“智慧政务”中的文本挖掘应用

摘要

近年来，随着众多的网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的一个重要渠道，各类社情民意相关的文本数据量不断增加，在当前大数据、人工智能流行的时代，建立基于自然语言处理及文本挖掘的智慧政务系统，对于提升政府的管理水平和施政效率具有极大的推动作用。

对于问题一：通过欠抽样对留言内容进行随机抽取，减少过多的某些类别样本量来保持数据分布平衡。利用数据清洗去除留言内容中一些重复内容，空白内容及脱敏处理后的 x 序列。利用 python 中文分词模块 jieba 对留言信息进行分词，并通过 TF-IDF 算法将留言信息转换为权重向量来从留言信息文本中提取特征。得到留言文本的向量表示后，通过 F-Score 来选择最终的最佳分类模型。最终选择用线性支持向量机来建立留言内容的一级标签分类模型。

对于问题二：对留言数据预处理后，中文分词和特征提取后。使用 DBSCAN 算法来对留言内容进行聚类，将同一问题归类到同一类别。建立热度指标因素， 利用 Topsis 算法根据热度指标因素对每一类问题进行热度指标评分并将其排序，取出热度前五的五类问题。再使用 Hanlp 提取地名人群完成热度问题表。

对于问题三：首先建立一套评价指标体系，通过分析设立几个指标作为评分的标准。再抽取一定数量数据进行类别的人工标注，分别使用机器学习分类法和综合权重法来构建评价体系。机器学习分类法使用最佳分类模型对所有留言回复进行预测；综合权重法利用机器学习中的随机森林根据人工标注的数据对几个指标的重要性进行排序得出各指标权重，最后通过该权重求得每条回复的综合得分并进行评价。

关键词:中文分词;TF-IDF 算法;SMOTE 算法;LDA;DBSCAN 算法;Topsis 算法;机器学习

目录

1.挖掘目标

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成为了社会治理创新发展的新趋势。

本次建模的目标是利用所收集的来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，在对留言记录及留言答复意见进行基本的预处理、中文分词、停用词过滤、数据分布平衡后，利用线性支持向量机、DBSCAN 算法、Topsis 算法机器学习来达到以下三个目标：

（1） 根据附件 2 所给数据，建立关于留言内容的一级标签分类模型，对留言内容进行分类；

（2） 根据附件 3 将某一时段内反映特定地点或特定人群问题的留言进行归类，定义合理的热度评价指标，得出排名前 5 的热点问题以及时发现热点问题并解决；

（3） 针对附件 4 相关部门对留言的答复意见，设计一套评价方案并对答复意见进行评价。

2.群众留言分类

2.1 流程图

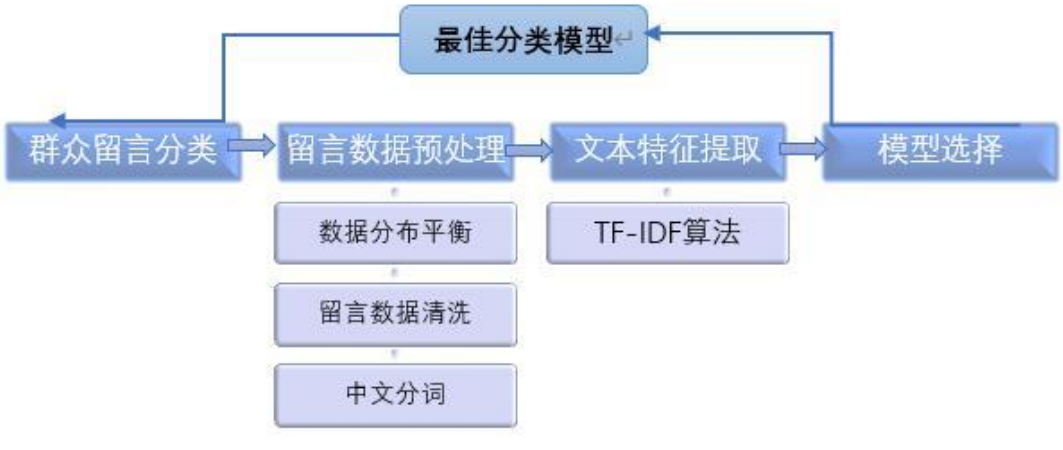


图 2.1 问题 1 流程图

2.2 数据预处理

2.2.1 留言数据分布平衡

首先根据附件 2 中的一级标签来查看其数据类别分布。数据类别分布图如下：

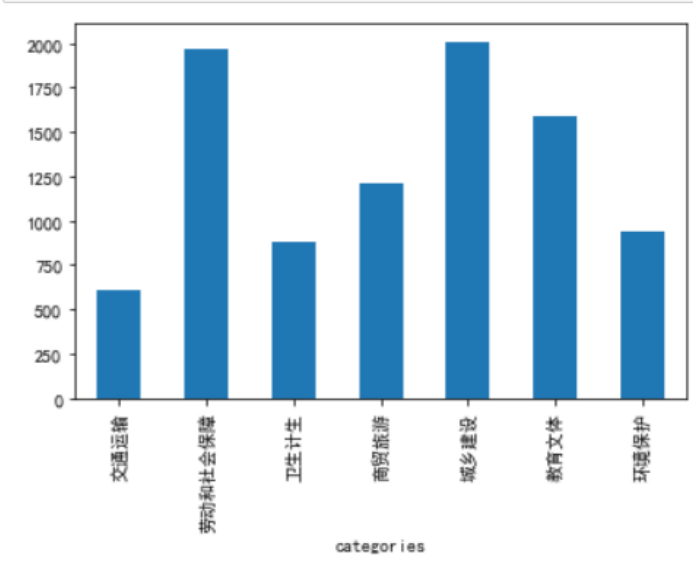


图 2.2 数据类别分布图

根据图 2.2 可以观察到每种类别的分布是不平衡的，数据类别更大程度上倾向于劳动和社会保障及城市建设。考虑到数据分布不均衡可能会导致样本量少的分类所包含的特征过少，并且得到的分类模型也容易出现过拟合的现象，预测结果偏向类别数更多的一类。当模型应用到新的数据集上，分类模型的准确性会大大降低。因此实验采用 SMOTE 算法，通过使用 imlbearn 库中上采样方法中的 SMOTE 接口来生成等量数据类。然后再使用类别分布平衡的新数据去进行后面的数据挖掘。

2.2.2 留言数据清洗

在赛题给出的数据中，考虑到留言信息中可能存在着重复内容，空内容，匿名隐藏修改的 x 序列（如手机号、银行卡号等序列被隐藏为 x 字符串）的问题，从而影响后续的数据分析，因此在进行分词前先对所给数据进行基本处理。

保留所需要进行分析的必要数据，增加一列整数化的一级标签方便后续分类使用。同时创建了几个字典对象保存类标签和一级分类的映射关系，以供后续使用。

2.2.3 对中文分词

在对留言信息进行分析时，分析的对象是中文文本的非结构化信息，在程序代码使

用中并不能够很好的被使用，因此第一步要把其中该文本信息转换为计算机能够识别的结构化信息。附件 2 以中文文本的形式给出了需要进行分析的数据。为了便于转换，要先对这些留言信息进行中文分词。这里采用了 Python 中的一个中文分词模块 jieba 对留言内容进行中文分词。分词过程中使用了自定义字典以保证分词时一些需要使用的词语组合不被分开。并且在分词之后还需过滤掉一些没有意义的停用词。其中分词所用到的自定义字典保存在 newdic1.txt 中；所用到的停用词字典保存在 stopword.txt 中。

2.3 文本特征提取

由于分类模型无法对文本的原始形式进行直接处理，其输入希望能是一个长度固定并且为数值型的特征向量。因此我们需要从留言信息文本中提取特征，这里使用了 TF-IDF 算法来将留言信息转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（TermFrequency）。

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数，即:

$$\text{词频(TF)} = \frac{\text{某个词在文章中的出现次数}}{\text{文章的总词数}}$$

第二步，计算 IDF 权重，即逆文档频率（InverseDocument Frequency）。IDF 越大，此特征词在文本中的分布越集中，说明该特征词在区分该文本内容属性能力越强

$$\text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数} + 1}\right)$$

第三步，计算 TF-IDF 值（TermFrequencyDocumentFrequency）

$$\text{TF} - \text{IDF} = \text{词频(TF)} \times \text{逆文档频率（IDF）}$$

TF-IDF 算法的主要思想是：如果某个单词在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。TF-IDF 值与一个词在留言信息表中文本出现的次数成正比，某个词文本的重要性越高，其 TF-IDF 值越大。通过计算数据集中每个词的 TF-IDF 值，得到每个留言信息的 TF-IDF 向量。

2.4 模型选择

当得到了留言文本的向量表示后，可以通过划分测试集和训练集来得到分类模型。并且尝试多种不同的机器学习模型，评估其性能以得到问题 1 的最佳分类模型。作为已经有标签的有监督学习分类问题，我们尝试使用以下五种模型。

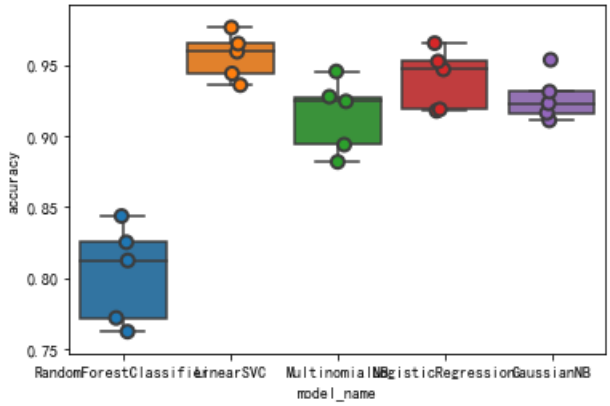
1）随机森林（Random Forest）

2）线性支持向量机（linearSVC）

3）多项式朴素贝叶斯（multinomialNB）

4）逻辑回归（logistic regression）

5）高斯贝叶斯（gaussianNB）




```
model_name
GaussianNB                0.926831
LinearSVC                 0.955915
LogisticRegression        0.940058
MultinomialNB             0.914459
RandomForestClassifier    0.802962
Name: accuracy, dtype: float64
```

图 2.3 五种模型准确率

由图 2.3，通过求每个模型的平均正确率可以发现线性支持向量机、逻辑回归、高斯贝叶斯平均正确率较高，因此在通过先前划分的测试集和训练集来验证这三者模型的各性能指标得分。根据赛题要求，使用 F-Score 来选择最终的最佳分类模型。F-Score 是精准度和召回率的加权调和平均。

2.5 最佳模型

通过最后三者模型的性能得分选择了线性支持向量机来建立留言内容的一级标签分类模型。线性支持向量机学习算法如下：

算法 7.3（线性支持向量机学习算法）

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中，$x_i\in\mathcal{X}=\mathbf{R}^n$，$y_i\in\mathcal{Y}=\{-1,+1\}$，$i=1,2,\cdots,N$；

输出：分离超平面和分类决策函数.

（1）选择惩罚参数 $C>0$，构造并求解凸二次规划问题

$$\min_{\alpha}\quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)-\sum_{i=1}^{N}\alpha_i$$

$$\text{s.t.}\quad \sum_{i=1}^{N}\alpha_i y_i=0$$

$$0\leqslant\alpha_i\leqslant C,\quad i=1,2,\cdots,N$$

求得最优解 $\alpha^*=(\alpha_1^*,\alpha_2^*,\cdots,\alpha_N^*)^{\mathrm{T}}$.

（2）计算 $w^*=\sum_{i=1}^{N}\alpha_i^* y_i x_i$

选择 α^* 的一个分量 α_j^* 适合条件 $0<\alpha_j^*<C$，计算

$$b^*=y_j-\sum_{i=1}^{N}y_i\alpha_i^*(x_i\cdot x_j)$$

（3）求得分离超平面

分类决策函数：

$$f(x)=\text{sign}(w^*\cdot x+b^*)$$

■

2.6 结果分析

探索最终所得到的最佳模型(LinearSVC)，并由测试集进行分类得到模型分类结果，查看其混淆矩阵。由图 2.4 可以看到预测结果绝大多数样本是位于混淆矩阵对角线上（即预测标签等于实际标签）。如表 2.6 所示：该模型的精准率为 0.96，召回率 0.96，F1 值为 0.96。但是由于每次测试集和训练集的随机划分以及模型使用的随机性使得该结果不为固定，但是每次得分能保证在 0.95 以上，因此该分类器具有一个良好的分类效果。

	precision	recall	f1-score	support
城乡建设	0.97	0.99	0.98	641
环境保护	0.96	0.96	0.96	677
交通运输	0.98	0.98	0.98	694
教育文体	0.95	0.94	0.95	684
劳动和社会保障	0.93	0.91	0.92	627
商贸旅游	0.97	0.97	0.97	666
卫生计生	0.97	0.99	0.98	652
accuracy			0.96	4641
macro avg	0.96	0.96	0.96	4641
weighted avg	0.96	0.96	0.96	4641

表 2.6 模型性能得分

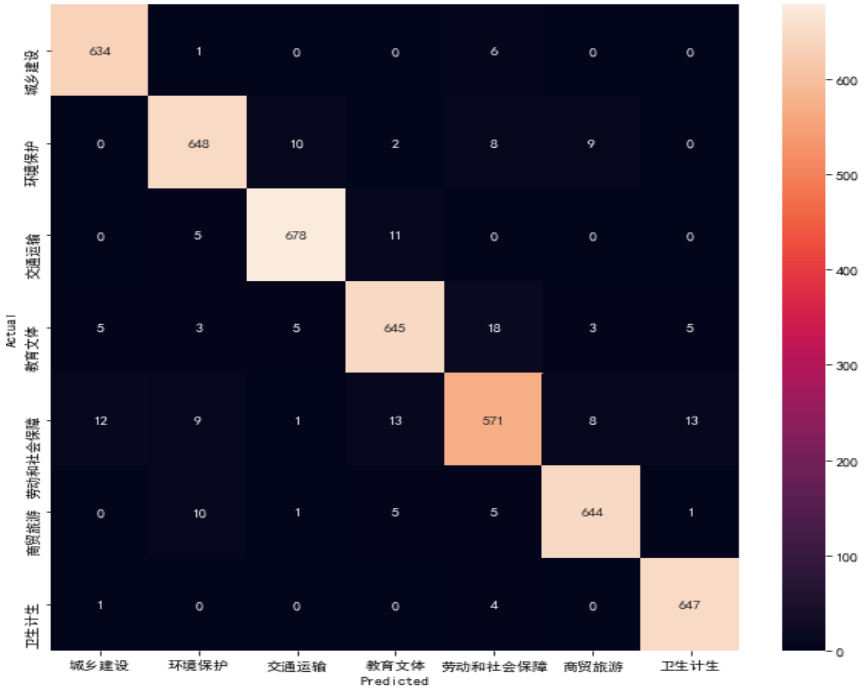


图 2.4 混淆矩阵可视化结果

3. 热点问题挖掘

3.1 文本预处理

同问题 1 中的数据处理操作对留言内容进行去重去空、去 x 序之后，使用 jieba 进行中文分词。

3.2 特征提取

在问题 2 中，采用同问题 1 一样的 TF-IDF 算法对留言内容进行文本表示，最终每个留言内容被转化为 4000 维的向量。并且利用 2-gram 模型来保证所提取的特征更具有留言内容代表性。

3.3 聚类算法

在通过 TF-IDF 算法获取了留言内容的向量化表示后，通过对文本聚类分析来得到热点问题的类别情况。常用的文本聚类方法包括 K-means、基于密度的 DBSCAN 算法等。下面简要介绍这两种算法。

1）K-means 算法

第一步，选取 K 个点做为初始聚集的簇心。

第二步，分别计算每个样本点到 K 个簇核心的距离，找到离该点最近的簇核心，将其归属到对应的簇；

第三步，所有点都归属到簇之后， M 个点就分为了 K 个簇。之后重新计算每个簇的平均距离中心，将其定为新的“簇核心”；

第四步，反复迭代 2 - 3 步骤，直到达到某个中止条件。

其主要思想是：以聚类簇个数 k 为输入参数，将 n 个数据对象合理的划分到 k 个类别中，使得同一类别的数据间隔较小，不同类别的数据间隔较大。

2）DBSCAN 算法

DBSCAN 是一种基于密度的聚类算法，与 K-means 不同的是，其并不需要输入聚类个数，而是通过参数(ϵ, MinPts)用来描述邻域的样本分布紧密程度。其中，ϵ 描述了某一样本的邻域距离阈值，MinPts 描述了某一样本的距离为 ϵ 的邻域中样本个数的阈值。DBSCAN 通过领域的概念描述样本的紧密程度，使得样本相连稠密被划分到同一类别中的。其算法如图 3.1 所示：

算法：DBSCAN，一种基于密度的聚类算法
输入：
D：一个包含 n 个对象的数据集
ε：半径参数
MinPts：领域密度阀值
输出：基于密度的簇的集合
方法：
1. 标记所有对象为 unvisited;
2. Do
3. 随机选择一个 unvisited 对象 p;
4. 标记 p 为 visited;
5. If p 的 ε－领域至少有 MinPts 个对象
6. 创建一个新簇 C，并把 p 添加到 C;
7. 令 N 为 p 的 ε－领域 中的对象集合
8. For N 中每个点 p'
9. If p'是 unvisited;
10. 标记 p'为 visited;
11. If p'的 ε－领域至少有 MinPts 个对象，把这些对象添加到 N;
12. 如果 p'还不是任何簇的成员，把 p'添加到 C;
13. End for;
14. 输出 C;
15. Else 标记 p 为噪声;
16. Until 没有标记为 unvisited 的对象;

表 3.1：DBSCAN 的算法流程：

3.3.1 K-means 算法与 DBSCAN 算法的分析与比较

1.k-means

在得到文本向量化表示后，考虑其特征稀疏高维，先使用 PCA 进行降维，再使用 TSNE。使用 K-means 模型后得到聚类大致情况如下图 3.1 所示。发现由于对留言内容聚类数目的不确定，主观选取 k 值后得到的聚类结果解释性差，无法很好将同等问题进行归类。

	theme	聚类类别
0	A3区 一米阳光 婚纱 艺术摄影 合法 纳税 座落在 A市 A3区 联丰路 米兰 春天 G2...	48
1	咨询 A6区 道路 命名 规划 初步 成果 公示 城乡 门牌 A市 A6区 道路 命名 规划...	4
2	A7县 春华 镇金鼎村 水泥路 自来水 到户 系 春华 镇金鼎村 七里 组 村民 不知 相关...	48
3	A2区 黄兴路 步行街 古道 巷 住户 卫生间 粪便 外排 靠近 黄兴路 步行街 城 南路 ...	44
4	A市 A3区 中海 国际 社区 三期 四期 空地 夜间 施工 噪音 扰民 A市 A3区 中海...	5

图 3.1 聚类情况

2.DBSCAN

考虑和 K-means 输入数据的不同，并且在 DBSCAN 中不能很好的反映高维数据，因此采取 LSA 降维，暂定为 15。同时在 DBSCAN 算法中，参数的选取十分重要，通过多次使用不同的参数组合进行聚类结果实验，发现当 eps=0.2， min_samples=4 时能够对聚类后结果有一个良好的解释，得到的聚类效果最好。

3.两者的比较。

由于留言内容具有文档数目大、特征稀疏、特征高维、簇形状不规则等特点。并且其数据分布也不是明确为典型的分布，因此不同的聚类算法对其的适用情况也有所不同，而根据资料显示这两种算法的比较如图 3.2：

	优点	缺点
K-means	✓ 算法实现简单、高效 ✓ 适用于处理大数据集 ✓ 当簇为高斯分布，效果表现好	● 不适用于发现非凸形状的簇 ● 适用聚 d 类中心不好确定，常收敛到局部最优 ● 对噪声和孤立点数据敏感
DBSCAN	✓ 可以对任意形状簇进行聚类 ✓ 对异常点不敏感 ✓ 聚类结果稳定，收敛全局最优	● 不适用于密度不均匀的样本 ● 大数据量情况下，收敛时间长 ● 对参数敏感

图 3.2 对比图

由图可知，这两种算法在留言内容的聚类上都有着各自的优缺点，但是根据我们无法很好的确定聚类数目，并且通过两种方法得到的聚类结果上看，DBSCAN 算法表现效果更好。因此我们选用 DBSCAN 聚类来完成对留言内容的分析。

3.3.2 LDA 主题模型

LDA(Latent dirichlet allocation)是由 Blei 等人在 2002 年提出的一种文档主题生成模型，通过无监督的学习方法发现文本中隐含的主题信息，目的是要以无指导学习的方法从文本中发现隐含的语义维度，使得文档中潜在的语义结构能更好的被发现。相较于传统 VSM 方法，其更适用于短文本内容少、特征稀疏的问题。

对留言内容文本进行主题抽取，构建语料，主题数目使用由 DBSCAN 聚类得到的聚类数，用 gensim 训练 LDA 模型。如图 3.3 为主题类别分布情况。

	留言主题	类别
0	A3区一米阳光婚纱艺术摄影是否合法纳税了?	60
1	咨询A6区道路命名规划初步成果公示和城乡门牌问题	55
2	反映A7县春华镇金鼎村水泥路、自来水到户的问题	63
3	A2区黄兴路步行街大古道巷住户卫生间粪便外排	15
4	A市A3区中海国际社区三期与四期中间空地夜间施工噪音扰民	61
...	...	...
4321	A市经济学院寒假过年期间组织学生去工厂工作	28
4322	A市经济学院组织学生外出打工合理吗?	28
4323	A市经济学院强制学生实习	25
4324	A市经济学院强制学生外出实习	25
4325	A市经济学院体育学院变相强制实习	28

图 3.3 主题类别分布图

3.3.2.1 主题模型可视化

pyLDAvis 是 python 中的一个对 LDA 主题模型进行交互可视化的库，它可以将主题模型建模后的结果，制作成一个网页交互版的结果分析工具。利用 pyLDAvis 在可视化上的优秀呈现来展现主题建模后的结果。结果如图 3.4 所示。

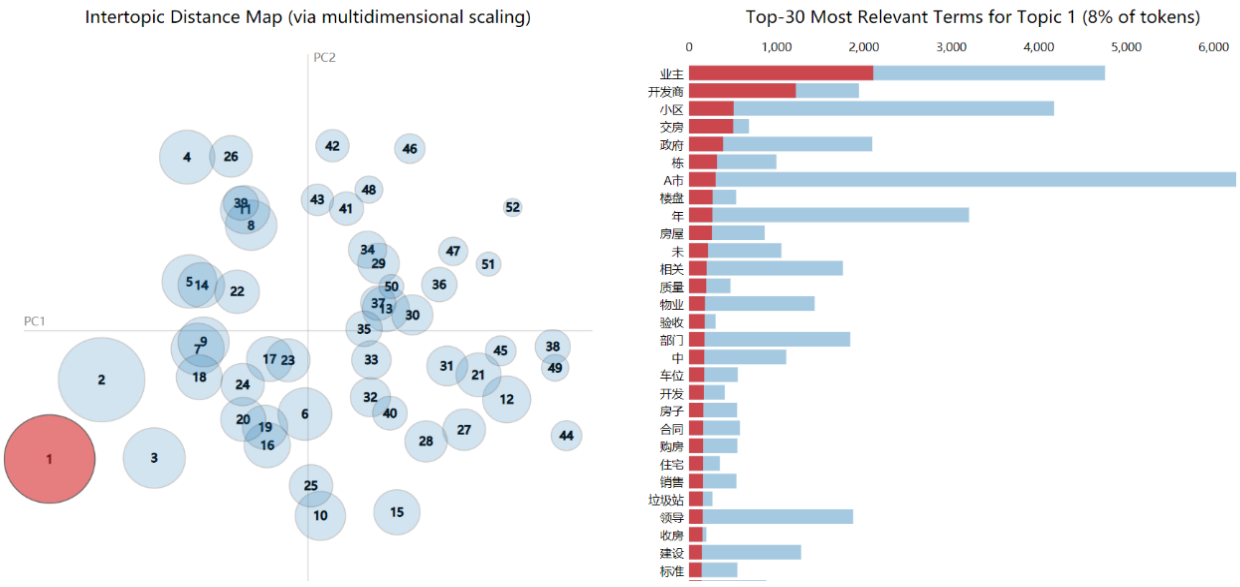


图 3.4 主题模型可视化

通过鼠标移动放置在不同主题气泡上呈现交互式页面。能够较好的解释每个主题中词语的比重分布、以及通过气泡大小来反映每个主题在总语料库中的比重情况、气泡之间的接触交叉则反应出这些主题之间的关联程度。

3.4 热度评价因素建立并 Topsis 排序

根据所给数据分析，设置两个热度指标：

①某一类问题总的点赞数+反对数在所有类问题点赞数+反对数中的比重；

②每一类问题出现次数占比（即该类问题出现总次数/所有问题总数）

并且使用 topsis 来对这两个指标进行综合评分，根据分数降序得到热点问题前 5 的留言信息。

第一步：Topsis 需要将所有的指标类型统一转化为极大型指标，而这两个指标都是越大热度越高，所以两个都是极大型指标，因此不需要进行转换。

第二步：正向化矩阵标准化。标准化的目的是为了消除不同指标量纲的影响，详细如下：假设一共有 n 个需要评价的对象，而这 n 个对象都有 m 个属性，那么原始的数据的矩阵为:

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

那么其经过归一化处理后的标准化矩阵 Z 中的每个元素为 $z_{ij}=\frac{x_{ij}}{\sqrt{\sum_{i=1}^{n}x_{ij}^2}}$

其中 Zij 为每一个元素/（√其所在列的元素的平方和）

第三步：计算得分且归一化。由上一步所得到的标准化矩阵为：

$$Z=\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix}$$

定义最优方案为 Z+ = (max {z11,z21,…,zn1}, max{z12, z22,…,zn2}, …,max{z1m, z2m,.., znm})= (Z1+ , Z2+,…,Zm+)(由 Z 中每列的最大元素构成)；定义最劣方案为 Z- = (min {z11,z21,…,zn1}, min{z12, z22,…,zn2}, …,min{z1m, z2m,.., znm})= (Z1- , Z2-,…,Zm-)(由 Z 中每列的最小元素构成)。

定义第 i(i=1,2,…,n)个评价对象与最大值的距离为 $D_i^+=\sqrt{\sum_{j=1}^{m}(Z_j^+-z_{ij})^2}$

定义第 i(i=1,2,…,n)个评价对象与最小值的距离为 $D_i^-=\sqrt{\sum_{j=1}^{m}(Z_j^-z_{ij})^2}$

那么就可以计算出评价对象与最优方案的贴近程度为 $C_i=\frac{D_i^-}{D_i^++D_i^-}$

可以看出，Ci 越趋近于 1 则其评价对象越好。Ci 即为综合得分，综合得分在附件中的热度得分表.xlsx 中，这个综合得分即为综合热度指数值，对其进行排序即可得到排名前五的热度问题。

	热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
0	1	1	0.35802	2019/01/01至2020/01/08	A市A3区	城市房屋建设存在安全隐患
1	2	0	0.29855	2018/11/15至2020/01/07	A市	社会保障补贴和劳动问题
2	3	2	0.14640	2019/01/01至2020/01/07	A市交通	交通建设和服务设施不完备
3	4	3	0.05751	2017/06/08至2020/01/06	A市学校学生	教育收费问题，学生权益未得到维护
4	5	10	0.02373	2019/01/02至2020/01/06	A市城市建设	希望公共设施建设加快

表 3.2-热度问题表

3.5 Hanlp 地名人群提取

HanLP 是一系列模型与算法组成的 NLP 工具包，由大快搜索主导并完全开源。其特点为：功能完善、性能高效、架构清晰、语料时新、可自定义等等。HanLP 从中文分词开始，覆盖词性标注、命名实体识别、句法分析、文本分类等常用任务，提供了丰富的 API。如下图所示：

```
text = 'A3区一米阳光婚纱艺术摄影是否合法纳税了？座落在A市A3区联丰路米兰春天G2栋320，一家名叫一米阳光婚纱艺术摄影的
print(HanLP.parseDependency(text))
```

```
1	A3	A3	ws	nx	_	2	定中关系	_	_
2	区	区	n	n	_	4	定中关系	_	_
3	一米阳光	一米阳光	nz	nz	_	4	定中关系	_	_
4	婚纱	婚纱	n	n	_	5	定中关系	_	_
5	艺术摄影	艺术摄影	nz	nz	_	8	主谓关系	_	_
6	是否	是否	v	v	_	8	状中结构	_	_
7	合法	合法	a	ad	_	8	状中结构	_	_
8	纳税	纳税	v	v	_	0	核心关系	_	_
9	了	了	e	y	_	8	右附加关系	_	_
10	？	？	wp	w	_	8	标点符号	_	_
11	座落在	座落在	i	i	_	20	定中关系	_	_
12	A	A	ws	nx	_	13	定中关系	_	_
13	市	市	n	n	_	15	定中关系	_	_
14	A3	A3	ws	nx	_	15	定中关系	_	_
15	区	区	n	n	_	16	定中关系	_	_
16	联丰路	联丰路	ns	ns	_	17	定中关系	_	_
```

在问题要求中需要我们提取出特定的地点/人群，而 Hanlp 的命名实体识别可以很好的帮助我们进行地名人群的提取。使用 python 直接调用 hanlp 的接口 pyhanlp 来进行使用。得到了排序前 5 的热点问题后，使用 Hanlp 命名实体识别对地名，词性进行整理，放入表 1-热点问题表中。而因为数据量小，问题描述则通过人工标注来简单实现。

4.答复意见的质量评价方案

4.1 设计评价指标体系

第一步：设置此评价体系为满分制。分数段为 0-20 的评价为非常差；20-40 为很差；40-60 为一般；60-80 为很好；80-100 为非常好。

第二步：针对附件 4 内相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度设置了以下五个指标。每个指标都具有不同的分值，最后根据所有指标的综合得分。以下为这五个指标的解释定义：

1）每条回复的字数。

从主观性出发，对于一个回复，字数的多少一般可以看出一个人在编辑这条回复时的用心程度。

根据对于所有附件 4 的观察，字数为 0-50 时大多数回复都无法解决问题，同时为避

免某些回复因为字数多而拉高评价对分数占比作了一个控制，以 0.17 为基底依次平均增加分数占比，如图 4.1。

字数	分数占比
0-50	0.17
50-100	0.40
100-200	0.63
200-	0.86

图 4.1 字数多少对应占比表

2）回复时间到留言时间的距离。

从自身出发，对于大多数群众而言，等待回复的过程往往是焦虑的，因为问题无法解决，同时对于不知什么时候能够解决问题的未知会使他们迷茫，所以一条及时的回复可以让人身心愉悦。

时间差值	分数占比
0-10080	1
10080-20160	0.9
20160-30240	0.8
30240-40320	0.7
40320-	0.6

图 4.2 时间差值对应占比表

一个星期的时间为 10080 秒，这里设置一周内回复为非常及时，依此占比减小，设置当回复时间超过四周时为非常慢，如图 4.2。

3）相似度。

从相关性出发，留言内容与回复的相似度可以从一定程度上说明回复与留言内容的相关性，可以说明这条回复确实有在分析留言内容，其占比为当前回复相似度/所有回复最大相似度。

计算相似度时利用余弦相似度，设定向量 a 和向量 b 是两个 n 维向量，其中 A=[A1,A2,⋯,An]，B=[B1,B2,⋯,Bn]，那么 A 和 B 的夹角 θ 的余弦值为：

$$cos\theta = \frac{\sum_{i=1}^{n}(Ai \times Bi)}{\sqrt{\sum_{i=1}^{n}(Ai)^2} \times \sqrt{\sum_{i=1}^{n}(Bi)^2}} = \frac{A \cdot B}{|A| \times |B|}$$

并且当这个余弦值越趋近 1，则其夹角越趋近于 0 度，也就是说这两个向量越相似，这就是“余弦相似性”。求得文本相似度的算法如下：

1. 使用 TF-IDF 算法，找出两篇文章的关键词；

2. 从每篇文章中各抽取若干个关键词，合并为一个集合，并计算每篇文章在这个集合中的词的词频（为避免文章长度的差异，可以使用相对词频）

3. 分别生成两篇文章的词频向量

4. 计算两个向量之间的文本相似度，值越大则这两个向量越相似

4）礼貌用语。

从主观性出发，生活中礼貌用语往往会让人愿意去查看或者听取别人的回复，并且可以使人舒适。具体礼貌词如下：

```
limaoci1 = ['您好','收悉','感谢','谢谢','祝您生活愉快','你好','祝您事事顺心','特此回复','欢迎','深表歉意','尊敬']
```

假设礼貌词有 L 个，当一条回复中礼貌词为 0 时，则这条回复礼貌词的分数占比为一半，那么依次每当多一个礼貌词，这条回复的礼貌用语占比则加上一个(另外一半占比/L)。

5）回复与留言内容中一致的非停用词。

从相关性出发，通过这个一致非停用词数量/该回复中总的非停用词数量作为占比。

根据附件 4 所给数据分别求出各个指标的分值，部分数据如图 4.4：

	words_num	time_distance	similarity	politeness	the_same
0	454	19484	0.394	2	0.344000
1	305	63930	0.150	1	0.085106
2	357	65350	0.500	2	0.302752
3	310	67332	0.356	3	0.241758
4	161	60491	0.627	3	0.333333
...	...	...	...	...	...
2811	34	5647	0.068	1	0.200000
2812	40	29090	0.000	1	0.000000
2813	637	47955	0.194	1	0.212435
2814	507	59178	0.182	1	0.338710
2815	248	44838	0.242	1	0.200000

图 4.4 指标分值部分数据

4.2 人工标注部分评价

由于不知道五个指标的权重占比，因此采用从原始数据中随机抽取 1/3 左右的数据进行人工标注回复质量情况。人工标注回复质量可见附近中人工标注表.xlxs，基本标注情况如图 4.2 所示。用机器学习的方法通过这个人工标注的数据来给指标的重要性进行排序。尽量使这个人工标注的数据中包含那些回复非常差的，以使后面进行的机器学

习能够更好地判断权重。

留言编号	留言用户	留言主题	留言时间	留言详情	答复意见	答复时间	len	人工标注
6556	UU0081320	役苗报销	018/3/20 15:19:4	请问领导，农合费用增加了，打狂犬役苗报销比例是多少。盼回音。 生谢了！	已收悉	2018/3/28 16:05:34	3	1
30019	UU008151	款购买二手	016/11/3 10:00:1	人在A6区准备全款购买二手房，房产局资金监管走哪家银行，需要手续费	已收悉	2016/11/22 12:25:56	3	1
114346	UU0081119	雍景园小	2013/6/3 13:05:4	商还有在合同上规定要业主缴纳。我认为这是开发商在业主当时不了解政策	已收悉	2013/7/5 16:47:46	3	1
25431	UU0081037	能否根据房	016/5/24 15:29:0	现在办理落户时无法办理。洛阳方面称无准迁证不能借出户口，而星沙这	2016年6月12日	2016/6/12 10:51:48	10	1
37459	A00039732	孩打疫苗	2019/1/13 1:56:01	请问，带小孩去打疫苗要带什么证件呢？是所有医院的都可以打吗？	2019年1月14日	2019/1/14 16:06:08	10	1
116958	UU0081337	考取健康管	2019/8/7 16:02:15	本人想考取"健康管理证"但是不知道向哪个有关部门咨询，还请告知部门，	请咨询K市人社部门。	2019/8/12 8:14:45	10	1
10924	UU0082420	路火车夜	015/10/30 22:48:1	天还能接受，最不堪忍受的就是晚上8、9点，深夜一、两点，凌晨四、五点	网友：您好！留言已收悉	2015/11/16 11:18:18	11	2

图 4.2 人工标注表

4.3 回复质量评价体系构建

针对上述定义的五个指标分别使用机器学习分类法和综合权重法对回复质量打分。

4.3.1 机器学习分类法

人工标注类别实现了给样本数据“打标签”的行为，将问题转化为有监督学习的文本分类问题，参考问题 1 的流程实现，将 TF-IDF 向量作为输入，标注的类别作为输出，训练分类模型，通过模型性能评估指标准确率、精准率、召回率、F1 值来综合考虑来选择最佳分类模型，然后对未人工标注类别的数据通过最佳模型预测出回复质量的类别。

通过该方法实现了对所有留言回复质量打分，评分结果见附件中最终评价表 1。

4.3.2 综合权重法

分别求出人工标注的数据所对应的五个指标的得分，将其根据 0.33 的比例划分为测试集和训练集，用机器学习中的 RandomForestClassifier 模型通过对指标分数及最后评价分析来显示出随机森林特征的重要性，根据每一个特征分类后的 gini 系数之和除于总特征的 gini 系数来计算特征重要性，并作出条形图。所得重要性排序如图 4.5 所示，由于每次训练集和测试集都是随机抽取的，因此每次的重要性排序都会有一点偏差，从中随机挑取一个合适的来作为评分权重。

使用综合权重法时，打分体系为：分数段为 0-20 的评价为非常差；20-40 为很差；40-60 为一般；60-80 为很好；80-100 为非常好。假设字数得分为 A，一致非停用词得分为 B，相似度得分为 C，时间差距得分为 D，礼貌度得分为 C。

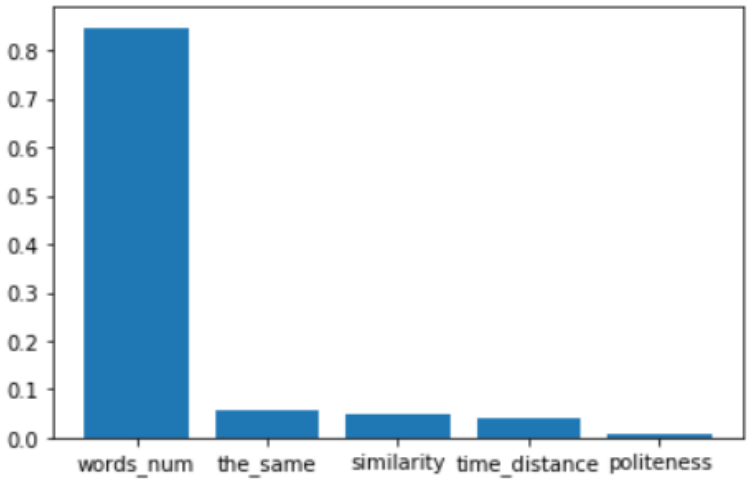


图 4.5 指标重要性排序

则每条回复的分数为(0.843*A+0.057*B+0.045*C+0.041*D+0.014*E)*100

最后根据这个分数来进行评判。评分结果见附件中最终评价表 2.xlsx 中。

4.3.3 两种评价体系的分析与比较

两种体系方法都需要人工对留言回复进行类别标注，而这一步骤的实现具有了不可规避的主观性。通过机器学习分类法所的得到的打分结果从模型的性能指标上，能够具有很好的分类效果，对回复留言质量进行较为精准评价（非常好、很好、一般、很差、非常差），而其缺点是不能十分完美的实现正确分类；而在综合权重法中，由于五个指标的特征重要性之间可能存在偏差，最后所确定的评分权重也有着主观因素，当质量评价指标十分具有代表性时，该方法可能更是用。

从最后两个评价表中发现，综合权重法的评价结果较机器学习分类法而言不够十分精准，因此在问题三中使用机器学习分类法来得到质量评价会更好。

5.小结

通过建立基于自然语言处理及文本挖掘的智慧政务系统，来实现对社会民意、民智和民气的了解。对政府在收集民意民情时能够带来极大的便利。通过对群众留言内容的分类来突显出在城市建设和各社会生活中，社会所聚焦的类别情况分布。本文采用了 DBSCAN 聚类和 LDA 主题抽取的方法实现对未进行标签的留言内容聚类，得出众多留言内容的主题内容。深入了解当下民众对当下政府的管理施政的各个方面的众多想法和反馈。

也由给与的数据集发现当下政府的热点问题集中于城市建设、社会保障和劳动问题上等。针对这些热点问题，政府应当如何去解决所面对的问题是一个十分重要的问题。除此之外，如何给予群众最为理想的留言答复也是解决民情的重要问题，文本通过机器学习来得到每一个回复质量指标的重要程度。答复内容也应当考虑到质量指标的问题从而实现最满意的答复内容。

6.参考文献

[1] Shih H S, Shyur H J, Lee E S. An extension of TOPSIS for group decision making[J]. Mathematical & Computer Modelling, 2007, 45(7):801-813.

[2] 杨凡. 基于 LDA 主题模型的在线评论聚类分析与推荐[D].大连理工大学,2018.

[3] 刘思宇. 基于聚类的短文本挖掘算法研究[D].

[4]赵星宇.基于相似性计算与半监督聚类方法的微博广告发布者识别研究[D].南京大学.硕士学位论文.2018

[5]赵丹.网络招聘信息的分析与挖掘.贵州财经大学[D].硕士学位论文.2017

[6] jessie_weiqing. 短文本聚类【DBSCAN】算法原理+Python 代码实现+聚类结果展示[EB/OL].https://blog.csdn.net/cindy_1102/article/details/95316841.2019/7/10

[7]普通攻击往后拉. python 中文短文本的预处理及聚类分析（NLP）[EB/OL].https://blog.csdn.net/weixin_43483381/article/details/85157579.2018/12/21

[8] 七八音. 调参必备--Grid Search 网格搜索 [EB/OL].https://www.jianshu.com/p/55b9f2ea283b. 2018/04/03

[9]nlpuser. 类别不平衡问题之 SMOTE 算法（Python imblearn 极简实现）[EB/OL].https://blog.csdn.net/nlpuser/article/details/81265614?utm_medium=distribute.pc_relevant_t0.none-task-blog-BlogCommendFromMachineLearnPai2-1&depth_1-utm_source=distribute.pc_relevant_t0.none-task-blog-BlogCommendFromMachineLearnPai2-1.2018/7/28

[10] 磐创 AI . 使用 scikit-learn 解决文本多分类问题[EB/OL].https://blog.csdn.net/fendouaini/article/details/82026817.2018/8/25