基于“智慧政务”的文本挖掘

摘要

本文基于自然语言处理及文本挖掘技术，针对智慧政务系统用户的留言信息等数据，建立了留言信息的分类模型，对数据进行归类，并挖掘出某一时间段发生的热点问题，针对回复信息建立了评价模型，对政府处理政务有提升作用。

针对问题一，通过对数据去重，文本分词和去停用词等操作对数据进行清洗后，对每一级标签对应的数据绘制词云，用 IF-IDF 算法将关键词向量化，构建朴素贝叶斯分类模型。

针对问题二，基于语义网络的评价分析进行初步数据感知，运用 **LDA** 模型进行主题分析，对主题的特征词出现频率量化，结合构建热度指标，将留言详情划分为 2000 个热点，构建热点评分指标，得到某一时段的热点问题。

针对第三问，对回复意见进行 jieba 分词，去停用词后对数据进行清洗，对清洗好的答复意见数据进行词频统计，并绘制词云图，通过词云图对答复意见给出评价意见。

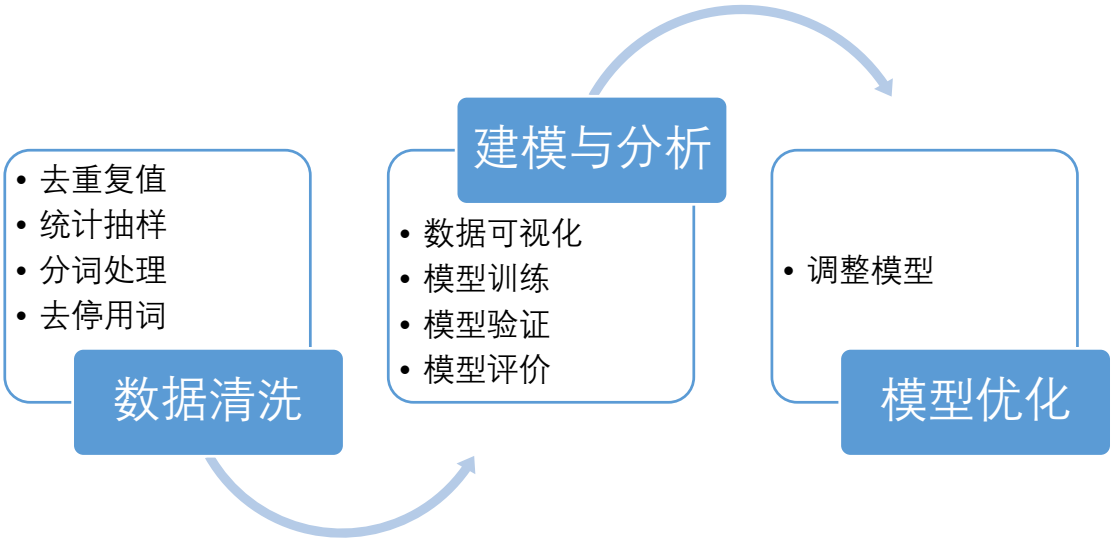
关键词：jieba 分词　IF-IDF 算法　朴素贝叶斯分类　LDA 模型

目录

1 挖掘目标

随着网络的不断发展，微信，微博等网络问政平台逐渐成为政府了解民意和凝聚民气的重要渠道，本次建模基于自然语言处理技术的智慧政务系统市民反映的问题及回应的信息等数据，采用数据挖掘技术，针对留言信息中的关键词，分析留言信息中不同权重大小，构建留言信息的分类模型，对留言信息归类，分析某一段时间内群众集中反映的热点问题，构建回复的相关性，完整性和可解释性评价体系，从而推动和提升政府的管理水平和施政效率。

2 分析方法与过程



本次建模主要包括以下步骤：

步骤一：数据预处理

步骤二：建模和分析

步骤三：模型优化

3 模型构建与分析

3.1 问题一分析方法与过程

3.1.1 数据去重与抽样

在题目给出的数据中，出现了重复的留言主题和留言详情数据。考虑到可能同一用户为同一问题没有解决而发送了多条相同留言主题而留言详情略有不同的情况，因此，只保留了一条留言主题相同的数据，见 data2.xlsx

将处理过的 Excel 文件导入到 python 中，对不同种类的一级标签数据数进行统计，统计结果如下表所示：

一级标签	数据数
城乡建设	1974
劳动和社会保障	1897
教育文体	1530
商贸旅游	1147
环境保护	909
卫生计生	856
交通运输	592

表格 1 一类标签数据统计

由于各类数据比例分布不均，构建分类模型分类性能会下降，通过减少多数类样本来提高少数类的分类性能，将每一类数据随机抽取 590 条，考虑到留言详情内容相同，留言主题不相同的情况，利用 python 对留言详情内容重复数据做删除操作。

3.1.2 jieba 中文分词

中文分词是指以词作为基本单元，使用计算机自动对中文文本进行词语的切

分，即使词之间有空格，这样方便计算机识别出各语句的重点内容。Jieba 分词中，首先通过对照词典生成句子的有向无环图，再根据选择模式的不同，根据词典寻找最短路径后对句子截取或直接对句子进行截取，对于未登录词使用 HMM 进行新词发现。最终获得较好的分词效果。

3.1.3 去停用词

由于一段话中会大量出现一些无用词或者标点符号，对模型的分析造成干扰，导入停用词文档，基于停用词库对分词结果去除停用词，同时对词频统计，不断完善停用词库，最终完成数据的过滤，将清洗好的数据输出为 data_yu.xlsx.

3.1.4 数据可视化

针对分词完成的留言详情数据，针对每一类的分词信息，对各类一级标签进行词频统计，根据词频统计结果绘制词云图，使数据更加直观，绘制的词云图如下：

图 1 城乡建设词云图

图 2 环境保护词云图

图 3 交通运输词云图

图 4 教育文体词云图

图 5 劳动和社会保障词云图

图 6 卫生计生词云图

图 7 商贸旅游词云图

3.1.5 文本的向量表示

对留言描述信息分词后，需要把这些词转换为向量，用来实现模型的构建使用，以下是采用的两种向量构建的算法：

CountVectorizer 算法

CountVectorize 算法是通过 fit_transform 函数将文本中的词语转换为词频矩阵，矩阵元素 a[i][j] 表示 j 词在第 i 个文本下的词频。即各个词语出现的次数，通过 get_feature_names()可看到所有文本的关键字，通过 toarray()可看到词频矩阵的结果。

TF-IDF 算法

权重策略文档中的高频词应具有表征此文档较高的权重，除非该词也是高文档频率。TF：Term frequency 即关键词词频，是指一篇文档中关键词出现的频率

IDF：Inverse document frequency 指逆向文本频率

$$TF = \frac{N}{M},$$

N 为单词在某文档中的频次，M 为该文档的单词数

$$IDF = \log(\frac{D}{D_w})$$

D 为文档数， D_w 为出现了该单词的文档数

$$TF - IDF = TF * IDF$$

实际分析得出 TF-IDF 值与一个词在留言信息描述表中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的职位描述表中文本的关键词。

3.1.6 构建朴素贝叶斯分类模型

朴素贝叶斯分类器建立在贝叶斯分类方法的基础上，其数学基础是贝叶斯定理，在贝叶斯分类中，确定一个具有某些特征的样本属于某类标签的概率，通常记为 P (L |特征)。

贝叶斯定理

$$P(AB) = P(A)P(B \mid A) = P(B)P(A \mid B)$$

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

当样本相互独立时有：

$$P(AB) = P(A)P(B)$$

基于贝叶斯定理，构建朴素贝叶斯分类模型如下：

$$P((\mid x) = \frac{P(y)P(x \mid y)}{P(x)} = \frac{P(y)}{P(x)}\prod_{i=1}^{d} P(x_i \mid y)$$

模型评价指标：

$$p = \frac{TN}{TN + FN}$$

P 为模型的准确度，TN 为被正确分类的样本数，FN 为错误分类的样本数，精确度越高，表示模型分类效果越好。

$$R = \frac{TN}{TN + FP}$$

R 为模型的召回率，FP 是指原来样本错误分类的的样本数，召回率越高，表示模型将误判的模型概率越低，模型效果越好。

$$F1 - SCORE = \frac{2PR}{R + P}$$

F1-SCORE 综合考虑精确度与召回率，是精确度和召回率的加权调和平均，F1-SCORE 越大，表示模型效果越好。

具体算法步骤如下

（1） 将处理好的留言信息据随机分成测试集和训练集；

（2） 将训练集进行模型构建，将测试集的维度转化成训练集的维度；

（3） 利用准确度对模型评价与优化。

3.1.7 模型分析与优化

通过 IF-IDF 算法将文本数据向量化，结合朴素贝叶斯分类构建了留言信息的分类模型，用测试集对分类模型进行验证，得到结果如下表所示：

	准确率
训练集	0.954
测试集	0.784

表格 2 模型精度

在模型的训练中，受样本个数限制，若某个属性值在训练集中没有与某个同类同时出现过，则连乘公式)则必为零，其他属性取任意值都不能改变这一结论，为了修正这种错误，采用了拉普拉斯平滑处理优化模型，计算公式如下：

$$\hat{p}(y)=\frac{|D_y|+1}{|D|+N}。$$

$$\hat{p}(x|y)=\frac{|D_{y,x}|+1}{|D_c|+N_i}$$

拉普拉斯平滑处理之后的模型训练集的精度明显提升，得到结果如下：

	准确率
训练集	0.929
测试集	0.885

表格 3 优化模型精度

再对优化后的模型输出混淆矩阵如图所示

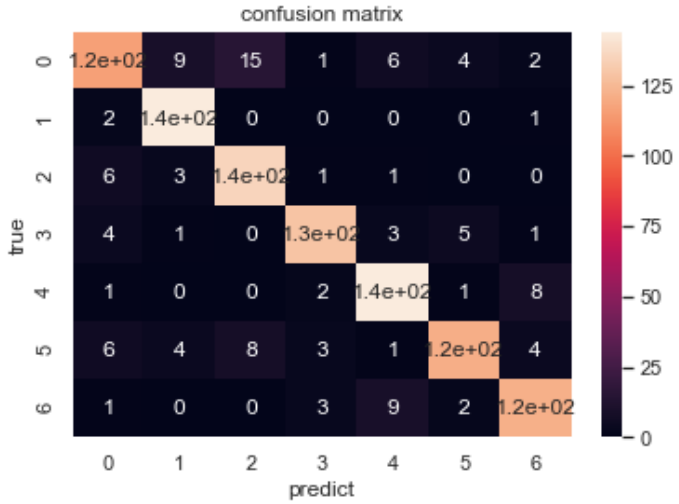


图 8 分类模型混淆矩阵

1.1 从上图我们可以看出，错误分类的数据较少，同时计算的得到的 R 和 F1-SCORE 值如下表所示，可以看出模型分类效果显著。

R	0.886
F1-SCORE	0.885

表格 4 模型评价指标

3.1.8 模型评价

（1） 算法逻辑简单,易于实现训练和预测的速度非常快，直接使用概率预测，通常很容易解释。

（2）实际上并不能做到属性之间相互独立，在属性个数比较多或者属性之间相关性较大时，分类效果不好。

3.2 问题二分析方法与过程

3.2.1 信息提取

根据题目所给数据信息，热点问题分析需要对地点人群进行信息提取，本文利用正则表达式，对留言信息中出现的地点进行提取，将提取出来的地点用列存

储，进而加入到 jieba 分词库里。

3.2.2 结巴分词

先将数据里文本内容进行去空格等处理，利用正则表达式提取到的地点信息，对 jieba 库进行扩充，利用扩充的 jieba 中文分词库，对数据中的留言信息进行分词，避免将地点分隔开，影响信息的完整性，从而造成信息丢失的情况。

3.2.3 去停用词

导入停用词表，利用停用词表对处理好的留言信息分词进行遍历过滤，清洗无用词及标点符号，保留我们能利用的数据，将清洗好的数据输出保存为 data3_yu.xlsx。

3.2.4 构建 LDA 主题模型：

LDA 是一种用于聚类离散数据集的概率模型，被广泛用于解决文本主题相关任务，主要应用于文本建模和文本分类。LDA 模型通常可以表示为一个包含词、文本和主题的 3 层贝叶斯概率模型，如图 9 所示.

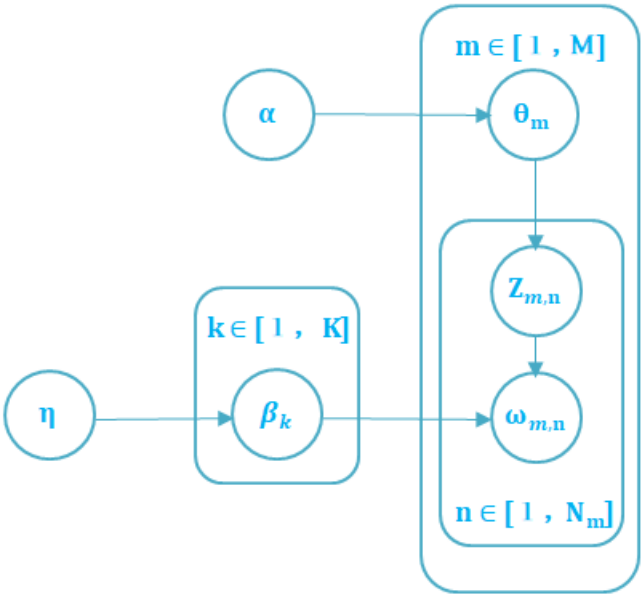


图 9 LDA 概率模型图

在 LDA 模型中， 从 Dirichlet 分布α中取样生成第 m 个文本的主题多项式分布θ_m；从θ_m中取样生成第 m 个文本的第 n 个词的主题$Z_{m,n}$；从 Dirichlet 分布η中取样生成主题$Z_{m,n}$对应的词语ω_m。其中$m \in [1, M]$，$n \in [1, N_m]$，$k \in [1, K]$，k，

M为待处理文本数，N_m为第m个文本的总词汇数，K为待分类主题数。

LDA 模型一般使用变分最大期望（variational EM）算法或 Gibbs 抽样方法进行参数估计，训练出上图中文本的主题分布概率θ_m，以及主题对应的词汇分布概率β_k。

定义词表大小为 V，由 N 个词构成的评论为$d=(w_1,w_2,.....w_n)$,假设留言信息集 D 有 M 条信息，记为$D=(d_1,d_2,.....d_m)$.M 条信息一共分类为 K 个主题。记α和β为狄利克雷函数的先验参数，θ为主题所在文档中的多项分布的参数，其服从超参数为α的 Dirichlet 分布，ϕ为词在主题中的多项分布的参数，其服从超参数为β的 Dirichlet 分布：

评价d_j条件下生成词w_i的概率表示为：

$$P(w_i \mid d_j)=\sum_{s=1}^{K} p(w_i \mid z=s)*P(z=s \mid d_j)$$

其中，$p(w_i \mid z=s)$表示词w_i属于 s 个主题的概率，$P(z=s \mid d_j)$表示第 s 个主题在评论d_j中的概率。

LDA 一般使用变分最大期望（variational EM）算法或 Gibbs 抽样方法进行参数估计，利用 Gibbs 抽样方法对 **LDA** 模型进行参数估计，依据下式：

$$P(z_i=s \mid z_{-i},W)\infty(n_{s,-i}+\beta)/(n_{s,-i}=s \mid z_{-i},\beta_i)*(n_{s,-i}+\alpha_s)$$

其中，$z_i=s$表示w_i属于第 s 个主题的概率，z_{-i}表示其他所有词的概率，$n_{s,-i}$表示不包括当前词w_i的被分配到当前主题Z_s下的个数，$n_{s,-i}$表示不包括当文档d_s的被分配到当前主题Z_s下的个数。

3.2.5　热度评价指标（rescore）：

$$rescore=\sum p(w_i \mid d_j)C_i$$

其中$C_i(i=1,2,...)$为每个主题内的关键词的词频。Rescore 值越大，反映出该问题越热门。

3.2.6 结果分析

对处理完的数据用 jieba 库进行分词后，统计词频，定义 rescore 评分指标，并加载 gensim 库中的 LDA 模型，利用 corpora.Dictionary 将数据做成词料包，训练 LDA 模型，分成 2000 个主题，并利用模型对原数据进行预测，将数据中前 5 个热点问题输出，结果如下：

热点问题 1		热点问题 2		热点问题 3		热点问题 4		热点问题 5	
d	w	d	w	d	w	d	w	教师节	w
工钱	0.114	A3 区	0.090	诈骗	0.100	A7 县	0.084	乱收费	0.114
A3 区	0.105	郝家坪	0.052	资金	0.077	人行道	0.057	物业	0.058
白泉村	0.073	地铁	0.045	西地省	0.071	交警	0.051	A 市	0.056
拖欠	0.062	暮云	0.033	开发商	0.067	停放	0.048	1000	0.050
坪塘镇	0.036	银杉路	0.021	投资	0.057	机动车	0.047	家长	0.043
领导	0.036	小学	0.021	刑法	0.056	罚款	0.044	补课	0.033
工人	0.027	扩建	0.020	经营	0.054	补偿	0.044	孩子	0.029
条例	0.022	增设	0.019	犯罪	0.038	冲突	0.037	一条	0.029

表格 5 排名前五的热点问题

从主题的关键词以及其权重可以分析出，前 5 热点问题分别是：

1.A3 区坪塘镇白泉村润泉山庄拖欠工钱

2.A3 区郝家坪小学何时扩建

3.西地省聚利人普惠投资有限公司涉嫌诈骗巨额资金

4.A7 县天华路凉塘路口出现交通事故

5.A 市外国语学校周日收费补课

前 5 热点问题的 rescore 分别为：

	热点问题 1	热点问题 2	热点问题 3	热点问题 4	热点问题 5
rescore	655.742	490.317	415.328	394.881	287.969

表格 6 热点问题 rescore 值

3.3 问题三分析方法与过程

3.3.1 数据预处理

针对题目所给数据，首先对留言内容和答复意见内容去除空格，接着利用jieba分词库对这两项内容进行分词处理，进而导入停词表对粉刺结果进行过滤，完成数据的清理。

3.3.2 回复意见分析

利用 jieba 库中的分析库，对处理好的数据进行词频统计，将回复信息绘制词云图，词云图如下所示：

图 10 回复信息词云图

从词云图中我们可以直观看出，政府回复留言信息的时候，大多会对每个网友提出的留言信息进行针对性的回复，相关性较高，回复问题也相对完整，但在探索数据的时候也能发现，有些信息是已转发字样，或者是设置自动回复信息，这是政府回复工作中需要提升的。

4 参考文献：

[1]王振振，何明，杜永萍. 基于 LDA 主题模型的文本相似度计算[C]// 全国智能信息处理学术会议. 2013.

[2]黄俊衡. 基于改进主题模型的微博短文本情感分析的研究[D]. 2017.

[3]杨成鹏，高占春，蒋研军. 基于朴素贝叶斯的文本挖掘算法的在 GPU 上的设计与实现[J]. 2013.

[4]姜宁. 贝叶斯层次聚类及其在文本挖掘中的应用[D]. 2001.

[5]王栋. 基于文本挖掘的短信分类技术的研究与实现[D]. 2013.

[6]许高建. 基于 Web 的文本挖掘技术研究[J]. 计算机技术与发展, 2007, 017(006):187-190.
[7]王珍珍. 关于文本挖掘中文本分类与文本聚类的研究[J]. 科技信息, 2007, 000(006):55.