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。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

针对问题 1，我们经过对数据的预处理后分词，通过 TF-IDF 提取特征，并使用传统机器学习分类方法朴素贝叶斯模型、逻辑回归模型、SVM、K 近邻以及深度学习方法 BP 神经网络、LSTM、CNN 等深度学习模型比较其 F1 值与时间等因素选择 SVM 为最优模型。

针对问题 2，我们通过计算文本相似度进而进行文本聚类，对相似文本归类，今儿设置热度指标对热点问题进行挖掘，并得出了排名前五的热点问题。

针对问题三，首先，我们建立 AS-模糊综合评价模型。本模型的重点是建立评论特征提取模型。首先，在对评论文本进行数据预处理之后，我们进行评论文本的数值化。而对于不符合这种结构的语言，我们基于潜在语义分析，对其进行词频矩阵降秩。从而，语义相近词语的相关性得到增强，反之亦然。最后，我们对词频矩阵进行 Logistic 回归，再利用回归后的数据进行聚类分析，可以得到 6 种聚集形式，即得到了等级分布。接着，我们开始建立评论-评级评分模型，我们运用支持向量机模型，绘制评论等级-评分等级分布的散点分布图。由该图可清晰的看出评论和评级之间的最佳测量度。而对于该散点图来说，在这个二维的空间中，随着横坐标增大，纵坐标增大，则该产品所代表的潜在成功就越高。因而该模型也能反映出潜在成功和潜在失败的关系。接着，我们引入相关系数模型。为探索特定星级与评论种类之间的关系，我们通过定性和定量两方面进行深入了解，在定性分析过程中，我们应用卡方检验得出卡方的值大于临界值，表明星级对评论有显著的影响。为定量分析，我们引入相关性分析，绘制相关系数列联表，可以得出：特定星级会引起更多的评论。同理，对于评论中的关键词和整体特征的关系，我们也利用相关分析模型，并发现一些评论中的关键词和评级有着密不可分的关系。

关键词： TF-IDF SVM 文本分类 聚类分析 Logistic 回归分析

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。Hu 等人提出了一个系统完整的评论挖掘过程，包括产品特征挖掘、主观句定位、情感分析、极性判定以及结果显示等部分，并在其论文中阐述了商品特征挖掘的方法.popescul5 等人将改进的 PMI(point mutual information)值引入特征词简直，查全率略有降低而查准率有了显著的提升。Scaffidi 等人开发了一种新的检索系统 red opal，它只识别单个词和二词短语作为特征，并对每个产品的每个特征都进行打分，输出结果按照产品特征的打分综合排序，该方法用打分的方式来区分特征的重要程度，但是在识别特征的过程中依然没有用更有效的剪枝来剔除冗余特征。我们发现，截至目前，少有研究深入考虑影响因素之间相互存在的制约关系。因此，基于题目背景，本文通过对各个特征关系建立模型得到其确定的制约关系，进而建立了一个反应满意度变化的评价-评级评分模型。

2.问题重述

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。

3.问题求解步骤

1.针对问题一，我们使用 F-Score 对分类方法进行评价，建立关于留言内容的一级标签分类模型。分以下步骤处理 1.预处理：去除文本的噪声信息、重复信息 2.中文分词：使用中文分词器为文本分词，并去除停用词。3.构建词向量空间：统计文本词频，生成文本的词向量空间 4.权重测量-----TF-IDF 方法：使用 TF-IDF 发现特征词，并抽取为反映文档主题的特征。5.分类器：使用算法训练分类器 6.评价分类结果：分类器的测试结果分析。

2.针对问题二。使用聚类的方法，将留言进行归类，然后根据热度计算方式计算热度，今儿获取排名，获得排名前五的热点问题。

5.问题模型的建立与解决

5.1 问题一

5.1.1 问题分析

问题一是对附件 2 中提供的三种数据材料进行预处理，以消除和合理处理不包含实际意义较小的数据。使用数学证据来识别，描述和支持有意义的定量和定向模式，从而有利于以便后续将群众留言分派至相应的职能部门处理。

5.1.2 基于神经网络的数据清洗及缺项有效剔除、插值

1.数据预处理

数据预处理是数据挖掘过程中的关键步骤。高质量的数据可以改善数据挖掘模型。数据质量可以快速有效地挖掘出有价值的信息。大数据环境中的数据结构越来越多，复杂且不断增加的数据量，并受到噪声数据，数据丢失和不一致的影响，通过上述方法收集的在线评论数据源通常是不规则且嘈杂的，不能直接用于挖掘在线评论信息首先需要对在线评论进行预处理。在线评论是非结构化的文本信息，中文在线注释的预处理主要包括垃圾邮件过滤，中文分词，属性扩展，减少属性，删除无用的单词以及词性标记。

在线商品评论质量特征	细分特征	特征描述
文本特征[7]	语法特征[9]	专有名词、数字、情态动词、感叹词等
	语义特征[10]	情感极性、评价对象等
	文体特征[11]	词汇和结构属性
元数据特征[8]	商品评论发布者与读者之间的互动属性[9]	评论“有用”票数、评论发表时间、商品评分等

图 1 数据剔除原理

2.TF-IDF 计算过程

TF-IDF(Term Frequency - Inverse Document Frequency)即词频-逆向文本频率，是一种用于信息检索和文本挖掘的常用加权技术。假设一个语料库包含多个文件，TF-IDF 则用于评估一字词对于一个语料库中的其中一份文件的重要程度。字词的重要性会随着它在文件中出现的次数成正比增加，但同时也会随着它在语料库中出现的频率成反比下降。

词频（term frequency）指的是某一个给定的词语在该文件中出现的频率。对于在某一文件里的词语来说，它的重要性可表示为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中，ni,j 是该词在文件 dj 中的出现次数，而分母是文件 dj 中所有字词的出现次数总和。

逆向文本频率（IDF）

逆向文件频率（inverse document frequency）是一个词语普遍重要性的度量。某一特定词语的 idf，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取以 10 为底的对数得到：

$$idf_i = lg\frac{|D|}{|j:t_i \in d_j|+1}$$

其中，|D| 是语料库中的文件总数，$|j:t_i \in d_j|$是包含词语t_i 的文件数目。

如果没有一个文件包含词语t_i ，那么导致分母为零，所以通常使用

$|j:t_i \in d_j|+1$.

3.支持向量机 SVM

支持向量机（support vector machines, SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM 还包括核技巧，这使它成为实质上的非线性分类器。SVM 的的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM 的的学习算法就是求解凸二次规划的最优化算法。

5.2 问题二

5.2.1 问题重述

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。

5.2.2 brich 层次聚类

BIRCH 算法利用了一个树结构来帮助实现快速的聚类，这个数结构类似于平衡 B+

树，一般将它称之为聚类特征树(Clustering Feature Tree，简称 CF Tree)。这颗树的每一个节点是由若干个聚类特征(Clustering Feature，简称 CF)组成

BIRCH 算法的流程。

1） 将所有的样本依次读入，在内存中建立一颗 CF Tree, 建立的方法参考上一节。

2）（可选）将第一步建立的 CF Tree 进行筛选，去除一些异常 CF 节点，这些节点一般里面的样本点很少。对于一些超球体距离非常近的元组进行合并

3）（可选）利用其它的一些聚类算法比如 K-Means 对所有的 CF 元组进行聚类，得到一颗比较好的 CF Tree.这一步的主要目的是消除由于样本读入顺序导致的不合理的树结构，以及一些由于节点 CF 个数限制导致的树结构分裂。

4）（可选）利用第三步生成的 CF Tree 的所有 CF 节点的质心，作为初始质心点，对所有的样本点按距离远近进行聚类。这样进一步减少了由于 CF Tree 的一些限制导致的聚类不合理的情况。

从上面可以看出，BIRCH 算法的关键就是步骤 1，也就是 CF Tree 的生成，其他步骤都是为了优化最后的聚类结果。

热度评价借鉴 reddit 热点排名算法：

Given the time the entry was posted A and the time of 7:46:43 a.m. December 8, 2005 B, we have t_s as their difference in seconds

$$t_s = A - B$$

and x as the difference between the number of up votes U and the number of down votes D

$$x = U - D$$

where $y \in \{-1, 0, 1\}$

$$y = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x = 0 \\ -1 & \text{if } x < 0 \end{cases}$$

and z as the maximal value, of the absolute value of x and 1

$$z = \begin{cases} |x| & \text{if } |x| \geq 1 \\ 1 & \text{if } |x| < 1 \end{cases}$$

we have the rating as a function $f(t_s, y, z)$

$$f(t_s, y, z) = \log_{10} z + \frac{yt_s}{45000}$$

5.2.3 结论

通过聚类排名我们分析得出热点问题随时间变化情况，由此我们得出热点问题留言明细表，并保存为附件“热点问题留言明细表.xls”。

7.参考文献

[1]https://www.jianshu.com/p/edad714110fb 简书博客

[2]https://blog.csdn.net/d__760/article/details/80387432 svm 原理