基于自然语言处理的热点问题挖掘

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。

针对问题一，我们首先对数据进行预处理，然后使用向量空间模型来转化精简表示留言，为了均衡不同的特征项对文档的重要程度和区分影响能力的强弱，对特征项进行权重计算,最后使用基于朴素贝叶斯的多项式文本分类模型将其进行分类。使用 F-Score 对该分类方法进行评价准确率为 92.31%，因此模型具有较高的可靠性。

针对问题二，我们延用第一问预处理后的用户留言的结果，并将处理后的数据进行密度聚类,然后跟据关联规则算法对可以代表热点问题的特征词进行挖掘，再通过热度计算公式来识别这些热点问题，得到“热点问题留言明细表”。

针对问题三，采用向量空间模型，计算留言答复的相关性，完整性、可解释性的基础是相关性较高，满足相关性较高也就基本实现完整性和可解释性。

关键词:向量空间模型　朴素贝叶斯 多项式模型　密度聚类　关联规则
相似度计算

Abstract

In recent years, with WeChat such as weibo mayor mailbox sunshine hotline network asked ZhengPing stage gradually become the government understand the importance of public opinion gathering intelligence condensed bull channels, all kinds of public opinion related text data quantity rising, leave a message to past mainly rely on artificial to divide and hot spots of relevant departments work has brought great challenge at the same time, as the big data cloud computing technologies such as artificial intelligence development, establish the wisdom of the e-government system based on natural language processing technology has is the new trend of development of social management innovation, to enhance the management level of government and governance efficiency has a great role in promoting

To solve the problem one, we first preprocess the data, and then use vector space model to transform and simplify the representation of messages. In order to balance the importance of different feature items on the document and the ability to distinguish the influence, we calculate the weight of the feature items, and finally use the polynomial text classification model based on Naive Bayes to classify them. The accuracy of F-score is 92.31%, so the model has high reliability.

To solve the second problem, we continue to use the result of the user's message after the first question preprocessing, and cluster the processed data with density, then mine the feature words that can represent the hot issues according to the association rule algorithm, and then identify these hot issues through the heat calculation formula to get the "hot issues message list".

To solve the third problem, the vector space model is used to calculate the correlation of message replies. The basis of completeness and interpretability is high correlation.

Keywords: Vector space model，Naive Bayes，Polynomial model，Density clustering，Association rule algorithm，Similarity calculation

目录

1.挖掘目标

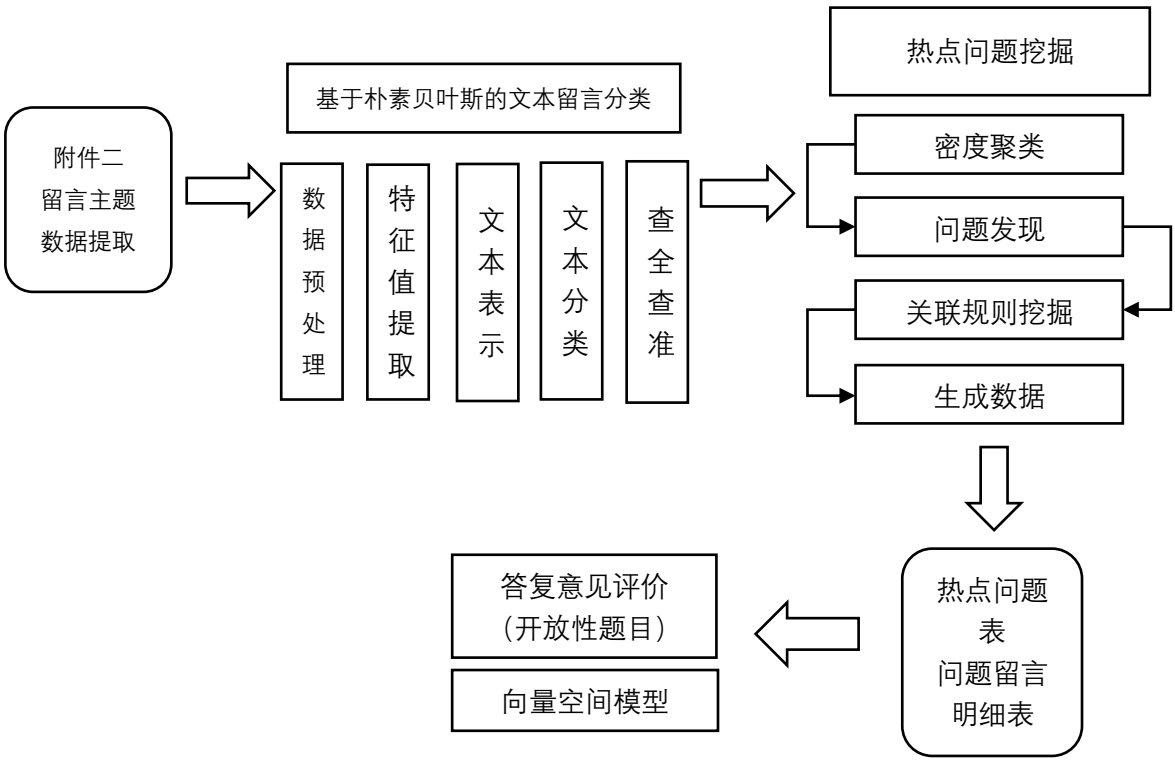
1.1 挖掘背景

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。 建立基于自然语言处理技术的智慧政务系统。

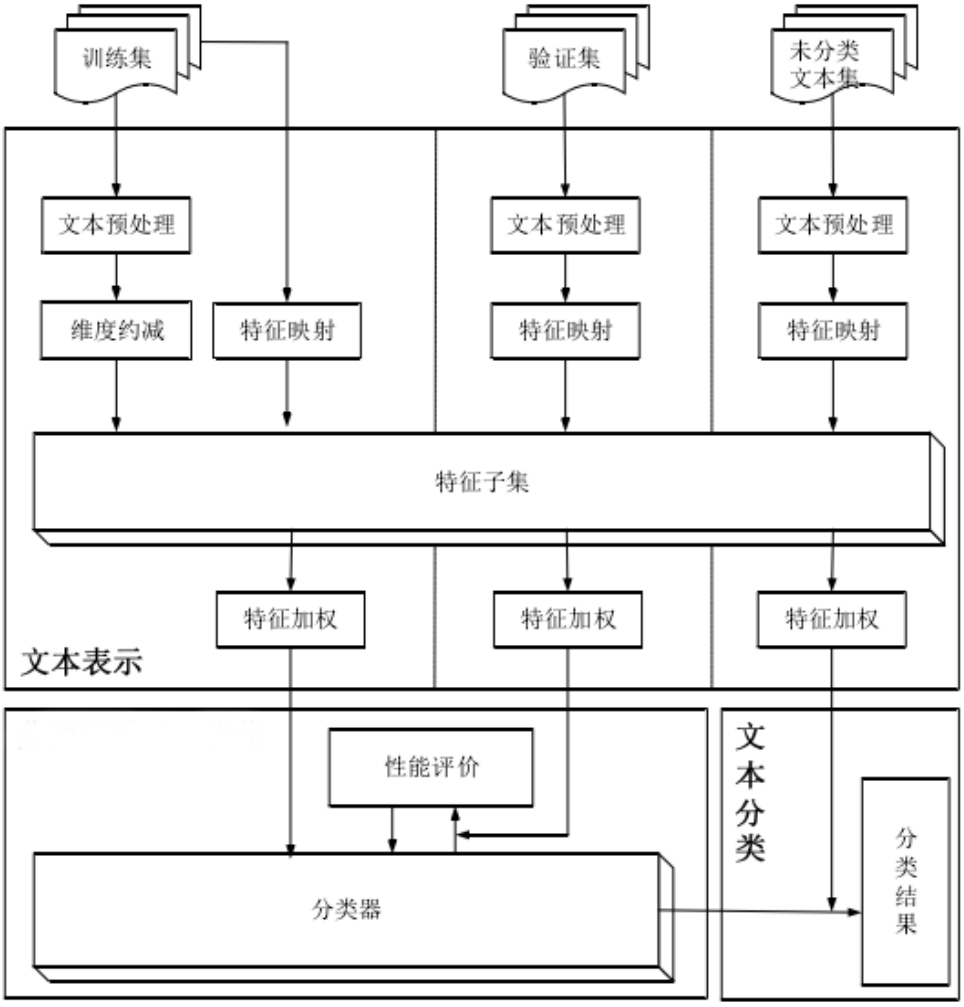
1.2 挖掘目标

在现有的政务收集信息系统下，设计对于留言的内容做到一级标签分类，对热点问题挖掘的模型，并将分类结果以 csv 格式存储，将热点问题另存为（只需要找出热点问题），并提出对于评价完善程度的综合评价模型。

2.全文脉络图



3.具体步骤



3.1 文本预处理

由于数据中留言主题更加明确清晰，并且文本较留言明细精炼，减少数据加快了运行程序，我们采用对留言主题进行数据预处理。

【读取的文本为】

镜中的她有着棕黄色的头发， 直挺的鼻梁，以及如同大理石般的白色肌肤……。皇后非常重视肌肤的保养，不惜重金从先进的法国买来护肤的配方，并使用各种草药制成油膏，每天早上都给肌肤做最好的按摩。 然而岁月不饶人，皇后的美貌、是有衰退的一天。 不知从何时开始，皇后的肌肤已逐渐松弛， 眼角出现了细纹，而国王也似乎不再那么的享受鱼水之欢了；看来，国王已经对皇后不再感兴趣。

皇后当然也听说过国王想在 贵族千金 中寻找宠妃的传闻，因为在不打仗的时候，即使留在城内，国王也把大部分时间都花在探访皇亲国戚上。

1）去除空格、标点等无关信息

文本预处理是将半结构化或非结构化的文本转换为适当的文本表示形式的

必经阶段。首先，删除文本中的所有特殊字符、标点符号、数字等字符。我们选择 Python 自带的 punctuation 包，可以消除所有的中文标点符号。我们再对文本进行去除空格、\t、数字。

【效果】

镜中的她有着棕黄色的头发直挺的鼻梁以及如同大理石般的白色肌肤皇后非常重视肌肤的保养不惜重金从先进的法国买来护肤的配方并使用各种草药制成油膏每天早上都给肌肤做最好的按摩然而岁月不饶人皇后的美貌是有衰退的一天不知从何时开始皇后的肌肤已逐渐松弛眼角出现了细纹而国王也似乎不再那么的享受鱼水之欢了看来国王已经对皇后不再感兴趣皇后当然也听说过国王想在贵族千金中寻找宠妃的传闻因为在不打仗的时候即使留在城内国王也把大部分时间都花在探访皇亲国戚上

2）文本分词操作

中文由于其特殊性，中文文本单词之间没有天然的分隔符，所以我们在提取文本特征之前，首先要对中文文本进行分词。分词的处理过程是中文信息处理中所特有的步骤。分词可以将连续的汉字序列按照一定的规则重新分割为词或词组。切分好的词或词组将会作为文本的特征用于留言分类分析过程，所以说能否高效、正确的对中文进行分词处理成为留言分类的重要任务。我们将采用中国科学院计算技术研究所专门开发的汉语语法分析系统 ICTCLAS（Institute of Computing Technology，Chinese Lexical Analysis System）。ICTCLAS 的主要功能包括中文分词、词性标注、新词识别、命名识别等功能，他的分词性能和分词精度都较高，是目前最受好评的汉语分词开源系统。

【效果】

镜中/的/她/有着/棕黄色/的/头发/直挺/的/鼻梁/以及/如同/大理石/般的/白色/肌肤/皇后/非常重视/肌肤/的/保养/不惜重金/从/先进/的/法国/买来/护肤/的/配方/并/使用/各种/草药/制成/油膏/每天/早上/都/给/肌肤/做/最好/的/按摩/然而/岁月不饶人/皇后/的/美貌/是/有/衰退/的/一天/不知/从/何时/开始/皇后/的/肌肤/已/逐渐/松弛/眼角/出现/了/细纹/而/国王/也/似乎/不再/那么/的/享受/鱼水之欢/了/看来/国王/已经/对/皇后/不再/感兴趣/皇后/当然/也/听说/过/国王/想/在/贵族/千金/中/寻找/宠妃/的/传闻/因为/在/不/打仗/的/时候/即使/留在/城内/国王/也/把/大部分/时间/都/花/在/探访/皇亲/国戚/上

3）去除停用词

用户留言所构成的中文文本中包含许多助词、虚词等词性的单词以及在中文文本中经常出现的高频词汇但是其本身对于留言分类意义不大，这些词汇我们称他们为停用词（Stopword）。停用词的存在不但会浪费存储空间，而且有很大概率形成噪声，影响留言分类的精度，在中文文本的预处理的过程中应该将文本的包含的停用词删除。但是随着文本分类数据集的扩大，停用词往往并不局限于停词表中的词汇，所以我们也可以通过比较摸个词汇的数量是否超过我们的阈值来

确定我们的词汇是否为停用词，但是仅仅通过词汇数量来确定停词方法在数据不平衡的情况之下，就会有可能删除关键词。例如在不平衡环境下，某个类可能包含数量较多的文本，因此代表该类的关键词出现在该类中的文本数量就可能较多，所以这些关键词不可以认定为停用词。

【效果】

镜中/有着/棕黄色/头发/直挺/鼻梁/如同/大理石/般的/白色/肌肤/皇后/非常重视/肌肤/保养/不惜重金/先进/法国/买来/护肤/配方/使用/草药/制成/油膏/每天/早上/肌肤/最好/按摩/岁月不饶人/皇后/美貌/衰退/一天/不知/皇后/肌肤/逐渐/松弛/眼角/出现/细纹/国王/似乎/享受/鱼水之欢/国王/皇后/感兴趣/皇后/听说/国王/贵族/千金/寻找/宠妃/打仗/留在/城内/国王/大部分/时间/探访/皇亲/国威/

3.2 文本表示

1）向量空间模型

由于中文文档是大量字符的集合，是非结构化或半结构化的数字信息，所以不能直接被任何分类器所识别，必须将其转换成为一个简洁的、统一的、能够被学习算法和分类器所识别的结构化形式，才能够进行进一步的研究和分析。目前文本表示通常采用向量空间模型（Vector Space Model，VSM）用 VSM 来进行中文文本表示，关系到特征项和特征权重两个概念。

特征项：特征项是向量空间模型中不可分的最小的语言单元。可以是字，词，句子。我们一般采用词作为特征项。一个文本的内容可以认为是它所含有的特征项所组成的集合，表示为 $D=(t_1,t_2,...t_m)$ 其中 t_k 是特征项，$1\le k\le m$

特征值权重：对于含有 m 个特征项的文档 $D=(t_1,t_2,...t_m)$，每个特征项 t_k 都根据一定的原则被赋予一个权重 w_k，表示表示它们在文档中的重要程度。这样一个文档 D 可用它含有的特征项及其特征项所对应的权重表示：$D=D(t_1,w_1;t_2;w_2;...;t_m,w_m)$，可简单记为 $D=D(w_1,w_2,...,w_m)$，其中 w_k 就是特征项 t_k 的权重，$1\le k\le m$ 称 $D=D(w_1,w_2,...,w_m)$ 为文档 D 的向量空间模型。

采用向量空间模型进行文本表示，必须经过两个关键步骤：首先根据训练样本集将文本表示成特征项序列 $D=(t_1,t_2,...t_m)$；然后根据文本特征项序列，对训练样本集中的各个文档进行权重赋值、规范化等处理，将其转化为所需的向量。图为向量空间模型下的文本表示，其中每行表示一个文本向量，每列表示一个特征

项， w_{ij} 表示第 j 个特征项在第 i 个文档的权重。

	t_1	…	t_j	…	t_m
D_1	w_{11}	…	w_{1j}	…	w_{1m}
…	…	…	…	…	…
D_i	w_{i1}	…	w_{ij}	…	w_{im}
…	…	…	…	…	…
D_n	w_{n1}	…	w_{nj}	…	w_{nm}

2）特征权重

文档使用向量空间模式表示后，为了权衡不同的特征项对文档的重要成度和区分影响力的强弱，需要对特征项进行权重计算。权重的调整一般从两方面考虑：一个词在某篇文档中出现的次数越多，则对识别文档的贡献越大；一个词在不同的文档中出现的次数越多，则它区分不同文档的能力越弱。权重计算的一般方法是利用文本的统计信息，主要是词频。在这里我们采用 TF-IDF 权重特征权重计算方法，其中使用变量的说明如下： w_{ij} 表示特征项 t_j 在文本 D_i 中的权重； tf_{ij} 表示特征项 t_j 在文本 D_i 中出现的频数； n_j 是训练样本集中出现特征项 t_j 的文档数；N 是训练样本集中总的文档数。

TF-IDF(term frequency－inverse document frequency)权重是一种非常常用的的计算权重的方法，即“词频与倒排文档频数”。权重与特征项在文档中出现的频率成正比，即特征项在文档中出现的次数越多就越重要；同时与语料库中含有该特征项的文档数成反比，即认为特征项在不同文档出现的频率越大，该特征项的重要性就越低。

$$w_{ij} = tf_{ij} \times \log \frac{N}{n_j}$$

当 $N=n_j$ 时，权重为 0，为此进行平滑处理，如下式所示：

$$w = \log(tf_{ij} + 1.0) \times \log(\frac{N + 1.0}{n_j})$$

3.3 基于朴素贝叶斯的多项式文本分类模型

1）朴素贝叶斯文本分类算法

朴素贝叶斯文本分类分类的任务就是将其表示成为向量的待分类文档 $D_i(x_1,x_2,...,x_n)$ 归类到与其关联最紧密的类别集合 $C=\{C_1,C_2,...C_m\}$ 中的某一类。其中 $D_i(x_1,x_2,...,x_n)$ 为待分类文档 D_i 的特征向量，$C=\{C_1,C_2,...C_m\}$ 为给定的文档类别集合。也就是说，求解向量 $D_i(x_1,x_2,...,x_n)$ 属于给定类别 $C_1,C_2,...,C_m$ 的概率值 $(p_1,p_2,...,p_m)$，其中 p_j 为 $D_i(x_1,x_2,...,x_n)$ 属于 C_j 的概率，则 $\max(p_1,p_2,...,p_m)$ 所对应的类别就是文档 D_i 所属的类别。假设 D_i 为任意一文档，根据贝叶斯分类器，文档 D_i 属于 C_j 的概率为：

$$P(C_j \mid D_i)=\frac{P(C_j)P(D_i \mid C_j)}{P(D_i)}=\frac{P(C_j)P(x_1,x_2,...,x_n \mid C_j)}{P(x_1,x_2,...,x_n)} \quad (1)$$

其中 $P(x_1,x_2,...,x_n)$ 对应所有类均为常量，所以只需估算 $P(C_j)P(x_1,x_2,...,x_n \mid C_j)$，求解式(1)的最大值可转化为如下公式：

$$c(D_i)=\arg\max P(x_1,x_2,...,x_n \mid C_j)P(C_j) \quad (2)$$

朴素贝叶斯文本分类的一个前提假设是：在给定的文档类别下，文档属性即特征项是相互独立的。即：

$$P(x_1, x_2, ..., x_n \mid C_j) = \prod_{i=1}^{n} P(x_i \mid C_j) \quad (3)$$

所以式(2)可简化为

$$c(D_i) = \arg\max P(C_j) \prod_{i=1}^{n} P(x_i \mid C_j) \quad (4)$$

朴素贝叶斯文本分类的关键就是计算 $P(C_j)$和$P(x_i \mid C_j)$ 的过程就是建立分类模型(或者说学习)的过程，通过计算 $P(C_j)$和$P(x_i \mid C_j)$，求出后验概率最大的类别。

根据 $P(D_i \mid C_j)$ 计算方式的不同，朴素贝叶斯分类方法可分为最大似然模型（Maximum Likelibood Model，MLM）、多变量伯努利模型（Multi-variate Bernoulli Model，MBM）、多项式模型（Multinomial Model，MM）、泊松模（Poison Model，PM）等。我们主要讨论多项式模型。

3）多项式模型

在多项式模型中，我们考虑了特征项在用户留言中出现的次数。用户留言被看做一系列单词序列，并且假定留言长度和类别无关，而且用户留言中出现的任何一个词与它在文档中的位置以及上下文无关。留言属于类C_j时特征词x_i出现一次的概率为$P(x_j|C_j)$，假定共有 n 个词，则$n = n_1 + n_2 + \cdots + n_k$，则有

$$P(D_i \mid C_j) = n! \prod_{i=1}^{n} \frac{P(x_i \mid C_j)^{n_k}}{n_k!}$$

在多项式模型中，$p(x_i|C_j)$再用词频估算：

$$P(x_i \mid C_j) = \frac{\sum_{k=1}^{|D|} N(x_i, D_k)}{\sum_{i=1}^{|V|} \sum_{k=1}^{|D|} N(x_i, D_k)}$$

其中，$\sum_{k=1}^{|D|} N(x_i, D_k)$ 表示特征项 x_i 在类 C_j 的各文档中出现的次数之和；

$\sum_{i=1}^{|V|}\sum_{k=1}^{|D|}N(x_i,D_k)$为类$C_j$中所有特征项的总次数。

为了避免出现零概率，通常加入平滑因子，其中 m=V，p=1/$|V|$，如下所示：

$$P(x_i \mid C_j)=\frac{\sum_{k=1}^{|D|}N(x_i,D_k)+1}{\sum_{i=1}^{|V|}\sum_{k=1}^{|D|}N(x_i,D_k)+|V|}$$

其中，V 是训练样本的单词表(即抽取单词，单词出现多次，只算一个)。

在多项式模型中，先验概率$P(C_j)$的计算如下：

$$P(C_j)=\frac{\text{类}C_j\text{中特征项总数}}{\text{训练样本的特征项总数}}$$

3.4.性能评价

我们使用 F-Score 对分类方法进行评价，其中P_i为第 i 类的查准率，R_i为第二类的查全率。

$$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$$

我们使用四个符号表示分类的所有情况：

- TP(真阳性):正样本被正确预测为正样本
- FP(假阳性):负样本被错误预测为正样本
- TN(真阴性):负样本被正确预测为负样本
- FN(假阴性):正样本被错误预测为负样本

说明：正样本为选定的第 i 类的留言，正确分类即为正确预测，负样本为第 i 类之外的留言，正确分类即为正确预测。将用户留言分类问题转化为二分类问题。

1）查准率

关注预测为正样本的数据(可能包含负样本)中,真实正样本的比例

计算公式

$$Precision=\frac{TP}{TP+FP}$$

2）查全率

关注真实正样本的数据(不包含任何负样本)中,正确预测的比例

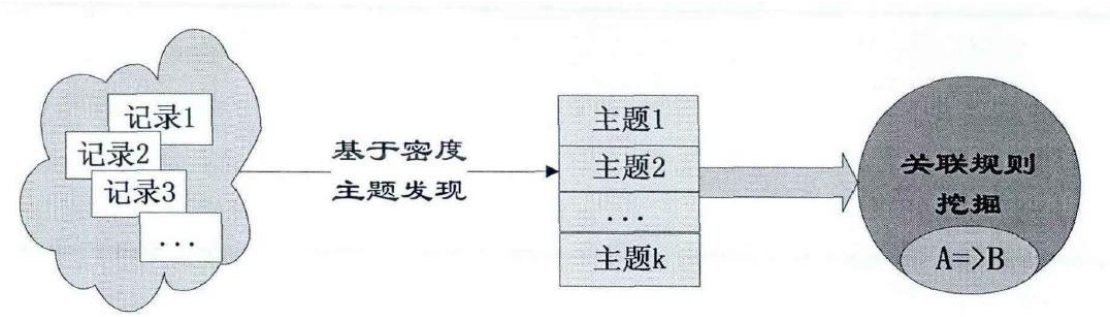
计算公式

$$Recall = \frac{TP}{TP + FN}$$

3.5 基于密度聚类和关联规则的热点问题挖掘模型

1）基于密度聚类的热点问题发现

对于未知的热点问题，需要先对用户留言样本进行聚类用来发现一些已经存在的热点问题，接下来对热点问题进行关联规则的挖掘。



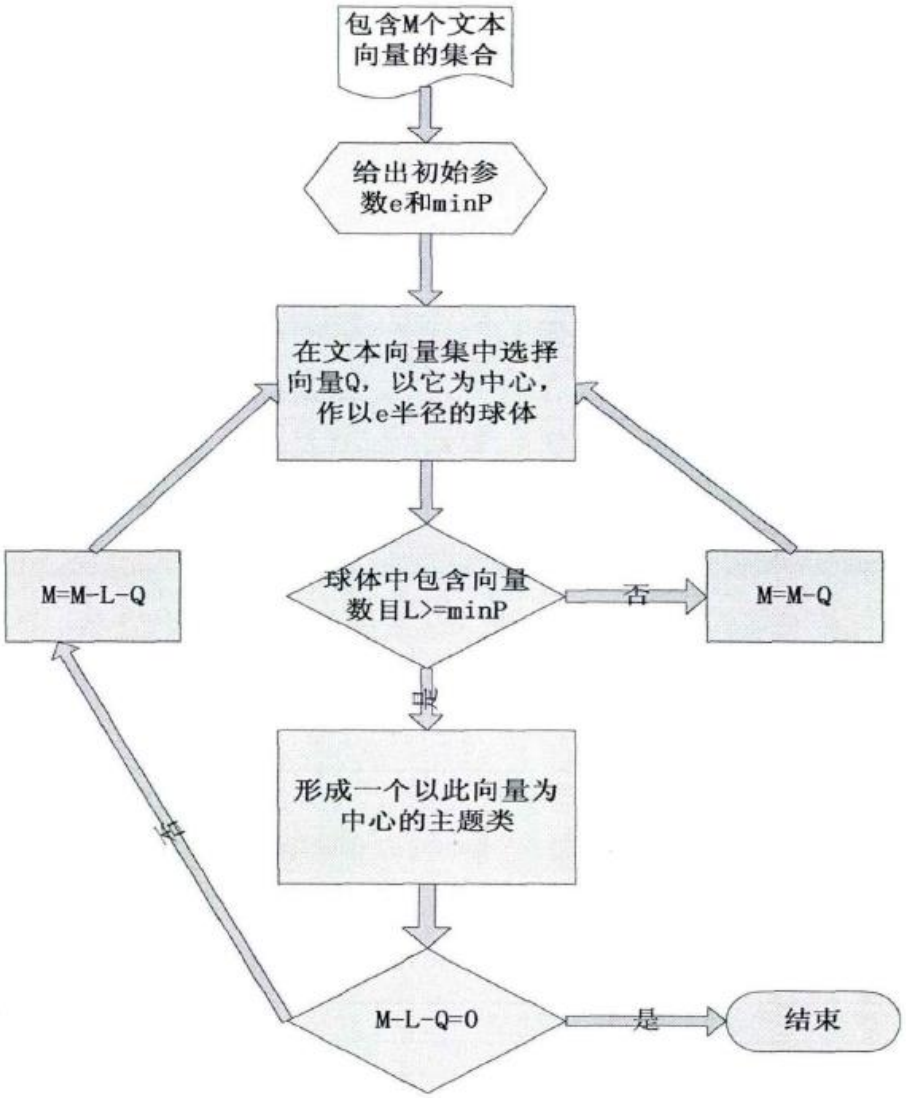
基于密度的聚类算法的特点：这种算法是以密度为基础来衡量，而不是距离，这里密度指的是单位空间内样本点的数目，它根据样本空间各处的不同密度，把相似密度的点聚集成一类。我们使用基于高密度链接区域算法。e表示聚类中领域半径的大小，minp表示以 e 为邻域半径的球体内应该包含的最少的点数。

算法描述：

如果用户留言主题集合中还有剩余的记录能被取出，就进入下一步，否则聚类完成，退出。

如果以某点为中心点，e 为半径的球体空间内包含了至少 minp 个样本点的话，就能够以此点为中心形成一个主题类，并返回第一步，如果没有发现这样的中心点，就进入下一步。

如果要判断的一点已经被划分到已经形成的主题类中就继续返回第一步，寻找下一个中心点。



2）关联规则挖掘

采用基于密度的聚类后，就对向量空间中的投诉文本进行了初步主题的划分，下一步是根据关联规则算法对可以代表热点问题的特征词进行挖掘，再通过热度计算公式来识别这些热点问题。

由于热点问题的形成需要一定的时间，所以此处引入时间度量，本文标记时间集合为$T=\{t_1,t_2,...,t_i...\}$，代表时间轴上的有序集合，在本文中用“天”作为时间的基本单位。

经过之前的聚类分析之后，将输入的投诉文本集合聚集为 k 个主题类别，先从中抽取出 m 个可能为热点问题的主题类别，命名为热点问题初始集合，标记为$CSet=\{C_1,C_2,...,C_m\}$，其中$C_j,1\le j\le m\le k$表示第 j 个可能的热点问题类别，每

个类别包含一定数目的样本记录，表示为$C_j=\{x_{j1},...,x_{jy}\}$,其中 y 表示该类别包含的样本数目，单个样本表示为$x_j=(w_{1j},w_{2j},...,w_{pj})$。加入时间度量之后第$t_i$日的热点问题初始集合为$CSet(t_i),m(t_i)$为初始集合包含的样本记录总数，$mC_j(t_i)$表示$C_j$中包含的样本数。

经过上一章对未知主题的热点问题特征分析可知，可通过考察热点问题初始集合随时间的增长速度来进行热点问题的识别，转化为可定量分析的数学模型如下:

(1)首先C_j需要满足一段时间内的连续可现性，这里的重现是指一段时间内存在与C_j紧密关联的相似类别，本文后面将使用关联规则来确定两个类别之间是否相关。

(2)在此基础上定义热点问题初始集合中某一个类别C_j的增长率为:

$$Ratio=\frac{mC_j(t_i)-mC(t_{i-1})}{mC(t_{i-1})},t_i,t_{i-1}\in T$$

经过连续时期内对同一个类别C_j进行考察，若增长率均大于规定的最小阈值$Ratio_{min}$时，可以认为C_j类别就是一个热点问题。根据上面的定义，可以将这个增长率指标 Ratio 定义为热度指标，因为正是对这一指标的考察确定了热点问题所属的类别。

假设从C_j的样本记录包含的特征词集合中，挖掘到热点相关的特征词集合为$\{Term_1,Term_2,...,Term_i,...,Term_y\}$，则可以用此集合代表该类热点问题。当已知一个热点问题类别C_j，如何判断其他的类别是否与C_j描述的热点问题相关联，我们采用关联规则来判断两个集合的关联性。

关联规则可以反映出数据集合的内在关系，它涉及两个基本的概念:

(1)对于两个类别的文本集合 A 和 B，支持度 support 定义为同时包含 A 和 B 文本的概率，如下式所示，其中 A 称为规则的前件，B 称为后件。

$$support(A\Rightarrow B)=P(A\cup B)$$

(2)置信度表示 A，B 相关联的可信度有多大，用公式表示是

$$Confidence(A \Rightarrow B) = \frac{P(A \cup B)}{P(A)}$$

定义最小支持度$\text{Support}_{\min}$，最大支持度$\text{Support}_{\max}$，最小置信度$confidence_{\min}$若支持度符合式((3-25),置信度满足式((3-26)，增长率满足式({3-27)

$$Support_{\min} \le Support(A \Rightarrow B) \le Support_{\max} \quad \text{式(1)}$$

$$Confidence(A \Rightarrow B) \ge Confidence_{\min} \quad \text{式(2)}$$

$$Ratio \ge Ratio_{\min} \quad \text{式(3)}$$

那么 A 和 B 同属于热点问题的集合C_j，并且 A，B 包含的特征词集可以代表该类热点问题，此外可以根据导出的关联规则$A \Rightarrow B$，来判断其他文本集合是否属于该类热点问题。

通过对热点问题初始集合中的类别进行增长率的比较分析，从中可以发现符合特定热点问题的类别集合，之后可以进行热点问题展示以及进一步的预警。

3.6 基于空间向量模型答复意见评价

答复意见评价的角度有三个：相关性，完整性，可解释性。类似于问答系统，需要满足回答的这三个性质，才能对用户中心问题进行反馈解决。

谈及相关性就不能不提相似性，在文本研究领域上文本相关性与相似性有很大不同，举个例子来说，餐桌与午饭相关，餐桌与书桌相似，但午饭跟书桌却不是一类。由于目前被广泛使用的方法大多通过比较相似程度来模拟相关性的度量，所以即使比较的最终目的是对相关性进行度量，在计算方法中北京场使用的仍然是相似一词。文本之间的相关性计算包含了词汇、语句以及文档等多个级别语言单位上的运算。

语句间相关性的度量机制往往依赖于对语句进行分析的层次和深度。在语句相关性计算中，对语句的分析深度，按照向量空间模型的思考方式，把句子看成是由若干词组成的线性序列，抛开语句的语法结构，忽略对其语法结构的分析，只能利用句子的表层信息，即组成句子的词的词频、词性等信息完成语句

间的匹配度量。向量空间模型具有直观、简洁、高效等的优点，单纯基于相同词型的方法和使用语义词典进行同义词替换的方法，不论是否对同义词的影响进行了考量，由于没有进行任何的结构分析，这些方法在计算语句之间的相关性时都没有考虑到句子结构对相关性的影响。这类方法所依赖的信息为留言问题，会达到较高的准确率。

在向量空间模型中，进行相似性比较的双方被表示为多次元向量空间，模型假设，两者间的相似程度，可以经由比较每个向量间的夹角偏差程度而得知。

在由 Salton，Wong and Yang 提出的古典的向量空间模型中，一个词在文档向量中的权重，为区域和全域参数的乘积，即所谓的 TF-IDF（词频-逆向文件频率 ），在之前问题分析中也用到了 TF-IDF，词频在相关性，相似性中发挥着重要作用。通过向量空间模型计算问答之间的相关性，相关性越高，对答复意见评价的指标也就越高。

考虑到对留言用户的答复为较权威机构的答复，可解释性是在相关性基础上的。那么，总体看来，留言满足完整性和可解释性的基础是相关性，所以满足相关性，或相关性较高，也就满足了完整性和可解释性，量化相关性、完整性、可解释性就是计算相关性。

4.结论

1）数据处理过程中发现，长文本的无意义表达太多，越庞大的数据系统运行越缓慢，要适当采用更加精简的数据，同时还会避免文本语义交叉带来的问题。所以我们采用了对附件二中的留言主题的数据处理。之后对文本数据提取特征值更加有利于后续有范围限制的热点问题的挖掘。标签分级关系对数据进行多项式朴素贝叶斯分类是最常见的一种分类模型。

2）在热点问题挖掘中，特征多，计算量大，就需要改进相似度计算方案，开始采用余弦相似度计算，之后我们改为基于密度机制的相似度计算，大大减少了数值计算，图形更加直观更易理解。对关联性的挖掘，识别热点问题，导出热点问题，再加入明细表中。另外，由于热点问题挖掘是在某一时间段内对某一特定地点和特定人群的识别，划分这个范围先由数据处理后的地点项提取在对其时间段具体划分，这样更加容易查找热点问题并且还降低查找的错误率。

3）对于最后一问中的相关性，其实向量空间模型还有很多不足之处，缺乏理

论基础；向量空间中各词彼此独立。而且，在自然语言处理的应用中，传统的相量空间模型的缺点还包括：词的顺序被忽略；一词多义和一义多词问题被忽略。可以考虑系统相似模型的搭配使用降低文本语义理解的错误率。这样依托系统相似理论，一条自然语言语句文本被看作是一个系统，语句中的词组或词语是系统的组成要素或元素，语句间的相关性计算就是两个系统间相似性的度量。

5.参考文献

[1]廖胜兰,吉建民,俞畅,陈小平.基于 BERT 模型与知识蒸馏的意图分类方法[J/OL]. 计算机工程:1-8[2020-05-08].https://doi.org/10.19678/j.issn.1000-3428.0057416.

[2]杨云龙,孙建强,宋国超.基于 GRU 和胶囊特征融合的文本情感分析[J/OL].计算机应用:1-6[2020-05-08].http://kns.cnki.net/kcms/detail/51.1307.TP.20200429.1704.010.html.

[3]Livia Celardo,Martin G. Everett. Network text analysis: A two-way classification approach[J]. Elsevier Ltd,2020,51.

[4]陈欢,黄勃,朱翌民,俞雷,余宇新.结合 LDA 与 Self-Attention 的短文本情感分类方法[J/OL]. 计算机工程与应用:1-8[2020-05-08].http://kns.cnki.net/kcms/detail/11.2127.TP.20200423.1322.025.html.

[5]张洋,胡燕.基于多通道深度学习网络的混合语言短文本情感分类方法[J/OL].计算机应用研究:1-7[2020-05-08].https://doi.org/10.19734/j.issn.1001-3695.2019.12.0616.

[6]周伟枭,蓝雯飞.融合文本分类的多任务学习摘要模型[J/OL].计算机工程:1-10[2020-05-08].https://doi.org/10.19678/j.issn.1000-3428.0057448.

[7]倪海清,刘丹,史梦雨.基于语义感知的中文短文本摘要生成模型[J/OL].计算机科学:1-9[2020-05-08].http://kns.cnki.net/kcms/detail/50.1075.TP.20200328.2133.006.html.

[8]宋国民,张三强,贾奋励,姜松言.中文文本中时间信息抽取及规范化方法[J].测绘科学技术学报,2019,36(05):538-544.

[9]杨锋.基于线性支持向量机的文本分类应用研究[J].信息技术与信息化,2020(03):146-148.

[10]张琳,李朝辉.文本分类中一种改进的特征项权重计算方法[J].福建师范大学学报(自然科学版),2020,36(02):49-54.

[11]范国凤,刘璟,姚绍文,栾桂凯.基于语义依存分析的图网络文本分类模型[J/OL]. 计 算 机 应 用 研 究 :1-5[2020-05-08].https://doi.org/10.19734/j.issn.1001-3695.2019.08.0522.

[12]李丽华,胡小龙.基于深度学习的文本情感分析[J].湖北大学学报(自然科学版),2020,42(02):142-149.

[13]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.

[14]薛金成,姜迪,吴建德.基于 word2vec 的专利文本自动分类研究[J].信息技术,2020,44(02):73-77.

[15]程勇,徐德宽,董军.基于语文教材语料库的文本阅读难度分级关键因素分析与易读性公式研究[J].语言文字应用,2020(01):132-143.

[16]毕硕本,徐瑞壮,万蕾,刘爱利,盛宇裕.基于多因素排序的南京市出租车乘客候车热点区域挖掘方法[J].中国科技论文,2020,15(01):23-30.

[17]邓朗妮,赖世锦,兀婷,廖羚,钟锰军.基于数据挖掘技术的 BIM 学术热点与学术趋势分析方法研究[J].土木建筑工程信息技术,2019,11(06):1-10.

[18]金吉琼,刘鸿,郑赛晶.基于在线评论文本挖掘技术的电子烟市场消费热点分析[J].烟草科技,2019,52(12):106-114.

[19]范彦勤,覃杨森,袁媛.基于主成分分析的贝叶斯网络在个人信用评估中的应用[J].桂林航天工业学院学报,2019,24(04):568-575.

[20]李学靖,张小艳,马睛雅,王斗,丛雪,曲畅,郝玉芳.基于数据挖掘及共词分析法的糖尿病中医护理研究热点分析[J].中华现代护理杂志,2019(26):3381-3385.

[21]张航. 基于朴素贝叶斯的中文文本分类及 Python 实现[D].山东师范大学,2018.

[22]方莹. 面向热点新闻话题的文本处理技术研究[D].北京理工大学,2015.

[23]王子慕. 一种利用 TF-IDF 方法结合词汇语义信息的文本相似度量方法研究[D].吉林大学,2015.

[24]樊小超. 基于机器学习的中文文本主题分类及情感分类研究[D].南京理工大学,2014.

[25]杨杰明. 文本分类中文本表示模型和特征选择算法研究[D].吉林大学,2013.

[26]时志芳. 移动投诉信息中热点问题的自动发现与分析[D].北京邮电大学,2013.

[27]李丹. 基于朴素贝叶斯方法的中文文本分类研究[D].河北大学,2011.

[28]赵玉茗. 文本间语义相关性计算及其应用研究[D].哈尔滨工业大学,2009.