摘要

本文在旨建立基于自然语言处理技术，通过对网络问政平台的群众留言进行分类处理，达到提升政府的管理水平和施政效率的目的。

针对任务一需要根据所给的附件1的一级标签分类对附件2给的留言内容建立一级标签分类模型，并使用 F-Score 对分类方法进行评价。由于分类方法众多，多分类问题往往无法保证准确度，为选出更好的分类方法，本文使用机器学习中文本挖掘常用的 KNN 算法、人工神经网络和随机森林进行监督学习，使用题目提供的 F-Score 评价三种分类方法，最终 F-Score 值最高的随机森林作为问题一的分类模型。

针对任务二，需要将附件 3 中的留言根据某一时间段内反映特定地点或特定人群问题进行归类，并定义热度评价指标，根据热度指标给出排名前 5 的热点问题，最后给出一定格式的归类评价结果。故本文考虑将稀疏主成分与 EM 聚类相结合，先用加权稀疏主成分对留言数据进行降维处理，稀疏主成分能够增加成分的可解释性。后采用了 EM 聚类对留言数据进行分类。构造热度指数，挑选热度排名前 5 的留言类。

针对任务三，本文从回复相关性、完整性、和可解释性来建立回复质量评价模型。相关性用文本相似度来计算；完整性选择了文本长度因子和文本内容因子来衡量计算；可解释性选择选择一元回归模型，选取留言中问题关键词频数作为自变量，回复中对应问题关键词频数作为因变量，用回归拟合 R 方来衡量可解释性因子。最后对三个因子进行加权计算出回复质量分数。

关键词：随机森林；稀疏主成分；EM 聚类；热度评估；回归分析

目录

一、问题分析

近年来，网络成为了政治活动的重要场地和媒介，社会公共事务交流与传播愈发网络化。同时，海量的文本信息使得政府管理者全面了解用户信息、真正解决公众所关心的问题变得愈发困难。故建立基于自然语言处理技术的智慧政务系统将会对政府管理和施政起极大作用。

问题给出了四个附件数据，附件 1 是留言内容分类的三级标签体系。附件 2 给出了部分留言数据，包括留言编号、留言用户、留言主题、留言时间、留言详情及一级分类这 6 个指标。附件 3 是包括留言编号、留言用户、留言主题、留言时间、留言详情、反对数及点赞数 7 指标的群众留言数据。附件 4 的重点是各条留言的答复意见，其含有留言编号、留言用户、留言主题、留言时间、留言详情、答复意见及答复时间 7 个指标。以下为需要我们解决的问题：

（1）任务一是根据所给的附件 1 的一级标签分类对附件 2 给的留言内容建立一级标签分类模型，并使用 F-Score 对分类方法进行评价。

（2）任务二是将附件 3 中的留言根据某一时间段内反映特定地点或特定人群问题进行归类，并定义热度评价指标，根据热度指标给出排名前 5 的热点问题，最后给出一定格式的归类评价结果。

（3）任务三是从相关性、完整性、可解释性等角度，设计一套评价方案，针对附件 4 有关部门对留言的答复意见的质量进行分析，并尝试实现此方案。

二、模型假设

为便于问题的分析研究，本文做出如下假设：

（1）假设问题提供数据真实可靠。

（2）假设相关部门的答复意见没有情感倾向。

三、任务一：群众留言分类

3.1 数据处理

目前，处理网络问政平台的群众留言，依靠人工根据经验处理大部分电子政务系统，存在工作量大、效率低，且差错率高等问题。因此，我们采用基于 R 软件进行文本挖掘、语义分析和文本分类来代替人工操作，以便后续将群众留言分派至相应职能部门处理，提高工作效率和质量。

分析问题一文本分类中可能存在的问题：“留言主题”信息太精简，而“留言详情”数据又过于冗杂；分类数目较多，而分类方法众多，多分类问题往往无法保证准确度；附件 2 中每种分类数目都不相同，存在数据不平衡问题。

针对这些问题为进行接下来的研究我们选择对附件 2 的留言详情处理：

（1）为使分词结果更准确合理，我们使用了搜狗细胞词库中的人民爆光网作为本文的分词词库，并使用 R 软件中 tmcn 包携带的停词字典进行停词处理，其包括语气词，常见的没有实际意义的词和一些标点符号等；

（2）针对文本语义带来的词语交叉问题，我们分析交叉的词语是作为修饰成分存在句中，通过去掉分词中的形容词和“的”前的词语来去掉该类修饰成分，同时也不会对句子主干造成影响；

（3）有些长文本的留言的无意义表达太多，我们将所有留言合并，分词后选取取出现频数大于 100 的词语作为考察的关键词，构建文本语料库及文档-词频矩阵。由于本文数据庞大，矩阵的阶数巨大且稀疏，故将其进行归一化作为本文主题分类的输入；

（4）由于每一个一级分类包含的样本量不同，为解决数据不平衡的问题，我们采用有放回的随机抽样从每一类别中抽取相同的样本量作为我们的训练集来解决该问题。

3.2 分类方法的介绍

由于分类方法众多，多分类问题往往无法保证准确度，为选出更好的分类方法，本文使用机器学习中文本挖掘常用的 KNN 算法、人工神经网络和随机森林进行监督学习，使用 F-Score 评价三种分类方法，最终使用 F-Score 值最高的分

类方法构建问题一的分类模型。

3.2.1 KNN 算法

KNN (k-nearest neighbor)[1]是一个简单而经典的机器学习分类算法，其基本思路是通过度量“待分类数据”和“类别已知的样本”的距离对样本进行分类。

本文使用 R 软件中 kknn()函数实现 KNN 算法,其具体实现的步骤为：

1) 对于训练样本集即文档-词频矩阵M,其共有 7 种类别即 7 个一级分类，$k_1, k_2, \ldots, k_7$,则样本数据集可表示为

$M=\{(x_1, y_1), (x_2, y_1), \ldots, (x_N, y_N)\}$ 其中 $x_i \in X \subseteq R^n$ 为实例的文本词频向量； $y_i \in C = \{k_1, k_2, k_3, \ldots, k_7\}$ 为实例类别。

2) 根据给定的距离的定义，计算待测样本集T中的向量t与训练样本集 M 中的向量$x \in X$的距离，本文距离采用余弦相似度作为度量:

$$sim(t, x) = \frac{t \bullet x}{|t| \times |x|}$$

3) 选取和待测样本距离最近的K份样本，其中所属最多的分类便为待测样本的所属类别。

3.2.2 人工神经网络

人工神经网络(artificial neural networks)是对自然神经网络的模仿，是最早的机器学习方法之一,它可以有效地解决很复杂的有大量互相相关变量的分类问题。

ＡＮＮ模型由输入层、隐藏层以及输出层组成，每层网络又包含若干数量的神经元。其中隐藏层的层数属于模型的超参数，可以人为设定。本文使用 R 软件中 nnet()函数进行人工神经网络的训练，由于训练样本较大，本文将隐藏层层数设置为１０层时，训练误差才足够小。故模型选定为 10 层隐藏层层数的人工神经网络模型。

3.2.3 随机森林

随机森林算法是基于 Bagging 算法发展起来的，Bagging 算法是从所有文本中重采样出 n 个文本，对其构建分类，然后重复 m 次此过程获得 m 个分类器，最后根据这 m 个分类器的投票结果决定文本属于哪一类。森林算法的基本构成单元是充分生长、没有剪枝的决策树，分类和回归问题可以用其解决。随机算法流程包括生成随机森林和进行决策两部分。随机森林不需要交叉验证且其不会导致过拟合[2]。随机森林算法流程如图 3.1 所示。

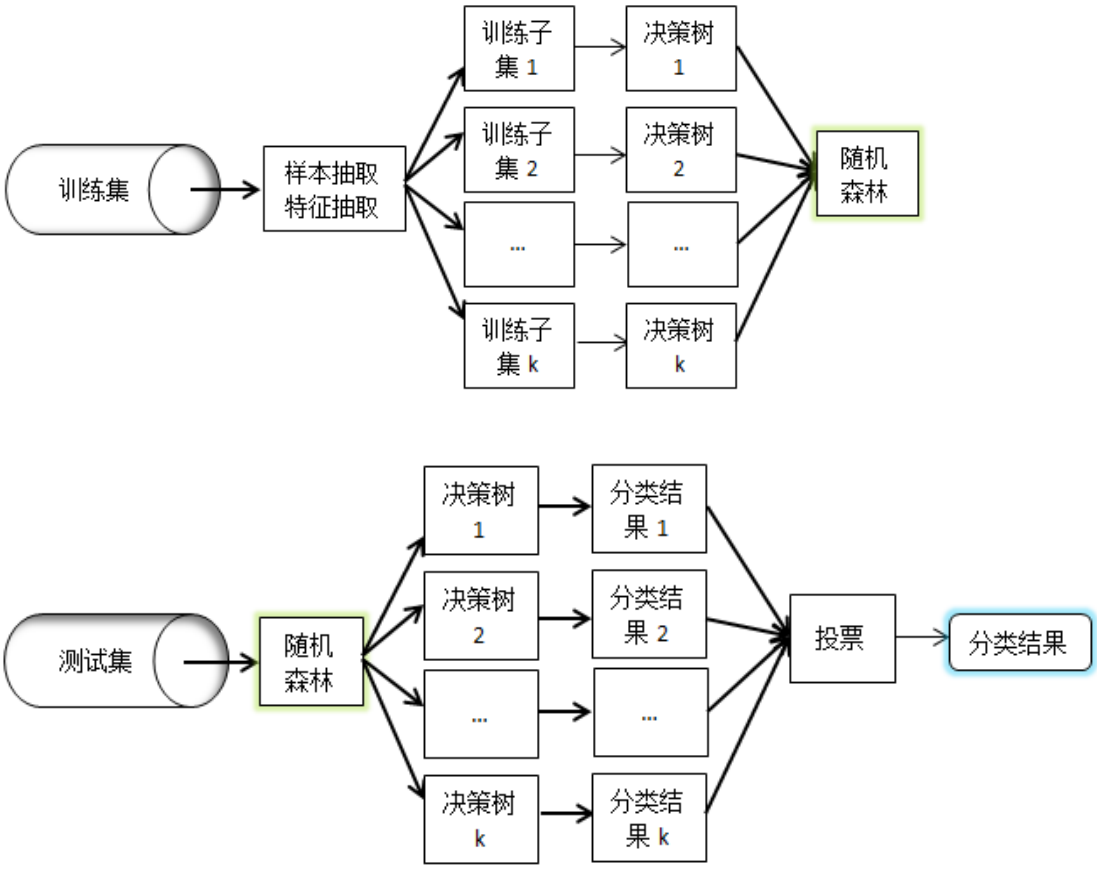

图 3.1 随机森林算法流程

本文使用 R 软件中 randomForest()函数构建随机森林模型，其具体的算法步骤如下：

(1) N 为原始训练集中的样本数量，M 为特征属性数量。采用 bootstrap 抽样方法从原训练集中随机选取 n 个样本形成训练子集。

(2)从原始训练集中的 M 个特征属性中抽取 m 个特征作为候选特征($m \leq M$)，在决策树的每个节点按照信息增益选择最优属性进行分裂，直到该节点的所有训练样例都属于同一类，过程中完全分裂不剪枝。

(3)重复上述两个步骤 k 次,构建 k 棵决策树,生成随机森林。

(4)使用随机森林进行决策,设 x 代表测试样本，h_i 代表单棵决策树，Y 代表输出变量即分类标签,I 为指示性函数,H 为随机森林模型,决策公式为:

$$H(\mathrm{x}) = \arg\max_{Y} \sum_{\mathrm{i}=1}^{k} I(h_i(\mathrm{x}) = Y)$$

汇总每棵决策树对测试样本的分类结果,得票数最多的类为最后的分类结果。

3.3 结果展示

分类模型的好坏一般从准确度和效率两个指标来判断，目前衡量准确度的优劣一般用查准率(Precision)，查全率(Recall)，F-Score 值三种评估指标。根据三种分类模型分类结果计算出平均查准率、平均查全率及 F-Score 值如表 3.2 所示：

表 3.1 模型评价表

分类方法	平均查准率	平均查全率	F-Score
KNN	0.736887957	0.756747843	0.7447912
ANN	0.845495886	0.870397929	0.8558229
随机森林	0.856868786	0.882718271	0.8665935

由表 3.1 可知随机森林分类的 F-Score 值最大，达到了 86.7%，且其平均准率、平均查全率均比 KNN、ANN 模型大，说明了三种方法中随机森林最优、分类效果最好，故分文任务一选定随机森林作为我们的分类模型。

四、任务二：热点问题挖掘

4.1 数据处理

任务二需要我们找出热点问题，即从主要表达意思入手，所以我们选择将附件 3 中留言主题内容进行分词、处理。

（1）为使分词结果更准确合理，我们使用了搜狗细胞词库中的人民爆光网

作为本文的分词词库，并使用 R 软件中 tmcn 包携带的停词字典进行停词处理，其包括语气词，常见的没有实际意义的词和一些标点符号等；

（2）将所有留言合并，分词后选取取出现频数大于 20 的词语作为考察的关键词，构建文本语料库及文档-词频矩阵 M。

4.2 留言归类

任务二中需要分类的留言共 4326 条，由于数据量较大，文本语料库及文档-词频矩阵 M 维度过高，为能够更准确地提取主要信息、使归类计算过程更简便，本文考虑将稀疏主成分与 EM 聚类相结合。具体步骤如下：

（1）为消除不同尺度带来的影响将文本语料库及文档-词频矩阵 M 进行标准化；

（2）KMO 检验和 Bartlett 球形检验

（3）主成分分析，确定提取主成分数量，查看主成分分析结果；

（4）根据确定的主成分数量进行稀疏主成分，由于各主成分对变量的解释性不同，且考虑分类结果各类数目存在差异，故将稀疏主成分根据方差贡献率加权，这也使得数量多的关键词占的比重更大，符合本文聚类目的；

（4）将稀疏后的主成分进行 EM 聚类。

4.2.1 主成分分析（PCA）

对标准化后的矩阵进行 KMO 检验和 Bartlett 球形检验。计算得 KMO 值为 0.5316945，Bartlett 球形检验显著，表明各指标间存在较高相关性，适合进行主成分分析。

计算方差贡献率可得前 137 个主成分对样本方差的累计贡献率为 66.69%，故本文取前 137 个主成分。

PCA 是一种线性无监督降维方法，通过线性转换，将拥有高维特征的原始数据投影到一个低维坐标空间中，得到一组拥有低维特征的新数据。这种算法在可接受的信息损失下，保留了原始数据的主要信息，且算法保证信息之间不相关，降低原始数据的冗余度，降低后续计算的复杂度。PCA 主要的计算步骤如下。

1）计算样本 Z 的协方差矩阵

$$S=\sum_{i=1}^{N}\left(X_{\mathrm{i}}-\bar{X}\right)\left(X_{\mathrm{i}}-\bar{X}\right)^{T}$$

式中 $\bar{X}$ 为总体样本均值。

2）根据协方差矩阵求特征向量 $u_1,u_2,...,u_N$ 和对应的特征值 λ_1，$\lambda_2,...,\lambda_N$。

3）根据降序排列求得的特征值并确定主成分。主成分的个数的确定，由累计贡献率 $D(m)$ 来确定的

$$D(m)=\sum_{i=1}^{\mathrm{m}}\lambda_i \Big/ \sum_{i=1}^{\mathrm{M}}\lambda_i$$

（a）主成分分析优点：

①利用降维技术用少数几个综合变量来代替原始多个变量，这些综合变量集中了原始变量的大部分信息。

②通过计算综合主成分函数得分，对客观经济现象进行科学评价。

③在应用上侧重于信息贡献影响力综合评价。

（b）主成分分析缺点：

①当主成分的因子负荷的符号有正有负时，综合评价函数意义就不明确。

②命名清晰性低。

基于主成分分析的缺点，为增强主成分对个关键词的解释性，使聚类效果更清晰，接下来将提取出来的主成分稀疏化处理。

4.2.2 稀疏主成分（SPCA）

将原来的主成分部分协方差阵 S 做参数变换: $A=QD$, 其中 Q 是列正交矩阵，且根据 A'A 的特征值分解来求得。这样将原来待求稀疏化 A 的问题转换成求解稀疏化 Q 的问题:

$$\underset{Q,D,\varphi}{\mathbf{argmin}}\left\{\left[\boldsymbol{log}\left|\boldsymbol{QD^2Q'}+\boldsymbol{\varphi}^2\right|+\boldsymbol{tr}\left(\boldsymbol{QD^2Q'}+\boldsymbol{\varphi}^2\right)^{-1}\boldsymbol{S}\right]+\boldsymbol{\mu P_\tau}\left(\boldsymbol{Q}\right)^{T}\boldsymbol{P_\tau}\left(\boldsymbol{Q}\right)\right\} \qquad (3)$$

其中罚项具体表示成：

$$\boldsymbol{P}_{\tau}\left(\boldsymbol{Q}\right)=\frac{\boldsymbol{q}_{\tau}\left(\boldsymbol{l}_{m}+\boldsymbol{sgn}\left(\boldsymbol{q}_{\tau}\right)\right)}{2}$$

这里的 $q_{\tau}=q_{11},\ldots,q_{m1}-\tau$，$q_i$ 表示矩阵 Q 的第 i 列。即按列对矩阵 Q 进行稀疏化。

（a）稀疏主成分优点：

①当变量特别多时，能对主成分进行合理的解释，增加主成分的可解释性。

（b）缺点：

①惩罚函数包含有稀疏度的信息，但不易找到最优惩罚函数。

4.2.3 加权稀疏主成分

WSPCA 是对 SPCA 在计算过程中进行加权处理，即引入权重 w_i，$i=1,2,...,p$，其中 p 为主成分个数。对稀疏化后的结果进行加权，得

$$S=\sum_{i=1}^{N}\sum_{\substack{i=1\\ j\neq i}}^{N}\mathrm{w}_{\mathrm{i}}\left(X_{\mathrm{i}}-\bar{X}\right)\left(X_{\mathrm{i}}-\bar{X}\right)^{T}\quad(4)$$

式中 $\bar{X}$ 为样本总均值，w_i 为权函数，本次研究取权重为主成分方差贡献率。

4.2.4 EM 聚类

实际的统计聚类算法通常以迭代求精优化过程为中心来计算局部最优聚类解，从而最大限度地适应数据。这些算法通常需要许多数据库扫描才能收敛，在每次扫描中，它们都需要访问数据表中的每条记录。对于大型数据库，扫描复杂得令人望而却步。

对于期望最大化（EM）算法来说，数据库社区已经将重点放在基于距离的聚类方案上，并且已经开发了对数值数据或分类数据进行聚类的方法。

与基于距离的算法例如 K-mean 不同，EM 为底层数据源构造了适当的统计模型，并自然地推广到包含离散值和连续值数据的聚类数据库。EM 算法与 k-means 聚类分析算法相比权，它有多个优点：

①该方法在有限的主内存缓冲区的范围内工作，最多需要一次数据库扫描。

②工作时不受内存(RAM)限制。

③能够使用只进游标。
④性能优于基于抽样的方法—即将传统内存中的实现“缩放”到大型数据库的直接替代方案。

4.3 留言热度评估

经过对留言按照“特定地点或特定人群问题”进行归类，每个“地点/人群”代表一个主题。对于网络问政平台的群众留言，参照麦林[4]提出的话题热度评估，这里考虑主题的热度取决于三个因素：关注度、认可度及时效性。

一、主题关注度

影响主题关注度的参数主要有留言居民数及留言数量。这里需区别居民数及留言数量：在所有留言中，可能存在某个居民多次进行留言的现象，而居民数只对该居民统计一次。

分别用 r, n 表示留言的居民数和留言数量，则主题关注度为：

$$Att = 0.6 \times r + 0.4 \times n$$

二、主题认可度

评价一条留言是否有热度，居民对此的态度至关重要，对此，引入主题认可度。主题认可度是指针对某位留言居民，其他居民对其留言内容的态度，包括支持和反对，分别为附件 3 中的点赞数及反对数。用 rec 表示主题认可度，用 sup 和 opp 分别表示点赞数和反对数，则主题认可度为：

$$rec = 0.8 \times sup + 0.2 \times opp$$

三、时效性

时效性是留言的一个重要因素，它决定了居民所反映的问题能否得到及时解决。留言的发布时间可以基本反映居民亟待解决问题的时间。一般来说，主题一旦发布，在短时间内受到的关注越多，该主题热度越大。但随着时间推移，主题热度会慢慢降低。因此，主题发布的时间可衡量该主题的热度。

将当前时间设为 t 为当前时间，t_0 为主题发布时间，则主题的时效性为：

$$Time = e^{-(t-t_0)}$$

其中，$t-t_0$ 为当前时间与主题发布时间的时间差。时间差越大，时效性则呈

现衰减的趋势。

4.3.1 主题热度

对以上计算出来的三个指标归一化处理，在计算主题热度之前，发现关注度和认可度这两个指标数量级相差较大，为使指标处于同一数量级，对其做归一化处理，适合进行综合对比评价。

由关注度、认可度和时效性，得到最终的主题热度评估公式为：

$$H_i = \left(Att + \mathrm{Rec}\right) \times Time$$

$$= \left(0.6 \times \mathrm{r} + 0.4 \times \mathrm{n} + 0.8 \times \mathrm{sup} + 0.2 \times opp\right) \times \mathrm{e}^{-(t-t_0)}$$

4.4 结果展示

5.5.1 热点问题留言分类

经过数据的处理，得到热点问题留言分类为 82 类，部分结果如下图：

63	190523	A00072847	A市丽发新城违建搅拌站，彻夜施工扰民污染环境	2019/12/26 13:55:15	严重；3、搅拌站几百米外	0	0
63	191943	A00038563	A市A2区丽发新城道路坑坑洼洼	2019/7/3 12:03:51	的道路坑坑洼洼，下雨泥	0	1
63	199379	A00092242	A2区丽发新城附近修建搅拌厂，严重污染环境	2019/11/25 10:17:56	因此还得了疾病住院，该	0	0
63	203393	A00053065	A市丽发新城小区侧面建设混泥土搅拌站，粉尘和噪音污染严重	2019/11/19 14:51:53	巨大的粉尘，严重影响居民	0	2
63	208714	A00042015	A2区丽发新城附近修建搅拌站，污染环境，影响生活	##################	质量和声环境质量急剧下降	0	4
63	213930	A909218	A2区丽发新城附近违规乱建混凝土搅拌站谁来监管？	##################	民强烈呼吁政府和有关职能	0	0
63	214282	A909209	A市丽发新城小区附近搅拌站噪音扰民和污染环境	##################	天天吵，烦死了不仅吵还臭	0	0
63	215563	A909231	A2区丽发新城小区旁边的搅拌厂是否合法经营	##################	生了噪音和灰尘。这给小	0	0
63	215842	A909210	A2区丽发新城小区附近太吵了	##################	个搅拌厂是怎么回事！下班	0	0
63	219174	A00081998	A2区丽发新城小区内垃圾站散发严重臭味	2019/7/3 23:27:02	有时候吃饭睡觉都能闻到	0	3
63	222831	A909228	噪音、灰尘污染的A2区丽发新城附近环保部门不作为	##################	成里能修改产生大量灰尘的	0	0
63	225217	A909223	A2区丽发新城附近修建搅拌厂严重影响睡眠	##################	建一搅拌站,每天尘土飞扬	0	0
63	231136	A909204	投诉A2区丽发新城附近建搅拌站噪音扰民	##################	。距离上次投诉已经过去	0	0
63	239336	A909213	A市A2区丽发新城小区遭搅拌站严重污染	##################	夜日运作，离居民区非常近	0	0
63	239648	A909211	A市A2区丽发新城小区附近搅拌站明目张胆污染环境	##################	尘颗粒！都不敢开窗透气	0	0
63	253040	A909202	投诉A2区丽发新城附近建搅拌站噪音扰民	##################	家里根本无法正常休息！	0	0
63	264944	A0004260	A2区丽发新城附近修建搅拌厂噪音、灰尘污染	##################	之修建搅拌厂，请问环保	0	0
63	267050	A909227	噪音、灰尘污染的A2区丽发新城附近已扰乱居民生活	##################	区居民不能正常休息，灰	0	0
63	268109	A909230	我要举报A市A2区丽发新城小区开发商违规建设搅拌站	##################	生的噪音污染小区居民生	0	0
63	268300	A909225	A2区丽发新城附近修建搅拌厂噪音污染导致生活不正常	##################	搅拌厂，请问是谁审批的	0	0
63	273282	A909226	A2区丽发新城附近修建搅拌厂烟尘滚滚，声音刺耳	##################	问政府，我们该怎么生活	0	0

图 4.1 部分分类结果

按照 2.4 给出的主题热度评估公式，计算各个分类的热度指数，结果如下图所示：

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	63	0.003919163	2019/6/19至2020/1/26	A2区丽发新城	小区附近搅拌站噪音扰民和污染环境
2	79	4.07331E-11	2019/1/23至2020/1/6	A市高新区	道路上长年泥土，工地通宵施工扰民
3	53	2.08418E-11	2019/1/9至2020/1/8	A3区梅溪湖街道区	街门面油烟扰民，业主私搭乱建
4	14	1.8526E-11	2019/1/4至2020/1/8	A7县北山镇	条件恶劣，出行困难，村民之间起矛盾
5	14	1.78257E-11	2018/10/27至2020/1/7	A市地铁	线路开工及建设问题

图 4.2 热度排名前五结果

由图 4.2 可知，热度排名第 1 的问题是 A2 区丽发新城小区附近搅拌站噪音扰民和污染环境；问题处于第 2 的位置，排名第 3 的是 A 市高新区道路上长年泥土，工地通宵施工扰民这一问题；A3 区梅溪湖街道区街门面油烟扰民，业主私搭乱建对居民也造成了比较严重的影响，排名第 4；排名第 5 的是 A7 县北山镇，该镇问题是条件恶劣，出行困难，村民之间起矛盾。

五、任务三：答复意见的评价

5.1 答复意见的质量评价方案

5.1.1 回复与留言的相关度

我们使用用户留言与有关部门回复内容的文本相似度来表示两者间的相关度。具体步骤为：

首先将文档用向量空间模型表示，即将文档 d 表示为 $d(t_1 \ w_1 t; {}_2 w \ ,\cdot_2 \ t_N; \ w_N)$，其中，$N$ 为特征项的总数，t_k 是第 k 个特征项，w_k 为特征项 t_k 的频率。在这里我们使用的是用户留言和相关部门回复文本分词结果的并集，这样可抓住文本主要信息，并且省略了去除各自的特征词的步骤。

然后通过计算两个文档向量之间的文本距离得到文本相似度。这里用向量间的夹角余弦值来计算留言文档和回复文档之间的文本相关度：

$$S_i = \frac{\sum_{k=1}^{N} \mathrm{w}_{1k} \times \mathrm{w}_{2k}}{\sqrt{\sum_{k=1}^{N} w_{1k}^2 \times \sum_{\mathrm{i}=1}^{N} w_{2k}^2}}, i = 1, 2..., m$$

其中 m 为用户留言或相关部门回复总条数。

5.1.2 回复的完整性度量

我们使用留言文本长度和回复文本长度对比结果和回复意见关键词提取来度量相关部门回复的完整性，具体步骤如下：

一、回复文本长度因子

考虑到回复文本的长度对完整性因子有影响，于是对留言和回复的文本长度进行对比分析。对留言文本和回复文本进行分词预处理，得到分词结果 $S1$ 和 $S2$，第 i 条留言和回复的分词结果的词汇总数分别为 $S1_{\mathrm{i}}$ 和 $S2_{\mathrm{i}}$，$\mathrm{i} = 1, 2, ..., m$，记 L 为回复文本长度变量，则有

$$L_i = \begin{cases} 1, \dfrac{S2_{\mathrm{i}}}{S1_{\mathrm{i}}} \geq 1 \\ 0.75, 0.5 \leq \dfrac{S2_{\mathrm{i}}}{S1_{\mathrm{i}}} \prec 1 \\ 0.375, 0.25 \leq \dfrac{S2_{\mathrm{i}}}{S1_{\mathrm{i}}} \prec 0.5 \\ 0, 0 \leq \dfrac{S2_{\mathrm{i}}}{S1_{\mathrm{i}}} \prec 0.25 \end{cases} \quad \mathrm{i} = 1, 2, ..., m$$

二、回复文本内容因子

分析回复文档的内容，我们认为一条完整的回复留言应该包括以下内容：

（1）问题处理因子：针对留言问题有展开相关调查；

（2）问题解决因子 N：对问题的解决方法或建议以及意见。

鉴于此分析，我们取关键词如下：若“已经”、“经”、“经过”、“通过”、“经调查”、“经过调查”、“通过调查”等问题处理关键词出现，则 $M_i, i = 1, 2, ..., m$ 取 1，否则取 0；同理，若“可以”、“建议”、“意见”等问题解决关键词出现，则

$N_i, i=1,2,...,m$ 取 1，否则我取 0。由于每条回复都会有“已知悉”或“已阅”等词汇，故将文本中的“已”停词去掉。对相关部门回复文档进行关键词出现情况进行统计，记第 i 条回复文本内容变量为 $C_i, i=1,2,...,m$，m 为回复留言总条数。则有

$$C_i=\begin{cases}0, M_i=0, N_i=0\\ 1, M_i=0, N_i=1 or M_i=1, N_i=0, \quad i=1,2,...,m\\ 2, M_i=1, N_i=1\end{cases}$$

综上所述，记完整性因子为 $Comp_i, i=1,2,...,m$ 可以用如下公式来衡量

$$Comp_i=\alpha L_i+\beta C_i, i=1,2,...,m$$

5.1.3 回复的可解释性

分析相关部门回复的内容，可解释性是指回复文本对留言文本中的相关问题的解释，基于此分析，我们采用一元回归分析来解决该问题，使用回归拟合 R^2 来衡量回复的可解释性。具体步骤如下：

首先对用户留言和相关部门回复文本内容进行分词预处理，提取留言里相关问题的关键词频数作为因变量 Y，提取回复里对应相关问题的关键词频数作为自变量 X，没有出现的问题关键词频数记为 0；然后对 X 和 Y 作一元回归分析，统计回归模型的 R^2 作为相关部门回复的可解释性因子，记可解释性因子为 $A_i, i=1,2,...,m$，则有

$$A_i=R_i^2, i=1,2,...,m$$

5.2 答复意见的质量评价指标

综上所述，我们从相关性、完整性、可解释性等角度，设计一套评价方案，对有关部门对留言的答复意见的质量进行分析，答复意见的质量评价分数 $Score_i, i=1,2,...,m$ 可用回复与留言的相关度因子 $S_i, i=1,2,..,m$，回复的完整性度

量因子 $Comp_i$，$i=1,2,...,m$ 和回复可解释性因子 $A_i, i=1,2,...,m$ 来度量，其中 m 为答复意见的条数，则有

$$Score_i = a_1 \cdot S_i + a_2 \cdot Comp_i + a_3 \cdot A_i, i=1,2,...,m$$

七、参考文献

[1] 李宏志,李苡兰,赵生慧.基于 Spark 的大规模文本 KNN 并行分类算法[J].湖南科技大学学报(自然科学版),2020,35(01):90-97.

[2] 刘希良. 酒店在线评论的情感倾向挖掘方法应用研究[D].广东工业大学,2014.

[3] 刘勇,兴艳云.基于改进随机森林算法的文本分类研究与应用[J].计算机系统应用,2019,28(05):220-225.

[4] 麦林. 虚拟社区热点话题意见挖掘模型研究[D].中国科学技术大学,2009.