C 题：“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的工作带来了极大挑战.留言文本具有信息内容短、信息量少的特征.我们可以利用数据挖掘技术从留言文本中挖掘出我们所需要了解到对我的问题.各类文本数据量的不断攀升给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.因此本题根据所给数据文本建立理想模型对所给附件进行文本数据分析.

问题一分类问题：根据附件 1 的留言分类，根据附件 2 建立关于留言内容的一级标签分类模型，利用贝叶斯公式，对文本数据类别和特征进行分类，由于贝叶斯公式存在的不足我们需要再利用四种传统的特征抽取方法词频法、互信息法、CHI 统计、信息增益法筛选出针对该类的特征集合对贝叶斯公式进行一定的优化.最后利用 F-Score 函数对我们所建立进行分类的模型进行评价，验证我们建立的模型是否准确.

问题二热点问题：利用群众所反映的问题建立问题热度指数的模型，根据热度指数越高其热度和频率也就越高，建立留言热度的聚类分析模型，利用文本数据的权重问题，根据聚类模型建立文本数据权重分析模型，利用 MATLAB 对模型进行分析定义出合理的热度评价指标，根据我们所建立的模型最后求解出“热点问题表”和“热点问题留言明细表”.

问题三对答复意见的评价：建立评价模型，通过答复内容与问题之间存在的相关性、完整性、可解释性各角度给出评价方案，利用信息价值体现三要素图对答复内容进行大概分析计算出特征值，再利用答复内容的相关性，建立相关性模型判定词汇间的集聚关系，建立多元线性回归模型从多个指标来判断答复意见的答复质量的准确性，再利用模糊建模方法对答复意见可解释性进行计算在各角度的研究下建立合理的答复意见评价方案.

关键词：贝叶斯公式、聚类分析、权重、MATLAB、多元线性回归模型、模糊模型、集聚关系、特征向量、热点话题

一、问题重述

1.1 问题背景

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.

1.2 目标任务

按照群众的留言我们根据分成不同的等级，我们所要研究的问题如下：

问题一：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对留言进行分类(c 参考附件一)，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且差错率高等问题.根据附件二给出的数据，建立模型对一级标签分类，并且使用 F-Score 对我们所建模型的分类方法进行评价.

问题二：某一时段内群众几种反映的某一问题可称为热点问题，根据附件三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，定义合理的热度评价指标，用问题一所给出的模型和所建模型给出对问题二的评价结果，并用 Excel 表给出表 1 排名前 5 的热点问题保存为“热点问题表.xls”给出表 2 相应热点问题对应的留言信息保存为“热点问题留言信息表.xls”.

问题三:针对附件四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.

二、问题分析

2.1 问题一的分析

本题问题是关于文本挖掘的问题，针对附件二给出的留言文本信息进行分析处理挖掘，问题一我们依据贝叶斯公式根据文本的不同特征并且每个特征都是相对独立的特点进行计算出特征概率，再利用四种传统的特征抽取方法词频法、互信息法、CHI 统计、信息增益法筛选出针对该类的特征集合，对于每一类，去除那些表现力不强的词汇，筛选出针对该类的特征项集合，提高分类精度.

2.2 问题二的分析

本题问题是关于热点问题评估的问题，通过特征权重计算的 TF-IDF 公式，用特征词的权重来表示文本向量从而得到特征词在整个文档集中的权重，话题通过权重计算之后，可得到一组用权重表示的话题向量，每个话题向量包含一个特征项序列，此时通过构建热度评估模型提取出一个时间段内的热点话题.通过提取留言特征参量计算用户留言评率，留言分布率，留言时长等问题，用户关注度可通过获取留言的点赞数等方法来构建热度指数模型.最后将计算出每条信息相应的热点指数排名前五的 Excel 表格排列出来.以及相应的所有问题的查询使用表格排列出来.

2.3 问题三的分析

本题问题是关于相关性的一个评价问题，针对附件三所给出的答复意见，我们选取其中的一类数据我们找出三个指标从三个指标的相关度来做出评价.利用集聚的方法找出各个词组的相关性运用多元线性回归模型从相关性，完整性和可

确定性找出线性关系，通过指标的系数来分析相关性，完整性和可确定性对答复意见质量的影响.然后，运用模糊综合评价模型分别按照相关性，完整性和可解释性对答复意见质量进行排名.相关性高的代表答复意见较为准确，相关系数低的代表答复意见出现的偏差较大.

三、问题假设

1.假设特征之间相互独立.
2.假设市民问题都为有效问题.
3.假设表中的问题有部分已经解决.
4.假设各答复之间因子没有相互影响.
5.假设市民点赞数存在不真实的情况.

四、符号说明

序号	符号	意义
1	N	文档总数
2	C	某一特定的类别
3	t	特定的词条
4	A	属于c类且包含t的文档频数
5	B	不属于c但是不包含t的文档频数
6	D	既不包含t也不属于c类的文档频数
7	$\|\mathrm{D}\|$	该类的训练文本数
8	$\|v\|$	总词数
9	$p(c_i)$	c_i类文档在语料中的概率
10	$\mathrm{p}(t)$	语料中包含词条t的文档的概率
11	$p(c_i/t)$	c_i文档包含词条t是属于类的条件概率
12	$p(\bar{t})$	语料中不包含词条t的文档的概率
13	$p(c_i/\bar{t})$	c_i文档不包含词条是属于的条件概率
14	n_{t}	在一段文本中出现的特征词的总数
15	$Miss_j$	文本的特征漏报率
16	FA_{j}	文本的特征误报率
17	P_{Miss}	文本集特征平均漏报率
18	P_{FA}	文本集特征平均误报率

五、模型的建立

5.1 问题一模型的建立

我们将附件2中的关于留言内容一级分类标签分类提取出，分别为城乡建设，环境保护，交通运输，教育文体，劳动和社会保障，商贸旅游，卫生计生，将七种分类看做七种类别，再把每个类别中的关键词也就是高频词作为特征.在我们所

分离出的文档特征和类别，利用以下公式进行分类:

贝叶斯公式：

$$P(A/B)=\frac{P(A)\ P(B/A)}{P(B)} \tag{1}$$

换个比较形象的形式也可如下

$$P(\text{类别}/\text{特征})=\frac{P(\text{类别})P(\text{特征}/\text{类别})}{P(\text{特征})} \tag{2}$$

$p(A)$是先验概率，表示每种类别分布的概率;

$p(B|A)$是条件概率，表示在某种类别前提下，某事发生的概率;该条件概率可通过统计而得出，这里需要引入极大似然估计概念，详见后文.

$p(A|B)$是后验概率，表示某事发生了，并且它属于某一类别的概率，有了这个后验概率，便可对样本进行分类.后验概率越大，说明某事物属于这个类别的可能性越大，便越有理由把它归到这个类别下.

特征类别往往是多维：

$$P(\text{feature}/\text{class})=P(f_0,f_1,...,f_n/c)$$

假设特征之间是独立的则有：

$$P(f_0,f_1,...,f_n/c)=\prod_i^n P(f_i/c)$$

利用贝叶斯分类器，计算分离出类别的概率，在计算所有特征条件概率

$$P(class/f_0,f_1,...,f_n)=\frac{P(class_i)}{p(f_0,f_1,...,f_n)}\prod_j^n P(f_j/c) \tag{3}$$

这样依据贝叶斯公式我们可找出一个最可能的类别.

贝叶斯公式有缺陷造成误差较大，所以运用以下方法对贝叶斯公式进行优化：

1. 词频法

词频分析法是文献计量学的重要分析方法之一,而确定高频词阈值是进行词频分析的必要前提,高频词阈值的选取不仅决定词频分析法的分析结果,而且对整个分析研究都有着极其重要的影响.我们定义 7 种类别分别计算其中高频出现的文本从而达到分类的标准.我们将某一个词组出现在文档中的频率称之为词频法也可称之为文档频率：

$$DF(t_k)=\frac{\text{出现词组}t_k\text{的文本数}}{\text{数据集文本总数}} \tag{4}$$

基于文档频率的特征选择一般过程：

(1)设定文档频率 DF 的上界阈值∂_u和下界阈值∂_l；

(2)统计训练数据集中词组的文档频率；

求解：(3) $\forall DF(t_k)<\partial_l$：由于词组$t_k$在训练集中出现的频率过低，不具有代

表性，因此在特征空间中去除词组 t_k；

(4) $\forall DF（t_k）<\partial_u$：由于词组 t_k 在训练集中出现的频率过高，不具有区分度，因此在特征空间中去除词组 t_k；

所以最终选取的作为特征的词组必须满足 $\partial_l \le DF（t_k）\le \partial_u$；

2.CHI 方法

对于每个词，计算词和类别的 CHI 统计为

$$\frac{N\times(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)}$$

词和类别的互信息：

$$\log\frac{P(W/C_j)}{P(W)}$$

$$P(W/C_j)=\frac{1+\sum_{l=1}^{|D|}N(W,d_l)}{|V|+\sum_{s=1}^{|V|}\sum_{l=1}^{|D|}N(W_s,d_i)} \qquad (5)$$

4.信息增量法

通过文档 c 出现和不出现词条 t 来表示 t 对文档的增益程度.词条的信息增量可用以下公式表示：

$$IG(t)=-\sum_{i=1}^{m}p(c_i)\log p(c_i)+p(t)\sum_{i=1}^{m}p(c_i/t)\log p(c_i/t)+p(\bar{t})\sum_{i=1}^{m}p(c_i/\bar{t})\log p(c_i/\bar{t})$$

符号说明：$p(c_i)$ 是 c_i 类文档在语料中的概率，$p(t)$ 是语料中包含词条 t 的文档的概率，$p(c_i/t)$ 是文档包含词条 t 是属于 c_i 类的条件概率，$p(\bar{t})$ 是语料中不包含词条 t 的文档的概率，$p(c_i/\bar{t})$ 是文档不包含词条是属于 c_i 的条件概率，m 是类别数.

F-Score 函数

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

我们可将此分类方法延伸到：

$$F-Score=(1-\beta^2)\cdot\frac{Precision\cdot Recall}{\beta^2\cdot Precision+Recall} \qquad (6)$$

Precision：准确率，又称为查准率，是分类器正确判断为该类的样本数与分类器判断属于该类的样本总数之比率，体现了系统分类结果的准确程度，计算公式为：

$$P=\frac{\text{分类的正确文本数}}{\text{实际分类的文本数}}\times 100\%$$

Recall：召回率，又称查全率，是分类器正确判断为该类的样本数与属于该类的样本总数之比率，体现了系统分类结果的完备性，计算公式如下：

$$R=\frac{\text{分类的正确文本数}}{\text{应有文本数}}\times 100\%$$

β值：用来衡量 precision，recall 在 F-Score 计算的权重，取值有下面三种情况：

如果取值为 1，表示 precision 与 recall 一样重要

如果取小于 1,表示 Precision 比 Recall 重要

如果取大于 1,表示 Recall 比 Precision 重要

5.2 问题二模型的建立

本题文本也可为话题文本问题，话题可将其聚类，聚类之后可得到一组用聚类中心表示的话题向量，每个话题向量包含一个特征项序列,通过热度评估模型提取出一个时间段内的热点话题.

根据本题需要，我们构建一个以下评价指标，因此我们用构建热度指数的模型，热度指数越高其热度和频率也就越高，根据留言热度的计算公式：

$$G1=\sum_{i_1=1}^{N_1}\mathrm{hi}_1 \tag{7}$$

其中 N_1 表示留言的总数，hi_1 表示第 i_1 条留言的计算得分，i_1 的取值 $1\sim \mathrm{N}_1$，hi_1 的计算公式为：

$$\mathrm{hi}_1=\mathrm{p}_1\bullet \mathrm{c}_1\bullet \mathrm{r}_1$$

其中，p_1 表示留言信息的权重，c_1 表示留言的点赞参数，r_1 表示留言时间的聚集参数；

其中 $p_1=\dfrac{e}{e*\lg N_a+1}$ （8）

Na 表示留言排名：

$$c_1=\lg C_1$$

C_1 表示留言的点赞数

$$r_1=0.5*\lg R_1$$

R_1 表示时间相近的信息数

模型的最终目标是实现一个有效的特征集,因此我们需要用权重的计算来保证目标的有效性，权重的计算能够确实代表目标文本内容的完备性,和能将目标文本与其他文本相区别的区分性.因为本题是从多个网站获取的市民意见，我们我们采用以上方法对文本热度进行权重计算以保证结果的准确性.

同样根据留言热度指数的计算公式可计算出留言热度的公式为：

$$G_2 = \mathrm{p}_2 * s$$

p_2 表示热度数量权重，s 为整体的留言数量得分

$$\mathrm{s} = (\frac{2}{1+\mathrm{e}^{\frac{\mathrm{S}}{\mathrm{N}_2}}} - 1) \bullet \mathrm{S}$$

s 表示整体的点赞数量，N_2 为常量

再根据热度指数的计算方法其热度的计算公式为

$$G3 = \sum_{\mathrm{i}_3=1}^{N_3} hi_3$$

其表示以及相应参数的计算方法与 G_1 的算法相同.

正指标：

$$y_{ij} = \frac{x_{ij} - \min(x_{ij})}{\max(x_{ij}) - \min(x_{ij})} \tag{9}$$

逆指标：

$$\mathrm{y}_{ij} = \frac{\max(x_{ij}) - x_{ij}}{\max(x_{ij}) - \min(x_{ij})} \tag{10}$$

式中 $\max(x_{ij})$，$\min(x_{ij})$ 分别表示指标评价值的最大值和最小值.

事件连续时间=事件从发生到最后一次发生所持续的时间（单位：天）

最后根据公式

$$H = G1 + G2 + G3$$

可得出留言的热度指数 H.最后热度指数越高则说明为更为热点问题，可对此进行热度的排名，排出前五的热点问题，用如图所给的表格填写.

评价体系指标建立的原则

(1) 科学性原则.评价指标体系的设计必须建立在科学的基础上，客观如实的反映住宅小区人居环境目标的构成，反映环境目标和指标的支配关系，而且指标体系的繁简也要适宜.评价指标不能过多过细，导致指标之间相互重叠，也不能过少过简，导致指标信息遗漏.

(2) 系统性原则.评价指标体系中所设置的每个指标项都应能独立地反映小区的某一个方面或不同层面的水平.各指标间相互独立，又相互联系，共同构成一个有机的整体，使评价结果可以全面的反映小区的整体效果及其综合效益.

(3) 可比性原则.评价指标应使用统一的标准衡量，尽量消除人为的可变动因素的

影响，使评价对象之间存在可比性，进而确保评价结果的准确性.
(4)可量化原则.体系中各评价指标都应定量化，对于评价指标中的定性指标，应该通过现代定量化的科学分析方法使之量化.这有利于衡量被评价对象实现目标的程度，也有利于运用计算机进行分析与处理.
(5)可行性原则.设置指标时应保证通过-般的比较简便的统计方法或者查阅资料就可以采集到确定指标值所需的数据，以便于在实践中的应用.
由上诉所说的我们可以画出如下框架图：

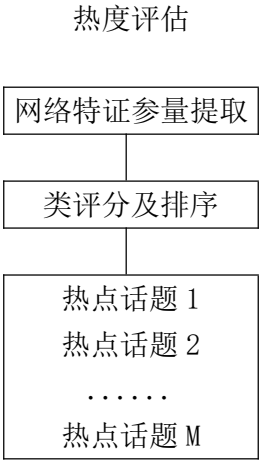

图 1：热度评估图

5.3 问题三模型的建立

问题三主要目标针对附件 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.评价指标体系是评价目标的具体化，行为化和可操作化，是进行评价工作的基本依据是评价方案的核心内容.我们在设计评价指标体系的时候采用因素分解法 ，把评价目标逐级分解.建立指标权重是一个表明该指标重要程度和作用大小的数字指标，在评价方案指标体系中各个指标权重是不一样的.由权重我们确定之间的相关性，完整性，可解释性.分为等级，略为重要，重要，同等重要.我们通过选取一部分的数据，其余的数据可以采用相同的方法.

信息的价值体现在三个方面，即包括三个要素:信息的完整性、信息的有用性和信息的可用性，我们可根据三要素画出信息价值三角形：

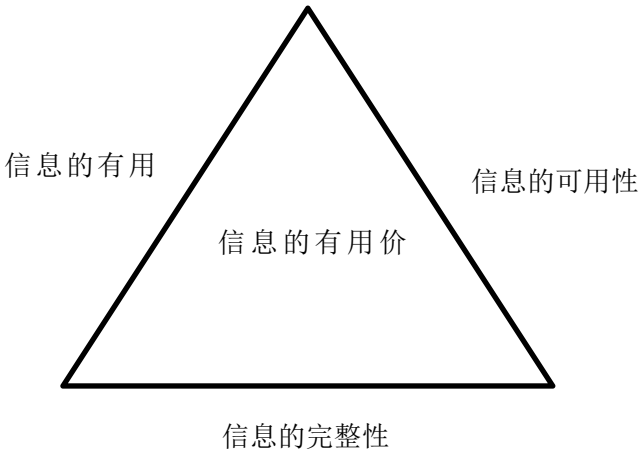

图 2：信息价值三角形

在文档相关性计算中，我们需要引入词汇集聚，集聚是指通过回指、省略、

链接以及语义联系等手段，把句子黏接在一起，使他们成为表达一定含义的整体的方法.词汇集聚将文本中存在语义关联的词构成一条链,使这些相关词保持语义上的连贯性.词汇集聚能把我们在答复中所需要的关键词串联起来，方便于研究和计算，从而获得答复中相关联系.

词汇集聚相关性计算我们应用 LCDRM 模型，在模型 LCDRM 中有一种算法 DRLC，用于判定词汇间的集聚关系：

设两次 w_1 和 w_2，$\{DEF(w_1)_i \mid i=1,2,3....,n,(n\geq 1)\}$ 与 $\{DEF(w_2)_j \mid j=1,2,3....,m(m\geq 1)\}$ 分别为 w_1 和 w_2 拥有的.其中 w_1 和 w_2 的语义连接关系包含以下 3 种关系：

（1）重复：$w_1 = w_2$

（2）直接关系：$w_1 \neq w_2$，存在 $I(DEF(\mathrm{w}_1)_i)_p$，使得 $R(I(DEF(w_i)_i)_p) \in DEF(w_2)_j$，$R$ 表示上下义、反义或对义等的一种关系.

（3）搭配关联：$\forall \mathrm{i}, j, i=1,2...n(n\geq 1), j=1,2....m(m\geq 1)$，$DEF(w_1)_i \neq DEF(w_2)_{\mathrm{j}}$，且 $\exists \mathrm{i}, j, i=1,2...n(n\geq 1), j=1,2...m(m\geq 1)$，使得 $DEF(w_1)_i \cap DEF(w_2)_j \neq \phi$

我们可以知道存在某些出现次数多的特征词比出现次数少的特征词更能体现一段回复的完整性，相关性，可解释性.我们构建每段回复的特征向量，除了用原来的特征词构成特征向量以外再将在文本中以高频出现的特征词补充到向量列表中.用一个特征词在一段文档中的绝度地出现次数来衡量显然是不合理的，我们通过下列式子计算每个特征词在一段文本中的相对出现频率从而计算结果评价：

$$TF_{\mathrm{i}}^{'} = \frac{\lg(TF_i + 1)}{\max\limits_{1\leq j\leq n_i} \lg(TF_j + 1)} \tag{11}$$

其中 $1 \leq \mathrm{i} \leq \mathrm{n}_t$，$n_{\mathrm{t}}$ 为在一段文本中出现的特征词的总数，TF_{i} 为第 i 个特征词在这段文本中出现的次数鉴于模糊特征的思想，将每个特征词的 $TF^{'}$ 值进行二值化，如果一个特征词在一篇文档中出现，则将其 $TF^{'}$ 值置为 0；如果一个特征词在一篇文档中以高频出现，则将其 $TF^{'}$ 值置为 1，设计如下：

$$TF^{'} = \begin{cases} 0 \\ 1 \end{cases}$$

在 $TF^{'}$ 为 0 的情况下 $0 < TF^{'} \leq 1$，在 $TF^{'}$ 为 1 的情况下 $T_{\max} < TF^{'} \leq 1$，我们设定一个阈值 $T_{\max}$，如果一个特征词的 $TF^{'}$ 值大于阈值则可以认为这个特征词是以

高频出现.

一段回复意见可以表示为：

$$Doc=\left\langle <Term_1,TF_1^{'}>,<Term_2,TF_2^{'}>,.....,<Term_n,TF_n^{'}>\right\rangle$$

模糊集合采用梯形隶属函数，假设$[m_i,n_i]$是输入x_i论域的范围，给定区间$[s_i,t_i]\subseteq[m_i,n_i]$其中$s_i$，$\mathrm{t}_i$分别是输入$\mathrm{x}_i$上相邻的两个分叉点的值，且$s_i<t_i$，$\mathrm{m}_i<n_i$，区间的长度为$\mathrm{w}_i=t_i-s_i$，区间$[s_i,t_i]$上的隶属函数定义如下：

若$s_i=\mathrm{m}_i$

$$\mu(\mathrm{x}_i)=\begin{cases}1 & x_i\in[s_i,s_i+0.8\mathrm{w}_i]\\ 1-\dfrac{x_i-(s_i+0.8w_i)}{0.4w_i} & x_i\in[s_i+0.8\mathrm{w}_i,\mathrm{t}_i+0.2w_i]\end{cases}\tag{12}$$

当$\mathrm{t}_i=\mathrm{n}_i$

$$\mu(\mathrm{x}_i)=\begin{cases}1-\dfrac{\mathrm{t}_i-0.8w_i-x_i}{0.4w_i} & x_i\in[s_i-0.2w_i,t_i-0.8\mathrm{w}_i]\\ 1 & x_i\in[t_i-0.8w_i,\mathrm{t}_i]\end{cases}\tag{13}$$

若$[s_i,t_i]\subset[m_i,n_i]$则有

$$\mu(\mathrm{x}_i)\begin{cases}1-\dfrac{\mathrm{x}_i+0.2w_i-x_i}{0.4w_i} & x_i\in[s_i-0.2w_i,s_i+0.2w_i]\\ 1 & x_i\in[s_i+0.2w_i,t_i-0.2w_i]\\ 1-\dfrac{x_i-t_i+0.2w_i}{0.4w_i} & x_i\in[t_i-0.2w_i,t_i+0.2w_i]\end{cases}\tag{14}$$

多元线性回归模型

由于答复意见可以从多个指标来判断其答复质量，为了研究答复意见回答的准确性我们建立多元线性规划.我们从完整性，相关性，可解释性三个指标来评价答复建议的准确性，设f_k，其中$k=1,2,3$分别表示完整性，相关性，可解释性得到的函数，我们可以建立以下方程：

$$\begin{cases}\mathrm{f}1=f(x_i,x_j,x_m)\\ f2=f(x_k,x_p,x_z)\\ f3=f(x_i,x_j,x_m)\end{cases}\tag{15}$$

通过评价指标可以知道，完整性，相关性，可解释性在整个回复意见指标中所占的权重是不一样的，我们定义各个权重为$\partial 1$，$\partial 2$，$\partial 3$，那么我们定义的评价

值 F 得函数为：

$$F = \partial 1\mathrm{f}1 + \partial 2 f2 + \partial 3 f3$$

我们通过 F 的值来判断回复的准确性的高低.
基于文档可利用聚类方法进行建模，所以基于聚类法建立其可解释性模型：

利用提出的含有熵的聚类有效性函数 $H(U,c,V)$ 来求得最佳聚类数 c^m，$H(U,c,V)$ 的定义如下：

$$H(U,c,V) = UVp(U,c) + (1-U)(1-\overline{E})$$

其中 $Vp(U,c)$ 是有效性函数，$\overline{E}$ 是聚类中心熵的平均值，U 是权重系数且 $U[0,1]$，Vp 定义如下：

$$V\mathrm{p(U,c)} = \frac{1}{N}\sum_{\mathrm{k=1}}^{N}\max_{i}(\mathrm{u}_{ik}) - \frac{1}{K}\sum_{i=1}^{c-1}\sum_{j=i+1}^{c}[\frac{1}{N}\sum_{\mathrm{k=1}}^{N}\min(u_{ik},u_{ik})] \quad (16)$$

$$K = \sum_{i=1}^{\mathrm{c\text{-}1}}\mathrm{i}$$

建立以上模型来求解问题.

六、模型的求解

6.1 问题一的求解

根据所建立模型利用词频法对“附件 2”中的城乡建设进行分类，根据数据出现次数较多的高频词为：安全（25），招标（8），违章（6），政策（10），垃圾（5）.

根据词频法的公式

$$DF(t_k) = \frac{\text{出现词组}t_k\text{的文本数}}{\text{数据集文本总数}}$$

可得出 DF（安全）=0.25，DF（招标）=0.08，DF（违章）=0.06，DF（政策）=0.1 可见其中大多以安全问题为主，安全问题出现的频率更高.

根据 CHI 方法，互信息方法，信息增量法，根据模型的信息，以及“附件 2”所给的信息代入进行求解可得到大致分类.

6.2 问题二的求解

根据问题二模型的建立我们从附件 2 和附件 3 中的数据求解出热点问题和其对应的回复.通过提取文本的特征参量计算留言频率，留言分布，留言时长等，由此可得留言的关注度的高低与留言特征参量的数值成正比.（详情见附录）

6.3 问题三的求解

我们根据附件 4 所给出的答复意见从其相关性、完整性、可确定性方面判断出，相关性越高越完整从而得出的，文本的留言的相关性、完整性在较大的程度

上影响着留言回复的准确性.

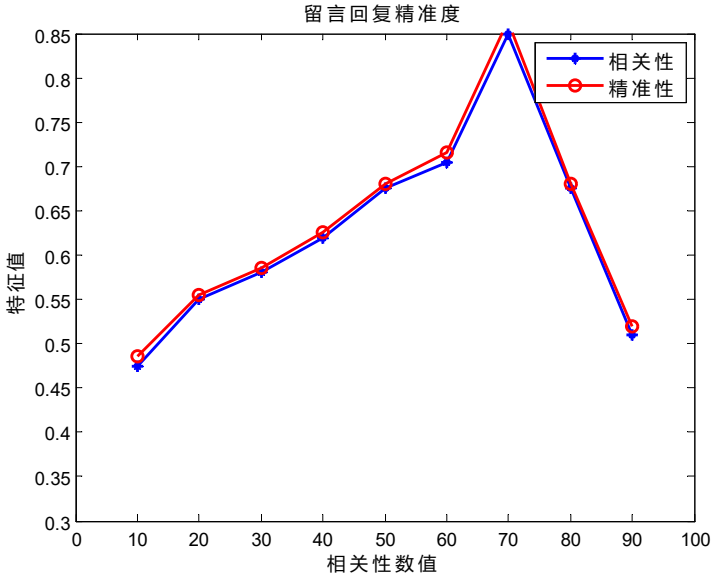

图 3：留言回复准确度

七、模型的检验

我们选取最为常用的基于 K-NN 的改进方法作为分类方法，采用 cosine 距离作为计算方法来检验我们的模型

$$W_{\mathrm{ik}}=\frac{tf_{ik}\times\log\left(N/n_{\mathrm{k}}+0.01\right)\times IG_k}{\sqrt{\sum_{k=1}^{t}\left(tf_{ik}\times\log\left(N/n_{\mathrm{k}}+0.01\right)\times IG_k\right)^2}} \tag{17}$$

$$sim\left(D_1,D_2\right)=\frac{\sum_{j=1}^{t}W_{D1j}\times W_{D2j}}{\sqrt{\sum_{i=1}^{t}\left(W_{D1j}\right)^2\times\sum_{i=1}^{t}\left(W_{D1j}\right)^2}} \tag{18}$$

Purity 方法是一种以为简单的聚类评价方法我们用来检验我们的模型，对于一个聚类结果 R_{h}，其中数量最多的类在聚类结果中的比例即为聚类的纯度. 对于一个聚类 i，首先计算聚类 i 中的成员属于类 j 的概率 P_{ij}，$P_{\mathrm{ij}}=\frac{\mathrm{m}_{ij}}{m_i}$. 其中 m_i 为在聚类 i 中所有变量的个数，m_{ij} 是聚类 i 中的变量属于聚类 j 的个数. 聚类 i 的纯度 P_{i} 定义为：

$$P_i=\max\left(P_{ij}\right)$$

整个聚类划分的纯度 Purity 为：

$$Purity=\sum_{i=1}^{k}\frac{m_i}{m}p_i$$

其中 K 是聚类的数目，m 是聚类划分所涉及的成员个数. 利用上述模型来检验我们

的聚类分析问题.

八、模型的评价

8.1 模型的优点

1.模型的建立大部分都考虑了特征词在每段文本总出现的词频率对类别信息的贡献.根据特征词在每段文本中出现的词频率标出以高频率出现的词，提高了文本分类性能.

2.采用模糊综合评估的方法,在评价留言回复意见的准确性时,确定了不同的评价因素的均衡分布和权重值,使得评估的结果更加全面,更加客观.更加准确的判断了留言的回复情况.

3.LCDRM 是一种适合比较严格文本的文档间相关性计算方法，将文档的字符表达抽象化，LCDRM 在一定程度上描述了文档的语义信息，带来了一定的主观色彩，我们采用所得出来的模型求解出我们的相关问题.

4.利用模糊系统具有最简约性、完备性、清晰性、一致性、紧凑性等特征能够更好地求解问题.

8.2 模型的缺点

1. 文字是一个比较精通的事物，我们所建立的模型考虑到的范围太小，太多的因素没有考虑到会导致我们的结果不够准确存在一定的误差.

2. 给出的数据较多我们所考虑的只是较少的一部分没有考虑全面，我们应该更加人性化.

8.3 模型的改进

对文本的表示是一个文档形式化的过程，其质量直接影响后续算法的执行精度和效率.其中，向量空间模型的方法是目前普遍认为效率较高的方法.特征空间表示为 $D=\{d_1,d_2,d_3,d_4....d_{|D|}\}$，每个文本表示为向量空间中的一个特征向量 $d_j=(w_{1j},w_{2j},w_{3j}....w_{|T|j})$，$w_{ij}\in[0,1]$表示第 j 段文本第 i 个特征项的权重，集合 T 包含了文档集中的所有特征.权重计算基于传统公式结合实际应用发展出了很多方式，如下：

$$w_{ij}=\frac{(1+lbtf_{ij})\times lb(N/n_i)}{\sqrt{\sum_{i=1}^{|T|}[(1+lbtf_{ij})\times lb(N/n_i)]^2}}$$

其中 $tf_{ij}=\mathrm{tf}(t_i,d_j)$表示词 t_i 在文档 d_j 中出现的频率，N 表示文档总数，n_i 表示文档集中包含特征 t_i 的文档数.

改进权重的计算的带的特征词更加接近我们所预想的，计算公式如下：

$$Miss_j=\frac{\text{未表示出的与文本}j\text{相关的特征数}}{\text{与文本}j\text{相关的特征总数}} \quad (19)$$

$$FA_j=\frac{\text{表示出的与文本}j\text{不相关的特征数}}{\text{与文本j不相关的特征总数}} \quad (20)$$

$$P_{Miss}=\frac{\sum_{j=1}^{|D|}Miss_j}{|D|} \quad (21)$$

$$P_{FA}=\frac{\sum_{j=1}^{|D|}FA_j}{|D|} \quad (22)$$

改良的特证权重的算法其平均漏报率和平均误报率都低于我们之前所用的公式，此算法在文本表示的准确度和相关性等因素方面都是较为完善的.

九、参考文献

[1] 张一文,齐佳音,方滨兴,李欲晓.非常规突发事件网络舆情热度评价指标体系构建[J].情报杂志,2010,29(11):71-75+117.

[2] 朱若初,张二伟,张星,李智.住宅小区人居环境评价指标体系研究[J].建筑管理现代化,2004(05):22-24.

[3] 刘星星,何婷婷,龚海军,陈龙.网络热点事件发现系统的设计[J].中文信息学报,2008,22(06):80-85.

[4] 王林芳.　桑蚕种品种纯度 DNA 检测方法的抽样检验方案设计[C].　中国蚕学会.中国蚕学会第四届青年学术研讨会会议论文集.中国蚕学会:中国蚕学会,2004:31-37.

[5] 赵玉茗,徐志明,王晓龙,朱鲲鹏.基于词汇集聚的文档相关性计算[J].电子与信息学报,2008(10):2512-2515.

[6] 罗海飞,吴刚,杨金生.基于贝叶斯的文本分类方法[J].计算机工程与设计,2006(24):4746-4748.

[7] 杨凯峰,张毅坤,李燕.基于文档频率的特征选择方法[J].计算机工程,2010,36(17):33-35+38

[8] 王晓兰,曾贤强,王文琰.一种基于模糊聚类的可解释性建模方法[J].甘肃科学学报,2006(04):75-79.

[9] 奉国和.文本分类性能评价研究[J].情报杂志,2011,30(08):66-70.

[10]张永.　基于解释性与精确性的模糊建模方法研究[D].南京理工大学,2006.

[11]陈莉萍,杜军平.突发事件热点话题识别系统及关键问题研究[J].计算机工程与应用,2011,47(32):19-22.

[12]鲁松,李晓黎,白硕,王实.文档中词语权重计算方法的改进[J].中文信息学报,2000(06):8-13+20.

[13]李霄野,李春生,李龙,张可佳.基于 LDA 模型的文本聚类检索[J].计算机与现代化,2018(06):7-11.

十、附录

```
x=10:10:90;
y1=[0.475 0.550 0.580 0.620 0.675 0.705 0.850 0.675 0.510];
y2=[0.485 0.555 0.585 0.625 0.680 0.715 0.865 0.680 0.520];
plot(x,y1,'-*b','LineWidth',2); %线形，颜色，标记
hold on
```

```
plot(x,y2,'-or','LineWidth',2); %线形，颜色，标记
axis([0,100,0.3,0.85])  %确定 x 轴与 y 轴框图大小
set(gca,'XTick',[0:10:100]) %x 轴范围 0-100，间隔 10
set(gca,'YTick',[0.3:0.05:0.85]) %y 轴范围 0-0.85，间隔 0.05
legend('相关性','精准性');   %图例
xlabel('相关性数值')
ylabel('特征值')
title('留言回复精准度')
```

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	5.1	2019/07/21-2019/09/10	A市A5区魅力之城小区	小区临街餐饮店油烟噪音扰民
2	2	4.8	2017/06/08-2019/11/05	A市经济学院学生	学校强制学生去定点企业实习
3	3	4.5	2019/04/05-2019/12-19	A市医疗	A市可不可以提高医疗门诊报销范围
4	4	4.3	2018/06/08-2015/11/05	K9县中医诊所	卖的药品是否合法
5	5	4.0	2019/06/08-2019/11/05	K市K1区外骨科医院	受伤赔偿