东化工学院学报,1989(01):25-32.

[9]Salih Güneş,Kemal Polat,Şebnem Yosunkaya. Multi-class f-score feature selection approach to classification of obstructive sleep apnea syndrome[J]. Expert Systems With Applications,2009,37(2).

[10]官琴；邓三鸿；王昊,中文文本聚类常用停用词表对比研究[J],2017

[11]李航,唐超兰,杨贤,沈婉婷.融合多特征的 TextRank 关键词抽取方法[J].情报杂志,2017,36(08):183-187.

[12]蒲梅,周枫,周晶晶,严馨,周兰江.基于加权 TextRank 的新闻关键事件主题句提取[J].计算机工程,2017,43(08):219-224.

[13]王石,曹存根,裴亚军,夏飞.一种基于搭配的中文词汇语义相似度计算方法[J].中文信息学报,2013,27(01):7-14.

[14]陈斌,基于 HLDA 与 CNN 相结合的短文本分类算法研究[D],南京航空航天大学,2018