“智慧政务”中的文本挖掘应用

摘要

政府的宗旨是以人民服务，所以政府需要及时对群众意见进行分析恢复，并解决问题，数据为工作人员带来很大困难。本文根据 logistic 回归模型，灰色关联分析法，相关系数，最后对数据分析处理，建立了分类模型，对分类进行评价，在及时回应群众的同时也能提高答复效率。根据模型找到最优的解决方案。

针对问题一，留言的一级标签分类模型，利用数据清洗来完成特殊字符等的去除，然后对基于大数据文本来进行一级标签的特征向量的选择与提取，可以利用信息增益算法，然后建立 logistic 模型，最后使用 f-score 模型对留言分类进行评价。

针对问题二，我们小组经过讨论后，先把某一时段内反映特定地点或特定人群的留言进行归类，并通过灰色关联分析法$r_i = \sum_{k=1}^{n} \omega_i \xi_i(k)$对归类的内容进行评价分析。

针对问题三需要对模型进行分析 数据的相关性，完整性， 可解释性 ，所以本文考虑采用相关系数对模型评价，改进 找到最优的方案提高效率。

关键词：Logistic 回归模型 灰色关联分析法 相关系数

一、 问题背景与问题的重述

1.1 问题的背景

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。

附件给出了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。请利用自然语言处理和文本挖掘的方法解决下面的问题。

1.2 问题重述

1.2.1 群众留言问题

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，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。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。

通常使用 F-Score 对分类方法进行评价：

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

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。

1.2.2 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，如“XXX 小区多位业主多次反映

入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。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。请根据附件 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，按表 1 的格式给出

排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题

对应的留言信息，并保存为“热点问题留言明细表.xls”。

表 1–热点问题表

热点排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	…	2019/08/18 至 2019/09/04	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	…	2017/06/08 至 2019/11/22	A 市经济学院学生	学校强制学生去定点企业实习
…	…	…	…	…	…

表 2–热点问题留言明细表

问题 ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	360104	A01247	A 市魅力之城商铺无排烟管道，小区内到处油烟味	2019/08/18 14:44:00	A 市魅力之城小区自打交房入住后，底层商铺无排烟管道，经营餐馆导致大量油烟排入小区内，每天到凌晨还在营业……	0	0
1	360105	A12036	A5 区魅力之城小区一楼被搞成商业门面，噪音扰民严重	2019/08/26 08:33:03	我们是魅力之城小区居民，小区朝北大门两侧的楼栋下面一楼，本来应是架空层，现搞成商业门面，嗓声严重扰民，有很大的油烟味往楼上窜，没办法居住……	1	0

1	360106	A2353 67	A 市魅力之城 小区底层商铺营业到凌晨，各种噪音好痛苦	2019/08/26 01:50:38	2019 年 5 月起，小区楼下商铺越发嚣张，不仅营业到凌晨不休息，各种烧烤、喝酒的噪音严重影响了小区居民休息……	0	0
…	…	…	…	…	…	…	…
1	360109	A0080 252	魅力之城小区底层门店深夜经营，各种噪音扰民	2019/09/04 21:00:18	您好：我是魅力之城小区的业主，小区临街的一楼是商铺，尤其是餐馆夜宵摊等，每到凌晨都还在营业，每到晚上睡觉耳边都充斥着吆喝……	0	0
2	360110	A1100 21	A 市经济学院 寒假过年期间组织学生去工厂工作	2019/11/22 14:42:14	西地省 A 市经济学院寒假过年期间组织学生去工厂工作，过年本该是家人团聚的时光，很多家长一年回来一次，也就过年和自己孩子见一次面，可是这搞……	0	0
2	360111	A1204 455	A 市经济学院 组织学生外出打工合理吗？	2019/11/5 10:31:38	学校组织我们学生在外边打工,在东莞做流水线工作，还要倒白夜班。本来都在学校好好上课，十月底突然说组织到外省打工……	1	0
…	…	…	…	…	…	…	…
2	360114	A0182 491	A 市经济学院 变相强制实习	2017/06/08 17:31:20	系里要求我们在实习前分别去指定的不同公司实训，我这的公司的工作内容和老师之前介绍以及我们专业几乎不对口，不做满 6 个月不给实训分，不能毕业……	9	0

1.2.3 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。

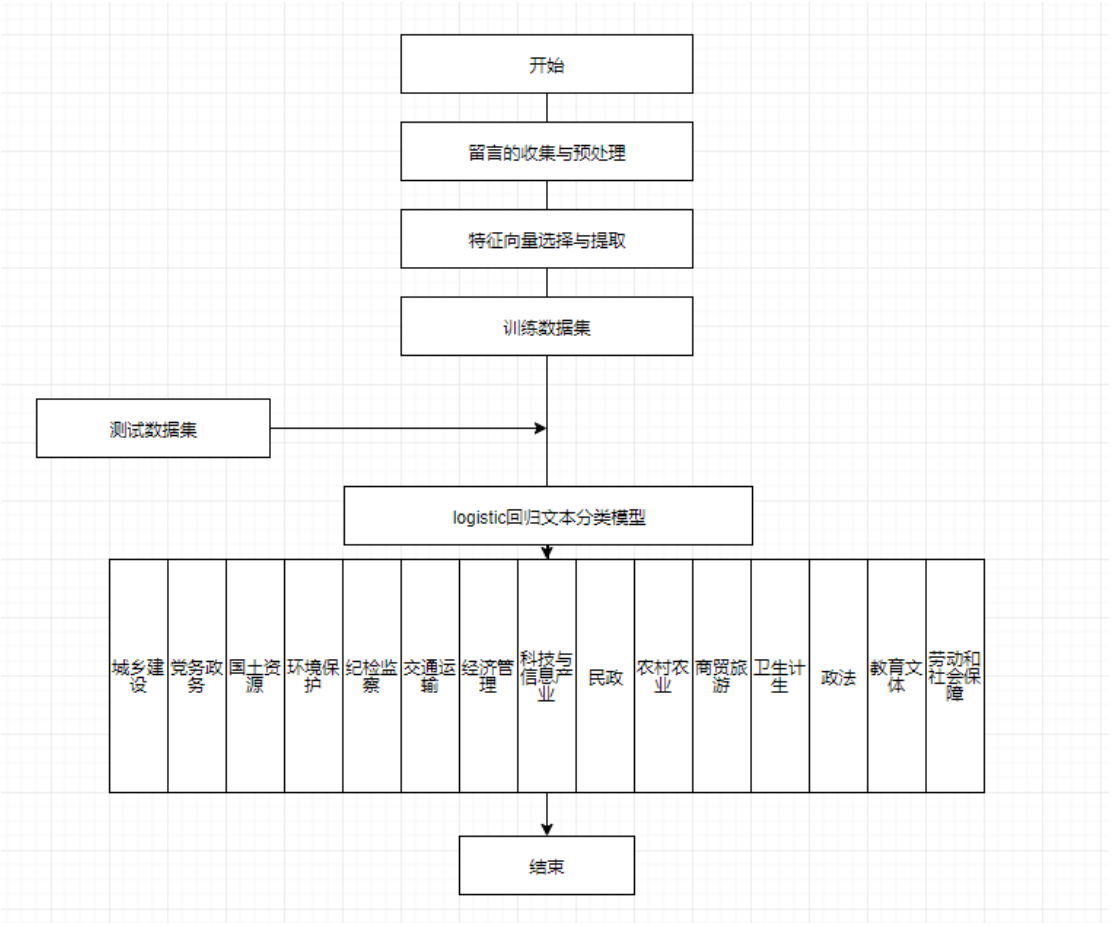
二、 问题分析

2.1 问题一的分析

针对于问题一的分析，我们考虑建立 logistic 回归模型，用 f-score 对分类

方法进行评价

问题解决流程图如下：



2.2 问题二的分析

问题二是让我们根据附件 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，按表 1 的格式给出排名前 5 的热点问题，按表二的格式给出相应热点问题对应的留言信息。首先，我们应该将附件 3 中的内容根据时间段进行划分。其次，再对特定地点或特定人群进行划分，这样就可以将数据分开进行处理。有了某一时间段内反映特定地点或特定人群的归类，我们就可以进行合理的热度评价。

2.3 问题三的分析

题目要求从完整性，相关性，可解释性等角度对答复意见进行评价，是否具有完整性即是数据的正确性，一致性；相关性，两个或两个以上的相关变量的统计分析；可解释性，数据具有可靠性，数据来源要准确；针对这个问题重要的是数据处理，我们采用相关系数的相似度量对相关性进行研究，完整性我们采用 Spearman 相关系数进行分析，可解释性根据其他两个数据分析结果进行综合分析。

三、 模型假设及符号定义与说明

3.1 模型假设

1.留言内容是根据真实情况填写，不存在乱留言现象。

2.留言时间准确。

3.2 符号定义与说明

$\xi_i(k)$：为比较数列x_i对参考数列x_0在第 k 个指标上的关联系数

ρ：分辨系数

$\max_s \max_t |x_0(t)-x_s(t)|$：两级最大差

$\min_s \min_t |x_0(t)-x_s(t)|$:两级最小差

r_{ik}：相似性度量

x_i:答复时间

y_i：随机抽出的 n 个样本

R_i:秩统计量

S_i:秩统计量

X,Y：样本观测数据

q_{xv}：Spearman 相关系数

P(c):所有训练数据集表示为某个一级标签 c 的概率

P(t,c):某个留言包含特征 t 在所有训练数据集的概率

P(t):所有训练数据集中包含特征 t 的留言的概率

μ:位置参数

λ：形状参数

W:标记一级标签的概率

四、 模型的建立与求解

4.1.1 问题一模型的建立与求解

4.1.1.1 留言的收集和预处理

基于群众留言，使用数据清洗技术来去除特殊字符等无关词，筛选有义的数据文本

4.1.1.2 特征向量的选择和提取

根据文本的属性，经过筛选的留言文本类别的特征，相对稳定且有区别的特征，出现的字或词和字或词的频度作为留言的文本特征和特征值

4.1.1.2.1 特征统计

可以利用统计软件进行特征值统计，将一级标签的 15 个类别进行特征值的统计

4.1.1.2.2 特征选择和提取

利用信息增益 IG 算法，基于特征关于一级标签的信息增益的计算公式 $IG(t_k,c_i) = \sum \sum P(t,c)\log \frac{P(t,c)}{P(t)P(c)}$

4.1.1.3 将训练数据集代入计算

4.1.1.4.建立 logistics 回归模型

4.1.1.4.1logistic 分布函数

设 x 是连续随机变量且 x 服从 logistic 分布，则 x 的 logistic 分布函数法 f（x）公式如下

$$F(x) = P(X \le x) = \frac{1}{1+e^{-(x-\mu)/\gamma}}$$

4.1.1.4.2logistic 回归模型应用

Logistic 回归模型的多分类概率如下公式

$$P(W=k|x) = \frac{e^{(w_k \cdot x)}}{1+\sum_{k-1} e^{(w_k \cdot x)}}, k = \{1,2,\ldots,k-1\}, x \in R^n, w_k \in R^n$$

4.1.1.5 测试分析

使用 f-score 对分类方法进行评价

查准率 P、查全率 R、已标记的样本数量 TP、被正确预测属于该一级标签的全部样本数量（TP+FP）、将 FP 定义为分类器把其他一级标签错误预测属于该常用标记一级标签的样本数量、（TP+FN）是该样本实际属于该标记一级标签的数量，FN 表示实际属于标记一级标签但被分类器错误预测属于其他一级标签的样本数量，公式如下

查准率：$P = \frac{TP}{TP+FP}$

查全率：$R = \frac{TP}{TP+FN}$

4.1.2 问题二模型的建立与求解

对留言信息进行量化与处理，然后确定比较对象（评价对象），我们假设评价对象有 m 个，评价指标有 n 个，参考数列为 $x_0 = \{x_0(k)|k=1,2,\ldots,n\}$，比较数列为 $x_i = \{x_i(k)|k=1,2,\ldots,n\}, i = 1,2,\ldots,m$.使用层次分析法确定各指标对应的权重 $\omega = [\omega_1,\ldots,\omega_n]$，其中 $\omega_k (k = 1,2,\ldots,n$ 为第 k 个评价指标指标对应的权重。

$$\xi_i(k) = \frac{\min_s \min_t |x_0(t)-x_s(t)| + \rho \max_s \max_t |x_0(t)-x_s(t)|}{|x_i(k)-x_0(k)| + \rho \max_s \max_t |x_i(t)-x_s(t)|} \quad (6)$$

$$r_i = \sum_{k=1}^{n} \omega_i \xi_i(k) \quad (7)$$

r_i 为第 i 个评价对象的灰色加权关联度，根据灰色加权度的大小，可以对各评价对象进行排序，我们按照降序排列，那么就可以得到排名前 5 的热点问题

4.1.3 问题三模型的建立与求解

本文提及相关性，所以我们考虑相似性度量，记答复时间为 x_i 的取值为 $(x_{1i},x_{2i},\ldots,x_{ni}) \in R^n (j=1,2,\ldots,m)$,则可以用两个变量 x_i 与 x_k 的样本相关系数作为相似性度量，即：

$$r_{ik} = \frac{\sum_{i=1}^{n}(x_{ij}-x_j)(x_{ik}-x_k)}{[\sum_{i=1}^{n}(x_{ij}-x_j)^2\sum_{i=1}^{n}(x_{ik}-x_k)^2]^{1}} \quad (8)$$

根据公式（1）判断 a.当$|r_{ik}| \le 1$ 对于一切 j,k;

b.$r_{ik} = r_{ki}$,对于一切 j,k;

$|r_{ik}|$越接近 1，x_i与x_k越相关或相似，$|r_{ik}|$越接近 0，x_i与x_k相似性越弱

本文提及完整性，我们采用 Spearman 相关系数进行分析，从附件 4 中随机抽出 n 个样本，记作$y_1,y_2,\ldots,y_n$顺序统计量是$y_{(1)},y_{(2)},\ldots,y_{(n)}$,如果$y_i = y_k$,这样 k 就是 y 样本中的秩，记为$R_i$

Spearman 相关系数：

$$q_{xy} = \frac{\sum_{i=1}^{n}(R_i-R)(S_i-S)}{\sqrt{\sum_{i=1}^{n}(R_i-R)^2\sum_{i=1}^{n}(S_i-S)^2}} \quad (9)$$

根据（2）的计算可以判断数据的平稳性，可判断是否具有一致性，对数据进行分性。

数据分析结果：

相关性：根据用户留言时间和政府答复时间可以看出政府处理事情的时间还是很快的，每天要处理那么多事情都还能在 15 天左右得到回复，可以说是真的全心全意为人民服务。对市民提出的留言没有一点拖延，给市民的感觉就是全心全意为人民服务。

完整性：从市民的留言和政府答复内容来看每个点都得到精确的答复，是从各个方面的因素来给市民最满意的答复。让市民有什么不懂的地方都能从政府得到相关的答复，有的地方还用到了相关的法律保证政府给出答复的可靠性。

可解释性：从解释性方面来说答复的内容都是从政府力所能及的来答复如果做不到他们就不会给出能做出来的，给市民一个很合理的答复，不会给出政府做不出的答复

五、 误差分析

1.一些留言内容不是特别清楚，会导致归类时出现错误。

2.间进行划分时不一定能将热点问题的时间段全部囊括，这样可能会导致热点排序出现错误。

3. 数据量太大，导致计算出现误差。

六、 模型的优缺点分析

Logistic 回归的优点是计算代价不高，易于理解和实现。Logistic 回归的缺点是

在面对多元或非线性决策边界时性能较差。若输入逻辑回归模型的特征数据存在量纲上的差异，则最好先将数据标准化处理，以获得更好的预测结果。因为模型一般采用梯度下降法，量纲差异可能导致模型无法收敛于最小值。

七、 模型的改进与推广

改进：模型建立的参数需要考虑全面，可以比较好的解决问题，误差的处理应该建立更完善的应对机制，尽可能使模型的应用有着更广的范围。

推广：模型若是能够很好的反映群众中存在的问题，那可以将其广泛应用到生活中。

参考文献

[1] 李新福，赵蕾蕾，何海滨等.使用 Logistic 回归模型进行文本分类.计算机工程与应用，2009.

[2] 杨杰明 . 文本分类中文本表示模型和特征选择算法研究 .吉林大学 ,2013.

[3] 司守奎，孙兆亮等.数学建模算法与应用.国防工业出版社，2019.