“智慧政务”中的文本挖掘

摘要

近年来，随着各类网络问政平台的不断应用，各类社情民意相关的文本数据量日益增多。因此，建立基于自然语言处理技术的政务系统对增强政府的管理水平和施政效率具有重大的意义。

针对问题一， 由于附件 2 数据分布不均衡，故通过欠抽样的方法来使数据均衡化，之后对处理过的数据运用 jieba 中文分词工具对留言详情进行分词及去停用词并将数据分为训练集与测试集。利用 TF-IDF 算法将训练集与测试集转化成词频向量。本文以支持向量机(Support Vector Machine，SVM)作为分类器，通过 SVM 算法建立模型，对训练集进行训练，对测试集进行预测。由此建立分类模型，通过模型分数对模型进行调整得到最终模型。

针对问题二，需要做到将特定时间、特定地点、发生的问题归并在一起，通过量化评价指标得出每个问题的热度值，再将其进行排序从而得到排名前五的热点问题。首先对附件 3 数据文本进行筛选，相对于数据库中结构化的数据而言，得到的文本是非结构化的，需要建立模型方便于计算机的处理，VSM 对词进行加权，去除一个词在不同语境下的不同意思，在降低维度的同时，解决近义词，同义词问题。然后进行文本相似度计算，K-means 算法为文本聚类提供了基础数据。由此定义一个热度指标定义热点问题。

针对问题 3，对附件 4 的留言和答复内容进行分析，提取我们需要进行分析的部分进行深度挖掘和分析。建立一套针对性回复的模型与评价方案。

关键词：中文分词 SVM 算法 TF-IDF 算法 文本相似度

1、挖掘目标

本次建模目标是利用网络问政平台的群众留言信息数据，使用 jieba 中文分词工具对留言详情进行分词、TF-IDF 权重策略的方法及 SVM 算法，达到以下三个目标：

1）利用文本分词和文本分类的方法对留言数据进行文本挖掘，按照附件 1 提供的三级标签体系建立留言分类模型。

2）根据附件 3 内容，按照特定时间特定地点发生的问题进行分类，并依据量化评价指标分析留言热度走向。

3）根据相关部分对留言的答复意见建立模型给出评价方案。

2 分析方法与过程

2.1 问题一分析方法与过程

2.1.1 数据预处理

2.1.1.1 数据均衡化

通过检查可得出题中所给数据出现数据不均衡现象，为了防止后续统计过程中出现数据过拟合，故使用欠抽样的方法对数据进行均衡化。欠抽样方法主要是通过随机的去掉某些多数类样本来达到降低数据不平衡的目的。[1]

2.1.1.2 对留言详情进行中文分词及去停用词

由于计算机不能识别非结构化的文本信息，故在对留言信息进行挖掘分析之前，需要先把非结构化的文本信息转化为结构化信息。而附件 2 中的数据以中文文本的方式进行展现，为了便与转换，先要对附件 2 中留言详情进行中文分词。本文采用 python 编程语言中的 jieba 库进行分词。jieba 库分词的原理是利用一个中文词库，将待分词的内容与分词词库进行对比，通过图结构与动态规划的方法，寻找出最大概率的词语。[2]

jieba 库对句子的分词有以下三种匹配模式：

（1）精确模式:根据相应的算法，可以将句子或者文本内的句子精准切分开来，在文本分析中得到广泛的应用。

（2）全模式:将本文或者句子内可以构成成词的词语全部快速扫描出来，但是在中间切分产生的词语可能会产生歧义的问题。

（3）搜索引擎模式:在精确模式产生的长词的基础上，将长词再次切分，这样做可以有效提高召回率，在搜索引擎分词中得到广泛的应用。[3]

本文采用精准模式进行匹配。

部分分词结果示例如图 2.1:

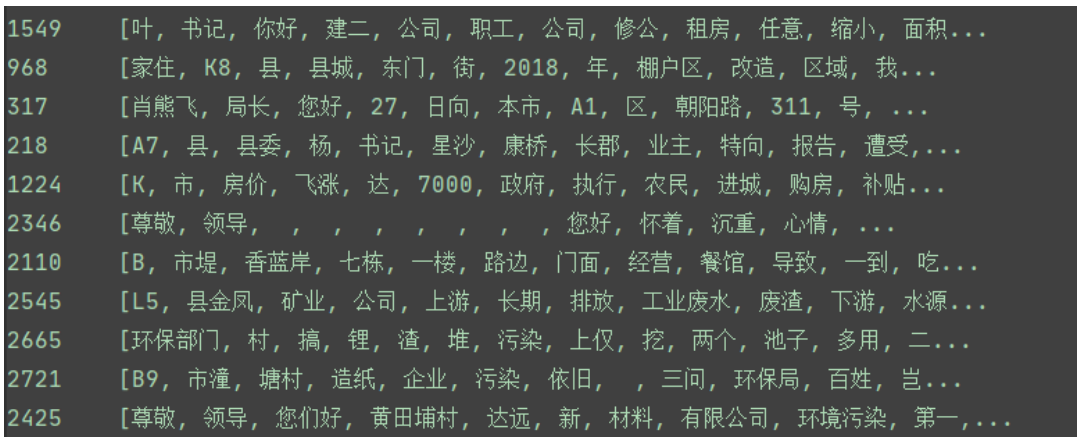

图 2.1 部分分词结果

由于分词完的数据中包含有许多与所需数据无关的字符串即 Stop Words（停用词），为了提高搜索效率和节省存储空间需要对数据进行停用词的去除，这里所用到的停用词表见附件 stop.list。

去除停用词后的部分结果示例如图 2.2:

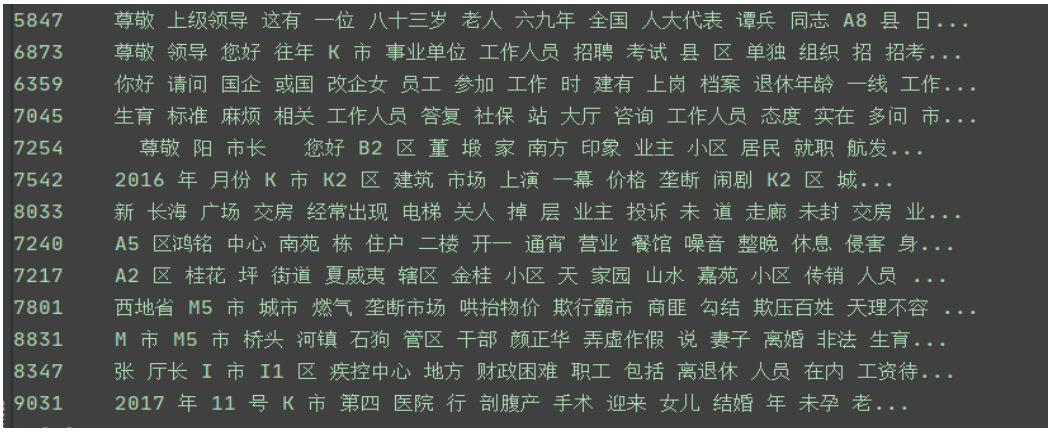

图 2.2 停用词过滤后分词结果

2.1.1.3 词频-逆文本频率（TF-IDF）算法

在对留言详情信息进行分词后，需要把这些语句转为向量，良好的词向量可以达到语义相近的词在词向量空间里聚集在一起，对后续的文本 1 分类提供了便利，以供挖掘分析使用。此处采用 TF-IDF 算法，把留言详情信息转化为权重向量。算法的具体原理如下：

TF 代表词频，表示某特征词在某文本文件中出现的次数，其定义为：

$$T_F = \frac{n_w}{\sum_{k}^{N} n_k} \qquad (2\text{-}1\text{-}1)$$

其中，n_w是特征词w在文本文件中出现的次数，N为该文本文件中特征词的数量，T_F为衡量特征词在该文本中重要程度的指标。

IDF代表逆向文件概率，其定义为：

$$I_{DF} = \log \frac{D}{1+Q} \tag{2-1-2}$$

其中，D是所有文本文件的总数，Q是包含特征词的文本文件数量，是衡量特征词在所有文本中重要程度的指标。

将式（2-1-1）和（2-1-2）相乘，可得文本特征值权重[4]

$$f_{TF \cdot IDF} = T_F \times I_{DF} \tag{2-1-3}$$

2.1.2 构建模型

2.1.2.1 支持向量机(SVM)

SVM 是一个线性分类器，具有简单高效的优点，本文采用 sklearn 中的 svm 算法作为分类器，下面介绍 svm 的相关理论知识：

考虑一个二进制分类问题的训练集$\{(x_1, y_1), (x_2, y_2), \cdots, (x_j, y_j)\}$，这里$x_i \in R_n$代表一个$n$维输入向量，$y_i \in \{-1, 1\}$代表样本点的类别。SVM 算法的目标是通过寻找下面最优二次规划问题的解来获得最优分类超平面:

$$\text{Min} \left(\frac{1}{2} w^T \cdot w + C \sum_{i=1}^{l} \varepsilon_i \right)$$

$$y_i (w \cdot \varphi(x_i) + b) \geq 1 - \varepsilon_i \tag{2-2-1}$$

$$\varepsilon_i \geq 0, i = 1, \cdots, l$$

（2-2-1）式中，ε_i松弛变量是错分代价；C是惩罚参数。

引入核函数，这个二次最优规划问题可以通过一个拉格朗日函数转化为如下的对偶问题:

$$\text{Max} \quad W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$\sum_{i=1}^{l} y_i \alpha_i = 0 \tag{2-2-2}$$

$$0 \leq \alpha_i \leq C^+, \text{当} y_i = +1$$

$$0 \leq \alpha_i \leq C^-, \text{当} y_i = -1$$

（2-2-2）式中，α_i是拉格朗日乘数，对应$\alpha_i > 0$ 的输入向量称作支持向量。于是，可以得到分划超平面:

$$g(x) = \sum_{x_p \in SV: y_p > 0} \alpha_p y_p k(x, x_p) + \sum_{x_n \in SV: y_n < 0} \alpha_n y_n k(x, x_n) + b \qquad (2\text{-}2\text{-}3)$$

（2-2-3）式中, SV 是支持向量的集合；α_p, α_n分别是正类支持向量和负类支持向量的系数。[1]

2.2 问题 2 分析方法与过程

随着互联网的迅速发展和移动互联网平台，自媒体的兴起，网络在社会影响日益扩大。新闻，微博，公众号等平台层出不穷，它们都包含了大量的，动态更新的信息，面对海量的信息，人工的力量很难进行全面的采集和整理，因此通过计算机去自动采集，整理，分析海量的信息并提取其中热点话题既有重要的研究意义。

2.2.1 数据筛选

对各个留言的概述进行计数，通过排序得出排序前 5 的问题，并定义为热点问题。热点问题即某一时段内群众集中反映的某一问题。通过排序归类得出各大热点问题的累积次数，并定义为热度评价指标，给出评价结果。

2.2.2 对热点问题的分析（热点问题的排名表，热点问题的比重）

通过对各项热点问题进行分析，得出工卫生计在热点问题占比，可以看出前 5 类问题在我们的生活中越来越多，当今社会发展之快我们有目共睹，但伴随的问题是否也会引起大家的注意，这些也是社会发展的需求，需要正确的解决方式，需要专业的团队进行处理，并促进社会的发展。

随着经济的发展，社会发展迅速，城市中的城乡建设，环境保护，交通运输，教育文本，卫生计生等在全国中的发展问题中处于引领的地位，各种问题在人们生活中带来了便利的同时也带来了各种各样的小烦恼。

2.2.3 文本聚类

面对互联网上海量的信息，构建一个完整的热点分析模型对于各类网络信息进行分类、筛选、分析、处理、组织呈现。整个系统能够运行，并具有时效性。在完成模型的过程中，进行聚类挖掘其中的热点问题，并且跟踪，记录信息最后展现模型运行结果。为了加快聚类速度采用单编聚类方法:

划分聚类是文本的集合划分到事先指定的类中，有迷糊划分和确定划分两种方法。K-means 是目前最为常规划分归类方法。它的核心思路是：首先将样本初

始划分成 K 类，计算 K 类各自的中心。然后每次迭代的样本分离到离中心点最近的那一类中，之后重先计算 K 个类的中心点，直到中心点趋于稳定，没有再发生变动的聚类终止。

可见，经过聚类，解析得到信息，热点问题的主体，来源，时间，详细内容等有效的区分开来，经正则表达式进一步的匹配，确定信息，完成对热点问题结构化提取。

2.2.4 文本相似度

对于提取后的文档，需要一个适合的模型计算和存储，在这部分我们对权重的计算，并为了方便模型的特征提取和降维，然后通过定义文本相似度，可以计算文本之间的关系，为下一步构建热点指标提供了基础。

2.2.5 命名实体识别

获取大量文本数据后，由于知识表示粒度不同、置信度相异、缺乏规范性约束等问题，出现命名实体表述多样、指代不明确等现象。故需要充分理解上下文语义来深度挖掘实体语义进行识别。可以通过实体链接、融合对齐等方法，挖掘更多有效信息和证据，实现实体不同表示的对齐、消除歧义，从而克服命名实体表述多样性和歧义性。[5]

2.2.6 模型构建

通过对热点问题的特征抽取，形成新的特征作为相似性的比较，特征权重主要是经典的 TF-IDF 方法及其扩展方法，主要一个词的重要度与在类别中的词频成正比，与所有类别出现的次数成反比。以初步构建热点指标。

加入时间持续因子，对于热点问题来说话题产生的时间长短是一个标准。如果在一小段时间里就消失了，那么这一类问题将不具备构成热点问题的基本条件，直接舍弃。

能过持续一段长时间的话题，我们将与新话题进行相似度比较-，将新话题倾向于长时间存在的问题，形成初步的热点问题。

在聚类的基础上我们定义 hot= A*该话题的数量+B*该话题单位时间的数量，（A,B 且 A+B=1 都为加权因子），每次聚类更新 hot，如果 hot 值大于指定阈值即为点问题，hot 为热点指标。

2.3 问题 3 分析方法与过程

2.3.1 答复意见的分析

依据留言与答复内容分析答复内容的相关性，完整性以及可解释性等相关问题。

相关性：根据留言与留言答复内容，检测两者是否有问答关系，是否存在答非所问的情况。

完整性：分析答复内容是依据某一标准，查看该标准。

可解释性：分析答复内容中由什么理由来支撑解决此问题。以及内容的相关解释，相关的法律法规。

2.3.2 构建指标来计算和评价

根据留言内容与答复信息提取文本内容，预设问题与答复信息。根据预测规则对所提取的文本进行识别，获取预设类别的留言信息，将信息与知识数据库中的预设问题进行匹配，判断知识数据库中是否存在所述留言信息对应的答复内容。若存在，根据留言信息针对性回复。[6]

3 参考文献

[1]吴敏,张化朋,李雷.欠抽样和 DEC 相结合的不平衡数据分类算法[J].计算机技术与发展,2014,24(04):110-113.

[2]荀雪莲,王晓宁.基于中文摘要关键词的毕业论文质量评价系统[J].廊坊师范学院学报(自然科学版),2019,19(04):30-32.

[3]邢彪,根绒切机多吉.基于 jieba 分词搜索与 SSM 框架的电子商城购物系统[J].信息与电脑(理论版),2018(07):104-105+108.

[4]但宇豪,黄继风,杨琳,高海.基于 TF-IDF 与 word2vec 的台词文本分类研究[J].上海师范大学学报(自然科学版),2020,49(01):89-95.

[5]陈曙东,欧阳小叶.命名实体识别技术综述[J/OL].无线电通信技术:1-11[2020-05-08].http://kns.cnki.net/kcms/detail/13.1099.TN.20200414.1436.002.html.

[6]中国联合网络通信集团有限公司.信息处理方法、客服服务器及主动式客服系统:CN201810757718.3[P].2018-12-25.