基于文本挖掘的网络问政平台留言分类及答复评估

摘　要：本文研究了智慧政务系统中群众问政留言的分类、热度评估、留言答复质量评估问题。

针对第一问，本文首先对文本数据进行预处理，基于 PYNLPIT 中文分词系统进行分词、文本去冗余、特征项初步筛选，建立向量空间模型，基于改进的基尼系数完成文本特征项的终选，并基于 TF-IDF 计算得到各终选特征项的权重计算。进而，建立基于 OAO-SVM 的文本分类模型，建立留言与一级标签之间的联系，并根据分类的正确率比对确定训练集、测试集的比率，最后根据 F1-Score 值确定可调参数，完成了分类的性能评价。

针对第二问，本文建立了基于文本聚类的热点问题动态挖掘。首先，本文结合正则表达式，基于留言标题涉及的地点信息（精确至区、县）将留言划分为 35 个初划分数据集，并按所含记录数量对初选数据集进行了筛选；进而，基于 PYNLPIR 分词系统对留言标题文本进行分词等处理，并将各数据集中记录按留言时间顺序排列，基于布尔模型建立文本表示模型，并基于词频对文本特征项进行筛选。最后，本文基于 Single Pass 算法对留言问题进行聚类，并基于各问题的留言总数、关注度、密度动态系数，建立了问题热度的动态评估系统，求得动态评估分数最高的五个热点话题。

针对第三问，本文首先分析了附件 4 的答复文本特征，进而确定答复质量评价指标为：相关性、完整性、时效性、可解释性、可行性，并从不同留言类型对答复各指标注重不同出发，建立答复质量评价模型。

最后本文进行了模型评价，并对建立网络行政平台提出了建立双向评价机制等建议。

关键词：网络问政平台，文本挖掘，PYNLPIT，OAO-SVM，Single Pass，密度动态系数

引言

随着信息技术的快速发展，互联网已经与社会生活密不可分。第 45 次《中国互联网络发展状况统计报告》显示，截至 2020 年 3 月 1 日，我国网民规模为 9.04 亿，互联网普及率达 64.5%，较 2019 年 6 月提升了 2.7 个百分点，增长迅速。

随着越来越多用户的加入，互联网也渐渐取代传统的文本数据保存、交流方式，成为人们获取信息的重要渠道，成为政府联系群众的重要纽带；利用信息技术推动电子政务成为实现国家治理体系和治理能力现代化目标的重要条件，信息化水平的提高也从一定程度上催生了“互联网+政务服务”模式的兴起。民众通过网络问政、网络建议、网络投诉与举报等方式反映诉求，政府也通过市长信箱、城市留言板、微博、微信等多种形式听取民声、收集民意，并及时通过网络回复、相关部门人员处理等措施及时解决群众问题；对群众留言进行分析研判，通过网

络问政于民、问需于民、问计于民，已经成为新形势的必然要求。然而，但在技术浪潮的推进下，信息爆炸，信息过剩，信息冗余等问题随之而来，依靠人工方式处理留言信息、划分类别、统计热点问题等工作已经面临出错率高、时效性差、工作量大等种种困难，利用自然语言处理和文本挖掘技术，更加快速高效准确智能地处理信息，具有重要作用。

文本挖掘技术主要涉及数据挖掘、模式识别、信息检索、自然语言处理等多领域的内容，国内主要应用中文自然语言处理技术和网络舆情分析模型，相关技术基本可以支持舆情分析中的中文分词、文本预处理、文本分类或聚类、计算语义相似度等。在舆情监控方面，多采用 K-means 或层次聚类等算法对文本进行聚类、识别热点事件等，但是自动文本分配还未能完全满足实际需求，存在很大的研究空间。

文章结构安排如下：第一部分，对附件 2 的留言文本进行分词、去停用词、降维等预处理，建立基于 OAO--支持向量机的留言文本分类模型并对模型分类效果进行评估；第二部分，结合附件 3 所给为本数据，分析评论分类与时间的关系，建立基于文本聚类的热点问题挖掘模型，总结得到排名前五的热点问题及其详细信息；第三部分，结合时效性、完整性等，建立答复质量评价模型，并针对案例开展分析。

一、基于 OAO-SVM 的留言一级标签分类

附件 1 显示了网络问政平台群众留言内容的一、二、三级标签划分体系，其中一级标签共包含城乡建设、党务政务、国土资源等 15 种；附件 2 共包含 9210 条记录，每个记录有 6 个字段，这些字段包括留言编号、用户、留言主题（文本信息）、留言时间、留言详情（文本信息）以及该条留言的一级标签分类。总体浏览两个附件可知，附件 2 的留言时间范围为 2010 年 11 月 2 日至 2020 年 1 月 8 日；留言涉及的一级标签划分有 7 种，相较于附件 1，缺少党务政务、国土资源、纪检监察、经济管理、科技与信息产业、民政、农村农业、政法 8 类一级标签。

1.1 文本预处理

计算机无法直接处理附件 2 所示的留言文本信息，由于文本的非结构化特点，在分类应将其转化成计算机可以处理的结构，这就需要将文本切分成语义单元。这些语义单元可以是句子、短语、词语或单个的字。

1.1.1 文本分词

文本分类过程中，关键技术的作用十分重要。分词是将连续的文字序列按照一定的规则重新组合词序列的过程[1]。英文文本各单词之间有空格，因而可自然得到英文文本的分词，中文文本无空格停顿，分界模糊，所以应该结合自身特点进行分词。国内的中文分词技术越来越成熟，其中，PYNLPIR 分词系统的前身是中科院 ICTCLAS 汉语分词系统，可以管理百万级别的词典知识库，其主要功

能包括中文分词、词性标注、命名实体识别、新词识别等，支持用户词典、繁体中文[2]；单机每秒可以查询 100 万词条,分词速度可达到 500KB/s，词精度精准到 98.45%，API 小于 100KB，种词典数据压缩后不超过 3M[3]。

综上，本文采用 PYNLPIR 分词系统，对附件 2 的各条留言的留言标题、留言详情进行初步分词，结果见附件一。

1.1.2 文本去冗余

Step 1：去停用词：语言学中，主、谓、宾是一个句子的主干，出现在文本中的副词、助词、连词、代词等，对于描述文本含义几乎无作用，称其为停用词(Stop Words)。例如：助词中的“的，呢，呀”；副词中的“很，极，都”；代词中的“你，我，她”等[2]。为了减少存储空间、降低停用词对分类的干扰，对停用词进行剔除。

首先利用 PYNLPIR 系统，标注分词词性，去除副词、助词、量词、代词等；进而，利用停用词表（见附件二）过滤其余停用词。

Step 2：文本特征项初选：去除停用词后，提取文本中的名词、动词、形容词，作为该条留言文本的特征项，结果见附件三。

1.2 留言文本向量空间模型

1.2.1 模型原理

预处理后，计算机仍无法直接处理文本信息。要使计算机能够有效地处理文本串，就必须找到一种理想的文本表示方法。文本信息的特征表示模型有多种，常用的有布尔逻辑型、向量空间型、混合型等，其中向量空间模型(VSM，Vector Space Model)是目前文本处理应用领域使用最多且效果较好的文本表示法。VSM 是 20 世纪 60 年代末期由 G. Salton 等人提出的，最早用在 SMART 信息检索系统中。向量空间用高维度空间对文本进行表示，其基本原理是：将每一文档都映射为由一组规范化正交词条矢量张成的向量空间中的一个点，向量之间的距离表示文档之间的相似度。对于所有的文档类和未知文档，都可以用此空间中的词条向量来表示[4]。

1.2.2 向量空间模型

向量空间模型中的基本定义有：

- 文档：泛指文本对象。
- 特征项：又称索引项，常指文档中的词或短语，是文档分类的重要依据。
- 特征项权值：表示特征项对文档分类的重要程度。

设文本集合 D 中共有 n 个文档，即 $D=(d_1,d_2,\cdots,d_n)$，特征空间有 m 个特征项，则对于第 i 个文档 d_i，可将其表示为：

$$d_i=(t_1,w_{i1};t_2,w_{i2};\ldots;t_m,w_{im}), \quad (1)$$

其中 $t_k(1\le k\le m)$ 是特征项，$w_{ik}(1\le k\le m)$ 表示该文本中第 k 个特征项的权重。

向量空间模型满足以下两条性质：

- 互异性：各特征项互异，即 $t_i \neq t_j (i \neq j)$；
- 无关性：文本各特征项都是互相独立的，与其在文本中的顺序、语义均无关。即对于文本 d_i，任意交换其特征项位置，仍表示原文本。

由向量空间模型的性质，可将文本集的所有特征项 $t_2, t_2, \cdots, t_n$ 看成 n 维坐标空间，而特征项权重 $w_{i1}, w_{i2}, \ldots, w_{im}$ 为相应坐标值。设 $\Re(r_1, r_2, \cdots, r_{15})$ 为一级标签向量，r_{ik} 表示文本 d_i 的一级标签分类为 r_k。因此，含 n 个文档的文本集合可由各文本各特征项权值坐标以及其一级标签分类共同表示为：

$$D = \begin{array}{c} \\ d_1 \\ d_2 \\ \vdots \\ d_n \end{array} \begin{array}{c} \begin{array}{cccc} t_1 & t_2 & \cdots & t_m \end{array} \quad \Re_{ik} \\ \begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{pmatrix} \begin{array}{c} r_{1k_1} \\ r_{2k_2} \\ \vdots \\ r_{nk_n} \end{array} \end{array}. \tag{2}$$

以上步骤建立的向量空间模型可将非结构化的文本集表示为数字矩阵形式，使得各种数学处理成为可能。因此，确定文本特征项、计算特征项权重，是建立向量空间模型的关键步骤。

1.2.3 基于改进的基尼指数的特征项终选

1.2 的 Step 2 处理过程已经初步确定了各留言文本的特征项，然而，初选特征项组成的特征向量具有极高的维数，直接进行分类处理则面临计算量大、算法复杂等缺陷；另外，特征项决定了分类器的输入内容，特征维数太高会造成系统响应时间过长，进而影响分类效率和准确率。为提高特征项对文本信息的代表性、缩减向量空间维度，本文引入改进的基尼指数算法，进行特征选择。

基尼指数 (*Gini Index*) 是应用在决策树算法中的不纯性分割方法。为了将基尼指数应用到文本分类中，Shang[5]等人提出了改进的基尼指数特征选择方法 (*Improved Gini Index*)，该特征选择方法认为：如果一个特征出现在所有文本中，则该特征只包含很少的分类信息，甚至不包含任何的分类信息；相反，如果一个特征只出现在某一个类别中，则该特征包含的类别信息最多[6]。因此，基尼指数是用来度量一个特征项包含某个类别信息的纯度，纯度越高，该特征项越能够代表该类别。改进的基尼指数的定义如下：

$$Gini(t_k) = \sum_{i=1}^{m} P(t_k \mid r_i)^2 P(r_i \mid t_k)^2, \tag{3}$$

其中 $P(t_k \mid r_i)$ 表示特征 t_k 出现在类别 r_i 中的概率，$P(r_i \mid t_k)$ 表示当特征 t_k 出现时，该特征属于类别 r_i 的概率。

计算各初选特征项的改进的基尼系数，并按从小到大顺序排序，选择前 500

个词作为特征项，完成文本特征项终选（终选文本特征项及其改进的基尼系数结果见附件四）。根据终选特征项，对前文得到的附件三进行再一次降维，由于算法连续性，降维结果所属附件在后文（3.3 末尾）予以说明。降维结果显示，经过文本预处理、文本特征项初选、文本特征项终选，留言文本向量空间的维度得到了极为有效的降低，为下一步特征项权重的确定、分类模型的建立奠定基础。

1.2.4 基于 TF-IDF 的特征权重计算

为了体现特征对类别的重要程度，对特征进行加权。特征选择时利用的特征之间相互独立的假设往往与实际有出入，特征向量对分类的有用程度不同，因此有必要对特征加权。常用的加权函数有布尔权重、词频权重、TF-IDF 权重等[7]。$TF-IDF$(*term frequency – inverse document frequency*) 是一种用于信息检索与数据挖掘的常用加权技术, 综合了绝对词频法和逆文档频率法，弥补了词频法忽略特征的类间分布和 IDF 忽视特征在文本中出现频率的不足。该算法原理为：特征在文档中出现次数越多则该特征越重要，整个文档集中，包含特征项的文档数越大，则说明该特征项在区分文档中越不重要。权重函数为：

$$w_{ik} = tf_{ik} \times idf_k, \tag{4}$$

其中 tf_{ik} 表示项 t_k 在文本 d_i 中出现的频率， idf_k 表示项 t_k 的反文档频率,是反映 t_k 在一个文档集中按文档统计出现的频繁程度的指标。该权重有多种计算方法，本文选择按如下公式计算各文本中各特征项的权重：

$$w_{ik} = tf_{ik} \times log(\frac{n}{n_k + 1}), \tag{5}$$

其中 n 的含义同上文，表示全体文本集数量； n_k 表示文本集中出现 t_k 的频数。公式可见， $TF-IDF$ 的值和特征项在单个文档中的出现次数成正比，与在文档集整体中的出现次数成反比。

考虑文本长度对文本特征权重的影响，进一步对式(5)做归一化处理[7]，将各特征选择权重规范至[0,1] 内：

$$w_{ik} = \frac{tf_{ik} \times log(\frac{n}{n_k + 1})}{\sqrt{\sum_{k=1}^{n} tf_{ik}^2 \times log^2(\frac{n}{n_k + 1})}}. \tag{6}$$

综上，本文建立了 910×500 的留言文本特征向量空间模型，算得留言文本各特征项的归一化权重，准备进行下一步文本分类。

1.3 OAO-SVM 分类模型

1.3.1 SVM 分类器原理

支持向量机(*Support Vector Machines SVM*)是最好的分类器，在各种分类任务中被广泛采用[6]。任何分类问题都可以看成二元分类问题。分类器学习的目标是要得到一个能在新的样本上性能最好的分类器。训练样本中，可能有很多的线性分类器能够划分这些样本，但是只有一个分类器能够最大化两类样本与分类超平面之间的边距，这个线性分类器被称为最优分割超平面。一般 SVM 算法结构为：

设 $D=\{(x_k,y_k)\,|\,k=1,2,\cdots n\}$，其中 $x_k\in R^n$ 为输入数据，$y_k\in R$ 为输出集。在权 w 空间中支持向量机问题可以描述为：

$$y_k=f(x_k)=\omega^T\varphi(x_k)+b,k=1,2,\cdots,n, \tag{7}$$

式中，$\varphi(*)$ 表示 $R\to R^m$ 的映射，将输入空间映射为高维特征空间函数；ω 为超平面的权值向量，$\omega\in R^n$，b 为偏置量。通过最小正则化风险泛函获得 ω 的解：

$$R(\omega)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}L(f(x_i),y_i), \tag{8}$$

式中 C 为惩罚函数系数，$L(*)$ 表示损失函数。最小化 ω，带入 $R(\omega)$：

$$f(x)=\sum_{i=1}^{N}(\alpha_i-\alpha_i^*)<\varphi(x_i),\varphi(x)>, \tag{9}$$

其中 $<\varphi(x_i),\varphi(x)>=K(x_i,x)$ 称作核函数，它是对称的正实函数，满足径向基核函数：$K(x_i,x)=exp[-|x-x_i|^2/2\sigma^2]$。模型中的惩罚系数 C 和核宽 σ 可通过交叉检验获得。

1.3.2 OAO-SVM 模型建立

分类模型的建立基于对题目留言信息分类的两条假设：

- 假设附件 2 的所有一级标签的划分是正确的；
- 假设附件 1 所给一级标签为网友意见中可能涉及的所有一级标签分类。

一对一 SVM 分类方法(*OVO, one – against – one*)是目前经常使用的多分类方法。对于本题的 n 类分类问题，在任意两样本类间构建一个 SVM 子分类器，共需构造 $n(n-1)/2$ 个二分类器，每个二分类器的生成只需要用到训练样本中的两类数据，实现 N 类问题中两类的区分，组合得到的所有子分类器就构成了一个 OAO 多类 SVM。[8]在训练 i, j 两类样本间的分类器时，需要解决如下最优化问题：

$$\min_{w^{ij},b^{ij},\zeta^{ij}} \frac{1}{2}\| w^{ij} \|^2 + C\sum_{t=1}^{l}\zeta_t^{ij}.$$

$$s.t.\begin{cases} w^{ij}\bullet\varphi(x_t)+b^{ij} \geq 1-\zeta_t^{ij}, y_t = i \\ w^{ij}\bullet\varphi(x_t)+b^{ij} \leq -1+\zeta_t^{ij}, y_t = j. \\ \zeta_t^{ij} \geq 0, \qquad\qquad i, j = 1,2,\cdots,l \end{cases} \tag{11}$$

同上，φ 是将 x_t 映射到高维特征空间的映射函数，C 为惩罚系数。最小化 $\frac{1}{2}\| w^{ij} \|^2$ 即实现最大化两样本类间的分类间隔。由式(11)易得 i, j 两类间分类器的分类函数：

$$f_{ij}(x) = (w^{ij})^T \bullet \varphi(x) + b^{ij} \tag{12}$$

分类器构造完毕，采用“投票法”进行类别的判断。方案为：对待测样本 x（即留言文本），依次用 $n(n-1)/2$ 个分类函数进行判别：

$$\begin{cases} i\text{类得一票，当} f_{ij}(x) \leq 0; \\ j\text{类得一票，当} f_{ij}(x) > 0; \end{cases}$$

投票结束后，根据票数最高的类别确定该留言信息所属的类别。

1.3.3 分类性能评价

分类正确率：

从处理好的留言文本集中分别随机抽取 75%、80%、85%、90%、95%的留言记录作为 OAO-SVM 分类模型的训练数据，以各文本信息特征项权重向量为输入，以留言对应的一级标签为输出，进行训练；将剩余留言记录作为测试数据，预测其一级标签分类，并统计预测的正确率，如表 1 所示：

表 1 不同训练数据选取比例对应分类的正确率

训练集比例	75%	80%	85%	90%	95%
测试集正确率	75.8142%	76.1129%	76.3386%	76.2215%	78.0911%

由表 1 可知，OAO-SVM 模型的分类正确率均位于 75%以上，分类效果较好；当训练数据占比为 85%时，测试集正确率为 76.3386%，高于占比 75%、80%、90%时的测试集正确率，说明训练集占比 85%、测试集占比 15%时模型分类效果较好；当训练集数据占比为 95%时，测试集正确率最高，这可能由于测试集数量较少，测试结果偶然性大。

综上，本文最终选择训练集占比 85%时得到的分类结果。附件五(1)(2)分别是在本节最终确定的划分比例下，附件三被随机划分为训练集(85%)、测试集(15%)；附件六(1)(2)为附件五(1)(2)经过 2.3 文本特征项终选降维后的结果。

分类方法的 $F_1-Score$ 值：

确定训练集占比 85%的情况下，通过计算分类结果的 $F_1-Score$ 值，确定使分类所得 F_1 分数最高的惩罚系数 C 以及 $gamma$ 数值。

F_1 分数（$F_1-Score$）是分类问题的一个衡量指标，它是查准率和查全率的调和平均数，计算公式为：

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

其中 P_i 为类别 r_i 的查准率（精确率）：$P_i=\dfrac{\text{被正确判定属于类别}r_i\text{的文本数量}}{\text{被判定属于类别}r_i\text{的全部文本数量}}$，

R_i 为类别 r_i 的查全率（召回率）：$R_i=\dfrac{\text{被正确判定属于类别}r_i\text{的文本数量}}{\text{实际属于类别}r_i\text{的文本数量}}$。

分别设定惩罚系数为 $C=\{0.70,0.75,0.80,0.85,0.90,0.95,1.00\}$，取 $gamma$ 数值为 $G=\{0.1,2.1,4.1,6.1,8.1,10.1,12.1\}$，求解 $C\times G$ 空间内点对应的 F_1 分数，并作出 F_1 分数分布的三维图：

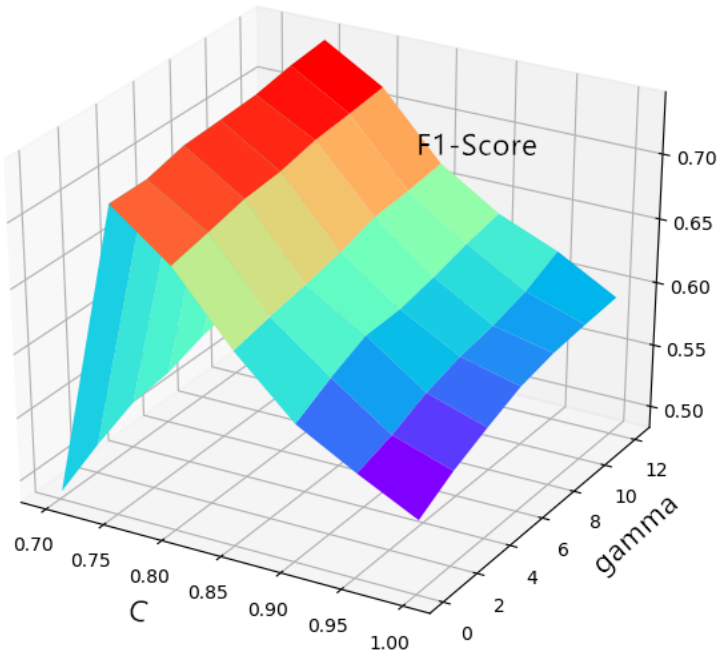

图 1 训练集占比 85%时 $F_1-Score$ 随惩罚系数 C 和 $gamma$ 数值变化三维图

由图 1 可知，随着惩罚系数 C 和 $gamma$ 数值的变化，F_1 分数波动幅度较大。当 F_1 近似取得最大值 0.7417 时有：$C=0.78,\ gamma=2.1$。

二、基于文本聚类的热点问题动态挖掘

附件 3 共包含 4326 条记录，每个记录有 7 个字段，这些字段包括用户留言编号、主题、时间、详情，以及留言收到的反对数和点赞数。浏览附件 3 可知，用户的留言时间范围为 2017 年 6 月 8 日至 2020 年 1 月 8 日。

2.1 文本预处理及基于问题地点的数据集初划分

美国耶鲁大学人工智能研究中心的“快速理解和记忆”（FRUMP）研究小组的研究表明，消息体裁的新闻报道大多数存在标题和导语，标题和导语是全文的中心，信息足够覆盖全文的内容。[9]由于附件 3 中绝大多数留言的标题能够很完整地描述话题发生的地点、人物、内容等主要信息，且若全面考虑留言标题和留言详情，得到的预处理结果往往维数过高、分布稀疏，这不仅会增加计算难度，还会增大话题分类的误差。因此，本文仅针对留言标题的文本进行研究。文本预处理以及基于话题地点的数据集初划分步骤如下：

Step 1：基于正则表达式匹配留言标题文本的地点信息。由题意，热点问题指某时间段内反映特定地点或特定人群的问题，故地点信息是初步划分热点问题的有力方式。本文依靠正则表达式匹配各留言标题的地点（精确至区、县），并根据不同地点将附件 3 划分为不同的数据集。结果显示，附件 3 共被划分为 35 个数据集，各数据集所含记录（即留言数量）的分布情况如表 2：

表 2 各初选数据集所含记录数量分布情况

数据集编号	所含记录数量属于该范围的数据集个数
数据集 1	432
数据集 2	142
数据集 3	678
数据集 4	233
数据集 5	183
数据集 8	82
数据集 9	43
其余数据集	≤10

基于留言话题涉及地点的初划分数据集信息详见附件七。

Step 2：基于地点划分的初划分数据集筛选。若某初选数据集所含记录数量很少，则说明群众对该地点各问题的关注度少，该初选数据集中的留言不会成为热点问题。故本文删除所含记录数量不超过 10 个的初划分数据集，将结果保存至附件八。此步骤共删除了 28 个初选数据集，保留 7 个。

Step 3：对留言标题文本进行分词等系列处理。类比一、1.1 的文本分词过程，基于 PYNLPIR 分词系统，对保留的 7 个数据集中的所有留言标题进行分词，保留名词、动词，去停用词，并删除 7 个数据集所有记录所属的区、县信息。

Step 4：分词后数据集记录按时间顺序排列。将分词后各数据集中的记录按照留言时间的先后顺序排列。

Step 5：文本表示模型选择。为简化模型处理步骤，本文选择布尔模型表示所得文本。布尔模型式向量空间模型的特殊形式，使用特征项向量表示文本信息，特征项的权重只有 1 和 0 两种。

Step 5：基于词频统计的文本特征项筛选。改进的 *Gini* 系数、TF-IDF 确定特征项权重的方法已经不适合当下情景，因此本文按照各特征项的词频多少，进行特征项筛选降维。本文选择频次大于等于 5 次的特征词，共 534 项。保留的特征项结果见附件九；筛选特征项处理后的数据集见附件九。

2.2 基于 Single Pass 算法的热点问题聚类

2.2.1 算法简介

Single Pass 是文本聚类和话题发现常用的一种增量聚类算法，通过比较当前文本和已有类簇的相似度来划分聚类。相比 *K – Means* 聚类算法，*Single Pass* 算法的优势在于：无需实现设定类目数量；各文档只需浏览一次。算法的核心思想如下: 在话题发现系统中，*Single Pass* 算法一次处理一篇文本。初始时将第一篇文本看作一个新的话题，构建它的表示模型，依次处理新到来的下一篇文本，将其与已有的话题模型进行相似度比较。如果相似度最大值小于初始设置的阈值 T，说明其不属于已有的任何一个话题，这是使用该文本创建新的话题模型；否则将该文本聚类到与之相似度最大的话题簇中[10]。*Single Pass* 算法过程简单，聚类速度快，时间复杂度低，效果明确而且可解释性强，适合用于动态发现新问题。

2.2.2 *Single Pass* 算法实现步骤

设 d_i 为第 i 篇文本，p_k 为第 k 类问题类簇，$Sim(p_i, p_j)$ 表示两个问题类簇的类间相似度。则 *Single Pass* 算法实现过程如下：

Step 1: 设定聚类相似度初始阈值 S，将第一篇文本 d_1 作为首个问题类簇。

Step 2: 加入文本 d_i，计算当前文本与已有全部问题类簇的相似度 $Sim(d_i, p_j)$。相似度计算采用类平均距离法，公式如下：

$$Sim(d_i, p_j) = \frac{1}{1 \times |p_j|} \sum_{d_j \in p_j} sim(d_i, d_j), \tag{13}$$

其中 $sim(d_i, d_j)$ 表示文本 d_i 和 d_j 之间的相似度，使用夹角余弦公式进行计算：

$$sim(d_i, d_j) = cos\theta = \frac{\sum_{k=1}^{m} w_{ik} \times w_{jk}}{\sqrt{\sum_{k=1}^{m} (w_{ik})^2 \times \sum_{k=1}^{m} (w_{jk})^2}}, \quad (14)$$

式中 w_{ik} 为文本 d_i 中第 k 个特征项， w_{jk} 为文本 d_j 中第 k 个特征项。

Step 3: 确定文本 d_i 的问题类簇划分

若存在类 p_j 使得 $Sim(d_i, p_j) \geq S$，则有：$d_i \in p_l : Sim(d_i, p_l) = \max_j Sim(d_i, p_j)$；

若对所有已有问题类簇，均有 $Sim(d_i, p_j) < S$，则创建新类簇，文本 d_i 属于该新类簇。

Step 4：所有文本处理完毕，算法过程结束。

以上过程可知，*Single Pass* 算法对文本输入的顺序相对敏感。而群众通过网络问政平台进行留言按照时间先后排列，故在留言问题的聚类中，算法本身带来的局限性不明显。

2.3 热点问题动态评估系统

2.3.1 基本假设

热点问题的评价基于以下假设：

- 假设不存在恶意刷点赞数、反对数的行为。
- 假设针对某条留言的点赞、反对投票同等程度地反映群众对该话题的关注度。
- 仅考虑网络问政平台中群众的关注度，忽略网络问政平台建立早期群众不熟练流程等对热点问题统计的影响。

2.3.2 动态评估系统建立

本文规定问题 p_i 在 t 时段的热度依靠三项热度评价指标：

- pos_num_{it}： t 时段内问题 p_i 的留言总数；
- vot_num_{it}： t 时段内问题 p_i 的关注度。由于本题仅考虑问题的热度，不考虑群众在留言或者点赞反对等行为中的情感因素，所以对每条留言的点赞数、反对数同等看待；群众为一条留言点赞或点击反对，即表示该群众关注该留言反映的问题，所以将点赞、反对行为等同于发布另一条相同问题的留言。综上所述， t 时段内问题 p_i 的关注度的定义式为：

$$vot\_num_{it} \triangleq vot\_for_{it} + vot\_against_{it}, \tag{15}$$

其中 vot_for_{it} 为 t 时段内问题 p_i 的总点赞数；$vot\_against_{it}$. 为 t 时段内问题 p_i 的总反对数。

- $\hbar_{it}$：t 时段内问题 p_i 的问题密度动态系数，其定义式为：

$$\hbar_{it} \triangleq \frac{pos\_num_{it} + vot\_num_{it}}{\max\limits_{t\in T, i\in I}(pos\_num_{it} + vot\_num_{it})}, \tag{16}$$

其中 T 表示所有问题的统计时段，I 表示问题的编号集合。

则问题 p_i 在 t 时段的热度评分为：

$$topic\_hot_{it} \triangleq \hbar_{it} \times (pos\_num_{it} + vot\_num_{it}) \tag{17}$$

综上，问题 p_i 在所统计的总时段内的热度动态评分为：

$$topic\_hot_i = \sum_{t=t\_start}^{t=t\_end} \hbar_{it} \times (pos\_num_{it} + vot\_num_{it}) \tag{18}$$

根据以上过程，计算 2.2.2 所归纳的各问题在总时段（即起始时段分别 2017 年 6 月和 2020 年 1 月）内的热度动态评分，其中时段 t 按月份计；得到热度动态评分排名前五的问题，归纳至附件十一，并将各热点问题包含的留言详情归入附件十二。

表 3 五个热点问题归纳

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	29.2	2019/7/28 至 2019/12/4	A 市 A5 区魅力之城小区	小区油烟扰民
2	2	21.88	2019/1/1 至 2019/7/8	A 市 A4 区	58 车贷案问题
3	3	18.28	2017/6/28 至 2019/11/27	A 市经济学院学生	学校强制学生去定点企业实习
4	4	11.23	2019/5/5 至 2019/9/19	五矿万境 K9 县	房屋存在质量问题
5	5	9.02	2019/4/11 至 2019/4/11	A 市金毛湾	配套入学的问题

三、政府答复质量综合评价

3.1 问题意义

网络问政平台的产生为政府了解民意、联系民众、回应民众需求提供了新思路，也提出了新的要求、带来了新挑战。如果问政平台的答复质量监督不到位，致使答复拖延甚至缺失、答复内容笼统或晦涩等问题频发，那么网络问政平台将成为一纸空谈。因此，建立网络问政平台政府答复质量的评价、监督体系，有助于促进网络问政良性健康发展、提高政府公信度、保障人民权益、推进民主政治生活建设。

3.2 分析原始数据集

题目附件 4 共包含 2816 条记录，给出了 2012 年 4 月 2 日至 2020 年 1 月 8 日中群众通过网络问政平台发送的留言信息，以及政府有关部门所做答复的意见和时间，答复时间范围为 2011 年 11 月 14 日至 2018 年 1 月 9 日。

求解附件 4 中各答复时间间隔，见附件十三。

3.3 答复质量评价模型建立

下面综合考虑答复的相关性、完整性、时效性、可解释性、可行性方面，建立答复质量的综合评价模型。

选取答复相关性、完整性、时效性、可解释性、可行性五个指标分别为$\mathbb{Q}_k$，权重分别为$\xi_k(k=1,2,3,4)$，定义每一条政府答复的评分为：

$$Score_i \triangleq \sum_{k=1}^{4} \xi_k \bullet \mathbb{Q}_k, \tag{19}$$

其中，$\mathbb{Q}_k \in \{1,2,3,4,5\}$即表示各指标得分为 1 至 5 分，低分表示该答复在该指标方面质量较低，高分则相反；五个指标的权重ξ_k满足：$\sum_{k=1}^{5} \xi_k = 1.$

网络问政平台中，群众反映的问题涉及生活中的方方面面，诉求不同，留言实际情况不同，对答复质量的评价方式也有差别，这反映在公式(19)中为：各指标权重ξ_k的数值分配不同。

咨询式留言：群众通过网络问政平台表达对某项政策的疑惑、对某领域规定的详细意义时，此时对留言答复的质量评价应更看重完整性、时效性、可解释性。

政府需要在较短的时间间隔内，为群众详细解答疑惑。

问责式留言：群众亲历或了解到执政过程中的弊端问题，急求政府做出合理的回应、解决办法，这类留言往往带有气愤的情感色彩。对于这类留言，政府相关部门不应一味追求缩短答复时间间隔，而是要落实责任，切实联系有关部门解决群众留言中反映的问题，并将具体的解决措施回复给该留言者。此时对留言答复的质量评价注重完整性、可解释性、可行性，三者的权重应调高。

建议式留言：群众向政府提出建设意见，答复的五个指标重要程度无明显差别，可均设为 0.2。

无论针对哪一类留言，都应重视答复内容与留言内容的相关性；对各类留言的答复时间限制应有具体的规定，咨询式留言和建议式留言的答复时间间隔应最多 7 天，问责式留言的答复时间间隔应控制在 14 天之内。

四、总结与展望

4.1 模型评价

在留言一级标签分类模型的建立中，本文在分类之前对特征项进行了两次降维，精简了文本的特征向量空间；借助分类正确率与 $F_1 - Score$ 值，实现了对分类模型的双重优化、评判。

在热点问题的发掘中，本文创新型地建立时间密度的动态因子，对问题热度进行打分并最终做出累加值，这种动态评价方法既考虑到整体问题热度水平，由避免了如专家打分法中出现的结果受赋分情况影响较大的问题。

然而，针对第三问，由于本文缺少充足的语义库等资料，所建立的模型较简单，由有较大提升空间。

4.2 未来展望

经过对数据的处理和分析，我们已经能够对留言分类，采集热点问题和评论答复意见，但问政工作也不能完全依靠数据来进行分析，还需要政府制度的支持和群众的参与。

政府制度方面可以建立完整的留言质量评判机制，通过设定答复时限、设定回复打分环节和追问、追评功能；强制答复率达到 95%及以上、按月记答复采纳率达到 90%及以上，并不断提升百分比。群众方面我们建议群众在留言标题就阐明事件的主要信息。

同时我们建议建立双向评价机制，根据留言的具体性、答复的时效性、完整性，群众与政府部门双向评价，并根据用户留言的挖掘出的有效信息，提高或降低留言权重，动态展示更有意义的留言信息；通过考察答复质量，鞭策刺激政府

部门提高网络问政效率，

在这里我们也呼吁大家理性留言，政府有权要求留言实名制，防止群众被欲谋不轨的留言者蛊惑。

我们相信通过政府制度的支持和人民群众的积极参与，我们可以提高网络问政效率，促进网络问政走向制度化道路。

参考文献

[1]费洪晓，康松林，朱小娟，等. 基于词频统计的中文分词的研究[J].计算机工程与应用,2005, 41(07): 68-70.

[2]朱美玲. 数据挖掘中的文本分类研究[D].燕山大学,2018.

[3]吴强. 基于统计与语法分析的关键词提取[D].辽宁科技大学,2012.。

[4]姚清耘,刘功申,李翔.基于向量空间模型的文本聚类算法[J].计算机工程,2008(18):39-41+44.

[5]Wenqian Shang,Houkuan Huang,Haibin Zhu,Yongmin Lin,Youli Qu,Zhihai Wang. A novel feature selection algorithm for text categorization[J]. Expert Systems With Applications,2006,33(1).

[6]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.

[7]张龙. 基于粗糙集和神经网络的中文文本分类研究与实现[D].西北大学,2008.

[8]周涛丽. 基于支持向量机的多分类方法研究[D].电子科技大学,2015.

[9]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.

[10]党燕,许志伟,刘利民,王宇,赵思远.基于 Single-Pass 算法的网络舆情文本增量聚类算法研究[J].内蒙古工业大学学报(自然科学版),2017,36(05):364-372.