“智慧政务”中的文本挖掘应用

摘要:随着各种网络问政平台的发展,各类社情民意相关的文本数据量不断上升,增加了政府相关部门的工作量。为了解决这一问题,本文利用文本挖掘工具对群众留言进行分类以便后续相关部门进行处理;通过数据挖掘找出热点问题进行集中处理,提高工作效率;最后给出一套评价方案对工作人员的答复意见进行评价。

针对问题 1,本文借助 PyCharm 程序先对所给附件 2 的数据进行预处理,进行词频统计画出词云图,接着利用 TF-IDF 权重策略将所处理好的数据用向量表示,然后建立朴素贝叶斯模型并利用 F-Score 评价法对模型进行评价,最终计算出交通运输、劳动和社会保障、卫生计生、商贸旅游、城乡建设、教育文体、环境保护的 F 值分别为 0.90、0.85、0.80、0.77、0.72、0.83、0.84,都有着较高的精确度。

针对问题 2,对附件 3 的数据进行预处理,将非结构化的文本数据转化为结构化的向量型数据,并采用了 K-Means 算法对文本进行聚类,结合合理的热度评价指标,得到前五名的热点问题分别为 58 车贷案件进展情况、项目强制捆绑销售车位、小区临街餐饮店油烟噪音扰民、附近修建搅拌厂污染环境影响生活和咨询 A 市住房公积金贷款问题。

针对问题 3,本文从相关性、完整性和可解释性三个方面来衡量答复意见的质量,针对这一类多要素评价问题,本文采用层次分析模型结合文本挖掘技术来求解。首先构建判断矩阵,用 matlab 计算得出各个指标的权重,用文本挖掘技术对各个指标进行量化分析,从而得出答复意见的评价函数。

关键词:文本挖掘;PyCharm 程序;朴素贝叶斯模型;K-Means 模型;层次分析法

1、问题重述

对于问题 1，为了方便后续将群众留言分派至相应的职能部门进行处理，我们先利用 PyCharm 程序将所给附件 2 的数据进行提取，并对这些数据进行处理，再根据这些数据建立关于留言内容的一级标签分类模型，然后使用 F-Score 对分类方法进行评价，不断调整分类方法，从而找出最优的分类模型。

对于问题 2，我们需要在众多的留言信息中挖掘出热点问题（热点问题：某一时段内群众集中反映的某一问题），及时地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。对此，我们根据所给附件 3 的留言信息将某一时段内反映特定地点或特定人群问题进行归类，定义合理的热度评价指标，并给出评价结果，得出排名前 5 的热点问题。

对于问题 3，我们需要针对附件 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。

2、问题分析与求解

问题 1：由于这些留言信息均为文本，如若要利用自然语言处理技术进行处理，首先要将这些文本数据转化为数量型的数据，因此笔者采用了 TF-IDF 权重策略将数据进行处理。而对文本分类有 k 临近、决策树、多层感知器、朴素贝叶斯等算法，本文采取朴素贝叶斯算法构建一级标签分类模型，再利用 F-Score 模型对分类方法进行评价（F 值越大，精确度越高）。

问题 2：对于问题 2，本文先利用欧氏距离计算出相似的留言，再将这些留言归类，把特定地点或人群的数据归并；然后建立合适的热度评价指标和计算方法。最后得出排名前 5 的热点问题。

问题 3：针对留言的答复意见，本文利用层次分析法并结合文本挖掘技术对答复意见的文本进行量化分析，从而给出答复意见在相关性、完整性、可解释性三个方面质量水平。

3、符号说明

符号	符号说明
TF	关键词词频
IDF	逆向文本频率
P_i	查准率
R_i	查全率
M	决策
A_i	各一级指标
CI	一致性指标
RI	随机一致性指标
CR	一致性比率
W	权重

4、模型的建立与求解

4.1 问题 1 模型的建立与求解

4.1.1 数据预处理

本文先将题目所给的数据导入到 PyCharm 程序中，一共 9210 份数据，其中城乡建设 938 份、交通运输 613 份、教育文体 1589 份、劳动和社会保障 1969 份、商贸旅游 1215 份、卫生计生 877 份，从全部数据里抽取各 600，再从这 7*600 份数据里抽 80%做实验数据，20%作为检验数据。然而有些留言的内容是近似或者完全相同的，所以利用文本匹配的方法将重复的内容删除。除此之外，对原始文本数据进行数据清洗、中文分词、去除文本数据的无用信息（即去停用词），最后得到有效的信息。为了将数据可视化，本文利用 PyCharm 程序进行词频统计，绘制了 7 个一级标签的词云图如下：

图 1：城乡建设

图 2：环境保护

图 3：交通运输

图 4：教育文体

图 5：劳动和社会保障

图 6：商贸旅游

图 7：卫生计生

从上面的词云图可以清晰地看到各个一级指标的关键词语。

4.1.2 文本向量的表示

由于计算器无法理解像英语单词、中文这类非结构化的数据，所以我们在建立模型前需要将非结构化数据转换为结构化数据，一般是用词向量矩阵来表示这些文本信息。考虑到词频信息和为了避免句子长度不一致的问题，本文采用了 TF-IDF 权重策略来提取特征项。TF-IDF 权重策略是对文本中的各项都赋予了一定的权重，由权重来表示该项内容对文本的重要性，其中 TF 是关键词词频，IDF 是指逆向文本频率，TF-IDF 的值越高，该特征项越能代表文本的内容特征。具体的计算公式如下：

$$TF = N / M$$

其中 N 为单词在某文档中的频次，M 为该文档的单词数。

$$IDF = \log\left(\frac{D}{D_W}\right)$$

其中 D 为总文档数，D_W 为出现该单词的文档数。则

$$TF - IDF = TF \times IDF$$

假设一个文本集合用 A 表示，包含多个文本数据 $A = \{a_1, a_2, a_3, ...a_n\}$，从 a_i 中提取表示该文本的特征项集合 $T = \{t_1, t_2, t_3, ...t_n\}$，计算对应的权重集合为 $Q = \{q_1, q_2, q_3, ...q_n\}$，组成一个特征向量，此时该文本可以表示为 $\vec{v}_{ai} = \{q_1, q_2, q_3, ...q_n\}$。类似地对其他文本数据进行向量化处理，得到整个文本数据的向量矩阵数据，如下表示

$$A=\begin{bmatrix}\vec{v}_{a1}\\ \vec{v}_{a2}\\ \cdots\\ \vec{v}_{an}\end{bmatrix}=\begin{bmatrix}q_{11} & \cdots & q_{1n}\\ \vdots & \ddots & \vdots\\ q_{m1} & \cdots & q_{mn}\end{bmatrix}$$

本文按照上述步骤，将附件 2 的文本数据全部进行向量化处理。

4.1.3 模型的建立与评价

对数据处理完后，开始建立文本分类模型。常用的文本分类模型有决策树算法、支持向量机算法、随机森林算法、贝叶斯算法等，本文采用了较为简单的朴素贝叶斯算法，其中包括了高斯朴素贝叶斯、多项式朴素贝叶斯和伯努利朴素贝叶斯算法。贝叶斯算法是以贝叶斯定理为基础，对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别，主要是对文本数据进行二分类。具体的公式如下：

$$P(B\mid A)=\frac{P(A\mid B)P(B)}{P(A)}$$

其中 $P(B\mid A)$ 是事件 A 发生的条件下事件 B 的概率， $P(A\mid B)$ 是事件 B 发生的条件下事件 A 的概率， $P(A)$ 、 $P(B)$ 分别是事件 A、B 发生的概率。

将贝叶斯定理运用于分类模型中：假设 $X=\{b_1,b_2,...,b_n\}$ 为一个待分类项，而且每个 b 为 x 的一个特征属性， $C=\{y_1,y_2,...,y_n\}$ 为所有类别集合。选取一部分已知分类的待分类项集合（即训练样本集），分别计算在各类别下各个特征属性的条件概率估计，即 $P(b_1\mid y_1),P(b_2\mid y_1),...P(b_n\mid y_1);...;P(b_1\mid y_n),...,P(b_2\mid y_n)$ 。假设各个特征属性是条件独立的，则根据贝叶斯定理有：

$$P(y_i\mid x)=\frac{P(x\mid y_i)P(y_i)}{P(x)}$$

又因为各特征属性相互独立，所以有：

$$\max h(y)=P(x\mid y_i)P(y_i)=P(b_1\mid y_i)P(b_2\mid y_i)...P(b_n\mid y_i)=P(y_i)\prod_{j=1}^{n}P(a_j\mid y_i)$$

由于原始的朴素贝叶斯只能处理离散数据，因此本文在此基础上使用了伯努利朴素贝叶斯算法完成分类任务。假设与每个类相关的连续变量的分布是基于伯

努利分布的，此时的伯努利朴素贝叶斯公式如下：

$$P(x_i = y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$

笔者借助 PyCharm 程序对附件 2 中的城乡建设、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 7 个一级标签分别建立伯努利朴素贝叶斯模型，再利用 F-Score 对此分类方法进行评价：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。

经过计算，得出 7 个一级标签的混淆矩阵如下图：

	0	1	2	3	4	5	6
0	103	1	3	2	8	0	2
1	0	107	5	2	5	4	2
2	0	11	88	3	8	4	2
3	4	2	1	86	15	5	4
4	1	2	1	6	95	7	7
5	1	5	2	3	7	99	1
6	1	0	4	4	7	2	94

图 8：一级标签的混淆矩阵图

并根据 F-Score 评价法得出各一级标签的准确率、召回率以及 F 值，实验结果如下：

	precision	recall	f1-score	support
交通运输	0.94	0.87	0.90	119
劳动和社会保障	0.84	0.86	0.85	125
卫生计生	0.85	0.76	0.80	116
商贸旅游	0.81	0.74	0.77	117
城乡建设	0.66	0.80	0.72	119
教育文体	0.82	0.84	0.83	118
环境保护	0.84	0.84	0.84	112
accuracy			0.81	826
macro avg	0.82	0.81	0.81	826
weighted avg	0.82	0.81	0.82	826

图 9：各一级标签的 F 值

4.1.4 结果分析

由上图可知，交通运输、劳动和社会保障、卫生计生、商贸旅游、城乡建设、教育文体、环境保护的 F 值分别为 0.90、0.85、0.80、0.77、0.72、0.83、0.84，其中交通运输、劳动和社会保障、环境保护、教育文体和卫生计生的精确度较高，其余两个较低，对此我们可以通过改变模型或增强数据不断提高精确度，进而更好地为政务工作者服务。

4.2 问题 2 模型的建立与求解

4.2.1 数据预处理与向量化表示

类似问题 1 的数据预处理，笔者将附件 3 的原始文本数据进行数据清洗、利用 jieba(结巴)库对其进行分词处理、去除文本数据的无用信息(即去停用词)。类似地，运用 TF-IDF 策略选取本文特征词、计算特征权重，从而得到文本的特征向量。前面的问题 1 已有对应的公式方法，在此笔者不再赘述。

4.2.2 相似度的计算

计算文本的相似度一般有内积、绝对值距离、欧式距离、切比雪夫距离、夹角余弦等方法，本文采用了欧式距离来计算文本的相似度。

$$sim(A_m, A_n) = \sqrt{\sum_{i=1}^{n} (w_{mi} - w_{ni})^2}$$

其中 $sim(A_m, A_n)$ 表示新文档和一个类别中的特征项的相似度，n 表示特征向量的维度，w_{mi} 表示新文档 m 的特征向量的第 i 个权重，w_{ni} 表示第 n 个话题特征向量的第 i 个权重。

4.2.3 模型的建立及求解

关于本文聚类的常用的方法有基于划分法的 K-Means 算法、K-MEDOIDS 算法、CLARANS 算法，有基于层次法的 BIRCH 算法、CURE 算法、CHAMELEON 算法等，还有流式数据聚类 Single-Pass 算法。由于 K-Means 算法是无监督的聚类算法，其实现起来较为简单且聚类效果也不错，因此本文采取了 K-Means 算法，对附件 3 中的 4326 条数据进行聚类。K-means 聚类算法又称 k 均值聚类算法，是一种基于距离的聚类算法，其采用距离作为相似性的评价指标，即认为两个对象的距离越近，其相似度就越大。

K-Means 算法的具体步骤：

（1）首先我们要确定一个 k 值，也就是希望将数据集经过聚类得到 k 个集合。

（2）从已有的数据集中随机选择 k 个数据点作为质心。

（3）对其他数据的每一个点，计算其与每一个质心的距离，计算出的距离离哪个质心近，就归属于该质心的集合。

（4）把所有数据归于 k 个数据集，再重新计算每个集合的质心。

（5）如果新计算出来的质心和原来的质心之间的距离符合我们的预想，即可认为聚类已经达到期望的结果，算法终止。如果新质心和原质心距离过大，需要重新按照执行（3）、（4）步骤，直至两者的距离符合预想。

上面的流程用下图表示为：

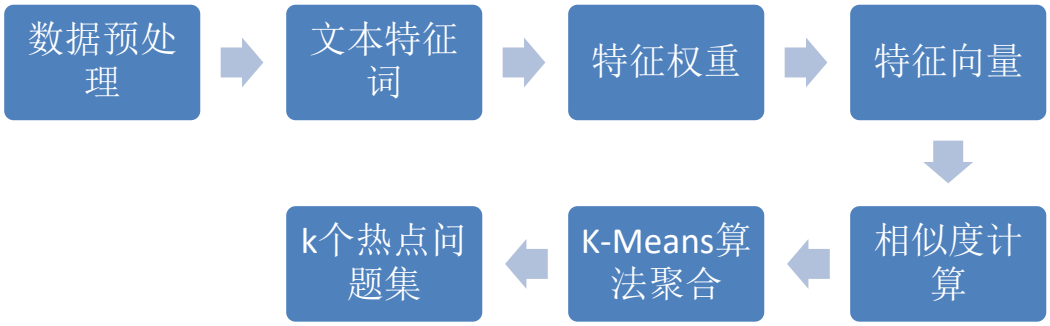

图 10:K-Means 算法流程图

本文先利用 PyCharm 程序先对数据进行处理，将文本数据转化为向量矩阵，为建立模型做好准备；接着调用 PyCharm 程序的 sklearn 过程中的 K-Means 模型对已处理好的数据进行分析，经过不断的实验，本文选取 k 为 100，将分类好的 100 个数据集导出，得到热点问题留言明细表。对于选出的 100 个热点问题，由于热点问题是某一时段内群众集中反映的某一问题，所以与问题的出现次数、点赞数和反对数相关，对此本文定义：

热度指数=留言数+总留言点赞数-总留言反对数

计算得出了前五名的热点问题分别为 58 车贷案件进展情况、项目强制捆绑销售车位、小区临街餐饮店油烟噪音扰民、附近修建搅拌厂污染环境影响生活和咨询 A 市住房公积金贷款问题，对应的热度指数如下表：

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	96	2411	2019/01/11 至 2019/12/17	A 市 A4 区	58 车贷案件进展情况
2	99	58	2019/07/07 至 2019/09/01	A 市伊景园滨河苑	项目强制捆绑销售车位
3	33	37	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
4	31	30	2019/07/03 至 2020/01/06	A2 区丽发新城小区	附近修建搅拌厂污染环境影响生活
5	73	25	2019/01/02 至 2019/12/26	A 市	咨询 A 市住房公积金贷款问题

表 1：热度指数

4.3 问题 3 模型的建立及求解

4.3.1 评价指标的选取

（1）因素的影响机理分析

本文选取了相关性、完整性和可解释性三个指标来衡量答复意见的质量。

a.相关性。相关性是指答复意见的内容是否与问题相关。其可以通过留言主题与答复意见文摘两个文本的相似度来衡量，两者的相似度越大，这说明该答复意见与留言问题的相关性较高；反之，相关性较低。

b.完整性。完整性是指答复意见内容是否完整、结构是否完整（比如开头是否有回应网友，中间是否有对问题进行解答，结尾是否有留下相应的联系方式等）。完整性越高，说明答复意见的质量越高。

c.可解释性。可解释性是指答复意见中内容的相关解释，即回复的内容是否有效，是否通俗易懂、是否简洁。可解释性越高，说明答复意见的质量越高。

当留言的答复意见做到与留言问题相关、内容完整且规范、有效且通俗易懂，则认为该答复意见的质量较优。

（2）建立指标框架

根据对题目及问题的深入分析，我们综合考虑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进行评价，并运用文本挖掘技术将数据化的答复作为评价指标，我们得到指标框架，如下表所示：

一级指标	评价指标
相关性	基于自动文摘技术的答复意见与留言主题两者间的文本相似度计算，完成其相关性的评价分析。
完整性	基于中文依存句法分析的答复意见，分析其句法结构与语义特征，总结和设计评价完整性的抽取规则。基于构建的情感词典，并辅以语义相似度设计内容完整和结构完整的计算规则，完成了完整性评价指标的测量。
可解释性	基于构建的通俗词典，对答复意见做中文分词、词性标注、实体名识别、关键词抽取，通过对其词频的统计分析，完成其可解释性的衡量。

表 2：指标框架

4.3.2 模型的建立

（1）数据标准化

实际应用时，需要简化数据，将数据进行归一化处理，数据的归一化使得不同来源的数据具有相同量纲和相同数量级，让每一个数据都有很规范的格式，即每一个输入源均值是 0，方差是 1,从而客观比较每一个数据的作用。

均值：$$\mu=\frac{\sum_{i=1}^{t}X_i}{t}$$

方差：$$\sigma^2=\frac{\sum_{i=1}^{t}\left(X_i-\mu\right)^2}{t}$$

归一后得到：$$Y_i=\frac{\left(X_i-\mu\right)}{\sigma}$$

（2）比较尺度

尺度 a_{ij}	含义
1	C_i与C_j的影响相同
3	C_i与C_j的影响稍强
5	C_i与C_j的影响强
7	C_i与C_j的影响明显的强
9	C_i与C_j的影响绝对的强
2,4,6,8	C_i与C_j的影响之比在上述两个相邻等级之间
1,1/2，…，1/9	C_j与C_i的影响之比为上面a_{ij}的互反数

（3）构造判断矩阵

对各指标进行两两对比之后,按9分位比率排定各评价指标的相对优劣顺序，依次构造出评价指标的判断矩阵L。

$$L=\begin{pmatrix} 1 & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & 1 \end{pmatrix}$$

a_{ij}要素 i 与要素 j 重要性比较结果，并存在以下关系：

$$a_{ij}=\frac{1}{a_{ji}}$$

（4）一致性检验,步骤如下：

a.用 λ-n 的大小来衡量 A 的不一致程度

$$CI=\frac{\lambda-n}{n-1}$$

b.随机一致性指标 RI 的数值

n	1	2	3	4	5	6	7	8	9	10	11
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51

表中 n=1,2 时，RI=0，是因为 1,2 阶的正互反阵总是一致阵

c.一致性比率 CR

对于 n≥3 的成对比较阵 A，将它的一致性指标 CI 与同阶（指 n 相同）的随机一致性指标 RI 之比称为一致性比率 CR,当

$$\mathrm{CR}=\frac{CI}{RI}<0.1$$

时认为 A 的不一致程度在容许范围内，可用其特征向量作为权向量。

d.组合权向量

$$w^{(s)}=W^{(s)}W^{(S-1)}\dots W^{(3)}w^{2}$$

e.组合一致性检验

$$CI^{(P)}=[CI_1^{(P)},\dots,CI_n^{(p)}]w^{(p-1)}$$

$$PI^{(P)}=[PI_1^{(P)},\dots,PI_n^{(p)}]w^{(p-1)}$$

$$\mathrm{CR}=\sum_{p=2}^{s}CR^{(P)}$$

当 CR<0.1 时，认为整个层次的比较判断通过一致性检验。

4.3.3 模型的求解

各一级指标之间的重要性比较：利用 matlab 软件计算出各一级指标（即相关性、完整性、可解释性）对答复意见的重要性，得到下面表格

M	A1	A2	A3	W	
A1	1	1/3	1/5	0.1047	CI=0.0193
A2	3	1	1/3	0.2583	RI=0.58
A3	5	3	1	0.637	CR=0.0332

表 3：各一级指标的相关权重

CR<0.1,通过一致性检验。

有上表可知,相关性、完整性、可解释性的权重分别为 0.1047、0.2583、0.637，由此看出可解释性对答复意见的质量评价占有较大的比例。因此，答复意见质量的评价函数为：

Z=0.1047 A1+0.2583 A2+0.637 A3

其中，A1 为相关性，A2 为完整性，A3 为可解释性。

5、模型的评价

对于问题 1 的伯努利朴素贝叶斯模型，其优点是有稳定的分类效率，算法也比较简单且能处理多分类任务；然而，由于朴素贝叶斯模型是建立在属性之间相互独立的假设之下的，这个假设往往在实际应用中并非总是如此，所以在属性较多或属性之间的相关性较大时分类效果不好。因此对于问题 1 可以进行多个分类模型的尝试以找出最合适的模型。

对于问题 2 的 K-Means 模型，其优点是原理较为简单且实现起来比较容易，算法的可解释性强；但是 k 值得选取不好把握，对数据的要求较高（如果各隐含类别的数据不平衡，则聚类效果并不明显），本文也因此聚类效果不太好，后期可以利用命名实体识别对模型进行改进。

对于问题 3 的层次分析法，能把主观的因素进行量化且操作起来比较简单方便，但是其主观性较强，往往是根据人们的经验确定其重要性，还存在选取的指标不全面的问题。在模型改进上，应该结合大数据，考虑更多的因素，进行更全面，更精确的评价。

6、参考文献

[1] 格桑多吉,乔少杰,韩楠,张小松,杨燕,元昌安,康健.基于 Single-Pass 的网络舆情热点发现算法[J].电子科技大学学报, 2015, 44(4):599-604.

[2] 马子岩.基于 Single-Pass 算法的微博舆情分析系统设计与实现[D].河北大学, 2019.

[3] 刘玲.基于数据挖掘的 S 市消费者投诉行为分析[D].福州大学, 2017.

[4] 裴倩雯.基于数据挖掘的电力客户投诉分析与预测研究[D].北京交通大学, 2019.

[5] 李其玲.基于双循环组织学习理论的政务社交媒体文本挖掘研究——以“平安武汉”为例[D].华中科技大学, 2018.

[6] 李少温.基于网络问政平台大数据挖掘的公众参与和政府回应问题研究[D].华中科技大学, 2019.

[7] 孙赫.基于微博的城市投诉文本的挖掘与分析[D].北京信息科技大学, 2015.

[8] 廉素洁.基干文本分类和情感评分的电信投诉文本挖掘研究[D].浙江工商大

学，2018.
[9] 吴柳,程恺,胡琪.基于文本挖掘的论坛热点问题时变分析[J].软件，2017,38(4).
[10] 候玉林.基于文本意见挖掘的快递服务质量评价研究[D].北京交通大学,2019.
[11] 刘动博.基于信息挖掘的旅游投诉特征研究--以人民网旅游3•15投诉平台为例[D].广西大学，2019.
[12] 李洋.文本挖掘视角下电子商务平台在线评论对大型家电销量的影响—以冰箱为例[D].合肥工业大学，2019.
[13] 姜玉坤.舆情热点信息挖掘技术的研究与应用[D].天津大学，2016.
[14] 薛彬,陶海军,王加强.基于文本挖掘的论坛热点问题时变分析[J].中国计量大学学报，2017，28(3).