“智慧政务”中的文本挖掘应用

摘要

“智慧政务”的思想是通过应用文本挖掘技术以及自然语言处理技术建立智慧政务系统，对留言进行类别划分以及热点整理，从而起到减轻相关部门工作难度的作用。

在数据预处理中，需要先对文本数据进行去重、去停，接着引入jieba库对文本内容分词解析，得到文本的词项，最后形成由多个字符串组成的集合。

针对问题一群众留言的分类，先对附件2的数据进行简单数据预处理后，通过观察数据的表现，进行不相关词的过滤，采用TF-IDF算法进行对文本特征值的提取和筛选，得到权值向量矩阵。最后应用线性SVM分类模型进行标签识别，得出F-score为0.91。

针对问题二热点问题挖掘，先对数据进行预处理，清洗、过滤后把文本数据转成词袋向量，引用TF-IDF算法进行赋权，采用主成分分析（PCA）算法对数据降维，接着采用层次聚类算法（Brich）进行聚类，对得到的聚类结果建立热度指数模型，取得热点问题类别。

针对问题三答复意见的评价，建立相关性、可解释性、完整性三个评价指标。通过计算留言详情及答复意见的Levenshtein编辑距离得出两者的相似度，通过对留言结构的识别得出答复意见的完整性，通过对答复意见文本长度以及权威理论的引用情况得出答复意见的可解释性。最后通过综合评价指标对三个标准进行综合得到最终的答复意见评分指数。

关键词：TF-IDF算法；SVM；PCA；层次聚类；Levenshtein编辑距离

“智慧政务”中的文本挖掘应用

一、背景与挖掘目标

1.1 背景

近年来，中国群众通过微信、微博、市长信箱、阳光热线等网络问政平台进行意见建议或投诉留言已成为政府了解民意的重要渠道，由此产生了海量的文本数据，如何从海量的数据中提取出有效的数据给相关部门带来了极大的挑战。同时，随着大数据、云计算、人工智能的不断发展，基于自然语言处理的文本挖掘技术给相关部门给予了巨大的帮助。利用这种技术，相关部门可以更有效率的从海量文本数据中获取有价值的信息，非常有益于政府部门及时了解群众民情民意并迅速采取措施解决群众提出的问题。

1.2 数据来源

数据由“泰迪杯”数据挖掘挑战赛组委会提供。附件 1 为三级标签体系，包含三级留言内容的分类标签；附件 2 为用户留言相关信息表，其中包含的信息有留言编号、留言用户、留言主题、留言时间、留言详情以及留言详情对应的一级标签；附件 3 除了用户留言相关信息外，还增加了对应留言主题的反对数以及点赞数两列数据；附件 4 除了留言用户相关信息外，还增加了对应主题的答复意见以及答复时间两列数据。

1.3 挖掘目标

问题 1：群众留言分类

根据附件 1 给出的三级标签体系，对附件 2 的留言内容建立一级标签分类模型并且利用 F-Score 对分类方法进行评价。

问题 2：热点问题挖掘

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，并给出评价结果。

问题 3：答复意见的评价

针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案，并尝试实现。

二、问题分析

2.1 问题一

数据预处理：导入附件 2 里的留言详情和一级标签，利用 Python 对留言详情的数据进行清洗、过滤，并引入 jieba 库对文本内容分词解析，得到文本的词项，形成由多个字符串组成的集合；针对一级标签我们用 Python 的 loc 函数通过行标签索引得到数据。

文本特征提取：留言内容是由自然语言构成的数据集合，每条留言由若干词项以一定语义逻辑组合而成，根据词项在留言中出现的频率及其表达的特定主题，使用 TF-IDF 算法对文本特征值进行提取和筛选，使文本转化为可量化表征的结构化数据进行特征挖掘，得到权值向量矩阵。

构建模型：针对得到的 TF-IDF 特征提取后的特征空间，对它进行训练集和测试集的分离，建立向量空间模型（Vector Space Model，VSM）对训练集进行无监督学习，最后对测试集进行预测。

模型评价：预测后可得到模型预测准确率，同时得到 F-Score 数值，通过查准率及查全率对模型进行综合性评价。

方案流程图:

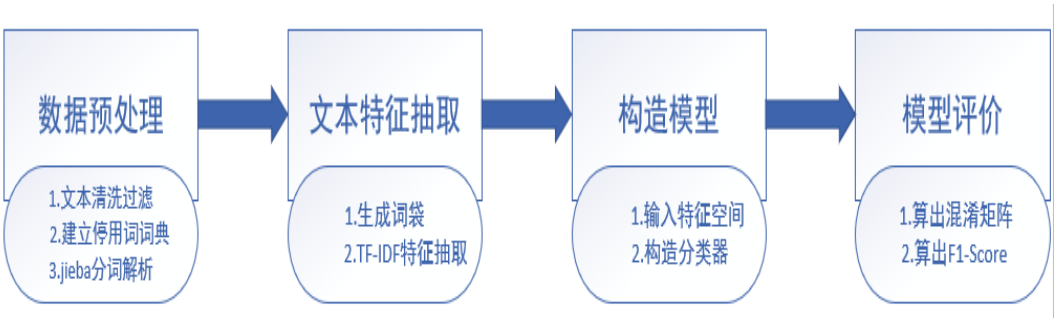

图 1　方案总体流程图

2.2 问题二

数据预处理:针对于附件三的数据,我们提取留言详情数据,对其进行清洗、过滤，分词解析，得到由多个字符串组成的词项集合。利用 TF-IDF 算法对词项集合进行文本特征的提取与筛选，对矩阵数据进行标准化，得到权值向量矩阵，对得到的权值向量矩阵进行主成分分析降维，根据效果调整维数。

构建模型:针对主成分分析降维后得到的特征空间，我们将建立层次聚类模型进行聚类，通过设定参数 threshold、n_clusters 的区间范围计算轮廓系数，根据数据规律和图像规律，找出最合适的阈值及聚类类别数后进行聚类，生成热点问题排序留言明细表。

热度指数模型:针对热点问题排序留言明细表，建立热度评价体系对留言进行热度分析,经过异常值处理及多重筛选和计算,最终得到较为准确的评判标准。

模型求解与评价:根据热度评价体系输出热度前五的问题类别，对其的结果进行评价。

方案流程图:

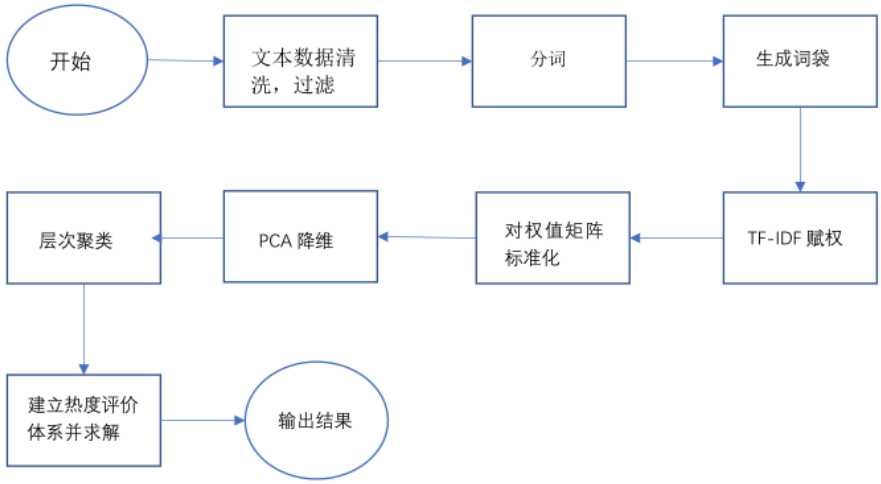

图 2　热点问题流程图

2.3 问题三

答复意见的评价，是对政务系统工作人员答复内容质量的一种检验，通过建立答复意见的评价方案，可以对政务工作质量进行反馈。我们针对该问题，从相关性、完整性以及可解释性对答复意见建立多角度评价模型，具体分析如下。

数据预处理：针对附件 4 的文本数据，我们对留言详情和答复意见进行清洗过滤、分词、去除停用词及保留新词，将每个文本处理为以空格间隔每个词语的语料库，以便后续使用。

答复意见的相关性：答复意见与留言详情的相关性，意指答复意见的内容与留言的内容是相关的，我们通过应用 Levenshtein 编辑距离算法对答复意见与留言详情进行文本相似度的计算，建立文本数据相关性的评价指标，进而转换为评价指数，得到答复意见的相关性评价。

答复意见的完整性：针对答复意见的完整性，我们基于答复意见结构的完整性给予完整性评价，而结构的完整性则由三部分构成，分别为开头部分、正文部分以及结尾部分，通过数据表现给予不同的权重，进而得到完整性的评价指标。

答复意见的可解释性：针对答复意见的可解释性，我们以留言用户的角度来进行分析。可解释性指答复意见内容的是否易于理解，是否具有理论知识。我们通过建立短文本识别模型，识别是否含有指定内容，进而建立基于文献支持度及文本长度的评价模型，通过数据表现，给予不同的权重，最后得到可解释性评价指标。

答复意见综合评分：通过对相关性、完整性以及可解释性进行综合评价，客观赋予不同权重，最终得到答复意见综合评分。

三、问题一：群众留言分类

3.1 数据探索和预处理

通过对附件 2 留言详情的观察，发现夹杂着较多的无用信息，应对其进行剔除。因此我们通过 jieba 库对留言详情进分词，根据 newci.txt 和停用词库 stopword2.txt，使用 jieba 分词工具和正则表达式对文本进行清洗、过滤，处理的文本中包含的数字、字母和符号等无意义字符，逐条解析留言文本，形成由多个词语构成的字符串集合，部分文本数据如图 3 所示，代码见附录 data_T1.py。

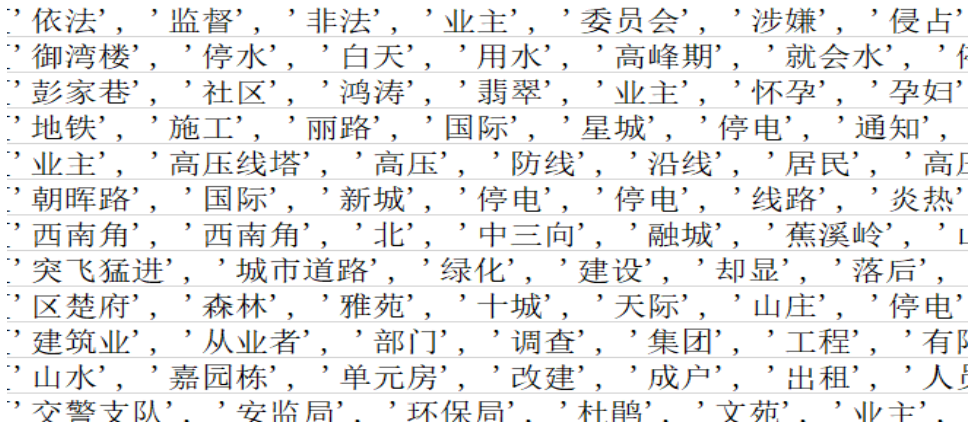

图 3　部分分词结果

3.2 挖掘建模

文本特征抽取是机器学习进行文本分类的重点和难点，特征选择是通过某种方式或算法选择出有益于分类的特征，去除那些无关或者关联性不强的特征，从而构造出更快，消耗更低的预测模型。我们处理的数据为中文文本数据，包含大量的词汇，术语和惯用短语。因此，我们应用无监督的特征抽取模型，将文本特征转换为数值矩阵的形式表示，以供机器学习。

3.2.1 文本特征抽取

TF-IDF（term frequency－inverse document frequency）是一种用于信息检索与数据挖掘的常用加权技术，用以评估字词对于语料库中的重要程度。TFIDF的主要思想是：如果某个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，则认为此词具有很好的类别区分能力，适合用来分类。

$$TF-IDF = TF * IDF = tf_i * \log\left(\frac{N}{df_i}+1\right)$$

其中，tf_i 为词项 i 在留言中出现的频率；df_i 为出现词项 i 的留言条数；N 为总留言条数。

（1） **词频（Term Frequency）**：特征词项在留言文本中出现的次数越多，就表示它和该留言内容主题相关度越大。TF=某词在留言文本中出现的次数，考虑到留言文本有长短之分，为了便于留言文本的比较，进行“词频”标准化。TF=某词在留言文本中出现的次数/留言文本的总词数。

（2） **逆文本频率（Inverse Document Frequency）：**留言集合中出现某个特征词项的留言文本数越多，则该词项的类别区分能力越差。IDF=log（语料库的留言文本总数/包含该词的留言文本数+1），如果一个词越出现的次数越多越常见，那么分母就越大，逆文档频率就越小越接近 0。分母加 1 避免了分母为 0 的情况。

我们将原始数据拆分为占比 0.8 的训练集数据和占比为 0.2 的训练数据，并基于 TF-IDF 算法的特征提取，计算出留言详情文本数据的特征权值向量，供下步分类预测模型使用。

3.2.2 线性 SVM 训练分类

我们采用支持向量机（Support Vector Machine, SVM）模型进行模型的构建，建立基于 SVM 的中文文本分类器。通过传统机器学习的方法，对训练数据集进行训练，同时对测试数据集进行分类预测，并输出分类结果。SVM 的特点是泛化能力比较强，另外，由于 SVM 算法是一个凸优化问题，因此局部最优解一定是全局最优解，可防止过学习。这些特点是其他学习算法，如神经网络学习算法所不及的。其算法如下：

（1） 假设样本空间为$\{(x_i,y_i)|i=1,2,...I\}$，其中$x_i \in R^n$为给定的样本集，即输入量，$y_i \in R^n$为目标输出量，i为训练样本个数。超平面可表示为

$f(x)=\omega^T gx+b$ ，当$f(x)=0$，x则是位于超平面上的点。

（2） 经过最大化支持向量间隔得到支持向量机的基本型：

$$\begin{cases} H(\omega,b)=\min\limits_{\omega,b}\dfrac{1}{2}\|\omega| \\ s.t.\ \ y_i(\omega^2 x_i+b)\geq 1,\ (i=1,2,...,n), \end{cases}$$

式中ω 为超平面方向参数；b为划分超位移参数；i为样本总数

（3） 由拉格朗日乘子法得到其对偶问题：

$$\begin{cases} L(\omega,b,\alpha)=\max\limits_{\alpha}\sum\limits_{i=1}^{n}\alpha_i-\dfrac{1}{2}\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n}\alpha_i\alpha_j y_i y_j {x_i}^T x_j, \\ s.t.\ \ \sum\limits_{i=1}^{n}\alpha_i y_i=0,\ \ \alpha_i\geq 0,\ (i=1,2,...,n), \end{cases}$$

式中α、α_i、α_j为拉格朗日乘子；x_i、x_j为 X 轴加速度，y_i、y_j为 Y 轴加速度；求解上述优化问题，即可求出α_i。

3.3 模型求解与评价

我们选用线性 SVM 作为文本数据一级标签的分类器，基于线性 SVM 的留言标签识别模型流程图如下，主要分为训练和预测两个阶段。

流程图思路：

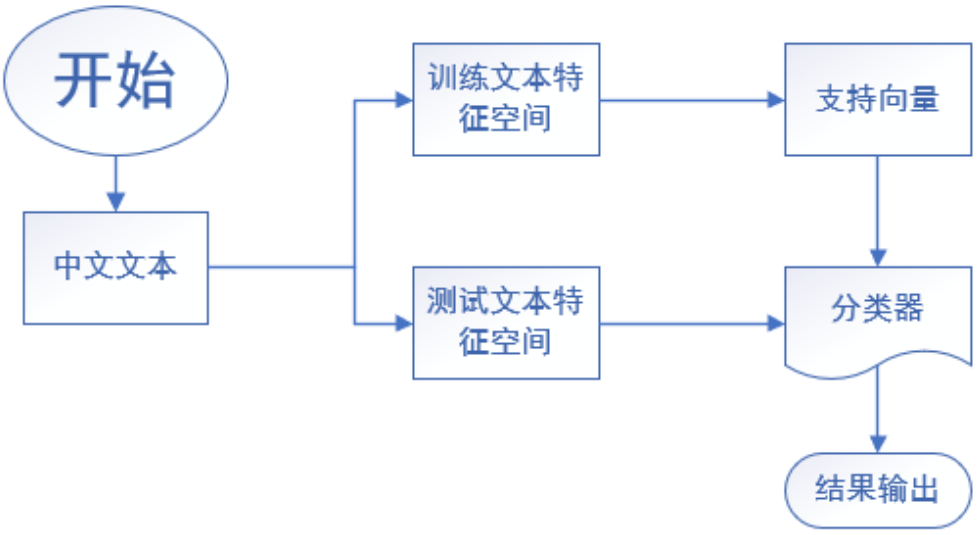

图 4　留言标签识别模型流程图

在训练过程中，将训练集生成的特征向量数据进行线性 SVM 训练，得到训练好的线性 SVM 分类器。在预测阶段，利用训练好的线性 SVM 分类器对测试集生成的特征向量进行多尺度、多窗口的分类预测。预测的具体步骤如下：

（1） 分别提取出附件 2 中分词去噪后的留言详情以及一级标签数据，并将其分为训练集和测试集，生成对应的 TF-IDF 特征权值向量。

（2） 对训练样本集进行训练，通过导入训练数据集和前面分好的训练标签集合，从而构造出一个适用于留言标签识别模型的线性 SVM 分类器。

（3） 用训练好的线性 SVM 分类器对测试集进行预测，输出留言分类预测结果，具体查看附件中留言分类预测结果表.xlsx。通过对分类结果的观察发现，该留言分类模型的应用效果相当不错，预测结果里没有出现缺失值。同时，

直观上看来，预测的准确率较高。随后，我们将预测值与真实值进行一一匹配，并生成混淆矩阵，如图 5 所示。其中包含 7 个子项，组成一个 7*7 的矩阵，一一对应七个一级标签，行元素为真实值，列元素为预测值，对角线则为预测值为真的个数。

```
array([[ 85,   1,   0,   2,  12,   0,   0],
       [  0, 360,   6,   1,   9,   2,   2],
       [  0,  11, 154,   7,   4,   1,   0],
       [  5,   6,   2, 232,  18,   3,   5],
       [  4,   5,   1,   5, 363,  10,   7],
       [  0,   1,   0,   4,   6, 312,   0],
       [  0,   0,   0,   0,   9,   1, 186]],
```

图 5　混淆矩阵

在模型的评价上，我们采用 F-Score 对留言文本标签识别模型进行评价，其公式如下：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i},\quad P_i = \frac{TP}{TP+FP}\quad ,\quad R_i = \frac{TN}{TN+FN},$$

其中，其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。TP 表示真正例数量，FP 表示假正例数量，TN 表示真反例数量；FN 表示假反例数量。通过使用 Python 计算得到模型的评价如图 6，其中 F-Score 为 0.91。

	precision	recall	f1-score	support
交通运输	0.88	0.81	0.84	117
劳动和社会保障	0.93	0.94	0.94	388
卫生计生	0.96	0.89	0.92	173
商贸旅游	0.91	0.83	0.87	251
城乡建设	0.84	0.94	0.89	421
教育文体	0.95	0.96	0.96	309
环境保护	0.96	0.89	0.92	183
accuracy			0.91	1842
macro avg	0.92	0.90	0.91	1842
weighted avg	0.91	0.91	0.91	1842

图 6　留言标签识别模型评价得分表

四、问题二：热点问题挖掘

4.1 数据探索和预处理

针对附件三的文本数据，通过观察发现文本内含有太多的无用信息，如果不处理对下面的分词与聚类都会产生负面效果，所以我们需要对这些无用信息进行剔除。

首先我们的数据预处理跟问题一方法一致，通过对数据的清洗得到剔除掉杂质的文本数据，然后通过分词进行词袋模型的构建得到向量空间模型，接着通过 TF-IDF 算法赋权得到权值向量矩阵，代码见附录 data_T2.py。

4.2 挖掘建模

4.2.1 PCA 特征提取

TF-IDF 权值向量空间矩阵，实际上是一个高维的，稀疏矩阵，为了提高聚类效果，使特征空间的表现形式在一定程度上考虑了语义、词语之间的相关关系，我们用主成分分析（Principal Component Analysis）对其进行降维，得到新的权值向量空间矩阵。

4.2.2 层次聚类分析

构建模型流程图如下图构建模型。

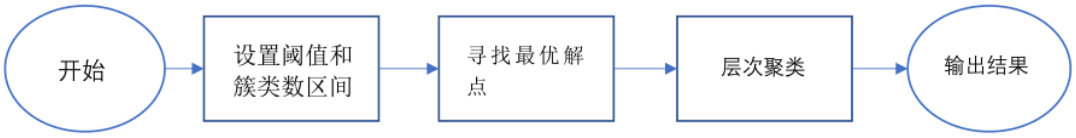

图 7　层次聚类流程图

层次聚类(Hierarchical Clustering)，通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树，在聚类树中，不同类别的原始数据点是树的最低层，树的顶层是一个聚类的根节点。本论文采用了层次聚类分析凝聚型聚类树对文本进行聚类。两个样本中簇与簇之间的相似度计算公式（欧式距离）为：

$$d_0 = \sqrt{\left(\overrightarrow{x_1} - \overrightarrow{x_2}\right)^2} = \sqrt{\left(\frac{\sum \vec{l}_1}{N} - \frac{\sum \vec{l}_2}{N}\right)^2}$$

在此题中，我们使用层次聚类的 BRICH 算法进行快速聚类，BIRCH 算法比较适合于数据量大，簇类数 K 较多且未知的情况，可适用于我们的数据。其中建模步骤为：

第一步：建立循环，设置参数 n_clusters 为（100，3000），其中步长为 100，参数 threshold 为（0.01，4.00），步长为 0.5，经过多次调试逐步缩小范围，最终找出最优值。

第二步：建立三维图散点图，找出最优值点，过程一部分图像如图所示，x 轴为 n_clusters 值，y 轴为 threshold 值，z 轴为轮廓系数。

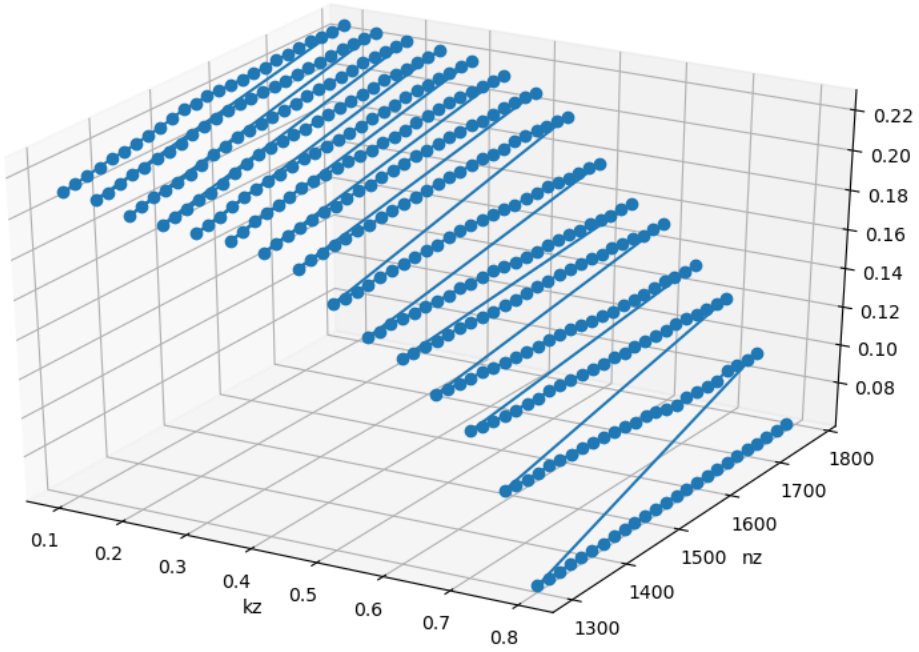

图 8　三维散点图

第三步：得到最优值点，对文本进行聚类，得到热点问题排序留言明细表。详情附录可见。

4.2.3 热度指数模型

在进行热度评价中，我们要考虑的不仅是文本内容的聚类结果，还有留言时间跨度，留言条数，以及留言的点赞数、反对数，我们将从这些因素对留言热度进行综合性评价。

去除无时间跨度聚类结果：对实际留言情况来说，如果存在一些没有时间跨度的留言，如聚类结果只有一条留言，那么此类留言只有一个日期，无时间跨度。一方面此类留言会严重影响到后续T_s值的计算，另一方面，经过对数据的观察，此类留言往往为低热度留言，因此对其进行筛除。

点赞异常值处理：在实际应用中，点赞数非常高的留言肯定是热点问题，此类留言只需要通过简单的降序排序即可得到，本题我们不对此类数据进行分析。经过对留言点赞数的观察，我们看出点赞数在 60-70 有明显的跳跃，因此我们将 70 作为临界点，点赞数高于 70 则视为点赞异常值，对于此类留言，我们另作处理，生成高点赞问题表，详情在附录可见。

量化热度指数一：

$$Score = 0.9*y*\log_{10}(z+1)+0.1*T_s,$$

其中，y 定义为：点赞数大于反对数即为正，反之为负； z 定义为$|$点赞数-反对数$|$；T_s定义为每个聚类时间的加权平均，权值由最大时间跨度决定，如时间跨度最大为 4 个月，即权值为 1，2，3，4，时间跨度越大权值越小。

去除无时间跨度聚类结果及点赞异常值后，对剩下的“正常留言”进行量化热度指数一，使用 log 函数能够有效的修正点赞数过大的影响。在T_s的计算中，将有效点赞数转换为重复留言数，并将重复留言数平均分配给此类留言的各个时间段，使得更大程度的减小误差。

在实际留言中，有效点赞数越大表示这条留言所表述的内容是群众更为关心的，所以对于公式我们将采用赋权处理，第一部分权重为 0.9，第二部分权重为 0.1。计算“正常留言”的热度指数，对其进行降序排序，取出热度排行前十五名的类，得到热点问题留言初表，详情可见附录。

去除杂质优化结果：由于每一类中可能会有聚类误差，因此对热点问题留言初表进行热点问题归纳，去掉杂质留言，剩下“有效留言”。

量化热度指数二：

$$Score = 0.8*y*\log_{10}(z+1)+0.1*T_s+0.1*P_k$$

其中， P_k 定义为该主题留言数/有效留言数 .

针对“有效留言”再次计算热度指数二，因P_k在一定程度上也能反应热度，所以此处增加P_k作为热度评价的指标，同时根据数据表现赋予各自特定的权重。

修正的热度指数公式后得到排名前五的热点问题留言明细表，详情可见附录。

4.2.3 模型求解

针对上述模型进行求解，我们得到热点问题留言明细表及热点问题表，描述性结果如下图 9 和图 10 所示，详情见附录。

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	反对数	点赞数
181	233542	A00080329	问A市经开区东六线以西泉塘昌和商业中心以南的有关规划	1/2 20:2	闲置的恒天九五重工原厂房和闲置的西地省源达置业	0	24
181	239670	A00080329	问A市经开区东六线以西泉塘昌和商业中心以南的有关规划	1/11 15:	闲置的西地省达源置业有限公司厂房， 有什么规划，	0	41
181	256358	A00080329	问A市经开区东六线以西泉塘昌和商业中心以南的有关规划	1/2 20:2	闲置的恒天九五重工原厂房和闲置的西地省源达置业	0	29
181	239595	A00057814	议A市经开区收回东六路恒天九五工厂地块，打造商业综合	11/8 15:	目前， 泉塘片区缺少大型商业，只有昌和，恒生，	0	44
181	285552	A00072434	7县将猎豹汽车所出让工业用地转商业用地规划建设大型购	12/26 18:	划为吉利代工长丰猎豹启动自救计划”的新闻，文章提	1	22
136	193286	A000103197	是住在地铁3号线A7县松雅西地省站西北方向10万居众的心声	4/17 11:	中心了解，开元路北侧穿越东即路和东即路西侧穿越开	0	32

图 9　热点问题留言明细表

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	181	1.798409	2019-01-02至2019-01-11	A市经开区	建议A市经开区打造商业综合体
2	136	1.638189	2019-04-17至2019-09-06	市地铁 3 号线松雅湖站点	A市地铁3号线松雅西地省站地下通道建设问题
3	309	1.518851	2019-03-24至2019-03-29	A3区郝家坪小学	A3区观沙岭郝家坪小学何时改扩建问题
4	1151	1.516015	2019-05-29至2019-09-15	A市三一大道	A市三一大道全线规划实施快捷快速化改造
5	592	1.471462	2019-10-29至2019-10-30	A市七中退休教师	区教育局落实发放原A市七中01年后退休教师的文明单位奖

图 10　热点问题表

五、问题三：答复意见的评价

5.1 数据探索和预处理

针对答复意见评价体系的创建，通过留言详情以及答复意见文本内容的观察，我们发现其中包含着太多的无用信息，同样，我们对其进行分词以及去除停用词操作。部分文本数据如图 11 所示，代码见 data_T3.py。

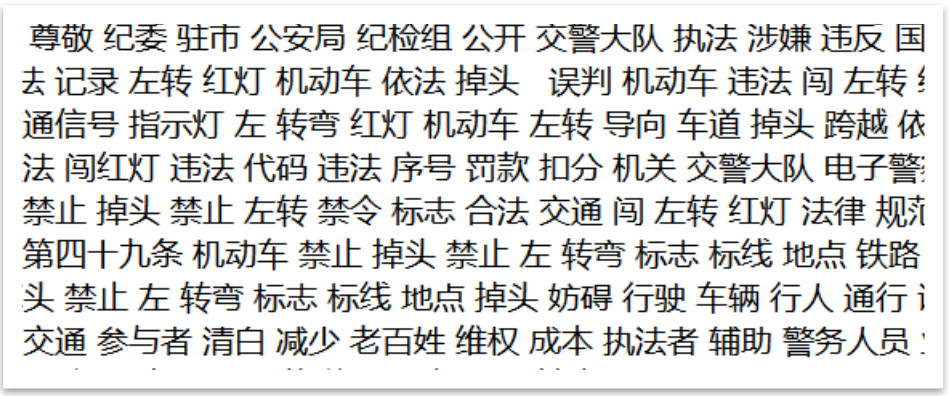

图 11　部分分词结果

5.2 挖掘建模

模型流程图如下：

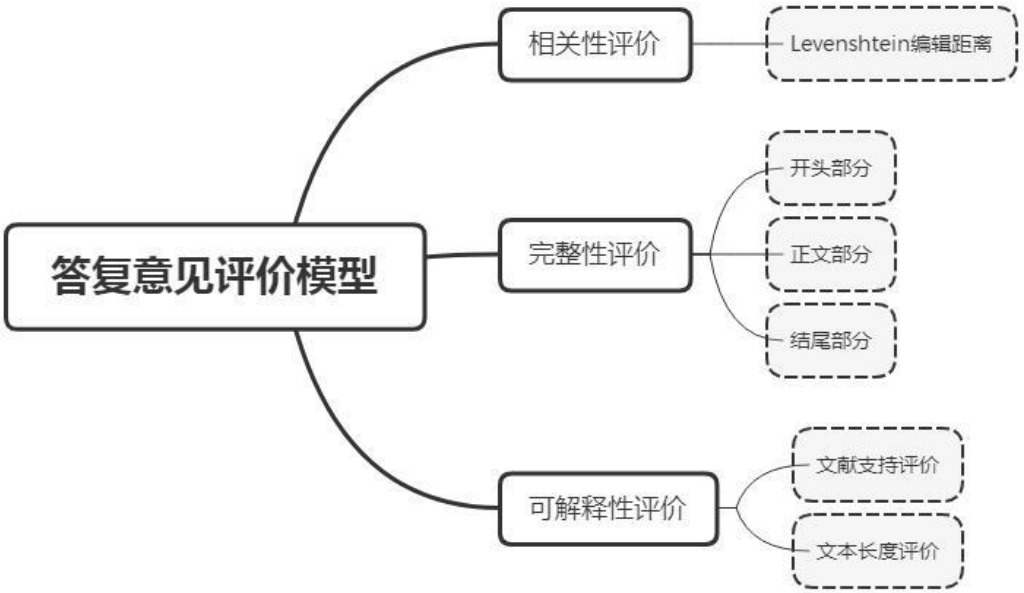

图 12　答复意见评价流程图

5.2.1 相关性

针对答复意见的相关性，我们从实际问题角度对相关性进行分析，意指答复意见的内容与留言的内容是相关的；从留言者角度来说答复意见内容应该为留言者想要获取到的相关内容，而不是答非所问、敷衍了事，内容的相关程度越高，则相似性得分越高。我们采用留言详情与答复意见的文本相似度来进行评价，文本相似度是两个及两个以上文本之间内容相关程度的一个度量，其取值的大小反应了文本相似程度的高低。取值越小，文本相似度越低；取值越大，文本相似度越高。我们通过应用基于编辑距离(Levenshtein Distance 算法)的文本相似度的计算进行相似性评价。

Levenshtein 编辑距离：Levenshtein 编辑距离算法是一种计算 2 个字符串之间的差异程度度量，编辑距离是通过改变源字符串到目标字符串的最小距离。这个距离包括插入、替换和删除等操作，利用动态规划的操作，对每个字符串进行顺序比较，其算法的时间复杂度是 $O(\mathrm{mn})$，空间复杂度是 $O(\mathrm{mn})$，m 和 n 分别表示源字符串 S 和目标字符串 T 的长度。

编辑距离 $D(S,T)$ 的计算方法如下，首先假设 $D_{ij}=D(S_0...S_i,T_0...T_i)$，$0\le i\le m$，$0\le j\le n$，$S_0...S_1$ 是源字符串，$T_0...T_j$ 是目标字符串，那么 $(m+1)\times(n+1)$ 阶矩阵 D_{ij} 可以通过下式计算得到：

$$D_{ij}=\begin{cases}0 & ,\quad i=j=0\\ \min(D_{i-1,j-1}+W_a,\ D_{i-1,j}+W_b,\ D_{i,j-1}+W_c), & otherwise\end{cases}$$

上式包含删除 W_a、插入 W_b、替换 W_c 这 3 种操作，W_a，W_b 和 W_c 分别表示每一种操作的权重。D_{ij} 是指从源字符串 S 到目标字符串 T 的最小编辑操作次数，目的是计算 S 与 T 的相似度，D_{ij} 随着 2 个字符串之间的相似度减少而增加。该算法从字符串左边第一个字符串位置开始比较，对已经比较过的编辑距离，继续计算下一个字符位置的编辑距离。矩阵能够通过从 D_{00} 逐行逐列计算获取，最终 D_{mn} 表示 $D(S,T)$ 的值，即 S 和 T 的最小编辑距离。

相似度计算：得到编辑距离结果后，需要进行相似度计算，相似度是 2 个字符串之间相似度的度量，基于编辑距离计算 2 个字符串相似的的公式为：

$$sim=1-\frac{ld}{m+n},$$

上式中，ld 表示 2 个字符串之间的 Levenshtein 距离；m 和 n 分别为 2 个字符串的长度；sim 值越大，表示 2 个字符串相似度越高；

5.2.2 答复意见的完整性

针对答复意见的完整性，我们以基于答复意见文本内容结构的完整性给予答复意见完整性评价，而结构的完整性则由三部分构成。

开头部分：首先，一个完整的留言答复在文本结构上需要一个良好的开头。答复人员以礼貌的态度对留言用户进行回复，并提及出所要答复的对象。通过对目标文本进行识别，若存在“您好”、“尊敬的某某”等此类情感用语时则标记为满足结构上具有开头部分。从整体结构上分析，开头部分的重要程度相对较低，通过主观赋权法，给予开头部分权重分 ω_1 为 2.5 和 0。在模型实现上，我们通过数据表现，总结得出开头部分的特征词，并形成词库。通过正则表达式对目标留

言内容进行识别，经过词库的逐一匹配，输出匹配结果。若匹配结果为是，则开头部分ω_1为2.5；若匹配结果为否，则ω_1为0。

正文部分：接着，更重要的是正文部分，正文部分是答复意见评价的重点部分，留言用户通过正文部分来获取有关留言的反馈信息。因此，正文部分的有无对整个留言的评分起到关键作用。对于该部分，同样对目标文本进行识别，若出现“答复如下”等此类正文部分结构词语，则标记为满足结构上具有正文部分；在模型实现上，我们通过数据表现，总结得出正文部分的特征词，并形成词库。通过正则表达式对目标留言内容进行识别，经过词库的逐一匹配，输出匹配结果。正文部分相对比较重要，因此赋予权重分ω_2 为5和1。若匹配结果为是，则正文部分ω_2为5；若匹配结果为否，则ω_2为1。

结尾部分：最后，一个完整的留言答复还需要一个完整的结尾，标志着答复的结束.同理，对于这部分的实现，对目标文本进行识别，若出现“特此回复”等结构词语，则标记为结构上具有结尾部分。从整体结构上分析，结尾部分的重要程度相对较低，通过主观赋权法，给予开头部分权重分ω_3为2.5和0。在模型实现上，我们通过数据表现，总结得出结尾部分的特征词，并形成词库。通过正则表达式对目标留言内容进行识别，经过词库的逐一匹配，输出匹配结果。若匹配结果为是，则结尾部分ω_3为2.5；若匹配结果为否，则ω_3为0。

答复意见完整性综合评价：针对上述三部分的得分，根据数据的表现，对其赋予0.3，0.4，0.3的权重，得到答复意见完整性的综合评价。

5.2.3 答复意见的可解释性

针对答复意见的可解释性，我们以留言用户的角度来进行分析，可解释性指答复意见内容的是否具有权威理论知识，是否引经据典。留言用户收到答复后，是否可以采用你的答复，又是否能够明白你的答复，便取决于你的可解释性强度。对于留言者来说，答复具有类似某某机构给出的文件说明，其可解释性更强；其次，留言答复的文本长度也有一定的参考价值，相比之下，答复意见的文本长度更长，说明答复的更详细，可解释性更强。

文献支持：针对文献支持评价，我们根据数据表现来进行分析，通过对答复意见的观察发现，部分答复意见具有引经据典的现象，根据留言详情，相关部门根据相关文件、法律、规定等，对留言用户进行相关答复，增强了答复内容的权威性，可解释性更强。在算法的实现上，我们通过对文本数据的总结归纳，形成文献、法律等的特征词库，应用正则表达式，对答复内容进行识别，从而评价答复意见的文献支持度。在评价上，给予文献支持评价以下的权重分：通过对答复意见的识别，若存在引经据典的词语，则给予权重分 Υ 为 8 分；若不存在此类词语，则权重分 Υ 为 2 分。

文本长度：针对文本长度的评价，句子的长短一定程度上可以反映的内容质量的不同，相比较之下，句子更长的阅读者可以获得的信息更多。因此，我们以答复意见的文本长短作为答复意见可解释性的一个因素。在算法实现上，我们对答复意见文本进行循环遍历，逐个计算文本长度，并对所有文本长度进行排序。通过对全部数据的表现，存在着文本长度相差较大的情况，针对这部分数据，做出以下调整：剔除在文本长度的前 10%，认为这部分数据是超大文本数据，内容很丰富；在剩余的数据中取其最大值为上限，内容最丰富，其余的递减。具体评价规则如下：对全部答复意见内容进行循环遍历，若该答复的文本长度为超大文本数据的区间值，则给予文本长度评价分满分 10 分；若该答复的文本长度为剩下的区间值，则以该分本长度与该区间最大文本长度的比值最为该答复的文本长度评价分，计算公式如下：

$$ls=\begin{cases}10\,,\ len_i\in\mathbf{Z}\\ \dfrac{len_i}{\lim\limits_{x\to\infty}\left(\bar{\mathbf{Z}}\right)}\times 10\,,\ len_i\in\bar{\mathbf{Z}}\end{cases},\ i\in[0,n]\,,$$

上式中 ls 表示文本长度评价分，len_i 表示每个答复的文本长度，$\mathbf{Z}$ 表示全部数据中认为是超大文本的区间值，而 $\bar{\mathbf{Z}}$ 则为剩余的区间，n 为总文本数据个数。

答复意见可解释性综合评价：对文献支持度得分及文本长度得分，根据数据结果的实际表现，我们对其分别赋予 0.4，0.6 的权重，得到答复意见可解释性的综合评价。

5.2.4 答复意见综合评分

综合评价（Comprehensive Evaluation，CE），也叫综合评价方法或多指标综合评价方法，是指使用比较系统的、规范的方法对于多个指标、多个单位同时进行评价的方法。它不只是一种方法，而是一个方法系统，是指对多指标进行综合的一系列有效方法的总称。综合评价方法在现实中应用范围很广。综合评价是针对研究的对象，建立一个进行测评的指标体系，利用一定的方法或模型，对搜集的资料进行分析，对被评价的事物做出定量化的总体判断。我们使用计分法对上述局部评分进行综合，将答复意见的综合部分分成三个子项，分别为相关性得分、完整性得分以及可解释性得分。其中相关性得分用前文的文本相似度直接代替；完整性得分由开头部分评价、正文部分评价及结尾部分评价三部分构成，取三部分的权重得分和作为完整性得分的评价指标；而可解释性由文献支持和文本长度构成，相比之下文本长度的权重更高。因此，可解释性得分为文献支持的分的 40%加上文本长度得分的 60%。最后，答复意见的得分则为相关性、完整性及可解释性三部分的权重得分组成。相比之下可解释性所占的权重比其他两部分应该略高一些，因此最终的答复意见得分为相关性得分的 30%和完整性得分的 30%以及可解释性得分的 40%之和，完整公式如下：

$$Score = 0.3 \times score_1 + 0.3 \times score_2 + 0.4 \times score_3 ,$$

$$score_1 = sim = 1 - \frac{ld}{m+n} ,$$

$$score_2 = \omega_1 + \omega_2 + \omega_3 ,$$

$$score_3 = 0.4 \times \Upsilon + 0.6 \times ls ,$$

其中 $score_1$ 为相关性得分，sim 为文本相似度得分；$score_2$ 为完整性得分；$score_3$ 为可解释性得分 Υ 为文献支持得分，ls 为文本长度得分。

5.3 模型求解

经过以上模型的构建，我们对附件 4 的数据进行模型求解,对答复意见进行综合性评价，并输出相关性得分、完整性得分、可解释性得分以及 Score 整体得分，部分数据如图 13，详细请看附件中答复意见评价表。

	留言用户	留言主题	答复意见	相关性得分	完整性得分	可解释性得分	Score
0	A0004558	A2区景蓉华苑物业	现将网友在平台栏目向胡华衡书	4.07	3.5	7.14	7.57
1	A0002358	A3区潇楚南路洋湖	网友“A00023583”：您好！针对	2.31	3.5	3.45	4.33
2	A0003161	请加快提高A市民营	市民同志：你好！您反映的“请	3.52	3.5	6.30	6.77
3	A0001107	在A市买公寓能享受	网友“A000110735”：您好！您	2.65	3.5	5.89	6.08
4	A0009233	关于A市公交站点名	网友“A0009233”，您好，您的	5.67	3.5	2.20	5.16
5	A0007753	A3区含浦镇马路卫	网友“A00077538”：您好！针对	3.43	3.5	2.81	4.48
6	A0001008	A3区教师村小区盼	网友“A000100804”：您好！针	3.35	3.5	5.33	6.06
7	UU00812	反映A5区东澜湾社	网友“UU00812”您好！您的留言	2.55	3.5	8.62	7.80
8	UU008792	反映A市美麓阳光住	网友“UU008792”您好！您的留	2.90	3.5	5.18	5.75
9	UU008687	反映A市洋湖新城和	网友“UU008687”您好！您的留	3.53	3.5	2.75	4.48
10	UU008220	反映A2区大托街道	网友“UU0082204”您好！您的留	2.87	3.5	5.05	5.65

图 13　答复意见评价表

六、结论

本文通过对“智慧政务”中的中文文本挖掘，运用到许多自然语言处理知识，包括对原始数据的去噪处理，去除停用词，分词，并将文本数据数字化。利用权值矩阵对文本进行分类，聚类。但要想得到最好的结果，就必须对模型进行多次的调参，优化，甚至寻找到最适用于题目的模型，期间要有足够的耐心与毅力。

在解决问题的过程中，本作品取轻避重，衡量利弊选择最优的方法。通过重重的筛选，多次的计算优化，得到的热点明细表趋于准确，答复意见评价方案也有理有据，符合数据的规律，理论和常识。

参考文献

[1]溪海燕，肖志涛，张芳.基于线性 SVM 的车辆前方行人检测方法[J].电子信息及自动化，2012，31(1):69-73.

[2]藏润强，孙红光，杨凤芹等. 基于 Levenshtein 和 TFRSF 的文本相似度计算方法[J]. 计算机与现代化，2018,(04):84-89

[3]占斌. 基于层次聚类算法的商业数据分析[D].沈阳师范大学,2019.

[4]金吉琼,刘鸿,郑赛晶.基于在线评论文本挖掘技术的电子烟市场消费热点分析[J].烟草科技,2019,52(12):106-114

附录清单

附件	附件名称	备注
1	data_T1.py	问题一标签分类完整代码
2	data_T2.py	问题二热度评价体系完整代码
3	data_T3.py	问题三答复意见评价体系完整代码
4	newci.txt	新词词典
5	stopword2.txt	停用词词典
6	高点赞问题表.xlsx	点赞数大于 70 的热点问题表
7	热点问题排序留言明细表.xlsx	问题二层次聚类的结果
8	热点问题留言初表.xlsx	量化热度指数一，得分前 15 的数据结果
9	**留言分类预测结果表.xlsx**	问题一测试数据预测表
10	**热点问题留言明细表.xlsx**	经量化热度指数二，得到排名前 5 明细表
11	**热点问题表.xlsx**	归纳后的 5 个热点问题
12	**答复意见评分表.xlsx**	答复意见综合评价表
13	附件 1.xlsx	三级分类指标
14	附件 2.xlsx	带有一级标签的留言表
15	附件 3.xlsx	增加点赞数反对数的留言表
16	附件 4.xlsx	带有答复意见的详情表