第八届泰迪杯

参赛赛题：C 题

2020 年 5 月

“智慧政务”中的文本挖掘应用

摘要

智慧政务即通过“互联网+政务服务”构建智慧型政府，利用云计算、移动物联网、人工智能、数据挖掘、知识管理等技术，实现由“电子政务”向“智慧政务”的转变。为了相关部门能够及时、准确地了解到市民的反馈意见和建议并及时解决群众问题，建立出基于自然语言处理技术（NLP）的智慧政务系统将具有极大的推动作用。

针对本次赛题提出的三个问题:

对于第一问，我们通过对比常见的 6 种文本分类算法，优先选取了逻辑回归算法和支持向量机算法进行文本分类，并建立准确率接近 90%的分类模型，最后使用 F-score 评分法进行评价。

对于第二问，我们使用 k-means 聚类算法和 DBSCAN 聚类算法对文本进行聚类，定义合理的评价热点问题的标准，得到热度前 5 的留言分类。

对于第三问，我们采用文本相似度对留言的回复情况作相关性的考察，相似度是比较两个事物的相似性。本题利用的是余弦相似度，即将文本映射到向量空间，再利用余弦距离进行相似度分析。运用相关性的结论以及一定的评价规则，进一步产生留言回复完整性的 11 类等级划分；综合对留言的回复相关性、完整性和可解释性这三方面对回复情况进行综合评价。

关键字：自然语言处理；逻辑回归；支持向量机；k-means 聚类；DBSCAN 聚类；余弦相似度

Abstract

Intelligent government is to build intelligent government through "Internet + government service ", using cloud computing, mobile Internet of things, artificial intelligence, data mining, knowledge management and other technologies to achieve the transformation from" e-government "to" intelligent government ". For the relevant departments to be able to timely and accurately understand the feedback and suggestions of the public and solve the mass problems in time, the establishment of a smart government system based on natural language processing technology (NLP) will have a great role in promoting.

Three questions in this competition:

For the first question, by comparing the common six text classification algorithms, we first select the logical regression algorithm and the support vector machine algorithm for text classification, and establish a classification model with an accuracy rate of nearly 90%. Finally, we use the F-score scoring method to evaluate.

For the second question, we use the k-means clustering algorithm and the DBSCAN clustering algorithm to cluster the text, define the reasonable criteria for evaluating hot issues, and get the message classification of the first 5 of the heat.

For the third question, we use text similarity to examine the response of the message, similarity is to compare the similarity of two things. This paper uses cosine similarity, that is, text mapping to vector space, and then using cosine distance for similarity analysis. Using the conclusions of relevance and certain evaluation rules, the 11 grades of message response integrity are further generated, and the three aspects of response relevance, integrity and interpretability are comprehensively evaluated.

Keywords: natural language processing; logical regression; support vector machine; k-means clustering; DBSCAN clustering; cosine similarity

目录

一、问题重述

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。

赛题希望答者对收集自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见，利用自然语言处理（NLP）和文本挖掘的方法解决下面的问题。

（1）**群众留言分类**：在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提供的内容分类三级标签体系）对留言进行分类，但存在工作量大、效率低，且差错率高等问题，请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。

并利用 F-score 对分类方法进行评分：

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

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

（2）**热点问题挖掘**：某一时段内群众集中反映的某一问题可称为热点问题。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。赛题要求根据附件 3 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，按一定的格式给出排名前 5 的热点问题。

（3）**答复意见的评价**：针对附件 4 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，并尝试实现。

针对问题一，我们要考虑长文本是否存在无意义的表达，是否需要转为短文本；要考虑数据不平衡带来的影响；要考虑文本语义带来的词语交叉。

我们使用计算机程序语言 Python 对导入的表格数据做预处理，预处理之后使用词频-逆文件频率（TF-IDF）这种统计方法对文本文档进行向量化，取特征

值，调用 Scikit-learn（sklearn）中的逻辑回归模型（LR）、线性判别分析（LDA）、K 近邻（KNN）、回归树（CART）、支持向量机（SVM）、高斯贝叶斯（NB）对文本进行分类，并通过比较这六种不同的分类结果，建立准确率较高的模型。

针对问题二，我们要找出某一时段内群众集中反映的某一问题，要做问题识别、问题归类、热度评价。核心的做法是文本聚类，按照统计频率的大小排序。做文本聚类时，我们使用了 K-means 聚类算法、DBSCAN 聚类算法对文本进行聚类，再将整个文本划分为 100 类。

关于热度的评价标准，我们考虑每条留言的点赞数和反对数，由于反对数在一定程度上也是在反应事件本身，于是我们对赞成数和反对数赋予不同的权重之后，再相加，与分类结果相结合，得到排名前 5 的热点问题。

针对问题三，我们要考虑如何将相关性、完整性和可解释性量化，要思考构建什么样的指标来计算和评价。在实际做题过程中，我们发现可解释性适合在相关性的基础上进行，于是我们将本题的评价顺序调整为：相关性、可解释性和完整性。

关于留言与留言回复之间的相关性，我们采用文本相似度对留言的回复情况作相关性的考察，相似度是比较两个事物的相似性，一般通过计算事物的特征之间的距离，如果距离小，那么相似度大；如果距离大，那么相似度小。经过欧几里得距离和余弦相似度的比较，我们选择了对本题适用性更强的余弦相似度。

可解释性和完整性都是在做完相关性的基础上，再加上自定义的规则进行的，后文将详细描述这两个自定义的规则。

二、群众留言分类

2.1 第一题思维导图

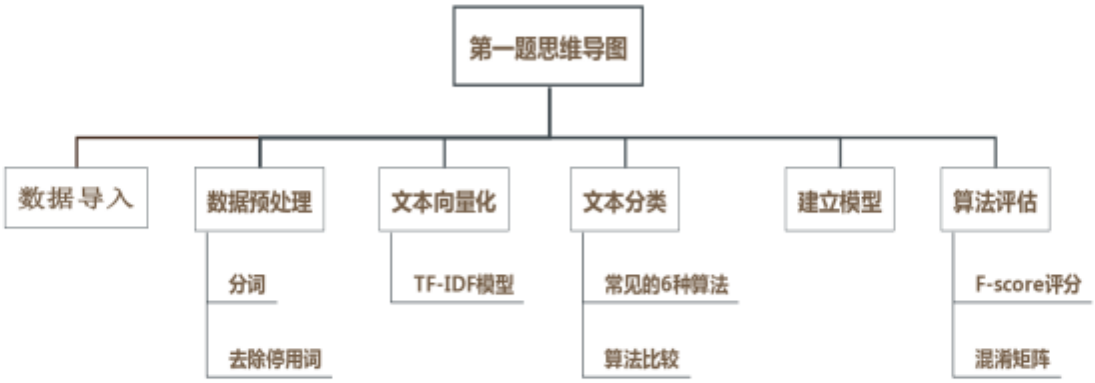

图1 第一题的思维导图

2.2 数据预处理

2.2.1 分词

对于题目所给的表格数据，提取出文本之后需要进一步做预处理。

汉语分词的主要任务是将汉语文本自动切分成词序列。由于词是自然语言中具有独立含义的最小的言语单位，而汉语文本中词与词之间有分隔标记，词语文本切分是汉语文本处理的第一步。

关于汉语言的分词方法，国内外有大量研究，最早起于词典的分词方法有最大匹配法、最短路径法，后来的基于n元语法的统计切割法，再到后来的由字构词的汉语分词方法等，人们先后提出过数十种切割方式。

本例题中，我们使用的是 jieba 分词。

Jieba 分词支持三种分词模式：

①精确模式：试图将句子最精确地切开，适合文本分析；

②全模式：把句子中所有的可以成词的词语都扫描出来， 速度非常快，但是不能解决歧义；

③搜索引擎模式：在精确模式的基础上，对长词再次切分，提高召回率，适

合用于搜索引擎分词。

2.2.2 去停用词

停用词（stop words）主要是指功能词（functional words），通常是指在各类文档中频繁出现的、附带极少文本信息的助词、介词、连词、语气词等高频词，如英文中的 the、is、that、which、on 等，汉语中的“的”、“了”、“是”等。为了减少文本挖掘系统的存储空间，提高运行效率，常常在文本表示时就自动将这些停用词过略掉。

这里我们利用一个常见停用词表，根据文件某些特性增添或删减部分词语，形成新的、适用于本题模型的停用词表。

预处理完成后的我们得到数据的词云图如下所示：

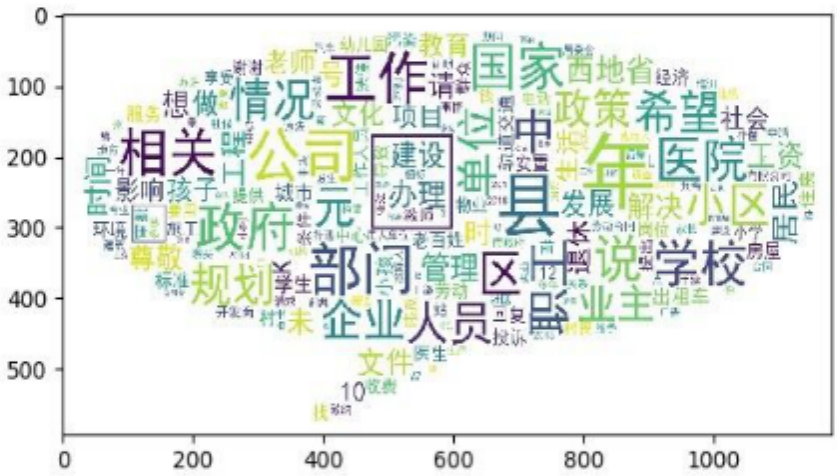

图 2 词云图

2.3 文本向量化

文本表示（text representation）：要想计算机能够高效处理真实文本，就必须找到一种理想的形式化表示方法。统计学习方法首先将输入的文本进行形式化，将其表示为向量或者其他形式，并基于形式化表示进行机器学习模型的训练和决策。这种将文本进行形式化的过程称为文本表示。

本文中，我们使用的是文本挖掘中最常见的一种文本表示模型——向量空间模型（vector space model，VSM）。

2.3.1 特征项

（1）特征项（feature term）

是 VSM 中最小的不可再分的语言单元，可以是字、词、词组、短语等。在 VSM 中，一段文字被看成是由特征项组成的集合，表示为：

$$(t_1, t_2, \cdots, t_n)$$

其中 t_i 表示第 i 个特征项。

（2）特征项的构造与权重

首先，向量空间需要一个特征集合：

$$(t_1, t_2, \cdots, t_n)$$

其次，定义特征项权重

$$(w_1, w_2, \cdots, w_n)$$

该权重为向量的每一个维度赋予一个值。

常见的特征权重包括下列几种：

①布尔（BOOL）权重：表示该特征是否在当前文本中出现，如果出现则记为 1，否则记为 0。

②特征频率（term frequency，TF）：表示该特征在当前文本中出现的次数。TF 权重假设高频特征包含的信息量高于低频特征的信息量，一次在文本中出现次数越多的特征项，其重要性越大。

③倒文档频率（inverse document frequency，IDF）权重：文档频率（document frequency，DF）表示语料包中包含的特征项的文档的数目。一个特征项的 DF 值越高，其包含的有效信息量往往就越低。IDF 是反映特征项在整个语料中重要性的全局统计特征，定义如下：

$$idf_i = \log \frac{N}{df_i}$$

其中 df_i 表示特征项 t_i 的 DF 值，N 是语料中的文档总数。

④特征频率-倒文档频率（TF-IDF）权重：定义为 TF 和 IDF 的乘积

$$tf\_idf_i = tf_i \cdot idf_i$$

TF-IDF 认为对区别文档最有的特征项应该是那些在当前文本中出现频率足够高，而在文本集合的其他文本中出现的频率足够小的词语。

在这个实例中，我们采用的便是 TF-IDF 的方法。

2.4 文本分类

2.4.1 六种常见的文本分类

一个文本经过文本表示和特征选择之后，就可以基于传统的机器学习算法进行文本分类。早期的文本分类模型包括相似度模型（如 Rocchio、K-近邻分类器）、决策树等，得到了广泛使用的文本分类算法，包括朴素贝叶斯模型、Logistic 回归模型、最大熵模型和支持向量机等。在这个实例中，我们使用了逻辑回归模型（LR）、线性判别分析（LDA）、K 近邻（KNN）、回归树（CART）、支持向量机（SVM）、高斯贝叶斯（NB）与这六种不同的分类方式，比较六个分类方式的准确率，以建立准确率较高的分类模型。

（1）逻辑回归模型

逻辑回归也被称为广义线性回归模型，它与线性回归模型的形式基本上相同，都具有 $ax+b$，其中 a 和 b 是待求参数，其区别在于他们的因变量不同，多重线性回归直接将 $ax+b$ 作为因变量，即 $y=ax+b$，而 logistic 回归则通过函数 S 将 $ax+b$ 对应到一个隐状态 p，$p=s(ax+b)$，然后根据 p 与 $1-p$ 的大小决定因变量的值。这里的函数 s 就是 Sigmoid 函数：

$$s(t)=\frac{1}{1+e^{-t}}$$

将 t 换成 ax+b，可以得到逻辑回归模型的参数形式：

$$p(x;a,b)=\frac{1}{1+e^{-(ax+b)}}$$

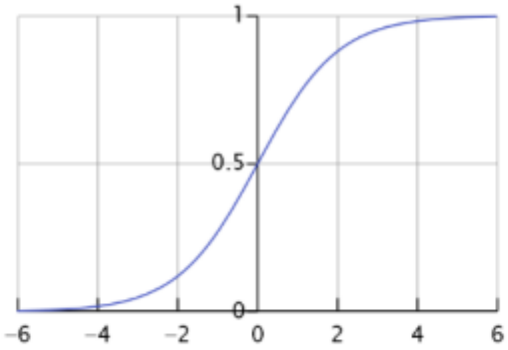

图 3 sigmoid 函数的图像

通过函数 s 的作用，我们可以将输出的值限制在区间 $[0,1]$ 上，$p(x)$ 则可以用来表示概率 $p(y=1|x)$，即当一个 x 发生时，y 被分到 1 那一组的概率。

（2）线性判别分析

线性判别分析（LDA）是对费舍尔的线性鉴别方法的归纳，这种方法使用统计学，模式识别和机器学习方法，试图找到两类物体或事件的特征的一个线性组合，以能够特征化或区分它们。LDA 是一种监督学习的降维技术，也就是说它的数据集的每个样本是有类别输出的。

以简单的二维数据为例，假设我们有两类数据分别为红色和蓝色，如图 4 所示，这些数据特征是二维的，我们希望将这些数据投影到一维的一条直线，让每一种类别数据的投影点尽可能的接近，而红色和蓝色数据中心之间的距离尽可能的大。

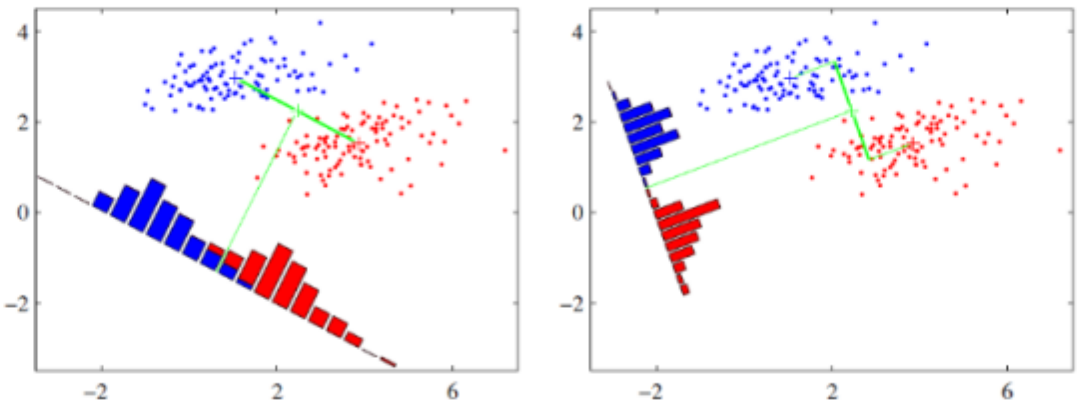

图 4 两种投影方式

（3）K-近邻

K 最近邻（k-Nearest Neighbor， KNN）分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：在特征空间中，如果一个样本附近的 k 个最近（即特征空间中最邻近）样本的大多数属于某一个类别，则该样本也属于这个类别。

如图 5 所示，有两类不同的样本数据，分别用蓝色的小正方形和红色的小三角形表示，而图正中间的那个绿色的圆所标示的数据则是待分类的数据。

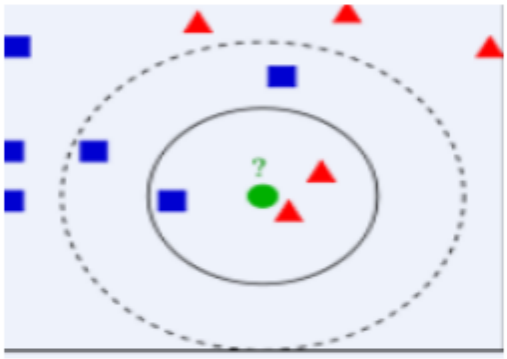

图 5 待分类小球

现在我们不知道中间那个绿色的数据是从属于哪一类（蓝色小正方形或者红色小三角形），我们就要解决这个问题就是给这个绿色的圆分类。

k 近邻算法使用的模型实际上对应于对特征空间的划分。k 值的选择，距离度量和分类决策规则是该算法的三个基本要素：

k 值的选择会对算法的结果产生重大影响。k 值较小意味着只有与输入实例较近的训练实例才会对预测结果起作用，但容易发生过拟合；如果 k 值较大，优点是可以减少学习的估计误差，但缺点是学习的近似误差增大，这时与输入实例较远的训练实例也会对预测起作用，使预测发生错误。

在实际应用中，k 值一般选择一个较小的数值，通常采用交叉验证的方法来选择最优的 k 值。随着训练实例数目趋向于无穷和 $k=1$ 时，误差率不会超过贝叶斯误差率的 2 倍，如果 k 也趋向于无穷，则误差率趋向于贝叶斯误差率。

该算法中的分类决策规则往往是多数表决，即由输入实例的 k 个最临近的训练实例中的多数类决定输入实例的类别

距离度量一般采用 LP 距离，当 $p=2$ 时，即为欧氏距离。在度量之前，应该将每个属性的值规范化，这样有助于防止具有较大初始值域的属性比具有较小

初始值域的属性的权重过大。

（4）线性回归

分类与回归树是分类数据挖掘算法的一种。它描述给定预测向量值 X 后，变量 Y 条件分布的一个灵活的方法。

该模型使用了二叉树将预测空间递归划分为若干子集，Y 在这些子集的分布是连续均匀的。树中的叶节点对应着划分的不同区域，划分是由与每个内部节点相关的分支规则确定的。通过从树根到叶节点移动，一个预测样本被赋予个唯一的叶节点，Y 在该节点上的条件分布也被确定。

（5）支持向量机

支持向量机（spport vetor machine， SVM）是统计机器学习领域富有盛名的分类算法。它的两个核心思想是：

①寻找具有最大类间距离的决策面；

②通过核函数在低维空间计算并构建分类面，将低维不可分问题转化为高维可分问题。线性回归是利用成为线性回归方程的最小二乘函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。

（6）高斯贝叶斯

贝叶斯模型属于生成式模型，它对样本的观测和类别状态的联合分布 $p(x,y)$ 进行建模。在实际应用中，联合分布转换为类别的先验分布 $p(y)$ 与类条件分布 $p(x|y)$ 乘积的形式： $p(x,y)=p(y)p(x|y)$ 。

前者可以分别使用伯努力分布和类别分布建模两类和多类分类的类别先验概率，但类条件分布 $p(x|y)$ 的估计问题是贝叶斯模型的难题。

朴素贝叶斯模型是一种简化的贝叶斯分类器，进行建模时利用观测向量 x 和类别 y 的联合分布：

$$p(x,y)=p(y)p(x|y)$$

2.5 算法评估

以上六种分类方式中，从准确率来看，准确率从高到低依次为：逻辑回归、支持向量机、K 近邻、回归树、高斯贝叶斯、线性判别分析。

逻辑回归和支持向量机的准确率很可观，在 90%左右徘徊；K 近邻的准确率在 80%-85%之间，略逊色于前两个，但是总体来看，用这种方法达到的效果是良好的。另外三个准确率比较低，而且线性判别的准确率相当不稳定，起伏过大，本题不宜使用。

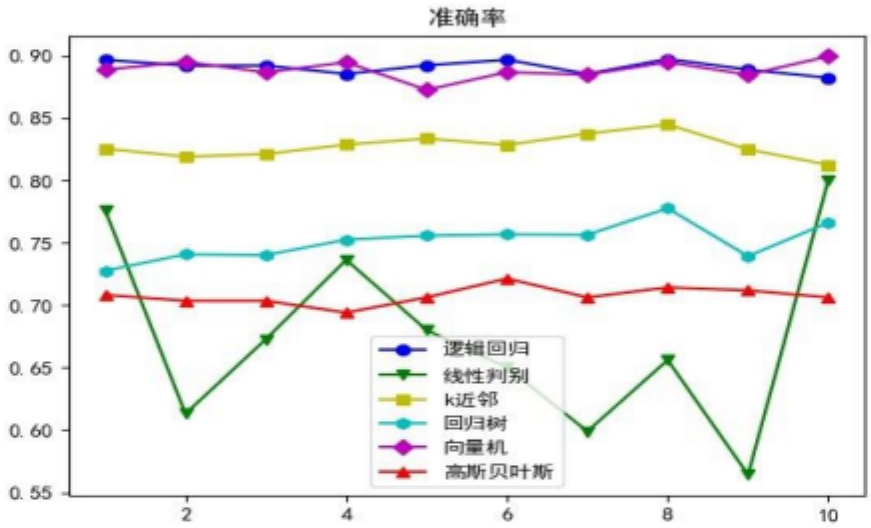

图 6 准确率

从时间复杂度来看，高斯贝叶斯运行时间最短；其次是逻辑回归；然后是 K 近邻、回归树、高斯贝叶斯、线性判别分析，这四类算法运行速度相近；最慢的是支持向量机。虽然支持向量机和逻辑回归的准确率不相上下，但是就运行速度来看，逻辑回归的远比支持向量机快，所以本例题中，逻辑回归是最适合用的模型。

2.6 分类结果对比

混淆矩阵（Confusion Matrix）：混淆矩阵也称误差矩阵，是表示精度评价的一种标准格式，用 n 行 n 列的矩阵形式来表示。具体评价指标有总体精度、制图精度、用户精度等，这些精度指标从不同的侧面反映了图像分类的精度。

在人工智能中，混淆矩阵是可视化工具，特别用于监督学习，在无监督学习一般叫做匹配矩阵。在图像精度评价中，主要用于比较分类结果和实际测得值，可以把分类结果的精度显示在一个混淆矩阵里面。混淆矩阵是通过将每个实测像元的位置和分类与分类图像中的相应位置和分类相比较计算的。

2.6.1 混淆矩阵

当以下四个指标汇聚成如图 7 的所示的矩阵时，我们便得到混淆矩阵。

混淆矩阵		真实值	
		positive	negative
预测值	positive	TP	FP
	negative	FN	TN

图 7 混淆矩阵

上表中，TP：真实值是 positive，模型认为是 positive 的数量；

FN：真实值是 positive，模型认为是 negative 的数量；

FP：真实值是 negative，模型认为是 positive 的数量；

TN：真实值是 negative，模型认为是 negative 的数量。

如图 8 所示，我们采用逻辑回归算法得到的混淆矩阵，左上到右下对角线上为预测正确的 对角线外面为错误。

```
confusion_matrix
 [[ 75   9   0  14  31   3   1]
 [  0 362  12   0   9   4   0]
 [  0  14 144   4   2   0   1]
 [  2   2   1 187  24   3   2]
 [  3   9   2   4 382   4   4]
 [  0  14   0   8  15 297   0]
 [  0   1   0   2  15   1 175]]
```

图 8 混淆矩阵

2.6.2 混淆矩阵标准化

有时候面对大量的数据，而混淆矩阵里面统计的是个数，只计算个数，很难衡量模型的优劣。因此混淆矩阵在基本的统计结果上又延伸了如下 4 个指标。

如此便可以将混淆矩阵中数量的结果转化为 0-1 之间的比率，如图 9。

	公式	意义
准确率 ACC	$ACC=\frac{TP+TN}{TP+TN+FP+FN}$	所有判断正确的结果占总观测值的比重
精确率 PVV	$PVV=\frac{TP}{TP+FP}$	模型预测是 Positive 的所有结果中，模型预测对的比重
灵敏度 TPR	$TPR=\frac{TP}{TP+FN}$	在真实值是 positive 的所有结果中，模型预测对的比重
特异度 TNR	$TNR=\frac{TN}{TN+FP}$	在真实值是 negative 的所有结果中，模型预测对的比重

图 9 混淆矩阵标准化

2.6.3 F-score 评分

便于进行标准化的衡量，我们得到 F-score 对分类方法进行评分：

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

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率（召回率）。程序运行结束得到评分如图 10 所示：

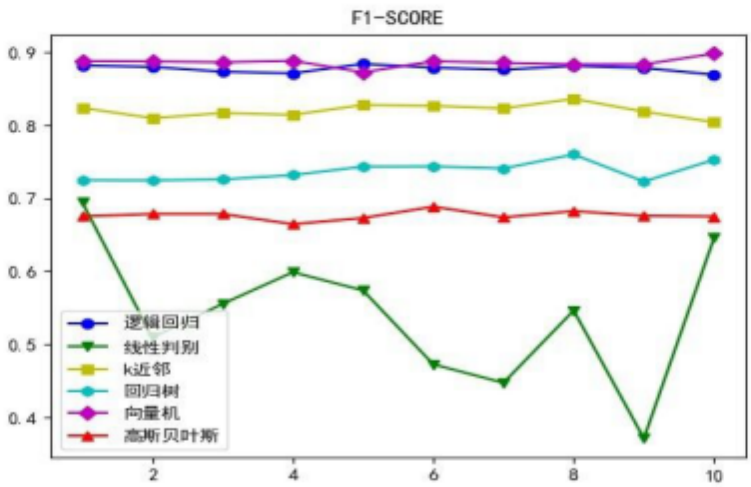

图 10 F-score

三、热点问题挖掘

3.1 第二题思维导图

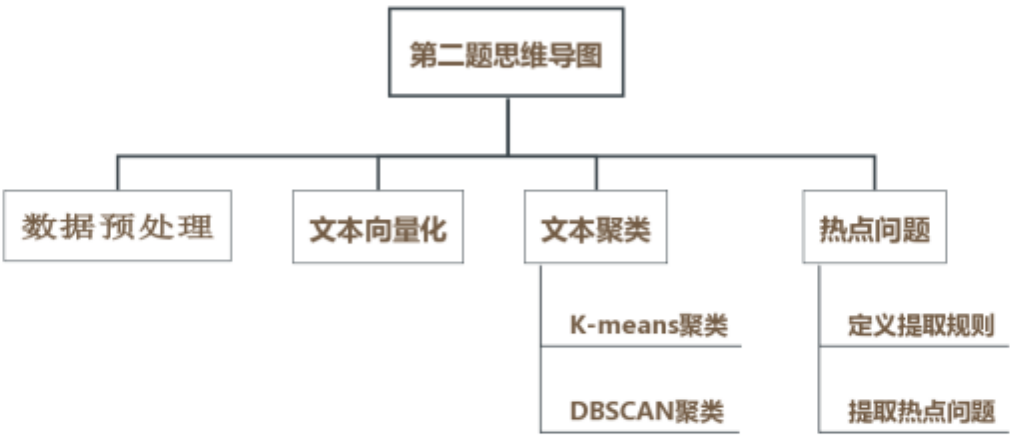

图 11 第二题思维导图

3.2 文本聚类

第二题的数据预处理与第一题相同，这里不赘述。

3.2.1 K-means 聚类算法

（1）概念：K-means 算法是最为经典的基于划分的聚类方法，是十大经典数据挖掘算法之一。K-means 算法的思想很简单，对于给定的样本集，按照样本之间的距离大小，将样本集划分为 K 个簇。让簇内的点尽量紧密的连在一起，而让簇间的距离尽量的大。

用数据表达式表示，假设簇划分为：

$$(c_1, c_2, \cdots, c_k)$$

找簇 c_i 的均值向量 u_i（也称为质心）：

$$u_i = \frac{1}{c_i} \sum_{x \in c_i} x$$

我们的目标是最小化平方误差 e：

$$e=\sum_{i=1}^{k}\sum_{x\in c_i}\|x-u_i\|^2$$

（2）算法描述：

①适当选择k个簇的初始质心；

②在第k次迭代中，对任意一个样本，求其到c个质心的欧氏距离或曼哈顿距离，将该样本归类到距离最小的质心所在的簇；

③利用均值等方法更新该簇的质心值；

④对于所有的c个簇质心，如果利用②③的迭代法更新后，当质心更新稳定或误差（误差平方和即簇内所有点到质心的距离之和）平方和最小时，则迭代结束，否则继续迭代。

（3）聚类结果

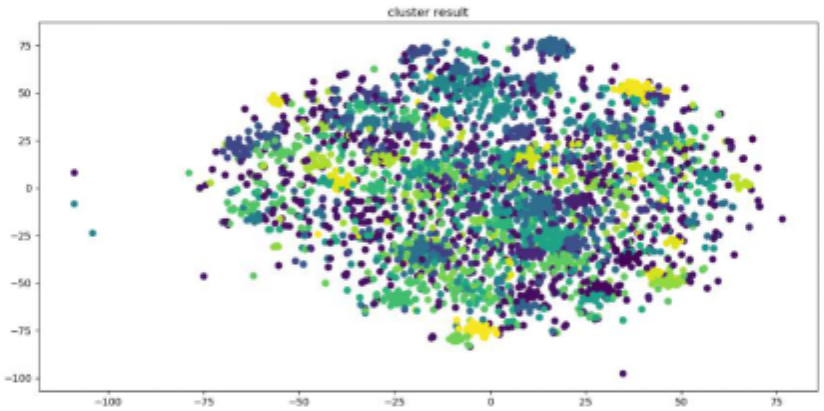

图 12 K-means 聚类结果散点图

在本题中，由于k值难以精确计算得出，根据附件 1 给出的标签类别，将k取值为 100，图 12 显示取得了较好的聚类效果，但是，聚类结果不是将同一地点的内容聚类，无法达到我们想要的效果。

3.2.2 DBSCAN 聚类算法

（1）概念：DBSCAN 是基于密度空间的聚类算法，在机器学习和数据挖掘领域有广泛的应用，其聚类原理通俗点讲是每个簇类的密度高于该簇类周围的密度，噪声的密度小于任一簇类的密度。簇类 A、B、C 的密度大于周围的密度，

噪声的密度低于任一簇类的密度，因此 DBSCAN 算法也能用于异常点检测，如图 13：

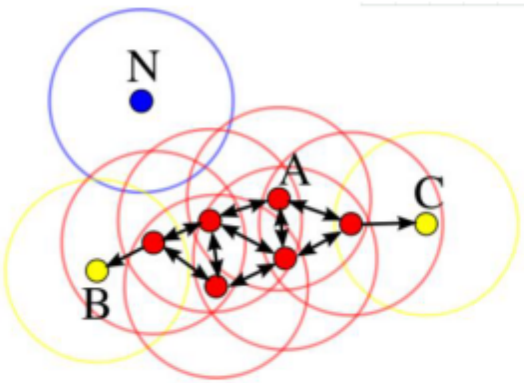

图 13 DBSCAN 算法示意图

DBSCAN 密度定义：DBSCAN 是基于一组邻域来描述样本集的紧密程度的，参数$(\varepsilon, \min\ pts)$用来描述邻域的样本分布紧密程度。其中，ε描述了某一样本的邻域距离阈值，$\min\ pts$描述了某一样本的距离为ε的邻域中样本个数的阈值。

（2）算法描述：首先给定数据集D中所有对象都被标记为“unvisited”，DBSCAN 随机选择一个未访问的对象p，标记p为“visited”，并检查p的ε-领域是否至少包含$\min\ pts$个对象。如果不是，则p被标记为噪声点。否则为p创建一个新的簇C，并且把p的ε-领域中所有对象都放在候选集合N中。

DBSCAN 迭代地把N中不属于其他簇的对象添加到C中。在此过程中，对应N中标记为“unvisited”的对象P^*，DBSCAN 把它标记为“visited”，并且检查它的ε-领域，如果P^*的ε-领域至少包含个对象，则P^*的ε-领域中的对象都被添加到N中。DBSCAN 继续添加对象到C，直到C不能扩展，即直到N为空。此时簇C完成生成，输出。

为了找到下一个簇，DBSCAN 从剩下的对象中随机选择一个未访问过的对象。聚类过程继续，直到所有对象都被访问。

（3）聚类结果

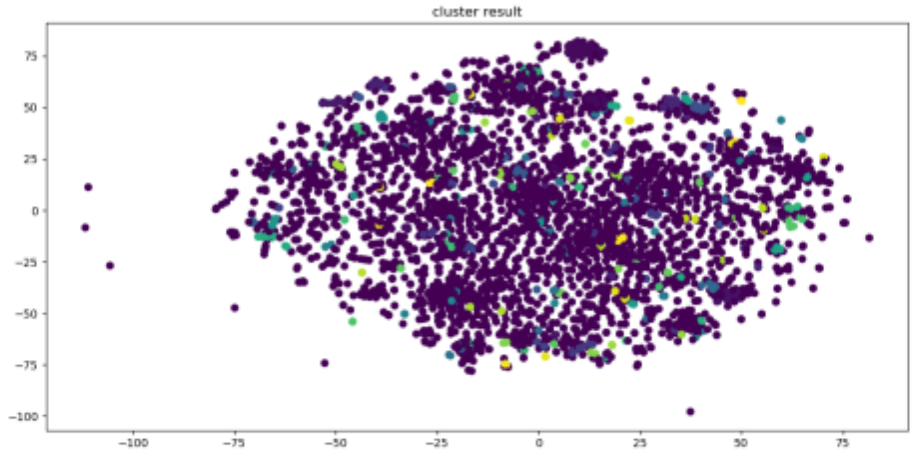

图 14 DBSCAN 聚类结果散点图

3.2.3 算法对比

（1）K-means 的主要优点有：

①原理比较简单，实现也是很容易，收敛速度快；

②聚类效果较优；

③算法的可解释度比较强；

④主要需要调参的参数仅仅是簇数 k。

（2）K-Means 的主要缺点有：

①K 值的选取不好把握；

改进：可以通过在一开始给定一个适合的数值给 k，通过一次算法得到一次聚类中心。对于得到的聚类中心，根据得到的 k 个聚类的距离情况，合并距离最近的类，因此聚类中心数减小，当将其用于下次聚类时，相应的聚类数目也减小了，最终得到合适数目的聚类数。可以通过一个评判值 E 来确定聚类数得到一个合适的位置停下来，而不继续合并聚类中心。重复上述循环，直至评判函数收敛为止，最终得到较优聚类数的聚类结果。

②对于不是凸的数据集比较难收敛；

改进：基于密度的聚类算法更加适合，比如 DBSCAN 算法。

③如果各隐含类别的数据不平衡，比如各隐含类别的数据量严重失衡，或者各隐含类别的方差不同，则聚类效果不佳；

④采用迭代方法，得到的结果只是局部最优；

⑤对噪音和异常点比较的敏感和传统的 K-Means 算法相比，DBSCAN 最大的不同就是不需要输入类别数 K，当然它最大的优势是可以发现任意形状的聚类簇。而 K-Means，一般仅仅使用于凸的样本集聚类。同时它在聚类的同时还可以找出异常点，这点和 BIRCH 算法类似。

（3）DBSCAN 的主要优点有：

①可以对任意形状的稠密数据集进行聚类，相对的，K-Means 之类的聚类算法一般只适用于凸数据集。

②可以在聚类的同时发现异常点，对数据集中的异常点不敏感。

③聚类结果没有偏倚，相对的，K-Means 之类的聚类算法初始值对聚类结果有很大影响。

（4）DBSCAN 的主要缺点有：

①如果样本集的密度不均匀、聚类间距差相差很大时，聚类质量较差，这时用 DBSCAN 聚类一般不适合。

②如果样本集较大时，聚类收敛时间较长，此时可以对搜索最近邻时建立的 KD 树或者球树进行规模限制来改进。

③调参相对于传统的 K-Means 之类的聚类算法稍复杂，主要需要对距离阈值，邻域样本数阈值联合调参，不同的参数组合对最后的聚类效果有较大影响。

3.3 确定评价规则

做完文本聚类以后，我们可以很直接的从看出出现频率最高的是哪些问题，但是我们不能只看到反映问题的频率，还应该考虑关注这件事情的其他人员和留言时间。比如编号为 191001 的事件提出时间为 2019 年 8 月 16 日，有 1 名反对者，12 名赞成者。

这种情况下，说明除了发言的这位群众，还有 13 个人关注这件事的人，虽然这些人里面有 1 人是反对的，但“反对”也是一种关注，值得相关人员去核实这件事。

我们定义将聚类之后的结果与点赞数、反对数做出来的排名综合起来，两者兼顾，形成最终的评价规则。定义为：

$$H_i = \sum_{j=1}^{n} T * (0.8 * Y_j + 0.2 * N_j + 1)$$

热度用字符 H(hot)表示，时间系数用字符 T(time)表示，n 表示归类为同一热点问题的个数，第 i 类热点问题的热度为 $H_i, i = 1,\ldots,100$，其中，第 i 类热点问题中第 j 个问题的点赞数为 Y_j，反对数为 N_j，时间系数定义为：

$$T = \begin{cases} 1, & 2020年 \\ 0.9, & 2019年 \\ 0.8, & 2018年 \\ 0.7, & 2017年 \end{cases}$$

3.4 提取热点问题

按照上述铺垫，我们得到所有热点问题中的前 5 个热点问题，如图 15。

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	1711.98	2019/1/8 至 2019/7/8	A 市	58 车贷诈骗案件进展缓慢
2	2	1521.9	2019/5/5 至 2019/9/19	A 市 五矿万境 K9 县	房屋存在质量问题
3	3	1270.44	2019/4/11	A 市 金毛湾学生	楼盘暂未解决配套入学问题
4	4	503.82	2019/8/23 至 2019/9/6	A 市 绿地海外滩小区	小区临近高铁噪音扰民
5	5	151.02	2019/3/26 至 2019/4/15	A6 区 月亮岛路	110kv 高压线杆的现状和规划问题

图 15 热点问题表

四、答复意见的评价

4.1 第三题思维导图

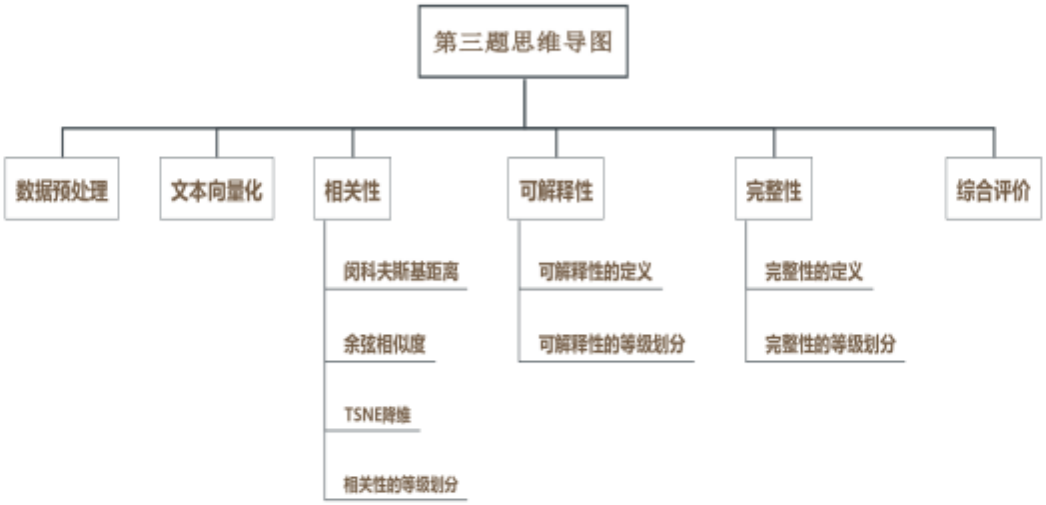

图16 第三题思维导图

数据预处理、文本向量化与第一题相同，本题不赘述。

4.2 相关性

相关性是数据属性相关性的度量方法，相似度是数据对象相似性度量的方法，数据对象由多个数据属性描述，数据属性的相关性由相关系数来描述，数据对象的相似性由某种距离度量。在本题中，我们采用相似度进行度量，选取了以下两种方法：

4.2.1 闵可夫斯基距离

（1）概念：闵可夫斯基距离不是一种距离，而是一组距离的定义。

$$x=(x_1,\cdots,x_n), y=(y_1,\cdots,y_n)$$

x 与 y 之间的闵可夫斯基距离公式 $d(x,y)$ 为：

$$d(x,y):=\left(\sum_{i=1}^{n}(|x_i-y_i|)^p\right)^{1/p}$$

当 p==1，“闵可夫斯基距离”变成“曼哈顿距离”；当 p==2，“闵可夫斯基距离”变成“欧几里得距离”；当 p==∞，“闵可夫斯基距离”变成“切比雪夫距离”。

4.2.2 余弦相似度

（1）概念：余弦距离，也称为余弦相似度，是用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小的度量。向量余弦值的范围在[-1，1]之间，余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，这就叫“余弦相似性”。

设向量：

$$a=(x_1,\cdots,x_n),b=(y_1,\cdots,y_n)$$

且 a，b 之间的夹角为 θ，则 a，b 之间的余弦 $\cos(\theta)$ 为：

$$\cos(\theta)=\frac{a\bullet b}{\|a\|\times\|b\|}=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}x_i^{\ 2}}\times\sqrt{\sum_{i=1}^{n}y_i^{\ 2}}}$$

（2）相似度等级：

相比于闵可夫斯基距离如欧氏距离，余弦距离更加注重两个向量在方向上的差异。

欧氏距离能够体现个体数值特征的绝对差异，所以更多的用于需要从维度的数值大小中体现差异的分析，如使用用户行为指标分析用户价值的相似度或差异。

余弦距离更多的是从方向上区分差异，而对绝对的数值不敏感，更多的用于使用用户对内容评分来区分兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题（因为余弦距离对绝对数值不敏感）。

因此，在本实例中，我们发现余弦相似度更能达到我们的需要，验证的结果更合理。由于向量维数较大，零元素过多，导致部分数据异常，因此我们进行了降维（TSNE）,然而降维后得到的数据出现负值，为了便于观察结果，我们将相关性等级定义为：

$$S=round(\frac{10*\cos\theta+10}{2})=round(5*\cos\theta+5)$$

其中 $\cos\theta$为降维后的余弦距离。

4.3 可解释性

4.3.1 自定义可解释性

可解释性是指答复意见中内容的相关解释，在解释中有没有引经据典，政策解读，法律引用等一定的理论支撑，来说明问题是否可以解决。因此，我们根据文本内容，来提取一些特有的特征指标，来衡量文本的可解释性。

4.3.2 回复的可解释性程度

通过对答复意见的提取与筛选，我们将答复意见的可解释性划分为四个等级，当答复意见中出现“《”或者“[”等特征时，我们定义得分为 9 分；当答复意见中出现网址链接，法律词汇等特征时，我们定义得分为 6 分；当答复意见对留言详情做到基本解释时，我们定义得分为 3 分；当答复内容与留言详情不相关时，我们定义得分为 0 分。得分越高，可解释性程度越高。

4.4 完整性

4.4.1 自定义完整性

（1）实体完整性规则

关系的主键可以表示关系中的每条记录，二关系的实体完整性要求关系中的记录不允许出现两条记录的主键值相同，既不能有空值，也不能有重复值。实体完整性规则规定关系的所有主属性都不能为空值，二不是整体不能为空值。例如，学生选课关系、学生选课、学号、课程编号、成绩中，学号、课程编号为主关键字，则学号和课程编号都不能取空值，二不是整体不能为空。

（2） 用户定义的完整性规则

不同的关系数据库系统更具其应用环境的不同，通常需要针对某一具体字段设置约束条件。例如，学生成绩字段的取值必须在 0-100 之间。参照完整性是相关联的两个表之间的约束，对于具有主从关系的两个表来说，表中每条记录外键的值必须是主表中存在的，如果两个表之间建立了关联关系，则对一个关系进行

的操作要影响到另一个表中的记录。

4.4.2 回复的完整性程度

文本向量化之后，我们使用答复文本中词的个数除以问题文本中词的个数的，得到阈值[0,10]的评分，划分成 11 个等级。在 4000 多条数据中，有 6 个数值超过 10 分的文本，我们不予以剔除的处理，而是直接替代为 10 分，作为正常数据使用。

4.5 综合评价

$$综合评分=相关性+可解释性+完整性$$

$$等级评价=\begin{cases}1 & 0 \le 综合评分 < 8\\ 2 & 8 \le 综合评分 < 11\\ 3 & 11 \le 综合评分 < 15\\ 4 & 综合评分 \ge 15\end{cases}$$

部分结果如图所示：

索引	相关性	完整性	可解释性	综合评分	等级评价
0	**5**	**1**	**9**	**15**	**4**
1	**2**	**2**	**3**	**7**	**1**
2	**7**	**2**	**9**	**18**	**4**
3	**5**	**4**	**9**	**18**	**4**
4	**3**	**4**	**3**	**10**	**2**
…	…	…	…	…	…
10	**6**	**2**	**3**	**11**	**3**

图 17 部分等级评价结果图

五、总结

为了相关部门能够及时、准确地了解到市民的反馈意见和建议并及时解决群众问题，基于人工智能的相关理论和实验，我们提出了智能政务模型，可以将依靠人工来进行留言划分和热点整理的相关部门的工作转为机器处理，使得相关部门能够更及时、准确地了解到市民的反馈意见和建议并及时解决群众问题。

在模型中，我们对不同算法模型作了详细解释以及结果对比，认真比对不同算法在本例题中的应用效果，在分类方法中选择了逻辑回归分类算法；在聚类方法中选择了 k-means 聚类算法；做文本相似度的时候利用改进过的余弦相似度对文本相关性进行评价，最终确定结果最优的算法模型。

在对赛题研究的基础上，我们根据研究思路撰写本论文，通过实验验证了本模型的可行性，基本实现了本赛题设立的目标。

致谢

在这篇报告完成之际，首先感谢指导老师在我们解题前期给予的的指导、帮助，感谢导师的监督和鼓励，老师工作繁忙但也从不忘对我们这次参赛进行细致的指导，为我们指明方向，提出宝贵的修改意见。

同时感谢大赛指导单位、主办单位、承办单位、协办单位用心筹划本次大赛。泰迪科技工作室为了顺应“互联网+”时代发展需求，帮助高校实现培养具有“ 创新、创造、创业、分享” 精神的数据人才的目标，以引导学生学习智能技术为导向，提高学生的创新能力和学习兴趣，举办本次大赛。这给了我们一次锻炼与学习的机会，给予我们创新与展示的平台，使我们在学习、思考、实践中提高能力，综合运用不同学科的理论，实现跨领域的融合，并不断在设计中创造。

最后感谢小组的每一位成员对此次比赛的付出，我们将继续努力。

参考文献

[1]邹俊杰. 受限域问答系统文本检索研究[D].昆明理工大学,2011.

[2]王欣.基于 Logistic 回归模型的多示例学习算法研究[D].大连理工大学.

[3]PouliakisA , Foukas P , Triantafyllou K , et al. Machine Learning for Gastric Cancer Detection: A Logistic Regression Approach[J]. 2020.

[4] 刘福刚.一种适用于中文博客自动分类的贝叶斯算法[J].长春师范大学学报,2019(12).

[5] L.W. D'Avolio， T. M. Nguyen, and L. D. Fiore. The automated retrieval console (arc): open source software for streamlining the process of natural language processing. In IHI, 2010. 4

[6]李欢.问答系统中的文本信息抽取研究与应用[D].中国科学技术大学,2009.

[7] https://www.cnblogs.com/always-fight/p/10159547.html

[8] https://www.cnblogs.com/always-fight/p/10159547.html

[9] M. Lux and S. A. Chatzichristofis. Lire: lucene image retrieval: an extensible java cbir library. In ACM Multimedia, 2008. 4

[10] L. Ma and Y. Zhang. Using word2vec to process big text data. 2015 IEEE International Conference on Big Data (Big Data), pages 2895–2897, 2015. 2

[11] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur. Recurrent neural network based language model. In INTERSPEECH, 2010. 2

[12] Shi M, Zhao Y, Yu W, et al. Enhanced performance of PAM7 MISO underwater VLC system utilizing machine learning algorithm based on DBSCAN[J]. IEEE Photonics Journal, 2019, PP(99):1-1.

[13] J. Ramos. Using tf-idf to determine word relevance in document queries. 2003. 2

[14] I. Rish. An empirical study of the naive bayes classifier. 2001. 2

[15] 范淼, 李超. Python机器学习及实践[M]. 清华大学出版社, 2016.