“智慧政务”中的文本挖掘应用

摘 要： 第一问，首先建立文本分类模型:在文本预处理、可视化、TF-IDF词频统计并检验相关性后,选取准确率最高的线性分类支持向量机模型对提取的特征进行训练，并通过F1值、混淆矩阵检验模型的优劣；第二问中，将留言详情文本预处理后，运用doc2vec模型训练出文档向量，通过文档向量来计算不同文本的文本相似度。然后根据文本相似度进行文档聚类，筛选出同一类里留言数量排名前10的问题。再结合同一类问题中的留言点赞数和反对数，得出热度指数，重新排名，取热度指数前五的问题；第三问，从答复的回答形式、回答内容以及回答效用三个方面提出了及时性、直观性、相关性、完整性、可解释性五个二级指标，并由此提出了政务答复意见的质量评价得分模型。

关键词：文本分类；文本相似度；doc2vec；答复质量评价

目录

0 绪论

近年来，随着网络问政平台逐步成为政府了解民意的重要渠道，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。本文从群众留言分类、热点问题挖掘、答复意见的评价三个问题入手，研究“智慧政务”中的文本挖掘应用。

1 群众留言分类

群众留言分类一直是困扰政府部门的一大难题，本文提出的留言文本分类方法[1]包括以下过程： ①导入给出的文本数据；②文本预处理，包括中文分词和建立停用词库； ③文本可视化，生成各类的词云；④利用 TF-IDF 提取特征值，并进行统计学检验，并选择分类器最终确定最优模型。具体过程如图 1-1 所示。

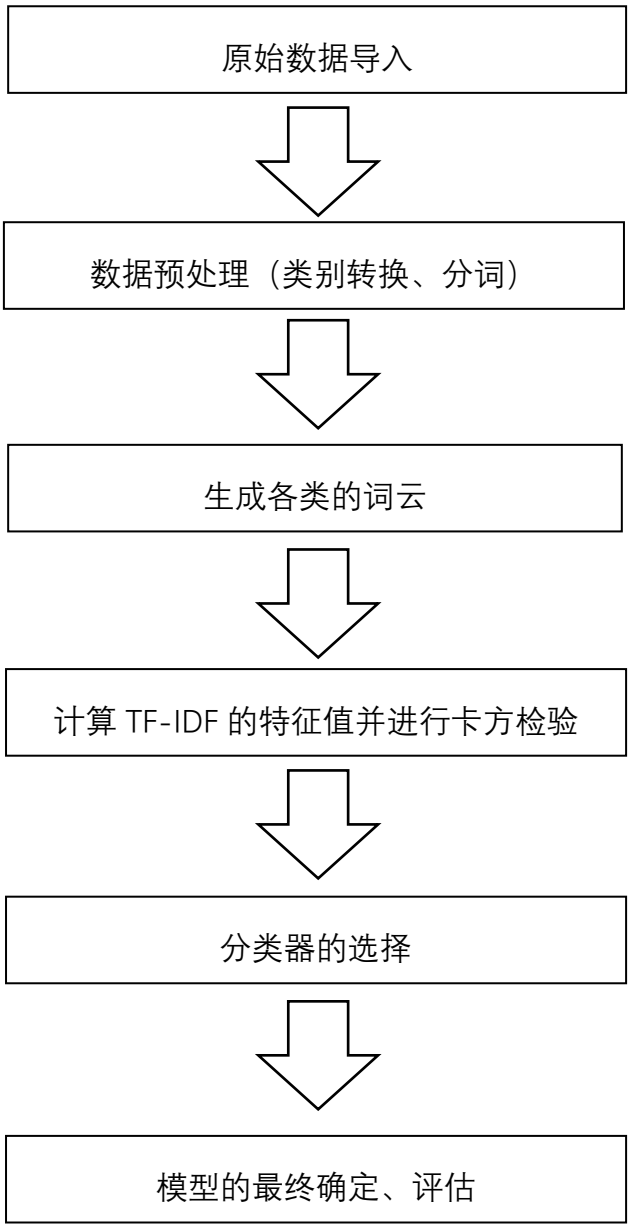

图 1-1 留言文本分类方法

1.1 文本预处理

文本预处理是文本分类中至关重要的一步，中文分词的结果以及停用词的存在都会直接影响特征提取的结果，进而影响文本分类的效果。

1.1.1 类别转换

将中文的类别转换为数字型，以便后续分类预测，具体转换方式如下：

Out[16]:

	cat	cat_id
0	城乡建设	0
1	环境保护	1
2	交通运输	2
3	教育文体	3
4	劳动和社会保障	4
5	商贸旅游	5
6	卫生计生	6

图 1-2 类别转换

1.1.2 数据的初步认识

本文使用的数据为群众留言文本及其分类，在第一问中相关的数据为群众留言及其分类所在列，因此提取该两列并按类别统计数据维度，如下图 1-3 所示：

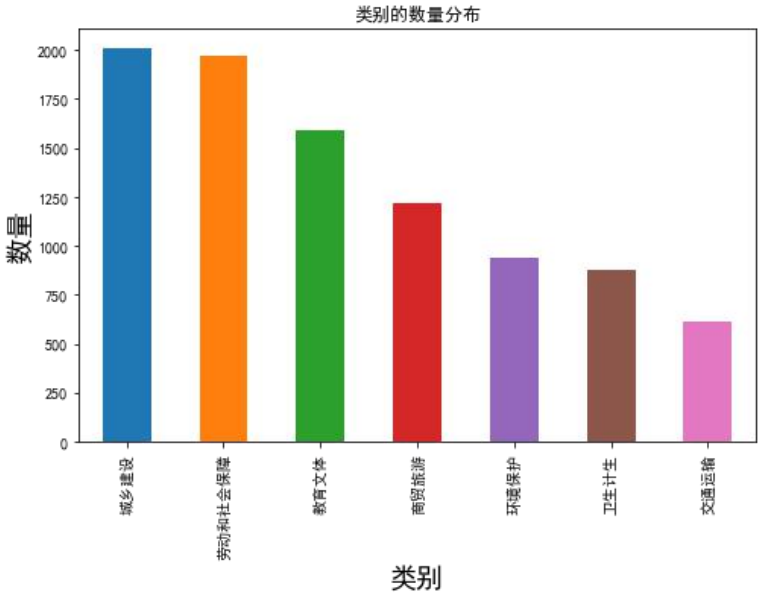

图 1-3 类别的数量分布

可以发现，数据中城乡建设、劳动和社会保障类所占比例较大，而交通运输类所占比例很小。初步认知数据对后续的数据处理有一定帮助。

1.1.3 中文分词并处理停用词

中文分词技术是自然语言处理领域中很多关键技术的基础，包括文本分类、信息检索、信息过滤等。现如今中文分词的方法很多，大多数研究者的研究重点都是按照提升算法的精度、速度来的，常用的算法可总结如下：字典分词方法、理解分词方法和统计分词方法[2]。本文采用业界比较知名的 Jieba 中文分词对留言文本进行分词，并去除掉标点符号。

在处理自然语言数据之前，为了节省储存空间、提高搜索效率，往往需要过滤掉某些经常出现却不能反映实际内容的常用词，这类词被称作 Stop Words（停用词）。通常意义上，停用词分为两类：一类是中文语言中的功能词，使用普遍却无实际作用，如“我们”、“在”等；另一类是表达情感的词语，在句子中只是起到加强语气的作用，如“得”、“地”等。因此通过对停用词的处理，能够显著提高特征词的文本质量，进一步提高后续分类的准确率。

考虑到如果直接使用留言内容作为语料库，可能会引起两个类别的留言分到同一类，即存在误分类情况。为了减少这种情况，本文根据换行符等特殊标记将整个留言内容进行分段，通过段落信息构建语料库最终得到的分词结果 cut_review 部分如下图 1-4 所示：

	review	cat	cat_id	clean_review	cut_review
0	A3区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项...	城乡建设	0	A3区大道西行便道未管所路口至加油站路段人行道包括路灯杆被圈西湖建筑集团燕子山安置房项目施工...	A3 区 大道 西行 便道 未管 路口 加油站 路段 人行道 包括 路灯 杆 圈 西湖 建筑...
1	位于书院路主干道的在水一方大厦一楼至四楼人为拆除水、电等设施后，烂尾多年，用护栏围着，不但占...	城乡建设	0	位于书院路主干道的在水一方大厦一楼至四楼人为拆除水电等设施后烂尾多年用护栏围着不但占用人行道...	位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电 设施 后 烂尾 多年 ...
2	尊敬的领导：A1区苑小区位于A1区火炬路，小区物业A市程明物业管理有限公司，未经小区业主同意...	城乡建设	0	尊敬的领导A1区苑小区位于A1区火炬路小区物业A市程明物业管理有限公司未经小区业主同意利用业...	尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 A 市程明 物业管理 有...
3	A1区A2区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味，大家都知道...	城乡建设	0	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 现在 自来水 龙头 ...
4	A1区A2区华庭小区高层为二次供水，楼顶水箱长年不洗，现在自来水龙头的水严重霉味，大家都知道...	城乡建设	0	A1区A2区华庭小区高层为二次供水楼顶水箱长年不洗现在自来水龙头的水严重霉味大家都知道水是我...	A1 区 A2 区华庭 小区 高层 二次 供水 楼顶 水箱 长年 不洗 现在 自来水 龙头 ...

图 1-4 中文分词及停用词处理后结果

1.2 词云的生成

本文经过分词以后生成了 cut_review 字段。在 cut_review 中每个词语中间都是由空格隔开，接下来要在 cut_review 的基础上生成每个分类的词云，在每个分类中罗列前 100 个高频词，并画出这些高频词的词云。词云图中占比最大的词与该类往往有密切的联系。具体的分类词云图见图 1-5：

图 1-5 各类的词云图

1.3 词频统计及卡方检验

为了进一步准确分类，本文选用 TF-IDF 词频统计方法抽取特征，并用卡方检验对所构建的特征进行拟合度、关联度检验，找出每个分类中关联度最强的两个词和两个词语对。这些词和词语对能很好的反映出分类的主题。

1.3.1 词频统计 TF-IDF

TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术。TF 意思是词频，IDF 意思是逆文本频率指数。

TF 指的是某一个给定的词语在该文件中出现的次数，这个数字通常会被归一化(一般是词频除以文章总词数)，以防止它偏向长的文件[3]。（同一个词语在长文件里可能会比短文件有更高的词频，而不管该词语重要与否）。

TF 的计算公式如下：

$$TF_w = \frac{N_w}{N}$$

其中N_w表示的是某一文本中词条 w 出现的次数，N 是该文本总词条数。

而 IDF 逆向文件频率反应了一个词在所有文本（整个文档）中出现的频率，如果一个词在很多的文本中出现，那么它的 IDF 值应该低。而反过来如果一个词在比较少的文本中出现，那么它的 IDF 值应该高。

IDF 的计算公式如下：

$$IDF_w = \log\left(\frac{Y}{Y_w + 1}\right)$$

其中 Y 是语料库的文档总数，Y_w是包含词条 w 的文档数，最终的 TF-IDF 就是：

$$TF - IDF_w = TF_w * IDF_w$$

如果单单以 TF 或者 IDF 来计算一个词的重要程度都是片面的，因此 TF-IDF 综合了 TF 和 IDF 两者的优点，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。上述引用总结就是,一个词语在一篇文章中出现次数越多，同时在所有文档中出现次数越少，越能够代表该文章，越能与其它文章区分开来。

本文使用 sklearn.feature_extraction.text.TfidfVectorizer 方法来抽取文本的 TF-IDF 的特征值。这里使用了参数 ngram_range=(1,2),这表示除了抽取评论中的每个词语外,还要抽取每个词相邻的词并组成一个“词语对”,如: 词 1，词 2，词 3，词 4，(词 1，词 2)，(词 2,词 3)，(词 3，词 4)。这样就扩展了特征集的数量,有了丰富的特征集才有可能提高分类文本的准确度。

以此方法构建出的特征数量达到 776013，表示 9210 条评论数据的所有词语和相邻两个单词的组合词语对的总数，这样就完成了特征的提取。

1.3.2 卡方检验

经典的卡方检验是检验定性自变量对定性因变量的相关性。假设自变量有 N 种取值，因变量有 M 种取值，考虑自变量等于 i 且因变量等于 j 的样本频数的观察值与期望的差距，构建统计量[4]，其中 A 为实际数，E 为理论值：

1.4 分类器的选择

本段中使用上述生成的 TF-IDF 向量，训练多种分类器（朴素贝叶斯、K 邻近、逻辑回归、线性分类支持向量机），通过对比准确率得出最优的模型，并确定该模型的 F1 值。

1.4.1 朴素贝叶斯

朴素贝叶斯的思想基础是这样的：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。朴素贝叶斯方法基于概率进行预测，将各属性作独立化处理[5]。算法的核心为贝叶斯公式，后验概率可以由先验概率和另一事件的可能性比率乘积表示：

$$P(B|A) = \frac{P(B) * P(A|B)}{P(A)}$$

即如果 A 类中某一特征极其明显（概率极大），那么这种方法就会将拥有此特征的数据判定为 A 类。朴素贝叶斯法逻辑简单且分类性能稳定，但如果特征之间相关性较大，分类的效果将大大降低,该数据集的特征就存在很大的相关性，最终的结果印证了这一点。

1.4.2 K-邻近

KNN 算法不同于上述算法，它并不需要进行训练，当新的样本输入后，直接在数据中找 k 个最近的样本，如果这些样本大多数属于某一类别，那么该样本被判定为这个类别[6]。具体的步骤如下：

（1） 计算测试的样本与各个数据之间的距离（一般使用欧式距离）。

（2） 选取距离最小的 k 个点（k 值需要指定，k 值的变动影响分类结果）。

（3） 返回 k 个点中出现频率最高的类别作为测试样本的类别。

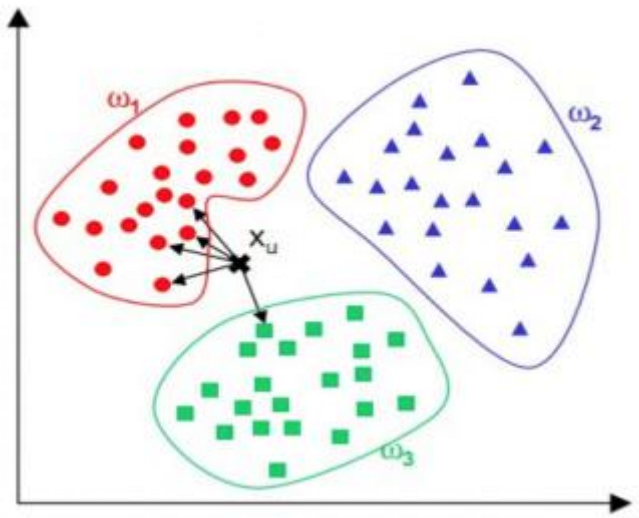

图 1-6 KNN 算法示意图

1.4.3 多分类 Logistic 模型

多类别逻辑回归模型，实际上是对于所有 K 个可能的分类结果，运行 K-1 个独立二元逻辑回归模型，在运行过程中把其中一个类别看成是主类别，然后将其它 K-1 个类别和所选择的主类别分别进行回归。通过这样的方式，如果选择结果 K 作为主类别的话，可以得到以下公式[7]：

$$P(y = K - 1|x) = P(y = K|x)e^{k_{n,1}x_n}$$

由于所有概率的和为 1，所以可以得到：

$$P(y = K - 1|x) = \frac{e^{k_{n,K-1}x_n}}{1 + \sum_{m=1}^{K-1} e^{k_{n,m}x_n}}$$

1.4.4 LinearSVC 线性分类支持向量机

支持向量机分类器是根据训练样本的分布，搜索所以可能的线性分类器中最佳的那个，决定分类边界位置的样本并不是所有训练数据，是其中的两个类别空间的间隔最小的两个不同类别的数据点，即“支持向量”。从而可以在海量甚至高维度的数据中，筛选对预测任务最为有效的少数训练样本。

1.5 模型的确定与评估

接下来尝试不同的机器学习模型,并评估它们的准确率，本文将使用上述介绍的四种模型并根据其准确率选出最优模型，并对模型进行评估。

1.5.1 模型的选取

依据上文介绍的四种机器学习方法：朴素贝叶斯、K-邻近、多分类 Logistic 模型以及线性分类支持向量机，结合箱线图得出对应模型的准确率。

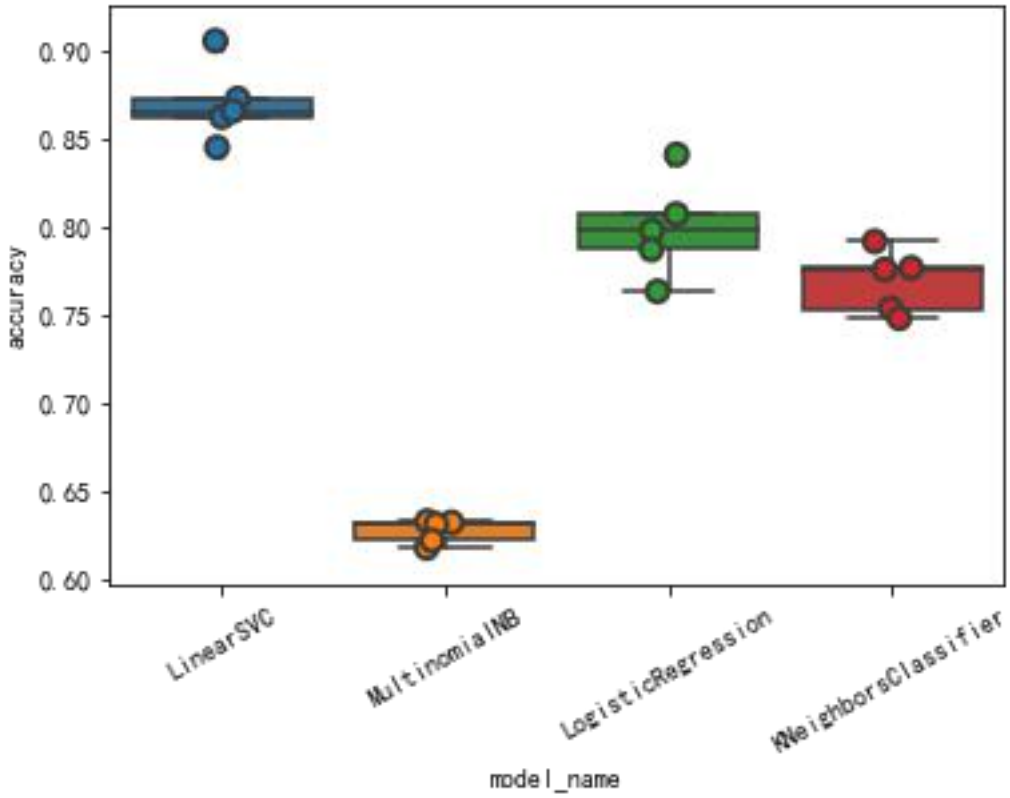

图 1-7 四种方法准确率对比图

从箱线图可知：由于特征的相关性，朴素贝叶斯模型的准确率最低，而线性

分类支持向量机平均准确率最高，具体的准确率如下表所示。

LinearSVC	0.870692
Logistic 回归	0.799571
KNN	0.769391
朴素贝叶斯	0.627143

表 1-2 四种方法的准确率

1.5.2 模型的评估

根据上述分析，本文确定一级分类的模型：线性分类支持向量机。下面针对平均准确率最高的 LinearSVC 模型，查看混淆矩阵，并显示预测标签和实际标签之间的差异。

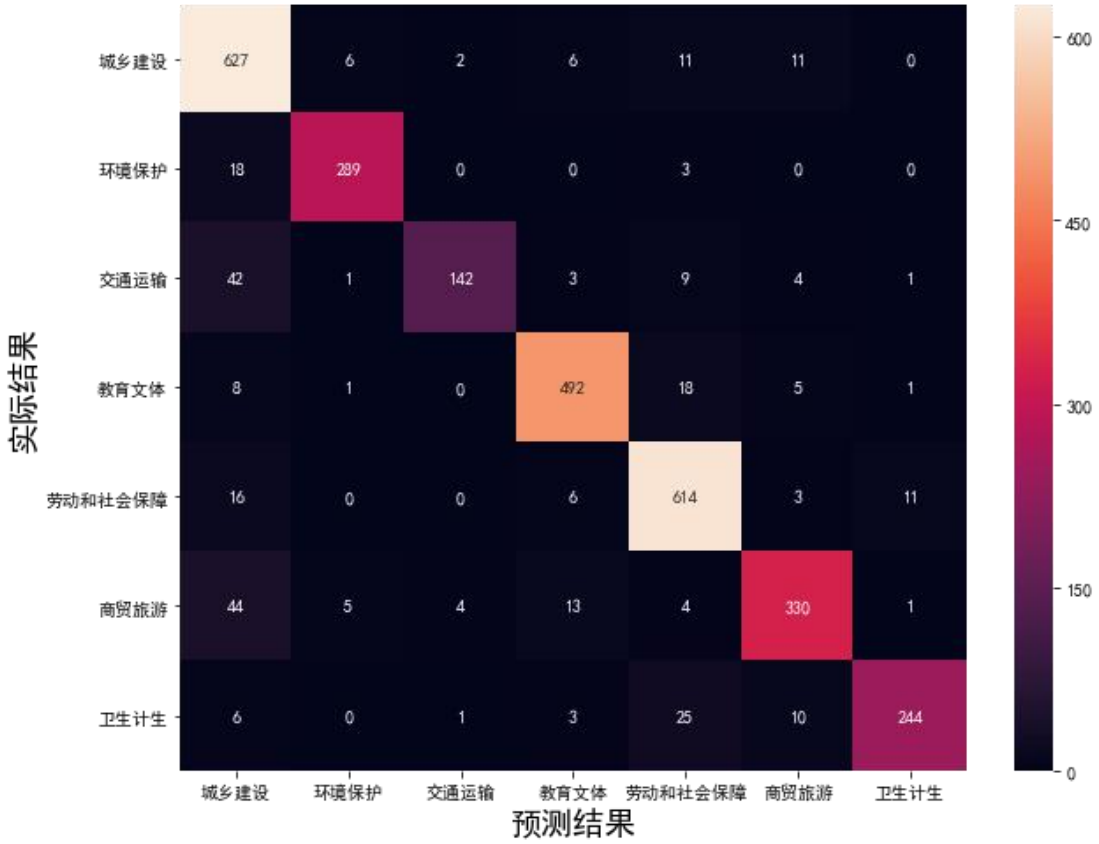

图 1-8 LinearSVC 的混淆矩阵

混淆矩阵的主对角线表示预测正确的数量，其余为预测错误的数量。然而多分类模型一般不采用准确度评估模型，因为准确度不能反映每一类的实际精度。因此需要借助 F1 分数来评估模型：

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

其中P_i为第 i 类的查准率，R_i为第 i 类的查全率。通过预测结果可知：模型的总平均 F1 能达到 0.90066，每一类对应的 F1-Score 如下图，可见模型能为群众留言分类提供一定的帮助。

accuracy 0.9006578947368421

	precision	recall	f1-score	support
城乡建设	0.82	0.95	0.88	663
环境保护	0.96	0.93	0.94	310
交通运输	0.95	0.70	0.81	202
教育文体	0.94	0.94	0.94	525
劳动和社会保障	0.90	0.94	0.92	650
商贸旅游	0.91	0.82	0.86	401
卫生计生	0.95	0.84	0.89	289
avg / total	0.90	0.90	0.90	3040

图 1-9 LinearSVC 的精度

2 热点问题挖掘

在对留言进行分类之后，为了找出某段时间内群众反映较多的热点问题，有利于政府更有针对性地满足群众的需求，更好地为群众服务。本文采用 doc2vec 技术训练句子向量模型，用句子向量来计算文本之间的相似性，用句子向量来对文本进行聚类分析，从而找出有关同一问题留言出现次数较多的一组文档，标注为热点问题。

热点问题挖掘过程如下：①导入给出的文本数据；②文本预处理，包括中文分词和建立停用词库，将文本转化为词列表形式； ③利用互信息和左右信息熵，对文本进行短语提取，优化分词效果；④将预处理好的文本输入到 doc2vec 模型中进行训练得到句子向量。⑤根据留言之间相似度对句子向量进行聚类。具体过程如图 2.1 所示。

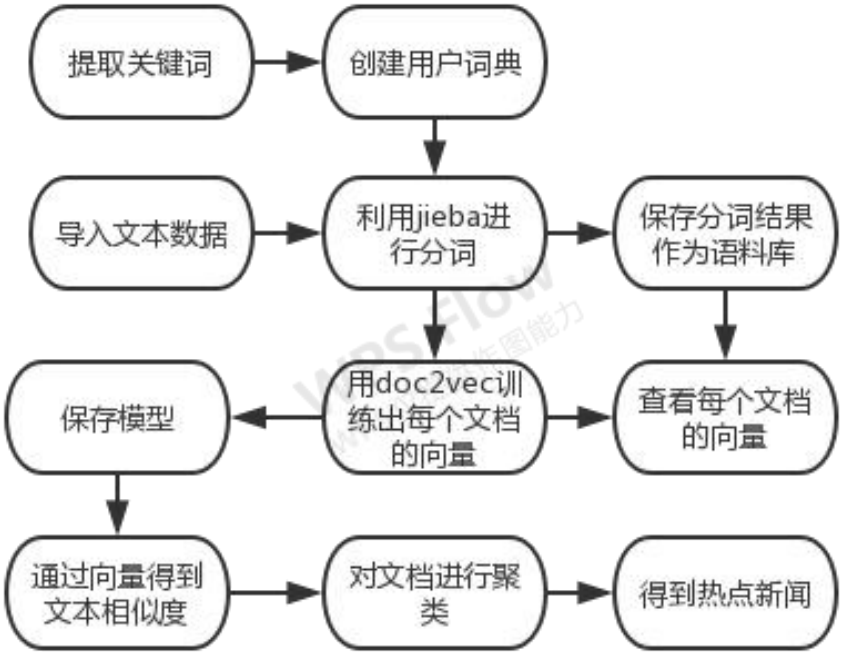

图 2-1 问题二处理流程

2.1 利用 TextRank 模型进行关键词提取

TextRank 是一种基于图的用于文本的排序算法，基本思想来自于 Google 的 PageRank 算法[9]。类似于网页的排名，对于词语可得到词语的排行，对于句子也可得到句子的排名，所以 TextRank 可以进行关键词提取，也可以进行自动文摘。通过词之间的相邻关系构建网络，然后用 PageRank 迭代计算每个节点的 rank 值，排序 rank 值即可得到关键词[9]。PageRank 本来是用来解决网页排名的问题，网页之间的链接关系即为图的边，迭代计算公式如下：。

TextRank 算法的主要步骤如下：

（1）预处理：分割原文本中的句子得到一个句子集合，然后对句子进行分词以及去停用词处理，筛选出候选关键词集。

（2）设置一个长度为 N 的滑动窗口，所有在这个窗口之内的词都视作词结点；则 TextRank 构建的词图为无向图。

（3）TextRank 的迭代计算公式如下：

$$WS(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum_{v_k \in Out(V_j)} w_{jk}} WS(V_j)$$

在用 TextRank 算法提取出留言中的关键词后，将关键词添加到分词字典中，这样能够提高分词的效率和准确率。

2.2 doc2vec 模型

Doc2vec 模型是 word2vec 词嵌入技术的拓展，训练句向量的方法和词向量的方法非常类似[8]。训练词向量的核心思想就是说可以根据每个单词的上下文预测，也就是说上下文的单词对是有影响的，是一个无监督框架，学习文本段落的连续分布向量表示，文本可以是可变长度的，从句子到文档。该方法可以应用于可变长度的文本，任何句子或大型文档。通过该方法将文本进行数字化表示，每个词语使用几百维的实数向量表示，然后使用欧式公式计算两个词语的语义相似性[8]。

Doc2vec 有两种模型：Distributed Memory Model of Paragraph Vectors(PV-DM) 模型 Distributed Bag of Words version of Paragraph

Vector(PV-DBOW)模型[8]。PV-DM 模型的目的是根据上下文周围的词语预测一个词语，通过训练语料，最后获得模型参数，得到词向量。后者正好相反，忽略输入的上下文，让模型去预测段落中的随机一个单词。就是在每次迭代的时候，从文本中采样得到一个窗口，再从这个窗口中随机采样一个单词作为预测任务，让模型去预测，输入就是段落向量。本文的实验采了 PV-DM 模型训练词向量。

每条留言详情进行分词处理后，映射到向量空间中，可以用一维向量来表示。每个单词同样被映射到向量空间，可以用矩阵的一列来表示。然后将段落向量和词向量级联或者求平均得到特征，预测句子中的下一个单词。

这个段落向量/句向量也可以认为是一个单词，它的作用相当于是上下文的记忆单元或者是这个段落的主题，所以我们一般叫这种训练方法为 Distributed Memory Model of Paragraph Vectors(PV-DM)

在训练的时候固定上下文的长度，用滑动窗口的方法产生训练集。段落向量/句向量在该上下文中共享。

在进行多次试验后，调整参数。将词向量维度设置成 150，句子中当前单词和被预测单词的最大距离即窗口大小设置成 20，将词频小于 5 的单词忽略，设置负采样为 5。训练得到的模型最符合数据真实情况。由此得到了不同的留言的向量形式，并将留言向量保存。

2.3 文本相似度的计算

本文将逆文本率与余弦相结合，在句子相似性匹配中，核心词起的作用要更大些。而逆文本频率就是衡量一个词的重要性，这样可将一个词的 idf 值作为权重参与计算[10]。这样计算词的 idf 和句子向量的公式为：

$$\mathrm{idf}(w)=\log(\frac{D}{D_w+1})$$

其中，D 为语料库中文档的矢量，D_{w} 为词出现的文档数量。

$$\mathrm{vec}tor(s)=\sum_{i}^{m}v(w_i)*idf(w_i)$$

其中，$v(w_i)$ 为留言的第 i 个词 w_i 的向量，$\mathrm{idf}(w_i)$ 为词 w_i 的 idf 值。

按上述方式求得句子向量后，按照余弦相似度计算方法：

$$\cos(\theta)=\frac{V_1 \bullet V_2}{\|V_1\|*\|V_2\|}$$

求得留言中的相似度。

2.4 文本聚类

求得留言与留言之间相似度后，设置一个相似度阈值，将与一条留言相似度大于该阈值的留言与该留言归为一类。这样就将留言分成了很多类，筛选出类中留言的个数排名前十的留言类。这样就挖掘出了热点问题。

经多次试验后，最终确定的相似度阈值为 0.65。将在该相似度阈值之上的留言归为一类后，得到一个留言类中同一个问题的可能性最大。

2.5 热度指数计算

在筛选出来的十个问题中存在相同留言数目的问题。为了得到准确的排名，就将该问题下留言的支持数和反对数纳入考虑范围。考虑到留言的支持数与反对数对于热度影响较大。将每个问题的留言数目加上该问题的支持数和反对数。将每个问题的该值联合起来形成一个十维向量，而后进行归一化运算。从而得到每个问题的热度系数，再将热度系数前五的问题挑选出来后，这五个问题就是最终的 5 个热点问题。

3 关于政务答复意见的质量评价得分模型

针对相关部门对留言的答复意见，本文从答复的回答形式、回答内容以及回答效用三个方面提出了及时性、直观性、相关性、完整性、可解释性五个二级指标[11]，并由此提出了政务答复意见的质量评价得分模型。如下图所示：

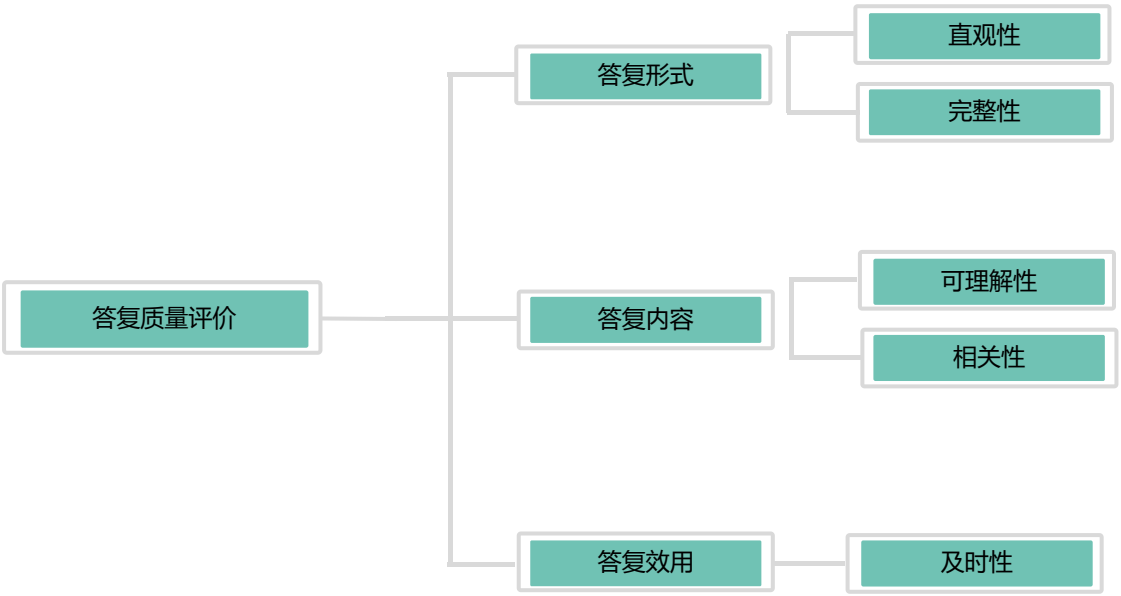

图 3-1 政务答复意见的质量评价得分模型

3.1 基于答复的及时性特征

留言与答复之间的时间间隔越短，回答的及时性越好。因此本文将及时性特征得分设置为减分项，由留言时间和答复时间的差值直接得出，部分数据如下所示：

及时性得分
-15.22550926
-14.73993056
-14.75636574
-14.77930556
-15.70012731
-31.05888889
-40.93443287
-28.52153935
-16.23244213
-16.23965278

图 3-2 答复及时性得分

3.2 基于答复的直观性特征

留言答复的文本长度越短、段落划分越清晰，直观性越好。本文赋予了不同文本长度和不同段落数对应的得分，并以此来体现答复的直观性特征。

3.3 基于答复的相关性特征

为了判断留言与答复内容之间的相关性，本文使用 TF-idf 构造出词袋模型，并对每一条留言与答复特征进行向量化数字映射，再对数字映射后的特征进行余弦相似度的匹配：

$$\cos(\theta)=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

由此使用 sklearn.metrics.pairwise 中的 cosine_similarity 来得到每一条留言与答复的文本内容相关性矩阵，并计算对应的分值。

3.4 基于答复的完整性特征

中文文本对于答复的完整性定义过于主观，经尝试发现对于完整性的打分难以契合文本本身的完整性，因此本文不再讨论答复的完整性。

3.5 基于答复的可解释性特征

答复中没有方言、网络语言、专业名词等，不同文化程度的用户都可以读懂，则可解释性高。

因此，本文将可解释性特征也设置成减分项。本文统计了一定数量的方言、网络言语以及专业名词，将其整合成一张表，并将分词后的答复与表进行匹配，若有匹配到，则按照一定的分数和对应的数量进行减分。

4 参考文献

[1] 刘鑫昊, 谭庆平, 曾平, 唐国斐. 几种基于 MOOC 的文本分类算法的比较与实现[J]. 软件, 2016, 37(09): 27-33.

[2] 李峰,柯伟扬,盛磊,陈雯,陈丙赛,罗韵晴.Doc2vec 在政策文本分类中的应用研究[J].软件,2019,40(08):76-78.

[3] 黄春梅,王松磊.基于词袋模型和 TF-IDF 的短文本分类研究[J].软件工程,2020,23(03):1-3.

[4] 胡思才,孙界平,琚生根,王霞,龙彬,廖强.基于扩展的情感词典和卡方模型的中文情感特征选择方法[J].四川大学学报(自然科学版),2019,56(01):37-44.

[5] 张航.基于朴素贝叶斯的中文文本分类及 Python 实现[D].山东师范大学,2018.

[6] 仲媛,杨健,涂庆华,李小舟.KNN-均值算法[J].现代计算机(专业版),2014:45-49.

[7] 李卓冉.逻辑回归方法原理与应用[J].中国战略新兴产业,2017(28):125-126.

[8]李峰,柯伟扬,盛磊,陈雯,陈丙赛,罗韵晴.Doc2vec 在政策文本分类中的应用研究[J].软件,2019,40(08):76-78.

[9] 徐立.基于加权 TextRank 的文本关键词提取方法[J].计算机科学,2019,46(S1):142-145.

[10]张旭,孙玉伟,成颖.不同特征对文本聚类效果的比较研究——以新闻文本为例[J].情报理论与实践,2020,43(01):169-176.

[11]袁红,张莹.问答社区中询问回答的质量评价——基于百度知道与知乎的比较研究[J].数字图书馆论坛,2014(09):43-49.