“智慧政务”中的文本挖掘应用

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在工作上带来了极大挑战。因此，运用数据挖掘技术对网络问政平台的信息进行分类和热点问题整理具有重大的意义。

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。

对于问题一，经过必要的数据清洗后，通过计算各类留言的 TF-IDF 权重向量，训练 SVM 支持向量机模型达到能够较精确分类各类留言问题。

对于问题二，对留言主题进行分词并提取关键词，得到每个留言的重要信息。通过 K-means 算法对留言的热点问题进行聚合，通过构建模型确定出热度评价指标来对所有问题进行热度排序，得到的热度最高的前五名即为热点问题。

对于问题三，将留言答复进行处理之后，我们定义出留言的相关性、完整性以及可解释性，通过对这些特性来评价留言答复，并通过星级打分的方式来显示出答复的好坏。

关键词： 智能政务，文本挖掘，SVM，K-means，TF-IDF

Abstract

In recent years, With WeChat, Weibo, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel to conquer and understand public opinion, gather people's wisdom and gather people's spirit. The increasing amount of text data related to social situation and public opinion has brought great challenges to the work of relevant departments that mainly rely on manual work to divide messages and sort out hot spots in the past. Therefore, it is of great significance to use data mining technology to classify the information of network politics platform and sort out the hot issues.

Therefore, it is of great significance to use data mining technology to classify and sort out the information of network politics platform. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government system based on natural language processing technology has become a new trend of social governance innovation and development, which is of great significance to improve the management level and governance efficiency of the government Big push.

For the first problem, the information in the message table is de duplicated by position ID to get the non repeated message information. The Chinese word segmentation tool of Jieba is used to segment the message information, and then the TF-IDF weight vector training classifier model of all kinds of messages is calculated to achieve a more accurate classification of all kinds of message problems.

For question 2, the same question 1 is to segment the message subject and extract the key words to get the important information of each message. Through the gensim algorithm to calculate the similarity of the message to determine the hot issues in the message, through the establishment of a model to determine the heat evaluation index to rank all the issues, the top five hot issues are hot issues.

For question 3, after processing the message reply, we define the relevance, integrity and interpretability of the message, evaluate the message reply by these characteristics, and show the quality of the reply by star rating.

Key word: Smart government affairs，text mining, SVM, K-means，TF-IDF

目 录

1. 挖掘目标 1
2 问题一的分析方法与过程 2
2.1 流程图 2
2.2 数据预处理 2
2.2.1 对数据去重、去空 2
2.2.2 去除无用字符及数字字母 2
2.2.3 分词处理 3
2.3 留言信息的分类 3
2.3.1TF-IDF 算法 3
2.3.2 生成 TF-IDF 向量 4
2.3.3 训练 SVM 分类器 5
3.问题二的分析方法过程 8
3.1 数据处理 8
3.2 相同留言归类 8
3.2.1 流程图 8
3.2.2 文本向量化 9
3.2.3 K-means 算法 9
3.3 热度评价指标 10
3.3.1 计算热度指数 10
3.4 提取热点问题 11
4 问题三的分析方法与过程 12
4.1 数据处理 12
4.2 答复评价指标 12
4.3 答复评价方法 12
4.3.1 答复及时性检测 12
4.3.2 答复完整性检测 13
4.3.3 答复相关性检测 13
4.3.4 答复可解释性检测 15
3.结果分析 17
3.1 问题一的结果分析 17
3.1.1 一级分类方法评价 17
3.1.2 问题分类结果 18
3.2 问题二的结果分析 19
3.2.1 聚类方法分析 19
3.2.2 热点问题表分析 19
3.3 问题三的结果分析 20
3.3.1 答复评价方法分析 20
3.3.2 问题三的改进 20
4.结语 22

1.挖掘目标

从20世纪末的“政务电子化”作为政府辅助优化决策的管理手段，到后来“政府再造”，政府机构使用相互独立的信息系统，再到如今建立起来的G2G、G2B以及G2C网上信息渠道，便捷、智能、高效的电子政务迅速发展起来[1]。

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量迅速攀升，以往的人工进行的文本分类、热点整理显得力不从心，耗时耗力。因此我们希望通过对互联网公开来源的群众问政留言记录进行文本挖掘，构建出一个模型达到对群众留言问题进行分类、热点提取等目的。

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，对于提升政府的管理水平也具有极大的推动效益。

在此我们的挖掘目标分为三个部分：

（1）在留言信息日趋增大的今天，由于人工分类的耗时耗力，导致人工分类的成本高、误差大。因此我们挖掘目标的第一部分为构建模型达到对于留言问题的一级分类。并且此模型能够达到一定的精确度，保证在数据量极大的情况下能够准确快速的完成文本的分类工作。

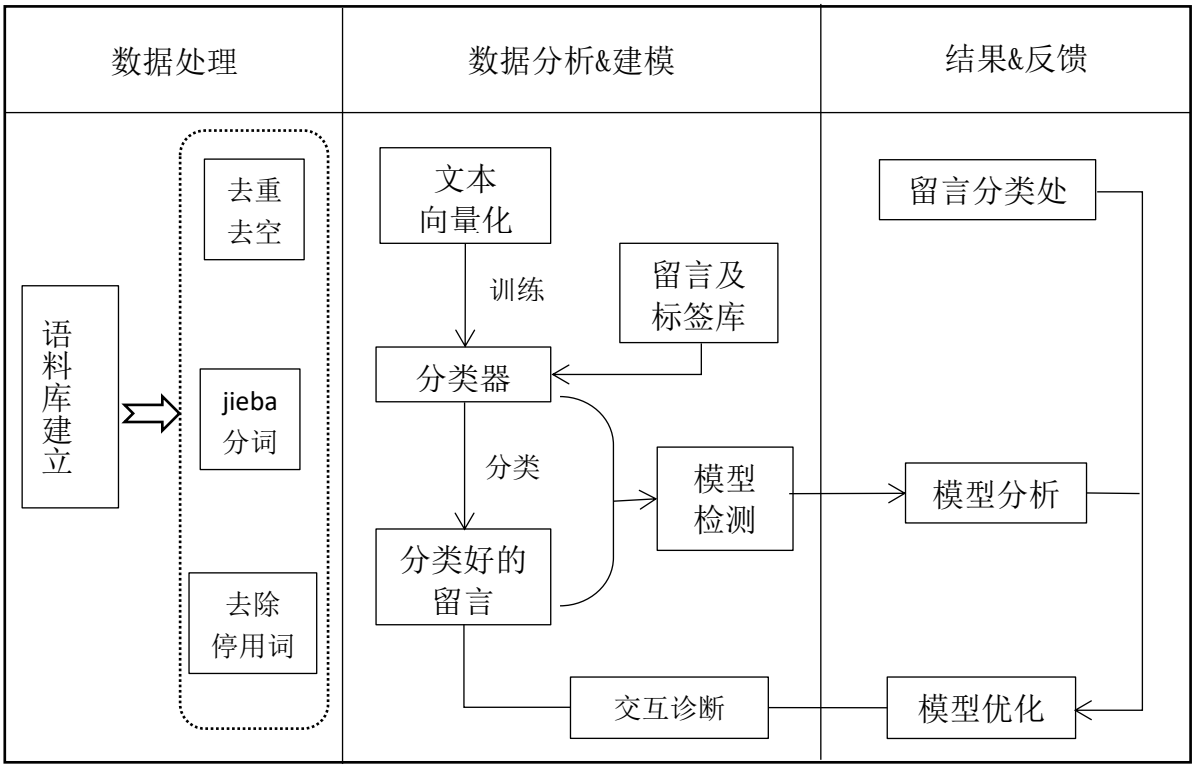
（2）当民众对于一个问题集中反馈时，此类问题我们可称为热点问题。及时发现热点问题可以提高相关部门的服务效率。我们尝试通过聚类分析的方式将某一时间段内反映特定地点或特定人群问题的留言进行归类，并通过评定给出排名前五的热点问题。聚类分析具有普遍意义，应用于很多领域，如图像处理、生物学、信息检索、数据挖掘、音频分析检索和统计学等[2]。我们在这里通过聚类分析的方式来得到可以检索出热点问题的模型。

（3）对于群众留言反映的问题是否解决，检测相关部门的服务效率，我们通过构建模型将相关部门对留言的答复意见从相关性、完整性、可解释性三个方面进行评测，达到检测相关部门的服务效率的作用。

我们通过构建模型的方法，从三个方面帮助网络问政平台对民众留言进行处理，通过数据挖掘的应用，提高网络问政的效率，提高政府处理民事的效率，使民众提出的问题能更快的解决。

2 问题一的分析方法与过程

2.1 流程图



2.2 数据预处理

2.2.1 对数据去重、去空

在题目所给的数据中，有部分重复或为空的留言，考虑到重复信息以及空白信息对模型自主学习的影响，我们通过 python 中自带语句进行去重操作。因为 python 中字典在保存数据时，key 相同的内容，value 取值为最后更新值，因此在读取数据时，按照时间升序将留言信息的位置 ID 作为 key，将整个留言内容作为 value 保存进 value 中。最后将字典中内容写入文本即可。对于文本内容为空的记录则直接删除避免影响。

2.2.2 去除无用字符及数字字母

在留言信息中掺杂有各类无用字符以及留言用户的隐私信息等等，例如联系方式。因此我们在进行后续操作前需要将文本数据中的\t 字符、\n 字符以及数字字

母等均转化为 x，目的是保护留言者个人隐私。通过 python 中的 lambda 函数，将用户信息中所有的\t 字符、\n 字符以及数字字母进行转化。

2.2.3 分词处理

因为中文文本是按句连写的，词与词之间没有界限标志[3]。在对中文文本进行数据挖掘时需要对中文文本进行拆分，将句子拆分为词语便于进行后续运算。中文文本分词问题是中文信息处理的一个重要问题，这个问题解决的好坏将直接影响后续数据挖掘的精确度以及模型的能力强弱。

一、jieba 分词

汉语语言的词语之间的关联度在全局上显示出高度的连接性，同时在局部具有高度的聚集性[4]。留言信息内容长、具有连惯性等问题使计算机在进行分类时难以操作。需要先把非结构化的文本信息转化为计算机可识别的结构化信息。在附件留言信息表中给出的中文文本数据，为了便于转换，需对这些信息进行中文分词处理。这里采用 python 自带的中文分词包 jieba 进行分词。

Jieba 采用的是基于前缀词典实现的高效词图扫描，生成句中汉字所有可能成词情况所构成的有向无环图（DAG），同时采用了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。对于未登录词，采用了基于汉字成词能力的 HMM 模型，使能达到更好的中文分词效果。

二、去除停用词

Jieba 分词后会将中文文本中类似“的”、“，”、“！”等字符单独分出成为一个词，但是此类字符并无含义，在问题分析时会增大计算量，提高误差，所以需要建立停用词表，将这些字符筛除。Jieba 分词属于词典切分法，是按一定的策略将汉子串与机器词典中的词条进行匹配，若在词典中找到某个字符串，则匹配成功[5]。

2.3 留言信息的分类

前期准备工作完成后，接下来对留言进行分类，在分类过程中我们通过 TF-IDF 将句子转化为词向量，之后通过构建 SVM 模型来对留言的一级分类进行划分。

2.3.1TF-IDF 算法

在对留言进行分词后，需要将分好的词转化为词向量，这里采用 TF-IDF 算法，将文本转化为权重向量。

TF-IDF(Term Frequency & Inverse Documentation Frequency)算法是Stilton提出的，主要思想是：如果一个词在特定的文档中出现的频率越高，TF值越大，代表它表达该文档内容的能力越强，应该被赋予较高的权重；如果一个词在一组文档中出现的范围越小，计算得到的IDF值越小，说明它区分文档内容的能力越强，应该被赋予较高的权重[6]。

TF-IDF算法的具体操作如下：

第一步：计算词频，即TF权重（Term Frequency）。

$$词频（TF）= 某个词在文本中出现的次数 \qquad （1）$$

考虑到文章有长短之分，为了便于不同文章的比较，进行“词频”标准化，除以文本的总次数或者除以该文本中出现次数最多的词的出现次数即：

$$词频(TF) = \frac{某个词在文本中出现的次数}{文本的总词数} \qquad （2）$$

或

$$词频(TF) = \frac{某个词在文本中出现的次数}{该文本中出现次数最多的词的出现次数} \qquad （3）$$

第二步：计算IDF权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$逆文档频率(IDF) = log\frac{语料库的文本总数}{包含该词的文本数+1} \qquad （4）$$

第三步：计算TF-IDF值（Term Frequency Document Frequency）。

$$TF-IDF = 词频(TF) \times 逆文档频率(IDF)$$

实际分析得出TF-IDF值与一个词在留言表中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF值越大。计算文本中每个词的TF-IDF值，进行排序，次数最多的即为留言的关键词。

2.3.2 生成TF-IDF向量

生成TF-IDF向量的具体操作如下：

（1）将所有关键词合并成一个集合，计算每个留言对于这个集合中词的词频，如没有则记为0。

（2）生成各类留言的TF-IDF权重向量，计算公式如下：

$$TF-IDF = 词频(TF) \times 逆文档频率(IDF)$$

2.3.3 训练 SVM 分类器

生成了各类留言的 TF-IDF 权重向量后，根据每种留言的 TF-IDF 权重向量，对留言进行分类。这里采用的是 SVM 模型进行计算。

SVM 的概念及计算：

从 Input 像 Feature Space 进行一个映射，且有超平面：$\omega \cdot x + b = 0$

$$f(x,w,b) = sign\big(g(x)\big) = sigb(\omega \cdot x + b)$$

而点到超平面的距离为：

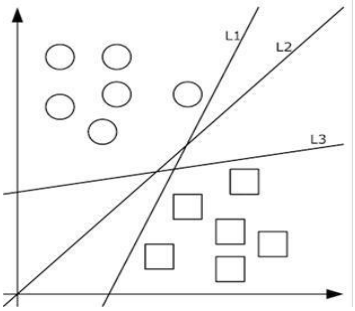
$$M = \|x - x'\| = \frac{|g(x)|}{\|\omega\|}$$

原点到平面的距离为$\frac{|b|}{\|\omega\|}$。

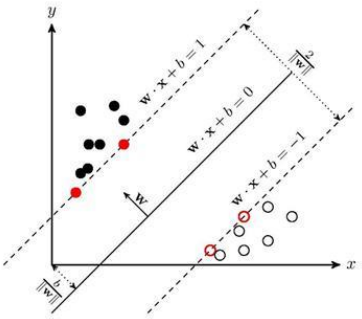
而在判断分界线的时候，应该尽量对空白区域分割公平。

定义间隔（Margins）：

Margins 的计算方法为：$M = \frac{2}{\|\omega\|}$



SVM 支持向量机分类操作



Margins 的引入

在 SVM 中，我们对两个样品进行分对：

$$\begin{cases} \omega \cdot x_i + b \geq 1 & , y_i = +1 \\ \omega \cdot x_i + b \leq -1 & , y_i = -1 \end{cases}$$

现最小化 Margins：

$$maxM = \frac{2}{\|\omega\|} \rightarrow min\frac{1}{2}W^T W$$

最小化$min\frac{1}{2}W^T W$且有$y_i(\omega \cdot x + b) \geq 1$。

通过拉格朗日法进行求解：

$$L_P = \frac{1}{2}\|\omega\|^2 - \sum_{i=1}^{l} \alpha_i y_i(\omega \cdot x_i + b) + \sum_{i=1}^{l} \alpha_i$$

将$\begin{cases}\frac{\partial L_P}{\partial \omega}=0\rightarrow \omega=\sum_{i=1}^{l}\alpha_i y_i x_i \\ \frac{\partial L_P}{\partial b}=0\rightarrow \sum_{i=1}^{l}\alpha_i y_i=0\end{cases}$带入上式中，得到：

$$L_D=\sum_i \alpha_i-\frac{1}{2}\sum_{i、j}\alpha_i\alpha_j y_i y_j x_i x_j=\sum_i \alpha_i-\frac{1}{2}\alpha^T H\alpha(\text{当}H_{ij}=y_i y_j x_i\cdot x_j)$$

Subject to：$\sum_{i=1}^{l}\alpha_i y_i=0$ and $\alpha_i\geq 0$

$$g(x)=\sum_{i=1}^{l}\alpha_i y_i x_i\cdot x+b$$

$$y_s(x_s\cdot\omega+b)=1$$

$$y_s\left(\sum_{m\in s}\alpha_m y_m x_m\cdot x_s+b\right)=1$$

最后得到

$$b=\frac{1}{N_s}\sum_{s\in s}\left(y_s-\sum_{m\in s}\alpha_m y_m x_m\cdot x_s\right)$$

引入惩罚量：

在这里我们为了得到的值更加合理,我们进行放宽条件的操作(Soft Margin)：

$$y_i(\omega\cdot x+b)-1+\xi_i\geq 0$$

并引入惩罚量：

$$\Phi(\omega)=\frac{1}{2}W^T W+C\sum_i \xi_i$$

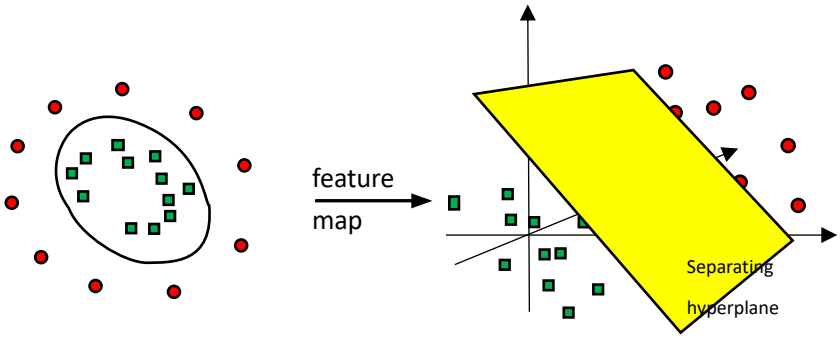
且有$\xi_i\geq 0$。

高维度映射分类：

在此问题中，由于文本分类的线性不可分性，我们通过将 x 映射到$\varphi(x)$的映射方式将文本的权重向量到更高维度进行分类。

在映射前，定义核函数（Kernel Trick），通过核函数的概念我们可以证明出低维运算操作的高纬度显示，以此绕过高纬度运算的计算量增大情况，减小计算量。在此我们使用 string kernel 进行映射。此部分运用 python 的软件包进行实现。

低维数据在高维的投射分类



低维状态分类情况　　高维状态分类情况

3. 问题二的分析方法过程

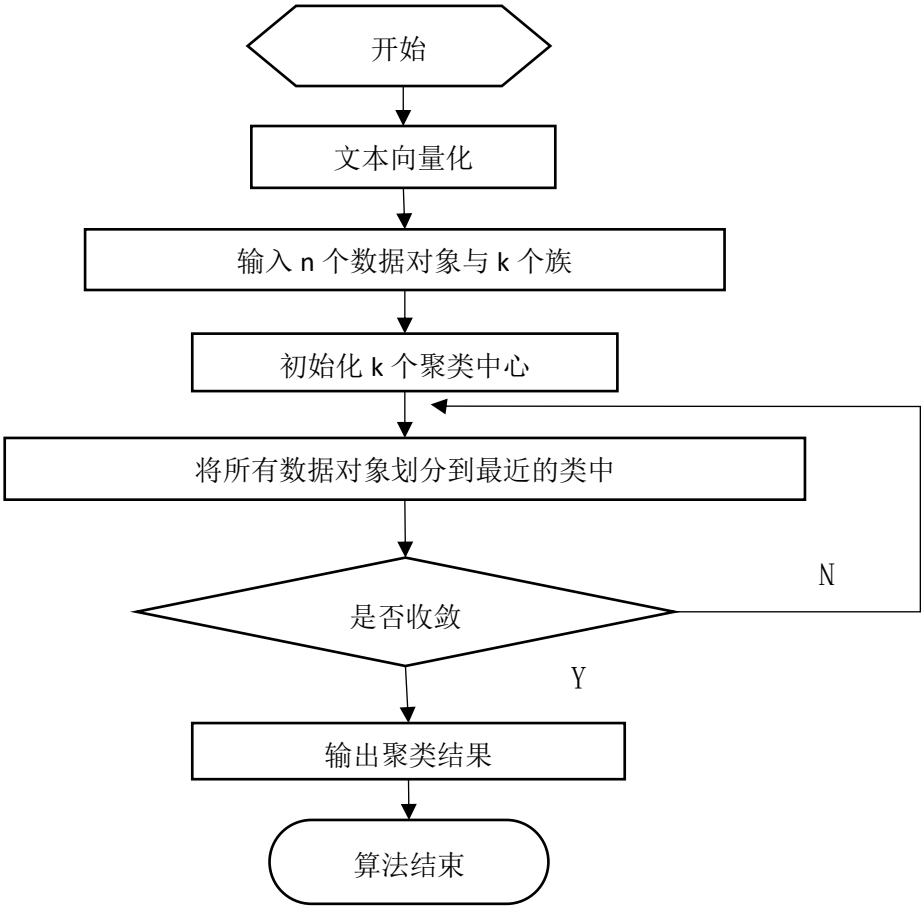
3.1 数据处理

在分析热点问题时，需要对留言信息进行数据预处理操作。即去除无意字符、分词以及去停用词等。

3.2 相同留言归类

对于已经处理好的留言，对其进行归类操作以此来分类出问题的种类。在对文本进行向量化操作之后，通过 K-means 算法，可以将相同种类的留言进行聚类操作，对同一聚类结果的文本打上同一标签，标记为一个事件。

3.2.1 流程图



3.2.2 文本向量化

为了提高聚类算法的效果，在尝试了 gesim、TF-IDF 等文本向量化的方式后，发现 TF-IDF 向量值在 K-means 聚类算法的表现最佳。故在本题中，同问题一的文本向量化方法，延续使用 TF-IDF 进行文本向量化。

3.2.3 K-means 算法

K-means 算法是一种被广泛应用的聚类分析算法，一般采用误差平方和准则函数作为聚类准则，在处理数据集时效率较高且聚类效果良好[7]。

K-means 算法的基本原理是设定参数K值，确定需要将数据集划分成K个类，从大量的数据对象中随机地选取K个数据对象作为初始聚类中心；然后根据距离公式计算剩下的所有数据对象到每一个初始聚类中心的距离，将计算好的数据对象划分到距离最近的初始中心，这样就可以形成以K个数据对象为初始中心点的聚类分布。

对划分好的初始聚类分布按照一定的规则（一般都是使用距离公式）重新计算每一类的中心点，以计算出来的点为中心形成新的类；如果计算出来的类中心点与前一次计 算的类中心点不同，则再次利用规则对数据集进行重新分配调整，如此循环往复，直到新的类中心点与前一次类中心点相同，所有数据对象都没有被重新划分，标志着该算法结束。K-means 算法就是通过迭代不断的变化中心点的位置，使得所有数据对象到其类的中心点的距离总和变成最小，这样就可以使得目标函数最小化。 K-means 算法就是通过不断地对数据对象进行迭代计算聚类中心点的过程。

其具体步骤如下：

（1）在数据集X中随机的选取k个数据对象，将这k个数据对象设定为初始聚类中心，即有C_1、C_2、 ……C_n个初始聚类中心点，这样可以确定数据集需要被划分成多少类；

（2）计算数据集中剩下的每一个数据对象到k个初始中心点的距离，将每一 数据对象划分到最近的类中，形成以k个初始中心点为中心的类。例如，数据对象X_p离中心点$C_i\ (i \le n)$最近，那么就将数据对象X_p划分到C_i类中；

（3）根据公式

$$C_i = \frac{1}{n_i}\sum_{x \in W_i} X$$

计算每一个聚类的中心点，即得到C_1^*、C_2^*、……C_n^*

（4）重复步骤（2）、（3），直到重新计算后的聚类中心点与计算前的聚 类中心点相同，没有发生任何变化， 说明聚类结果已经达到收敛；(离小于某一个设置的阈值或者达到设定的迭代次数)

（5）输出聚类结果。

3.3 热度评价指标

在对同一问题进行了聚合之后，即可对相似文本进行热度评价。这里利用了点赞数、反对数以及同一事件数作为评价指标。

3.3.1 计算热度指数

在此结合点赞数与反对数进行热度评价，具体操作如下：

第一步：计算出每个项目的“热度指数”，将同一标签A下的问题A_1、A_2……A_n计算热度指数η_A，其中n表示同一标签下的项目数量，即

$$\eta_A = \frac{A_1、A_2、\ \ldots\ldots A_n\text{点赞数之和} + n}{A_1、A_2、\ \ldots\ldots A_n\text{点赞数与反对数的总数和} + n}$$

第二步：计算每个项目热度指数的置信区间（以95%的概率）为了避免小样本带来的不准确性，引入威尔逊空间对其进行修正。

一、置信区间对热度指数的修正

由于部分样本的数据量可能相对较小，与其余数据量大的样本相比可能会引起偏差，我们通过引入置信区间对热度指数进行一个修正作用。

假设θ是总体的一个参数，其参数空间为 Θ，x_1，x_2，…，x_n是来自该总体的样本，对给定的一个$\alpha(0 < \alpha < 1)$，若有两个统计量$\theta_L = \theta_L(x_1, x_2, \ldots\ldots, x_n)$和$\theta_U = \theta_U(x_1, x_2, \ldots\ldots, x_n)$，使得对任意的$\theta \in \Theta$，有$P_\theta(\theta_L \le \theta \le \theta_U) \ge 1 - \alpha$则称随机区间$[\theta_L, \theta_U]$为$\theta$的置信水平为$1 - \alpha$的置信区间，$\theta_L$和$\theta_U$分别成为$\theta$的置信下限和置信上限。

二、威尔逊置信区间的引入

威尔逊置信区间的计算依赖于留言的热度指数、同一标签下的项目数，它的计算公如下：

$$\frac{p+\frac{1}{2n}z_{1-\alpha/2}^2 \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}+\frac{z_{1-\alpha/2}^2}{4n^2}}}{1+\frac{1}{n}z_{1-\alpha/2}^2}$$

其中p表示热度指数;n表示同一标签下的项目数量;$z_{1-\alpha/2}$表示对应某个置信水平的z统计量，在95%的置信水平下，z统计量为1.96

威尔逊区间的下限值为:

$$\frac{p+\frac{1}{2n}z_{1-\alpha/2}^2 - z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}+\frac{z_{1-\alpha/2}^2}{4n^2}}}{1+\frac{1}{n}z_{1-\alpha/2}^2}$$

根据威尔逊置信区间的计算公式，可以确定置信区间的宽度为:

$$\frac{2z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}+\frac{z_{1-\alpha/2}^2}{4n^2}}}{1+\frac{1}{n}z_{1-\alpha/2}^2}$$

我们得到，当n足够大时，该均值趋向于p，即表明当同一标签下的项目数足够大时，置信区间越窄。此时它的下限值接近于热度指数，此时留言的排名主要根据热度指数高低；相反，当n小于一定值时，该均值远小于p，置信区间越宽，它的下限值与热度指数的差也越大，此时热度指数对热点问题排名的影响削弱。

通过上面的分析，可以得到以下结论:当同一标签下的项目数n足够大时，可以直接依据留言的热度指数p作为热点问题排名的标准。但对于小样本情形，置信区间跨度较大，同一标签下的项目数n越小，威尔逊区间置信下限与热度指数的差就越大，使用热度指数进行排名其可信度就会存在问题。因此，采用威尔逊区间的下限值来代替热度指数p[8]。

3.4 提取热点问题

对于已经计算好的各项目的置信区间，我们对于每个置信区间均取其下限值用来代表此项目的热度，下限值越高则表示该类问题的热度指数越高，选出排名前五的五个项目作为热点问题，将这五个热点问题中对应的留言信息按照热点问题 ID 进行排序归类后可得“热点问题留言明细表”。将这些项目中所有留言的时间取出最早与最晚时间，即为时间范围，并提取出地点、人群等关键词与问题描述，最后可得“热点问题表”。

4 问题三的分析方法与过程

4.1 数据处理

对于附件四中的数据，需要对于留言回复进行评价。在进行评价前，对于留言回复进行数据预处理，对于留言回复为空的部分不进行删除，而是直接给与 0 分评分。而对于其余留言回复内容进行分词等数据预处理操作。

4.2 答复评价指标

对于留言回复的评价标准将从答复的及时性、完整性、相关性可与解释性进行评价。对于所有的答复进行评分模式，将答复从四方面进行给分，四个方面均为 100 分，计算后将答复四方面得分进行加权平均，并进行评星。

方面	及时性	完整性	相关性	可解释性
加权值	0.2	0.2	0.3	0.3

得分情况	对应星级
0——20	★
21——40	★★
41——60	★★★
61——80	★★★★
81——100	★★★★★

4.3 答复评价方法

4.3.1 答复及时性检测

对于答复的及时性，通过留言时间与答复时间之间时间的长短进行评价。

48 小时内答复	100 分
72 小时内答复	80 分

4 天内答复	70 分
5 天内答复	60 分
10 天内答复	50 分
20 天内答复	40 分
30 天内答复	30 分
30 天之后	0 分

4.3.2 答复完整性检测

对于答复的完整性，通过规定一个答复模板，将答复与答复模板进行对比，来判断答复的完整性，搜索答复中是否有关键的词句来得到答复的完整程度，并对其进行评价，每个关键词句得分为 10 分。

答复模板：

网友/同志***，您好/你好！您反映的问题已收悉/留言已收悉，现将有关事项回复如下/答复如下：*******。如果您还有其他疑问，请咨询*****，联系电话/咨询电话。感谢您对**工作的支持和理解，祝您生活和工作愉快、****年**月**日

关键词句：

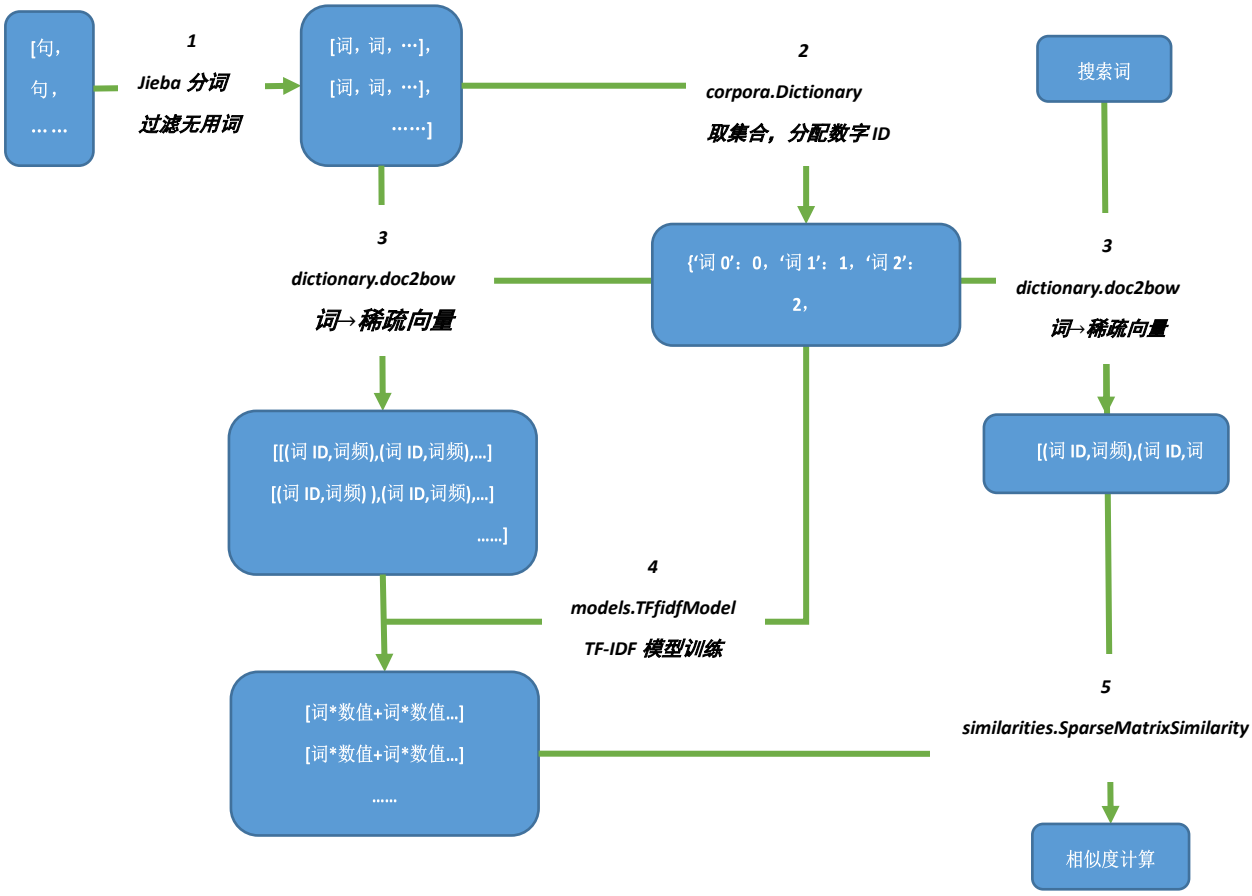
网友/同志	10 分
您好/你好	10 分
问题已收悉/留言已收悉	10 分
现将有关情况回复如下/回复如下/答复如下	10 分
感谢您对/感谢你对	10 分
工作的支持/工作的理解/工作的关心	10 分
如果您还有其他疑问，请咨询	10 分
祝您生活和工作愉快	10 分
联系电话/咨询电话	10 分
年月日	10 分

4.3.3 答复相关性检测

对于答复的相关性，首先需要考虑留言内容与答复意见之间的关系。我们认为，相关性可以解释为留言文本与答复文本之间存在内容上的关联。既然这种关联存在，则二者必然存在大量的重叠词。经过分析，我们使用了 gensim-word2vec 对留

言与答复文本之间的关联度进行了提取。

一、gensim 流程图



二、建立语料库

基于已经分词好的留言答复文本集，通过 python 自带语句建立词典，并计算出词典的特征数，词典特征数为词典中词的个数。后通过 doc2bow 函数对词典中单词分配 ID。

三、doc2bow 函数

Doc2bow 函数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。bow 使用一组无序的单词(words)来表达一段文字或一个文档。

Doc2bow 函数操作大致分为两步：

（1）将所有单词取【集合】，并对每个单词分配一个 ID 号

（2）转换成稀疏向量

例如对于['东京', '啊', '东京', '啊', '东京']，对单词分配 ID：东京→0；啊→4。变成：[0, 4, 0, 4, 0]，其中 0 有 3 个，即表示为(0, 3)；4 有 2 个，即表示为(4, 2)，最终结果：[(0, 3), (4, 2)]。

四、word2vec 文本向量化

在计算句子之间的文本关联度前需要将文本转化为文本向量。在 gensim 算法中转化为词向量的方法是 word2vec 词向量模型，word2vec 中包含了 2 种神经网路，分别是 CBOW 与 Skip-Gram。

CBOW 根据上下文预测当前词语的概率，公式如下：

$$P(\omega|c)=\frac{\exp\left(e'(\omega)^T x\right)}{\sum_{\omega'\in V}\exp\left(e'(\omega)^T x\right)}$$

式中：(ω,c)为从语料中抽取的n元短语，$P(\omega|c)$即词ω在文本c中出现的概率，V为训练语料，e'为更新权重，x为输入矩阵。

Skip-Gram 则是根据当前词语预测上下文概率，其中D为词集合，计算公式如下：

$$\max\left(\sum_{(\omega,c)\in D}\sum_{\omega_j}\left(\log P(\omega|\omega_j)\right)\right)$$

最后，将该句子中的所有含单词转化为词向量，再求出词向量的总和，将该值作为图的一个顶点计算。类比 PageRank 网页重要程度公式，文本图 G 的权重计算公式为：

$$WS(V_i)=(1-d)+d\times\sum_{V_j\in in(V_i)}\frac{\omega_{ji}\times WS(V_j)}{\sum_{V_K\in Out(V_j)}\omega_{jk}}$$

五、相似度计算

计算得到留言答复与留言的稀疏向量集后，通过构建 TF-IDF 模型来达到对比两段文本，TF-IDF 是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。通过 python 自带语句计算出 keyword 稀疏向量以及 TF-IDF 的稀疏向量集，后通过对应量的稀疏向量集和 keyword 稀疏向量的乘积累加得到文本相似度。

4.3.4 答复可解释性检测

对于答复的可解释性，我们判断的依据答复是否满足文本的可读性。

我们考虑了目前使用较多的机器学习法。机器学习法将文本的复杂特征表示成数据,计算各种与文本可读性相关的指标，再应用机器学习中的分类或者回归方法,训练得到可读性分类器,然后利用分类器判定新文本所属的可读性级别]。

支持向量机 SVM(Support Vector Machine)是在可读性预测领域应用较多的

分类方法。Schwarm 和 Ostendorf 将 SVM 与三元语言模型结合起来,用于预测英语新闻文章的可读性。实验证明 SVM 对于文本可读性的预测效果明显好于实验中用到的传统公式方法。

朴素贝叶斯是另-个应用较多的分类方法。朴素贝叶斯方法是一种利用先验概率计算后验概率的学习算法。在预测新文本的可读性级别时,应用最大似然估计法或条件概率计算出该文本属于各个可读性级别的概率。

多元线性回归是传统公式法的一个拓展。它将文本特征向量中的某些指标作为自变量将文本所属的可读性级别作为因变量通过求解多元线性回归方程获得可读性级别的预测函数。孙刚提出了基于线性回归的中文文本可读性预测方法。研究表明，采用机器学习方法来计算、预测文本的可读性,效果好于传统公式但目前的研究成果较少[9]。

3. 结果分析

3.1 问题一的结果分析

3.1.1 一级分类方法评价

对于问题一分类后得到的结果，我们通过 metrics 模块对模型性能进行评估分析，通过调用 Classification metrics 来计算分类的准确率。通过外部指标（将聚类结果与实际结果进行比较）得到分类的准确度，并在正式使用模型时通过内部指标来判断模型在正式使用时的正确率情况。

我们首先给出定义：

数据集：$D = \{x_1, x_2, \ldots\ldots, x_m\}$

聚类结果：$C = \{c_1, c_2, \ldots\ldots, c_k\}$

参考模型：$C^* = \{c_1^*, c_2^*, \ldots\ldots, c_s^*\}$

集合：

$$a = |SS|, SS = \{(x_i, x_j) | \lambda_i = \lambda_j, \lambda_i^* = \lambda_j^*, i < j\}$$

$$b = |SD|, SD = \{(x_i, x_j) | \lambda_i = \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j\}$$

$$c = |DS|, DS = \{(x_i, x_j) | \lambda_i \neq \lambda_j, \lambda_i^* = \lambda_j^*, i < j\}$$

$$d = |DD|, DD = \{(x_i, x_j) | \lambda_i \neq \lambda_j, \lambda_i^* \neq \lambda_j^*, i < j\}$$

（集合SS包含了在C中属于相同簇，且在C^*中也属于相同簇的样本对。集合SD包含了在C中属于相同簇，但在C^*中属于不同簇的样本对。）

一、外部指标

外部指标（external index）用于对建模时的模型好坏进行评价，外部指标包括：

Jaccard 系数（Jaccard Coefficent，JC）：

$$JC = \frac{a}{a + b + c}$$

FM 指数（Fowlkes as Mallows Index，FMI）：

$$FMI = \sqrt{\frac{a}{a + b} \cdot \frac{a}{a + c}}$$

Rand 指数（Rand Index，RI）：

$$RI = \frac{2(a+d)}{m(m-1)}$$

上述外部指标结果均在[0,1]区间，且越趋近于 1 值越好。

二、内部指标

内部指标（internal index）用于检测模型在实际运用的时候其性能好坏，内部指标有：

紧密性（Compactness，CP）：

$$CP_i = \frac{1}{|C_i|}\sum_{x_i \in C_i} dist(x_i, \mu_i)$$

$$CP = \frac{1}{k}\sum_{i} CP_i$$

紧密性是用来形容各类样本到聚类中心的平均距离。

间隔性（Separation，SP）：

$$SP = \frac{2}{k(k-1)}\sum\nolimits_{1 \le i \le j \le k} dist(\mu_i, \mu_j)$$

间隔性是用来形容各类中心间的平均距离。

DB 指数（Davies-Bouldin Index，DBI）：

$$DBI = \frac{1}{k}\sum_{i=1}^{k} \max_{j \ne i}\left(\frac{avg(C_i) + avg(C_j)}{d_{cen}(\mu_i, \mu_j)}\right)$$

Dunn 指数（Dunn Index，DI）：

$$DI = \min_{1 \le i \le k}\left\{\min_{j \ne i}\left(\frac{d_{min}(C_i, C_j)}{\max_{1 \le l \le k} diam(C_l)}\right)\right\}$$

其中，$avg(C)$：簇 C 内样本间的平均距离。$diam(C)$：簇 C 内样本间的最远距离。$d_{min}(C_i, C_j)$：簇C_i与簇C_j最近样本间的距离。$d_{cen}(\mu_i, \mu_j)$：簇μ_i与簇μ_j中心点间的距离。

内部指标中，DBI 指数的值越小越好，DI 指数的值越大越好。

三、模型评价结果

通过 metrics 模块的引用，我们计算得到问题一的分类模型在进行测试时准确率为 91.4412%，这表明模型在进行分类时的准确率较高，可以进行实际应用。

3.1.2 问题分类结果

在问题一中，我们通过 SVM 支持向量机对留言进行了分类，使用 SVM 支持向

量机进行留言内容的分类。但是现有的 SVM 算法在计算上存在这一些问题，包括：训练算法速度慢、算法复杂而且难以实现及检测阶段运算量大等等[10]。因此，本题中模型在运算量等方面依据存在驳杂、繁多等问题。

对于问题一算法可优化方面即为优化 SVM 算法部分，使其精确度更高、计算量更少。在此方面最突出的应用研究是贝尔实验室对美国邮政手写数据库（USPS）进行的试验，采用三种不同核函数的 SVM 方法的识别率均在 96%以上，优于决策树方法和五层神经网络的识别率[11]。

对于此类原因，我们可引入 SVM 的改进方法 RM-SVM 进行修改。RM-SVM 主要分为以下几个步骤：首先是计算出两类样本在映射空间中的中心点，然后我们在原有 SVMs 优化目标中添加平衡项，用来调整原有的决策面，最后我们得到解[12]。

详见附件——过程数据

3.2 问题二的结果分析

3.2.1 聚类方法分析

在对于聚类算法的选择中，我们排除了多种算法选择了在试验中发挥最好的 K-means 算法，k-means 算法简单，收敛速度快，可扩展且效率高，然而该算法存在聚类个数确定困难，初始聚类中心选取不准确导致聚类结果易陷入局部最优 解等缺陷[13]。对于 K-means 算法在实际运用中的表现是否依旧能够达到预期的标准暂无实验案例支撑。而在对簇数进行定量的时候，经过多次选择与验证，最终将簇数定为了 1500。从最后的聚类结果看来，簇数的确定还需不断改进。由于运行时间的问题，我们最终停在了 1500 簇，相信在有高性能计算机的情况下能更快的找到更有效的簇数。

在 K-means 算法进行聚类的基础上还可以做出改进，通过计算聚类结果总的轮廓系数 S_t来选取优k值。轮廓系数是聚类效果好坏的一种评价指标，聚类结果总的轮廓系数越大，聚类效果越好[14]。

3.2.2 热点问题表分析

“热点问题留言明细表”的提取涉及到问题 ID、留言编号、留言用户、留言主题、留言时间、留言详情以及点赞数和反对数。我们将 K-means 的聚类后的留言进行了热度指数排名后，将热度值排名前五的留言按顺序置于表中，并聚合了每条留言的相关数据。

但对于“热点问题表”的提取，其涉及地点、人群和事件的提取，较为复杂。在

我们看来，热点问题表的提取在已知热点问题明细的情况下，又鉴于五个热点问题的数量并不多，完全可以人工提取。所以我们对热点问题明细表中的数据进行了简单的人工再提取，即可得到“精确的”热点问题表。

在热点问题表的提取部分尚有不足，对于热点问题明细已有的情况下，需要少量的人为参与进行热点问题表的较精确提取，依旧未完全实现自动化。

详见附件——过程数据

3.3 问题三的结果分析

3.3.1 答复评价方法分析

对于答复的评价方法，从完整性、可解释性、相关性对其进行解释以外，还增添了及时性，及时性的处理过程较易，仅需对留言时间与答复时间之间的差值进行判断。此数据可以作为检验答复是否及时，同样可以判断出相关部门的工作效率。

而在对于完整性的定义上，我们给出的定义为对答复的内容进行规范，即必须出现规定内容。仅需在答复文本中寻找关键词句。此定义通过预先设定好的留言模板对其进行评价。优势在于操作简单，可以将模板下发给各负责组织，提高留言答复的完整性。缺点则在于中文中同意义的词句偏多，可能出现误判等操作，所以在后续的实际使用过程中需要人为添加完整性评价的规范语句。

对于答复的相关性，通过 gensim、TF-IDF 构建模型对留言文本与答复文本进行系统计算得到相关度，通过相关度的大小对答复的相关性进行了评定。

可解释性即答复文本的可读性，由于能力有限加之理论部分没有攻关，可解释性简单定义为了上述的相关性。

3.3.2 问题三的改进

在对于可解释性上明确的定义答复文本的可读性，在对于答复的可解释性判断上应该联系答复的文本长度、是否解决留言提到的问题、若问题尚未解决是否告知何时可以解决等内容，通过合适的算法进行计算得到可解释性的标准。

在相关性方面则需要解决留言内容较少答复内容较多情况的出现，想到方法可以与问题二中相同，通过威尔逊置信区间进行修正，但依旧存在些许问题，在后续操作中需要改进。

完整性方面效果较好，但是对于关键词方面同样出现了误判等操作，对于模板需要进行优化，同时也要提高关键词的数量等，构建出合理的完整性判断方法。

详见附件——过程数据

4. 结语

本文所构建的模型以及得到的成果与预期的目标基本相符，但尚有一些问题值得更近一步的探究和讨论。

对于问题一，我们在处理的时候运用到了 jieba 分词对留言文本进行分词，进行了数据的预处理后，通过 TF-IDF 将文本转化成了词向量的模式，通过词向量的计算来摆脱了文本数据无法进行数据计算的问题。之后通过 SVM 支持向量机的方法，构建出了留言的一级分类标签模型，在后续对模型的检测中达到了 91.4412% 的正确率。但同样也有一些可以进一步优化的地方，在算法方面，SVM 并非最适合的算法，后续 SVM 的优化方法中，RM-SVM 的添加平衡项的优化方案有待实施。对于分词操作，也有着更加优秀的分词系统。

对于问题二，在经过多次试验与碰壁，对于留言问题热点的挖掘，我们使用了大量的算法进行计算，最终选出了效果最好的 K-means 算法进行。同时对模型进行了优化，通过威尔逊空间进行修正，使得到的结果更准确。同样，问题二的分析也尚有缺陷，K-means 算法的缺陷很明显，聚类个数难确定，初始聚类中心无法指定。对于热度指数考虑到的参数也有待改进。

对于问题三，留言的答复从及时性、相关性、完整性以及可解释性四方面对答复进行评估。及时性最通俗易懂，即留言时间与答复时间的时间差。在相关性中，通过答复与留言的相似度、关键词的重复等等方面来进行判断，完整性则是通过与我们预设的模板进行比较查找是否完整，而可解释定义较模糊，在后续的探究中，对于可解释性的定义需要更加精确，完整性的对比还可以通过答复之间的相互比较进行。

就最后结果而言，三个问题均基本解决，但依旧有很大的进步空间，模型的实用性尚未得知，能否投入实用还需要进行更多更广泛的验证。

致　谢

在此，首先我们要感谢主办方给予我们这个比赛的平台，正是因为主办方在这种特殊时期依旧开展本次泰迪杯数据挖掘大赛，我们才有机会参与这次的比赛。其次，我们要感谢我们的指导老师。她在我们尝试解题的过程中给与了我们莫大的帮助，在我们一筹莫展的时候给我们提供了解题的思路，让我们能够更加顺利的完成本次比赛的内容。然后,我们要感谢学院老师们的支持，他们为我们参赛的后勤做出了保障。在这个特殊时期我们能够顺利参加本届泰迪杯数据挖掘大赛实属不易。在最后，我们要感谢我们自己小组的成员，在这个特殊时期我们相互帮助完成了本次的比赛内容。

参 考 文 献

[1] 刘伟.我国电子政务绩效评估方案的综合研究[J].中国行政管理，2013（2）

[2] 李芸娆.K-means 聚类方法的改进及其应用[D]：[硕士学位论文].保存地点：东北农业大学，2014 年.

[3] 许林杰.中文文本分词研究[D]：[硕士学位论文].保存地点：山东师范大学，2003 年.

[4] 赵 鹏.一种基于复杂网络特征的中文文档关键词抽取算法[J].模式识别与人工智能,2007,20(6):827-831.

[5] 李 华,陈 硕,练睿婷.神经网络和匹配融合的中文分词研究[J].心智与计算,2010(2).

[6] 杨凯艳.基于改进的 TFIDF 关键词自动提取算法研究[硕士学位论文].保存地点：湘潭大学，2015 年.

[7] 梁 彦.基于分布式平台 Speak 和 YARN 的数据挖掘算法的并行化研究[D]：[硕士学位论文].保存地点：中山大学，2014 年.

[8] 徐林龙，付剑生，蒋春恒，林文斌.一种基于威尔逊区间的商品好评率排名算法[J].计算机技术与发展，2015 年，第 25 卷（第 5 期）：168-171.

[9]周东杰,郑泽芝.可读性研究综述[J].泉州师范学院学报,2020,38(01):55-63.

[10] 束诗雨.基于集成学习的支持向量机预测优化算法及其应用[D]:[硕士学位论文].保存地点：东华大学，2015 年.

[11] V.N.Vapnik.The nature of statistical learning theory.Springer:New York,1995.

[12] 郭 骏.基于支持向量机和深度学习的分类算法研究[D]：[硕士学位论文].保存地点：华东师范大学，2015 年.

[13] 张宜浩，金 澎，孙 锐，等. 基于改进 k-means 算法的中文词义归纳[J]. 计算机应用，2012，32(5)：1332-1334.

[14] 李亚，刘丽平，李柏青，易俊，王泽忠，田世明.基于改进 K-Means 聚类和 BP 神经网络的台线损率计算方法[J].中国电机工程学报，2016 年，第 36 卷（第 17 期）：4543-4552.