“智慧政务”中的文本挖掘应用

摘要

随着大数据和机器学习的不断发展，基于自然语言处理技术的智慧政务系统已成为社会治理发展的主趋势。

针对问题一，首先对附件 2 中的文本进行预处理，构建停用词词库并进行 jieba 分词，筛选出聚类关键词，然后通过 TF-IDF 算法构建词向量，将文本信息转换为机器语言。比较随机森林、线性支持向量机、多分类 logistics 回归等多个模型精度，并绘制分类准确率的玉玦图，得到线性支持向量机模型预测结果最优，精度达 90%以上，F-score 加权平均值最高为 90%。依据此方法将留言具体划分为：城乡建设类 663 条、劳动和社会保障类 650 条、教育文体类 525 条等。

针对问题二，首先将附件 3 中的留言主题分类，并构建新的停用词词库，首先筛选地名，避免同一地区的不同问题被聚为一类，然后构造词频矩阵，使用改进 K-means 法对留言主题进行聚类，最后定义关注度、留言热度值等指标，建立热度评价模型，得到热度值前五的问题，由高到低依次为：A4 区 58 车贷案问题、丽发新城搅拌站污染扰民问题、伊景园捆绑销售车位问题、五矿万境水岸住问题、A 市外来人才补贴问题。

针对问题三，要求对附件 4 中的留言答复情况从及时性、相关性、完整性和可解释性四个维度建立回复评价模型。本文定义了答复时间间隔、答复与留言相关度、规范性用词、答复文本长度、标点符号使用合理性 5 个指标，通过归一化与主成分分析法得到各指标权重分别为：0.186383、0.161758、0.214514、0.232524、0.204822。最终得到每条留言综合得分，并绘制留言综合得分随时间变化曲线。可明显看出从 2011 年到 2019 年，回复的综合得分由 0.436 增加到 0.502，有明显的提升，但在 2019 年，回复的完整性这一指标呈下降趋势，即回复中忽略了礼貌用词和日期等规范性用词，应加强政务素养规范。

关键词：TF-IDF 算法，线性支持向量机，K-means 聚类，主成分分析

Abstract

With the continuous development of big data and machine learning, smart government systems based on natural language processing technology have become the main trend of social governance development.

For problem one, build a stopword vocabulary and perform jieba word segmentation, filter out clustering keywords, and then build word vectors through TF-IDF algorithm to convert text information into machine language . Compare the accuracy of multiple models such as random forest, linear support vector machine, and multi-class logistics regression, and draw the classification accuracy rate Yuyu graph to obtain the best prediction result of the linear support vector machine model, with an accuracy of more than 90% and the message is specifically divided into: 663 urban and rural construction categories, 650 labor and social security categories, and 525 educational styles.

For problem two, first classify the subject of the message in Annex 3, extract the location information separately and construct a new stopword lexicon to avoid different problems in the same area being grouped together, then construct a word frequency matrix and use improved K- The means method clusters the message topics. When the total number of messages is divided into 30 categories, the F-score weighted average is up to 90%. Finally, define the indicators such as attention degree and message heat value, and establish a heat evaluation model to get the top five questions of heat value, in order from high to low: 58 car loan in A4 area, pollution and nuisance in Lifa New City Mixing Station, Yijing The problem of bundled parking spaces in the park, the problem of water and housing in Minmetals, and the subsidy for foreign talents in city A.

In response to question three, firstly, a reply evaluation model was established from the four dimensions of timeliness, relevance, completeness and interpretability of the reply to the message in Annex 4, defining the response time interval, relevance of the reply and message, and standard terms The length of the reply text and the use of punctuation symbols are reasonable indicators. The weights of the indicators obtained through normalization and principal component analysis are: 0.186383, 0.161758, 0.214514, 0.232524, 0.204822. Eventually get the comprehensive score of the message, It can be clearly seen that from 2011 to 2019, the comprehensive score of the reply was born from 0.436 to 0.502, Sexuality is in a downward trend, and attention should be paid to them.

Keywords: TF-IDF algorithm, linear support vector machine, K-means clustering, principal component analysis

目录

1. 问题背景

近几年来，随着互联网的发展，很多线上问政平台如微博、微信、市长信箱等已经逐渐成为政府知民意、汇民智、聚民气的重要渠道，不同类型的相关文本数据量正不断上升，这给传统的主要以人工进行留言分类和热点数据整理的有关部门带来了极大的工作量。同时，随着大数据及机器学习等技术的不断发展革新，基于自然语言处理技术的智慧政务系统已成为社会治理发展的主趋势，对政府管理水平的提高及政务处理效率有极大的推动作用。

2. 待挖掘的目标

本次研究的目标是利用互联网公开来源的群众问政留言记录及相关部门的答复意见，利用自然语言处理和文本挖掘等方法，解决以下三个问题：

1、利用文本分词和文本聚类的方法对文本数据进行计算机结构化处理，采用 jieba 中文分词提取出留言的关键词，TF-IDF 构造词向量，根据现代化线性支持向量机聚类结果，对问政留言进行分类，以改进传统的人工分类。

2、对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，结合点赞数及反对数等社会共性反映指标，分析出目前存在的热点问题，以便政府部门能够有的放矢。

3、根据政府相关部门的答复意见，从政务素养规范、答复及时性与相关性、民众理解程度等多个维度对政务答复系统给出合理的评价。

3. 问题分析

随着大数据及机器学习的广泛应用，线上政务处理平台已成为政府部门知民心、聚民气的重要渠道。因此，运用网络文本分析和数据挖掘技术对政务平台留言的研究有重要的意义。

3.1 问题一

该问题要求建立一个分类模型，对留言内容进行一级标签处理，建立如查准率、查全率等相关性指标对给出的分类方法做模型信度检验，以便能确实解决人工分类工作量大、效率低且出错率高的问题。为达到较高的精度，可在多项式朴素贝叶斯、多分类

logistics 回归、随机森林和线性支持向量机等多个分类模型中选出精度最高的，结合题中所给的 F-score 的精度指标，对附件 2 中的留言进行一级标签分类。

3.2 问题二

该问题要求对给出的附件 3 中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，找出其中的热点问题，并给出相应的评价结果。初步分析该部分留言文本数据包含了留言编号、留言时间、留言详情、点赞数和反对数等多项指标。其中留言主题及详情最能直接反应待解决的问题，而点赞数和反对数可以侧面反映留言的关注度。本文运用 K-means 聚类法对留言主题进行聚类，使相似留言聚为一类，便于观察出现次数较多的问题。最后，综合考虑问题的留言次数和点赞数以及反对数，筛选出热点问题。

3.3 问题三

该问题是典型的评估问题，要求在相关部门给出留言的答复意见之后，从答复的相关性、完整性、可解释性四个维度对答复意见的质量给出一套合理的评价方案。即给出的答复意见是否能够切实可行地解决民生问题，措辞是否合乎规范、内容是否通俗易懂等。本文建立问答相关度、答复间隔时间、答复文本长度等指标，建立主成分分析权重修正的层次分析模型，计算出四个维度的权重，得出答复得分。

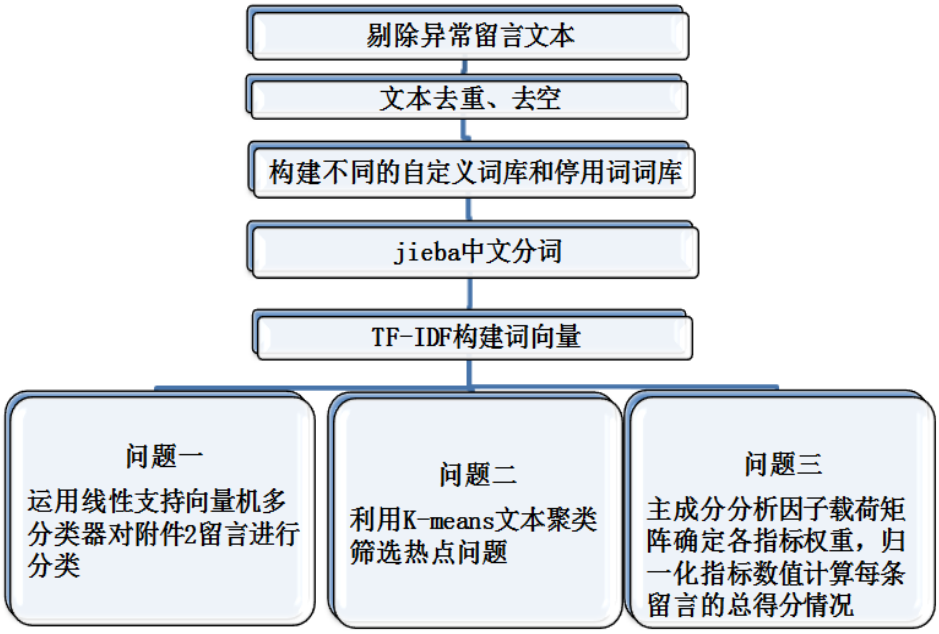


图 1 思路展示流程图

4. 符号说明

符号	含义
$M(c)$	表示文本i中特征词总数
P_i	第i类的查准率
R_i	第i类的查全率
$J(l_k)$	类内所有的样本点到聚类中心的距离平方和
D	指标总得分

5. 数据预处理

5.1 无效留言信息处理

“去空”——留言信息中难免存在空格或其他特殊字符等无用符号，干扰了文本长度等问题的分析，采取字符串过滤的方法，将留言处理为纯文本。

5.2 构建停用词词库和自定义词库

停用词是在文本处理中为了提高搜索效率或当研究更侧重某一问题时，可以暂时将某些字或不能反映信息的连接词等加以过滤或不予考虑，这些词会构成一个停用词库，便于以后的分词。但是并没有一个明确的停用词词库可以适用于所有情况，故问题一、二构建了不同的停用词库，下有具体体现。

自定义词库的构建类似于停用词词库的构建，对于 jieba 自带词库中不包含、又希望以某种固定格式输出的词可构建自定义词库。如可将“A 市”定义为一个词，当出现“咨询 A 市落户的问题”时，会分为“咨询、A 市、落户、的、问题”，而不是“咨询、A、市、落户、的、问题”。

5.3 jieba分词

在对留言信息进行分析之前，需要先把留言的文本信息转化为计算机可以识别的机器语言。应对留言文本信息进行 jieba 分词。jieba 工具包有属于自己的前缀词典，当输

入一个中文句子，会自动进行高效词扫描，采用动态规划最大概率路径输出最大词频，输出句子中所有可能的组词情况。如输入“咨询 A 市落户的问题”，会将其分为“咨询、A、市、落户、的、问题”。其中“问题”即为所有可能的组词情况，不会得到“的问”这样的分词结果。

在用 jieba 完成中文分词后，为进行深度数据挖掘，需要先把分词结果转化为词向量，下面采用 TF-IDF 算法实现词向量的建立。

5.4 TF-IDF算法构建词向量

TF-IDF（词频率-逆文档频率），其基本思想是如若选定某一词作为分类的关键词，则该词应在一篇文章中出现次数较高，而在其他文章中出现次数较少。TF 即为所说的词频，也就是某一给定词 c 在该文本中出现的频率，这个容易理解，IDF 即为所说的“逆文本频率”，也就是某一给定词 c 在所有文本中出现的频率。算法如下：

第一步，计算词频-TF：

$$TF(\mathrm{c}) = \frac{m(c)}{\mathrm{M}(c)} \tag{1}$$

其中 $m(c)$ 表示文本 i 中特征词 c 出现的次数， $\mathrm{M}(c)$ 表示文本 i 中特征词总数。

第二步，计算 IDF 权重：

$$IDF(\mathrm{c}) = \log\frac{N}{N(c)} \tag{2}$$

N 代表语料库中文本的总数，$N(c)$代表语库中包含词 x 的文本总数。

但是在某些特殊情况下，如某一生僻词在文本库中没有，则此时分母 N(c)就为 0，这样没有意义，故我们需要对上面的 IDF 公式做适当的平滑，平滑之后的公式为：

$$IDF(\mathrm{c}) = \log\frac{N+1}{N(c)+1}+1 \tag{3}$$

第三步，计算 TF-IDF 值：

$$TF-IDF(c)=TF(x)*IDF(c) \tag{4}$$

通过分析得出 TF-IDF 的值会随着某一词在文本中出现次数的增加而变大，即两者是正相关关系。计算出政问留言中每个词的 TF-IDF 值进行排序，次数最多的也就是留言中的关键词。可以由上面的 TF-IDF 数值排序的结果，找出排名靠前的关键词,如若将

其以集合的形式或者行向量的形式给出，当留言中有这个关键词，则记为 1，如果没有记为 0，则可以得到一个 0-1 矩阵，运用 TF-IDF 的定义式，计算出权重向量，生成词向量由此建立。

6. 模型的建立与求解

6.1 模型一：基于线性支持向量机模型的留言分类问题

6.1.1 文本的预处理

（1）构建自定义词库和停用词词库

问题一要求对附件 2 中的留言进行一级标签分类。附件 1 给出的有医疗卫生、公共建设、党务政务等一级指标，仅仅涉及留言内容分类，不同地区的同类指标应归为同一类。地名如“A 市”分词结果为“A、市”因此首先应自定义地名关键词，将“A 市、A1 县”加入自定义词库，其次将“A 市”等地名信息、连接词、语气词等加入停用词词库。

（2）对留言信息进行分词

通过构建停用词词库，可对留言进行 jieba 分词。在排除掉地名、连接词等信息之后，以第二条留言为例，“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围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市 A 市，尽快整改这个极不文明的路段。”分词后为“大道、未管所、路口、加油站，人行道、路灯杆、西湖建筑集团、燕子山、安置、房、项目、施工、围墙、上下班、人流、车流、安全、隐患、整改、不文明、路段”。

（3）计算一级分类标签的关键词词频

对分词结果进行词频统计，出现次数较多的词语可作为一级指标的分类依据，如“学校”、“教育”、“学生”等可归为教育文体类，“医生”、“医院”、“医疗”等可归为卫生计生类，其他类同理。为初步了解各个一级指标内的关键词信息，直观认识其内关键词的分布情况，可以通过绘制关键词词云图，将占比较大的关键词突出显示，形成“关键词云层”，从而过滤到其他不重要的或词频较低的文本信息，使得通过查看词云图就可直接观察到各个一级分类标签中的主要信息。如下图所示：

图 2 一级分类标签词云图

从图 2 可直观看到各个一级分类标签内的关键词，这些关键词可作为未知类别归类的依据。如在社会保障中，关键词为“单位”、“工资”、“劳动”等；在交通运输中关键词为“租车”、“司机”、“收费”等。这样不仅可以直观看到类内问题，而且进一步证实了分词结果的准确性。

（4）TF-IDF建立词向量矩阵

由分词的结果，计算 TF-IDF 值，可构建关键词矩阵，包含某一关键词记为 1，不包含记为 0，就转化成了词向量矩阵。如构建的关键词矩阵为 [公司、部门、整改、污染、社会保障、不文明、路段]，以第一条留言为例：“请求整改不文明路段”,分词结果为“请求、整改、不文明、路段”，则对应的词向量矩阵为[0，0，1，0，0，1，1]，其他同理。

（5）一级分类标签分布情况

为初步了解一级分类标签中各类的总数量，可通过绘制玫瑰图得到各类占比情况：

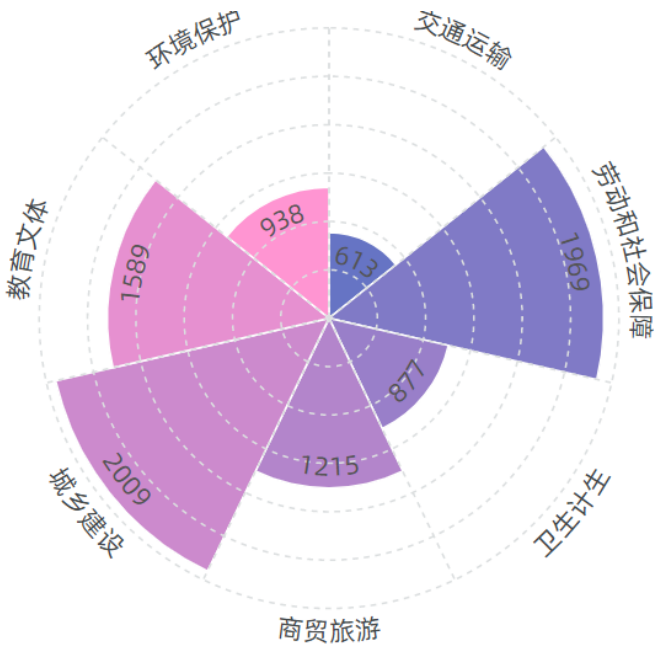


图 3 一级指标分类玫瑰图

从图 3 可以看出，在 7 项一级指标分类标签中，城乡建设和劳动社会保障留言总数较多，多达 2000 条左右；教育文体类次之，在 1000-2000 条；交通运输和卫生计生相对较少，留言在 1000 条以下。说明附件 2 留言大部分是城乡建设和社会保障类，政府部门可通过各指标分布情况初步了解群众留言趋向。

6.1.2选择分类模型

经过文本预处理，得到留言的有效信息，可通过不同的分类方法划分一级指标。常见的分类方法有较传统的多项式朴素贝叶斯分类、二分类和多分类的 logistics 回归模型、现代算法线性支持向量机、随机森林等。下面通过计算各种方法的准确率、召回率等指标选择最优的分类方法。

- 多项式朴素贝叶斯分类

常见的机器学习朴素贝叶斯分类有高斯朴素贝叶斯、多项式朴素贝叶斯、伯努利朴素贝叶斯。其中多项式朴素贝叶斯是基于贝叶斯算法和特征独立性假设的一种分类方法，基本思想是计算条件概率，该模型常用于文本分类，特征值是单词的出现次数。

本文首先对数据进行预处理，即生成相应的词向量，其次将 67%的数据作为训练集，33%的数据作为测试集，运用训练集生成训练模型后，调用相应的概率值实现测试集的分类，程序结果如下：

```
In[13]:
   ...: from sklearn.model_selection import train_test_split
   ...: from sklearn.feature_extraction.text import CountVectorizer
   ...: from sklearn.feature_extraction.text import TfidfTransformer
   ...: from sklearn.naive_bayes import MultinomialNB
   ...:
   ...: X_train, X_test, y_train, y_test = train_test_split(df['cut_留言详情'], labels, random_state=0)
   ...: count_vect = CountVectorizer()
   ...: X_train_counts = count_vect.fit_transform(X_train)
   ...:
   ...: tfidf_transformer = TfidfTransformer()
   ...: X_train_tfidf = tfidf_transformer.fit_transform(X_train_counts)
   ...:
   ...: clf = MultinomialNB().fit(X_train_tfidf, y_train)
In[14]: clf
Out[14]: MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True)
```

图 4 朴素贝叶斯算法实现

从图 4 可以看出，在 sklearn 中，MultinomialNB()类的 partial_fit()方法可以进行训练。这种方式比较适合于训练集大到内存无法一次性放入的情况，可得到分类准确度为 68%，模型效果一般。下面分析随机森林的精度。

- 随机森林

随机森林（Random Forest，简称 RF）是一种由多颗决策树集成的分类方法，可以集成所有的分类结果，将分类结果概率最大的类作为最终的分类结果输出。下面给出随机森林模型中树的个数不同情况时模型精度的变化情况：

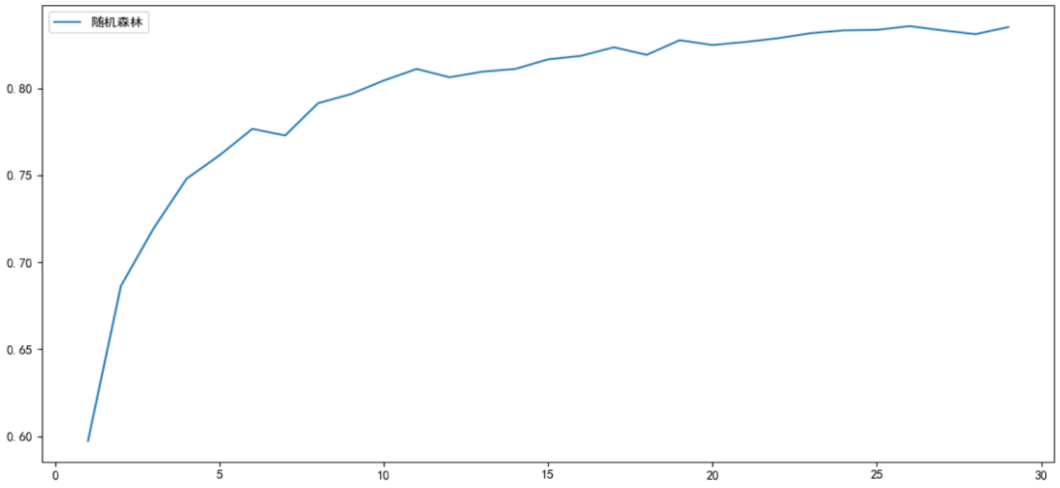


图 5 随机森林模型中树的个数及对应精度

从上图可以看出，树的个数越多，模型精度越高，但相应的训练时间也会增加。当树的个数大于 10 时，模型精度增加缓慢，故可取树的个数为 10，此时精度为 83.2%，模型的各个参数取值如下表所示：

```
In[15]: rfc = RandomForestClassifier(random_state=10)
   ...: rfc.fit(X_train_tfidf,y_train)
Out[15]:
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=None, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=None,
            oob_score=False, random_state=10, verbose=0, warm_start=False)
```

图 6 随机森林参数取值及含义

从图 6 的程序结果可以看出，warm-start 取 False 说明随机森林采用有放回且从头开始训练的模式，每次随机产生器产生随机种子数为 15，当树的个数为 10 时，计算过程耗时大，但模型精度只达到了 83%，模型效果一般。

- 多分类 logistics 回归

常见的是二分类 logistics 回归模型，分为{事件发生}和{事件不发生}，则可运用同样的思想拓展为多分类问题，将 67%的数据作为训练集，33%的数据作为测试集，建立多分类的 logistics 回归模型，求得最优参数如下表所示：

```
   ...: logistic_model = LogisticRegressionCV(multi_class='multinomial', solver='lbfgs')#构造模型实例化逻辑回归模型对象
   ...: logistic_model.fit(train_X, train_y)#进行拟合回归
   ...: predict_y = logistic_model.predict(test_X)#利用数据进行测试预测
   ...:
   ...: pd.DataFrame(confusion_matrix(test_y, predict_y), columns=labelEncoder.classes_, index=labelEncoder.classes_)
   ...: #绘制对应的混淆矩阵
Out[26]:
          交通运输  劳动和社会保障  卫生计生  商贸旅游  城乡建设  教育文体  环境保护
交通运输       114         4     4    27    55     4     2
劳动和社会保障      3       560    23     3    21    15     2
卫生计生         0        27   247    16     5     4     1
商贸旅游        17         5     5   292    50    12     2
城乡建设        18        22     3    33   538    13    21
教育文体         0        23     3    16    19   474     4
环境保护         1         1     0     5    38     1   287
```

图 7 多分类 logistics 回归模型结果

从上图程序结果可以得到，对于附件 2 中的留言处理，考虑到留言样本数据较大，为多分类问题，故分类方式（multi_class）应选用 multinomial，即 many-vs-many（MvM）多元逻辑回归，对应算法（solver）可用 lbfgs 拟牛顿算法，通过利用损失函数二阶导数矩阵来多次迭代优化损失函数，可得到分类精度为 85.3%，模型效果一般。

本文将样本数据进行交叉验证，分为 10 个训练集，得到以上 4 种模型精度如下图所示：

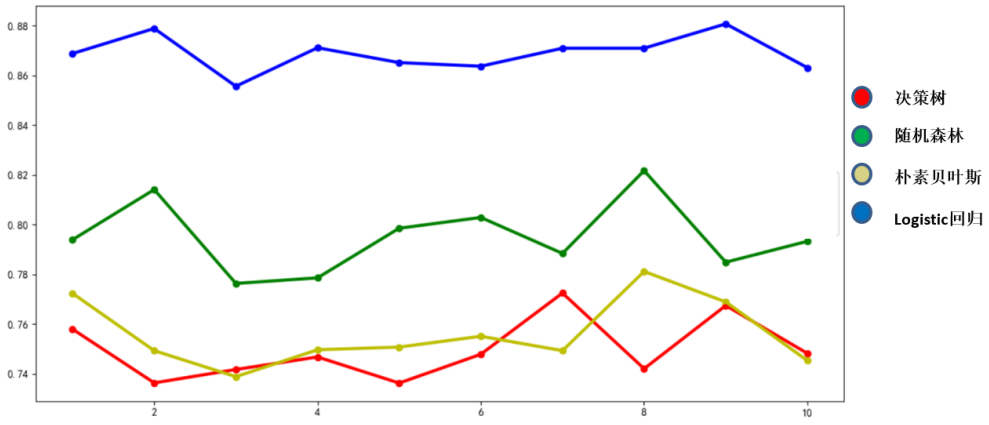


图 8 各模型精度对比图

由图 8 可以看出，在多次交叉验证中，多分类 logistics 回归模型的精度远远高于其它三类模型。但模型准确率仍有改进的空间，故下面探索现代算法线性支持向量机模型。

- 线性支持向量机

线性支持向量机是一种现代化分类算法，已经由经典的二分类问题改进到多分类问题，本文通过组合多个二分类问题并采用向无环图（DAG）算法实现多分类，可保证训练数据和测试数据都有较好的分类效果。

通过多次调整训练集参数比例，选用 67%的样本数据作为训练集，33%为测试集，模型精度达 91%，模型效果较好，故选用此方法进行一级标签的分类，具体操作如下。

6.1.3 线性支持向量机处理留言问题

（1）线性支持向量基本思想：

线性支持向量机是在 VC 维理论和风险最小理论基础上创建的，在样本线性可分的情况下寻找最优分类面，现已由最经典的二分类问题扩展到了多分类问题，分类模型的精度主要由分类精度和分类速度来评判。对于机器学习处理留言问题，应该在保证精度较高的情况下尽快实现分类，故选择多线性支持向量机分类模型。

设线性可分的样本集为(x_i, y_i)，则对应的分类函数可写为$g(x)=w\cdot x+b$，则分类面的方程为$g(x)=0$，则先进行判别函数归一化，使得待分类的样本都满足$|g(x)|\geq 1$，保证离分离面最近的样本满足$|g(x)|=1$，则最优函数为

$$g(x)=\mathrm{sgn}[\sum_{i=1}^{n}\alpha_i^* y_i(x_i\cdot x)+b^*] \qquad (5)$$

其中 sgn(·) 为符号函数，$(x_i \cdot x)$ 为内积运算，α_i^* 和 b^* 为最优分类面的参数。则分类面可作下列说明：

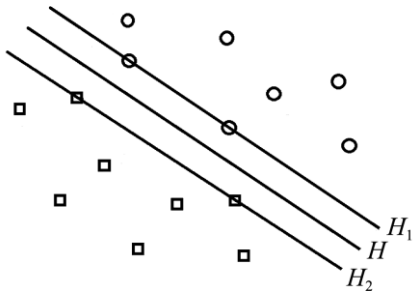


图 9 线性支持向量机分类面说明

经典问题中无法解决多分类问题，则通过优化和组合多个二分类问题实现，经典算法有 One-Versus-One(1-v-1)、One-Versus-Rest(1-v-r)、二叉树、Directed Acyclic Graph(DAG)等，1-v-r 算法和 1-v-1 算法都需要构造每个二分类器进行计算，如果样本数据过大或类别过多，计算速度就会很慢，故在面对大数据下的留言问题时，选用有向无环图（DAG）算法实现分类。

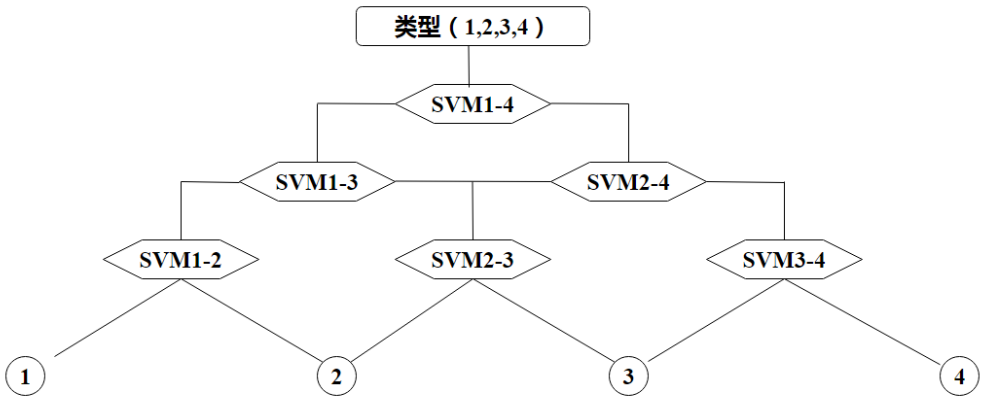


图 10 完全二叉树构造示意图

DAG 算法构造所有的二分类器，作为有向无环图的节点，底层的“叶”由 N 个类别组成，最底层有 N 个叶节点，如下图所示，按照“由上到下”的原则，从顶部分类器开始，根据底部的分类结果，判断采用下一层的左节点还是右节点，直到底层的某个“叶”为止并编号，对应“叶”的序号即可定义为类别。采用排除法对测试样本进行分类识别，将样本输入到已构造的子二分类器中，排除掉不可能的类别，最终得到底层的某个“叶”所对应的类别。

（2）划分训练集

为了保证分类模型具有较高的泛化能力，同时不失准确率，可将样本数据划分为训练集和测试集。本文选用 67%的样本数据作为训练集，33%为测试集。

（3）由预测函数得到分类结果

通过 predict 得到了类别预测函数，当输入留言内容时，可自动回复留言归属的一级指标。下面给出部分程序运行截图：

```
In[19]: myPredict("L9县一幼建设教学楼不要与L9县十三五规划修缮熊公馆相背离")
城乡建设
In[20]: myPredict("何时发布2019年A7县普惠性幼儿园清单及收费标准")
教育文体
In[21]: myPredict("L6县企业社会保险管理局提前内退人员对政策执行表不满")
劳动和社会保障
In[22]: myPredict("L5县医院的药品以次充好流入县内各大医院")
卫生计生
In[23]: myPredict("投诉D11市信美物业公司联合电信公司恶意垄断小区宽带")
商贸旅游
```

图 11 预测函数示例

则通过预测函数，可得到所有留言的一级指标。分类结果如下图所示：

图 12 一级指标分类情况

由上图可知，“城乡建设”类、“劳动和社会保障”类留言数量较多，占比 20%以上；“交通运输”类数量最少，仅为 6.64%。值得注意的是，附件 1 中一级分类标签包含“医疗卫生”、“公共建设”、“党务政务”等 15 类，但附件 2 中留言分类结果只涉及上述 7 个类别，这可能是因为附件 2 给出的仅是部分留言信息。

6.1.4 分类结果评价

对于留言分类效果的检验，国际上通用的有查准率、召回率、F-score 等多项指标。下面对各项指标给出具体解释：

- “**查准率** P_i”：指的是检出的相关信息量在文献总量的占比。
- “**召回率** R_i”：召回率又称为查全率，指的是检出的相关信息量在系统中相关信息总量的占比。

当数据样本量较大时，查准率和召回率是相互制约的，即查准率高时召回率低，在实际情况中应在保证较高的召回率下，尽量提高准确率，故定义一个综合性指标：

- “**F-score**”：是查准率和召回率的调和值，计算公式为：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i} \tag{6}$$

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

（1）各个模型精度比较

通过计算可得，各个模型的相关精度指标如下：

```
In[30]: for model in models:
   ...:     model_name = model.__class__.__name__
   ...:     accuracies = cross_val_score(model, features, labels, scoring='accuracy', cv=CV)
   ...:     for fold_idx, accuracy in enumerate(accuracies):
   ...:         entries.append((model_name, fold_idx, accuracy))
   ...: cv_df = pd.DataFrame(entries, columns=['model_name', 'fold_idx', 'accuracy'])
   ...: print(cv_df[cv_df['fold_idx']==3])
                 model_name  fold_idx  accuracy
3   RandomForestClassifier         3  0.406301
8                LinearSVC         3  0.906573
13           MultinomialNB         3  0.676263
18      LogisticRegression         3  0.854427
23  RandomForestClassifier         3  0.406301
```

图 13 多线性支持向量机运行结果

由图 13 可知，各个模型中准确率（accuracy）最高的是线性支持向量机（LinearSVC），运用此方法对附件 2 中的留言进行分类，

朴素贝叶斯、多分类 logistics 回归模型都是传统分类模型，在面对大数据文本时计算量大，效率低；随机森林对低维数据处理效果较好，在面对高维数据时计算量较大且精度不高。相比较起来，线性支持向量机不仅计算速度快，而且达到了较高的精度。以下绘制不同模型精度的玉玦图以直观认识：

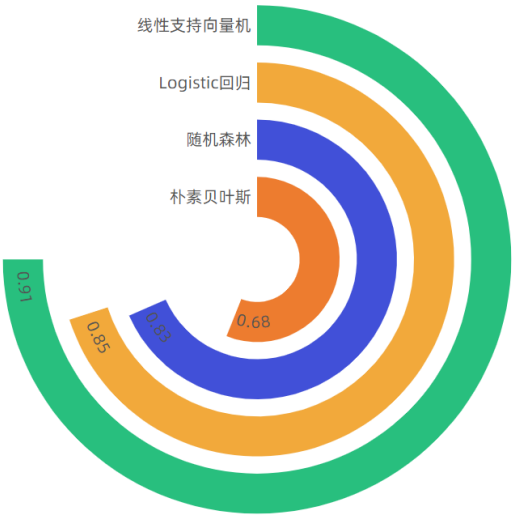


图 14 不同模型分类精度的玉玦图

上述玉玦图中弧长表示模型精度，其越接近圆，表示模型准确率越高。可以得到多项分布朴素贝叶斯分类模型准确率最低，仅为 68%，这是因为该方法较适用于二分类问题，在多分类问题中精度较差；另外三个模型：多分类 logistics 回归、随机森林和线性支持向量机精度相对较高，均在 80%以上。其中线性支持向量机准确率最高，达 91%，故选择线性支持向量机方法对附件 2 留言进行一级标签分类。

（2）线性支持向量机模型精度

表 1 线性支持向量机分类模型结果及精度

指标	类别									
	城乡建设	环境保护	交通运输	教育文体	劳动和社会保障	商贸旅游	卫生计生	微平均	宏平均	加权平均
查准率	0.82	0.96	0.94	0.94	0.91	0.90	0.94	0.99	0.92	0.90
召回率	0.95	0.92	0.70	0.94	0.93	0.83	0.86	0.90	0.88	0.90
F-score	0.88	0.94	0.80	0.94	0.92	0.86	0.90	0.90	0.89	0.90
计算用时	共耗时 1.980178 秒									

由上表可知，分类计算总用时为 1.98 秒，计算用时少；查准率、召回率和 F-score 加权平均值为 90%，模型精度高。各个类别中的 F-score 均在 80%以上，说明分类既能

保证较高的查全率，又能保证较好的准确率。部分召回率较低如“交通运输”类，这可能是由于该类样本量较少所致。由查准率可知部分样本存在着误判的情况，为进一步研究哪些类别之间出现了误判，可绘制类别混淆矩阵，得到样本分类的真实情况。

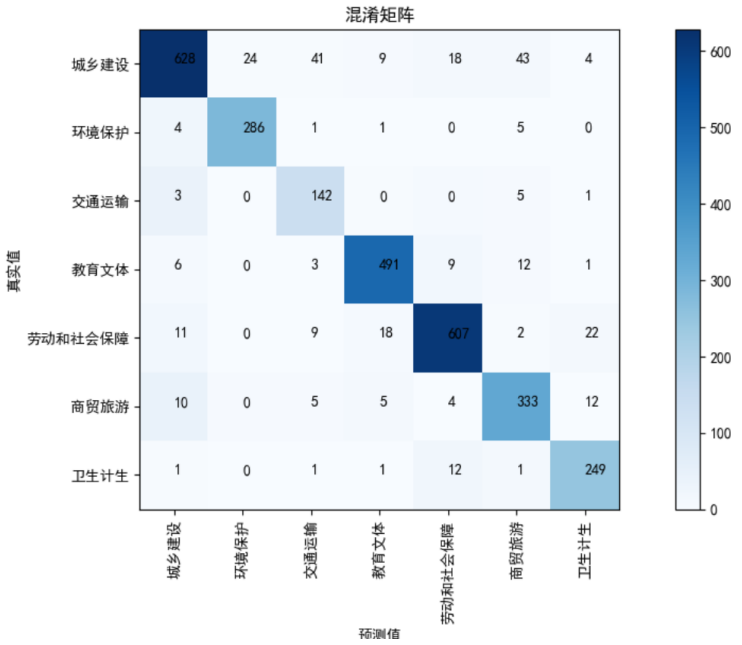


图 15 不同类别混淆矩阵

混淆矩阵的每一行表示预测为该类别的真实样本数量，每一列代表了被预测为该类别的样本数量。混淆矩阵的主对角线元素表示正确的分类结果，如第一行中的 628 表示有 628 个样本数据被正确预测到了“城乡建设”类，24 表示有 24 个样本数据本应该属于“城乡建设”类，却误判到了“环境保护”类，其他同理。同时，可以看出除了交通运输类的计算精度较低之外（原因可能在于该类样本数量较少），其他类都保证了较高的精度，说明模型选用合适，分类效果较好。为具体了解误判情况，下给出部分误判具体留言：

表 2 部分一级指标误判情况

劳动和社会保障 预测为 教育文体 ： 9 例.

	一级标签	留言详情_处理后
5433	劳动和社会保障	现在孩子都是被爷爷奶奶宠大的不管在哪只要受一点点的委屈就会大夸其词我孩子上一年级每次放学回来...
6450	劳动和社会保障	贺厅长您好我是今年5月份参加A市心理咨询师二级考试的考生7月15号省人社厅网站上公布了考试成...
6667	劳动和社会保障	我有2个小孩全家刚从外省回C市一个3岁一个5岁户口都属于C1区居民作为小孩的家长请问我应该去...
5287	劳动和社会保障	你好彭厅长我是J市技师学院的一名学生2010年就读于J市技师学院现第三个学期已接近尾声而我们...
7027	劳动和社会保障	我是于2005年在外省毕业中医学类的中专学生因为这么多年都没有参加助理医师考试现在还可以参加...
6592	劳动和社会保障	I6市市2017年教师绩效工资百分之三十奖励部分已经到位其实周边县市及外地县市都是半年内至少...
6297	劳动和社会保障	A8县教育局在给教职工生育保险缴费方面一直欠缴整个教育系统这么多人欠费不是一两个月而是一直都...
6021	劳动和社会保障	我们是A8县电影发行放映公司以下简称电影公司聘用的12名农村公益数字电影放映员我们从1993...
6134	劳动和社会保障	敬爱的领导你们好我是15年毕业并参加护士资格考试的一名护理专业学生在A市卫生职业学院毕业已经...

通过上述留言的具体分析， “劳动和社会保障类”类中编号为 5433 的留言，因为涉及“学校”、“学生”等关键词，涉及到关键词信息的相似或重叠，故被误判到“教育文体”类。综上可知，线性支持向量机理论已达到较高的精度，仅存在部分误判的情况。

6.2 模型二: K-means文本聚类模型

6.2.1 文本预处理

问题二要求提取特定地点或特定时间的热点问题，首先应提取“地名”信息，如“A1”、“A2”等，可判别是否为同一地区的问题；其次考虑到同一地区存在不同类型的问题，且地名重复率较高，会影响到聚类结果，故先将地名加入到停用词词库中，再对留言中的“事件”信息进行聚类，会达到更好的聚类效果。如留言主题为“咨询 A 市落户的问题”时，可提取出地名信息为“A”，同时得到分词结果为：“咨询、落户、问题”。

6.2.2 构建留言主题的词频矩阵

群众留言主题相当于是留言详情的中心句，故可对留言主题进行处理，这样不仅能够保证信息的完整性，而且计算量较小。对留言主题进行分词后，可根据分词结果构建词频矩阵。

如编号 264944 留言“A2 区丽发新城附近修建搅拌厂噪音、灰尘污染”、编号 268300 留言“A2 区丽发新城附近修建搅拌厂噪音污染导致生活不正常”、编号 272361 留言“丽发新城小区旁搅拌厂严重扰民”，若分词结果为[A2 区、丽发新城、附近、修建、搅拌厂、噪音、灰尘、污染、导致、生活、不正常、严重、扰民]，则以上三条留言对应的词频矩阵如下表所示：

表 3 词频矩阵

留言编号	留言主题	丽发	新城	附近	修建	搅拌厂	噪音	灰尘	污染	导致	生活	不正常	严重	扰民
264944	A2区丽发新城附近修建搅拌厂噪音、灰尘污染	1	1	1	1	1	1	1	1	0	0	0	0	0
268300	A2区丽发新城附近修建搅拌厂噪音污染导致生活不正	1	1	1	1	1	1	0	1	1	1	1	0	0
272361	丽发新城小区旁建搅拌厂严重扰民	1	1	1	0	1	0	0	0	0	0	0	1	1

每行元素表示每条留言的词频矩阵，如第一行元素表示编号为 264944 的留言分别出现了 1 次 “丽发”、“新城”、“附近”、“修建”、“搅拌厂”、“噪音”、“灰

尘”、“污染”，没有出现“导致”、“生活”、“不正常”、“严重”、“扰民”，其他同理。

6.2.3 对留言主题进行K-means文本聚类

生成留言主题的词频矩阵之后，根据每条留言对应词频矩阵，对留言进行分类。常用的聚类分析方法有很多种，如 K-means 聚类分析（适用于样本聚类）、层次聚类（适用于对变量聚类）、基于密度的聚类算法等等，本文选取 K-means 聚类对留言进行分类。

（1）K-means 聚类的原理

若数据包含 n 个 d 维样本点，$Y=\{y_1,y_2...y_i...y_n\}$，其中 $x_i \in R^d$，聚类将数据集 Y 进行 k 类划分 $L=\{\ l_k,i=1,2...k\}$，每个划分代表一个类 l_k，每个类 l_k 有一个类别中心 φ_i，选用欧式距离判别法和距离最近分类原则，计算类内所有的样本点到聚类中心的距离平方和

$$J(l_k)=\sum_{y_i\in l_k}\| y_i-\varphi_k \|^2 \tag{7}$$

聚类的目标是保证 $J(l_k)$ 最小

$$J(L)=J(l_k)=\sum_{y_i\in l_k}\| y_i-\varphi_k \|^2=\sum_{k=1}^{K}\sum_{i=1}^{n} I_{ki}\| y_i-\varphi_k \|^2 \tag{8}$$

其中 $I_{ki}=\begin{cases}1,\mathrm{y}_i\in l_k\\0,\mathrm{y}_i\notin l_k\end{cases}$，故由最小二乘法可知聚类中心 φ_i 应该为类别 l_k 内各个样本点的平均值。

（2）K-means聚类的步骤：

1）从样本数据 Y 随机选取 k 个元素做为初始聚类中心。

2）分别计算各个类中其他元素到 k 个类中心的欧式距离，将这些元素分别划分到距离最近的类。

3）根据聚类结果，重新计算 k 个类中心，计算方法是取类中所有元素各自维度的算术平均数。

4）将 Y 内全部元素按照新的类中心重新聚类。

5）多次重复第4步，直到聚类结果不再发生变化。

对留言主题进行 K-means 聚类，得到分类结果，下面仅给出热点问题的的误判的情况，总分类结果见附件：

表 4 部分热点问题误判情况

留言编号	留言问题	原类别	改为类别
2707、68等	A2区丽发新城附近建搅拌站噪音扰民问题	27	1
2381、4320	A市魅力之城小区底层商铺深夜营业噪音严重	9	19
4317、2440	A市魅力之城小区底层商铺深夜营业噪音严重	10	19
4315、4187、58	A市魅力之城商铺无排烟管道小区内到处油烟味问题	18	19
3375	反映A市人才补贴问题	14	25
2745	询问A市人才新政落户人员的档案管理问题	18	25
3420、1751、1154、469	咨询A市新引人才奖励政策相关问题	26	25
2356、689、2229、999	伊景园捆绑销售车位问题	12	28
3403、2950、1297	伊景园捆绑销售车位问题	18	28
1951、4147	伊景园捆绑销售车位问题	24	28
3269	A市丽发新城三期违建是否会影响交房问题	1	31
3244	A市丽发新城旁商品交易市场影响家长送孩子们上学	1	31
1353	A市贷款购房与购房资格的相关问题	25	33

以第一项为例，留言编号为 2707、68 的留言详情为“A2 区丽发新城附近建搅拌站噪音扰民问题”，分类结果为 27 类，但实际应属于第 1 类（仅给出聚类号，不定义聚类名），可能是因为产生了关键词相似或重叠。对此，可由后期人工筛选实现精确分类。

6.2.4 筛选热点问题

对于群众留言中的热点问题，不仅要考虑某一问题的留言总量，还应结合点赞数和反对数综合考虑，其侧面反映了社会关注度。通过定义“关注度”、“热度值”等指标，确定出最终的热点问题。

- “**关注度**”：群众对政务问题的关心程度。定义 1 个点赞数或 1 个反对数为 1 个关注度，即

$$\text{关注度}=\text{点赞数}+\text{反对数} \tag{9}$$

- “**热度值**”：可直观反映热点问题的指标，是综合考虑留言总数和关注度的量化值。因部分问题存在留言数较少而关注度较高的情况，故应统一数量级，本题定义一个留言为 50 个热度值、1 个关注度为 1 个热度值。即

$$热度值=留言总数*50+关注度 \quad (10)$$

其他定义情况在灵敏度分析中给出。

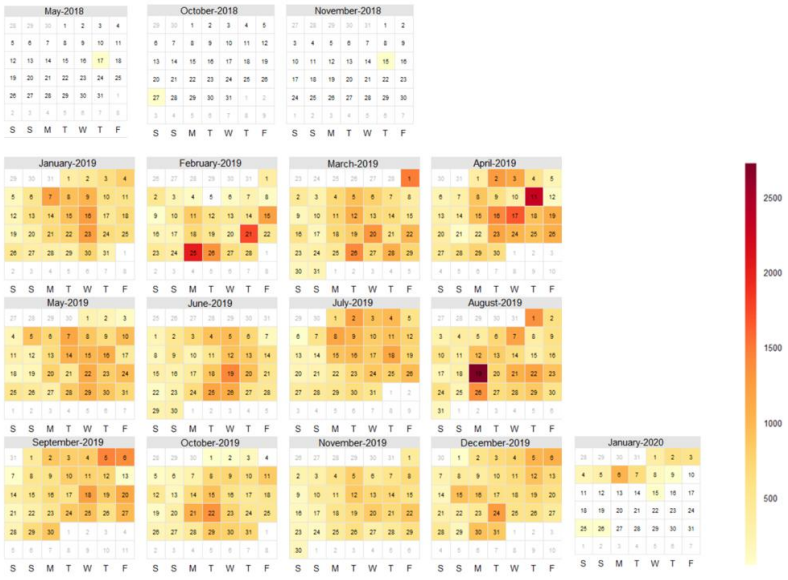


图 16 附件三留言热度值日历图

从上述日历图中可以看到，随着颜色由浅变深，热度值逐渐增加。总体来看，留言问题主要集中在周一到周五的工作日，且 2018 年留言热度较低，仅仅存在于 5 月、10 月和 11 月；2019 年 1 月到 9 月，热度值普遍较高，最高在 2019 年 8 月 19 日、2019 年 2 月 25 日、2019 年 4 月 11 日，且在 2019 年 10 月以后，热度值有所降低，说明政府的相关部门在逐渐改善各方面问题，到 2020 年 1 月，热度值有了明显的下降。

本文筛选出留言总数排名前 4、关注度排名前 5 的问题，通过计算热度值指标，排序选出了排名前 9 的问题，如表 5：

表 5 排名前 9 的热点问题

留言总数	留言编号	问题详情	点赞数	反对数	关注度	热度值换算结果	热度最终排序
14	268251、220711等	A4区58车贷非法经营	2386	0	2386	3086	1
53	208285、208714等	丽发新城搅拌站污染扰民严重	48	2	50	2700	2
51	191001、214975等	伊景园售房捆绑销售车位	25	1	26	2576	3
8	208636、234086等	五矿万境水岸存在住房问题	2106	0	2106	2506	4
26	247736、225657等	外来人才补贴政策问题	28	4	32	1332	5
21	272122、236798等	魅力之城小区商铺污染扰民	18	18	36	1086	6
6	263672、202575等	A4区绿地海外滩存在住房问题	691	0	691	991	7
1	193091	A2区碧云街道丽发新城小区物业存在问题	242	0	242	292	8
2	284571、253369等	长楚高速存在噪音扰民问题	109	0	109	209	9

上表中的数值随着颜色由蓝变红而逐渐增大，通过观察可发现热度值最终排序前 5 的分别是留言总数和关注度较高的问题，这与实际情况相符合，即每一类问题的留言总数和关注度都是民意的反映方式，故从两方面综合考虑较为合适。整理得出排名前五的热点问题如下：

表 6 排名前五的热点问题

热度最终排序	热点问题ID	热度值换算结果	时间范围	地点/人群	问题描述
1	1	3057	2019/1/11至2019/7/8	西地省58车贷案受害者	A4区58车贷非法经营
2	2	2700	2019/11/2至2020/1/26	丽发新城小区居民	丽发新城搅拌站污染扰民严重
3	3	2576	2019/7/7至2019/9/1	伊景园小区居民	伊景园售房捆绑销售车位
4	4	2506	2019/1/15至2019/9/19	五矿万境水岸小区居民	五矿万境水岸存在住房问题
5	5	1332	2018/11/15至2019/12/2	A市区外来人才	外来人才补贴政策问题

由上表得知，排名前五的问题为“A4 区 58 车贷案问题”、“丽发新城搅拌站污染扰民问题”、“伊景园捆绑销售车位问题”、“五矿万境水岸住房问题”、“A 市外来人才补贴问题”，是综合考虑留言总数和关注度得出的结果，政府相关部门可对这些问题给出相应举措。

此外，为进一步了解政府部门是否切实解决了群众的问题，可通过绘制留言总数排名前四的留言问题日历图，了解留言问题的时间跨度，明确问题的处理情况。留言问题日历图如下：

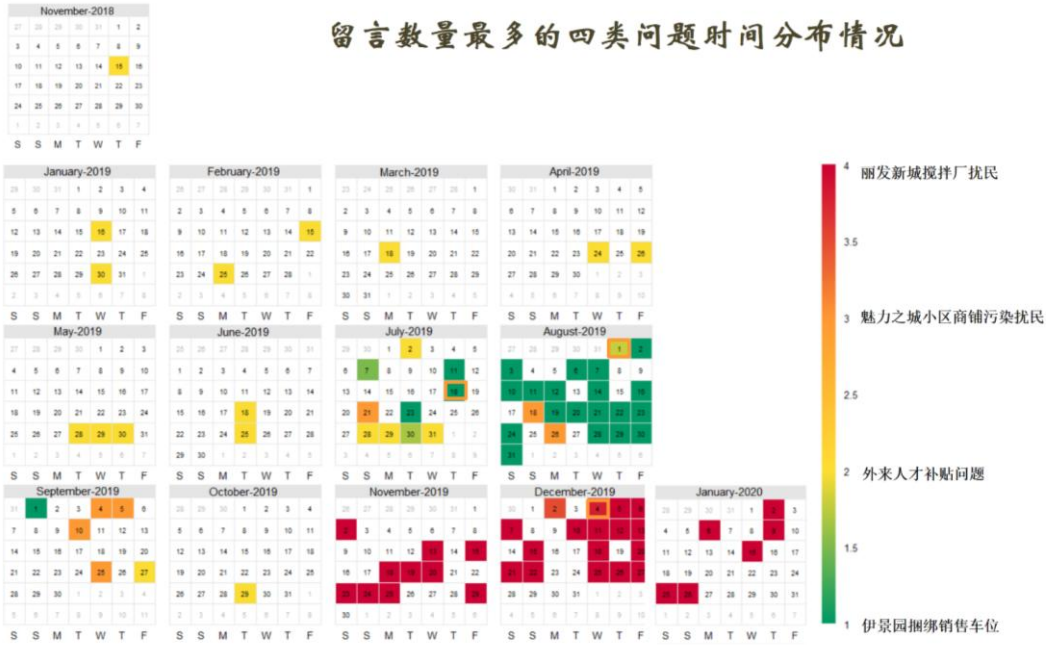


图 17 热点问题日历图

从图 17 可以直观看出，不同颜色分别表示不同类型的问题，总体来看热点问题主要集中在周一到周五的工作日内，符合问政系统运行时间。按照时间顺序来看，首先出现的黄色热点“外来人才补贴问题”时间跨度较长，分布散乱，说明这是一个长存在待解决的问题；绿色热点“伊景园捆绑销售车位”主要集中在 2019 年 7 月-8 月，9 月仅为个例，说明该问题是短期存在且政府部门已经给出了相应的解决措施；橙色热点“魅力之城小区商铺油烟、噪音扰民”仅在 2019 年 8 月、9 月出现个例；红色热点问题“丽发新城搅拌机扰民”反映主要集中在 2019 年 11 月-2020 年 1 月，时间跨度最长，热点度极高，但在 2020 年初，数量有了明显的减少。

综合来看，即使是留言数量最多的问题，但随着时间变化其数量都在逐渐减少，这与政府的及时解决的高效率是密不可分的。

6.3 模型三：基于主成分分析权重修正的层次分析模型

考虑到评价指标涉及到定性和定量变量，难以量化，故采用基于主成分分析法权重修正的层次分析模型对留言回复给出评价。

6.3.1 确立指标体系

政问留言系统优越性应基于能否切实解决以下问题：（1）留言回复是否及时？（2）留言答复是否答非所问，与所提问题的相关性如何？（3）政问留言回复是否规范性用词？（4）对于不同年龄层次、不同学历层次的留言者能否保证其理解满意等多个问题，综合考虑以上情况，本文定义以下指标：

“答复时间间隔”：附件 4 中群众留言时间和政府部门答复时间的时间差值。注意到在答复意见末尾也有一个时间，将这部分作为政府部门拟定留言的时间，最后一列给出的留言时间是在网络上公布留言的时间，考虑到政府内部运行机制，对留言意见末尾的时间不予考虑。

“答复与留言相关度”：可理解为两个样本之间的相似度或相关性。以附件 4 第一条留言为例:

群众留言“A2 区景蓉华苑物业管理有问题”;留言回复“您好，首先感谢您对我们工作的信任和支持，关于您在平台栏目给胡华衡书记留言，反映“A2 区景蓉花苑物业管理有问题”的情况已收悉。现将我们调查处理情况答复如下……”

（1）首先应对文本进行分词处理，构造包含语气词、连接词等的停用词库，使用 jieba 对文本分词，得到两个列表:

A=[A2 区 景蓉华苑 物业 管理 有 问题]

B=[您好 感谢 信任 支持 平台栏目 给 胡华衡 书记留言 反映 A2 区 景蓉花苑 物业 管理 有 问题 情况 已 收悉 现 调查 处理 情况 答复 如下…]

（2）TF-IDF 构建词向量，且将两个列表涉及到的所有的分词结果放于一个列表 C 中，即为：

C=[A2 区 物业 管理 有 经理 问题 您好 景蓉华苑 感谢 信任 支持 平台栏目 给 胡华衡 书记留言 反映 情况 已 收悉 现调查 处理 情况 答复 如下…]（TF-IDF 算法构造词向量顺序是随机性的，这里仅仅是举例说明）

（3）对位置列表进行 oneHot 机器编码，计算 C 列表中每个分词在 A、B 列表内出现的次数。得到编码后的列表为：

AoneHot=[1，1，1，1，0，1，0，1，0……]

BoneHot=[1，1，1，1，0，1，1，1，1……]

AoneHot 中第 1 个元素“1”表示 C 列表中第 1 个关键词“A2 区”在 A 列表中出现了 1 次；第 5 个“0 表示”C 列表中第 5 个关键词“您好”在 A 列表中出现了 0 次，BoneHot 同理。

（4）计算余弦相似度

得到两个文本的词频向量矩阵后，可计算这两个向量之间的夹角余弦值：

$$\cos(\theta)=\frac{x_1y_1+x_2y_2+\ldots x_ny_n}{\sqrt{x_1^{\,2}+x_2^{\,2}+\ldots x_n^{\,2}}*\sqrt{y_1^{\,2}+y_2^{\,2}+\ldots \mathrm{y}_n^{\,2}}} \tag{11}$$

假设两个词频向量均为 n 维，x_i 表示第一个样本 X 的 i 个元素，y_i 表示第二个样本 Y 的 i 个元素。则 AoneHot、BoneHot 的余弦相似度为：

$$\cos(\theta)=\frac{(1\times1)+(1\times1)+(1\times1)+(1\times1)+(0\times0)\ldots}{\sqrt{1^2+1^2+1^2+1^2+0^2+\ldots}\times\sqrt{1^2+1^2+1^2+1^2+0^2+\ldots}}=0.239275277 \tag{12}$$

“规范性用词”：留言回复信息中是否存在“您好”、“你好”、“谢谢”、“感谢”等礼貌性用词和留言日期。该指标可以反映出行政系统的规范性和政务素养。

“答复文本长度”：去除空格、无效字符串等之后的文本长度。文本长度较短，可能存在回复敷衍的情况，该指标可以反映回复的认真程度。

“标点符号使用的合理性”：在政问留言回复信息中，是否存在标点符号使用过少或过多的情况。在回复中标点符号使用较少，则可能造成句子冗余，难以理解；标点符号使用过多，可能存在不认真回复的情况，该指标可以反映出留言回复的可解释性。将各指标情况汇总如下：

表 7 政问回复评价指标

目标层	准则层	指标层
政府回复留言评价指标体系 A	及时性 B1	答复时间间隔 C1
	相关性 B2	答复与留言相关度 C2
	完整性 B3	规范性用词 C3
		答复文本长度 C4
	可解释性 B4	标点符号使用的合理性 C5

如若各个指标之间存在着相关性，则可能会产生多重共线性等问题。通过绘制各指标相关图，分析指标定义是否合理：

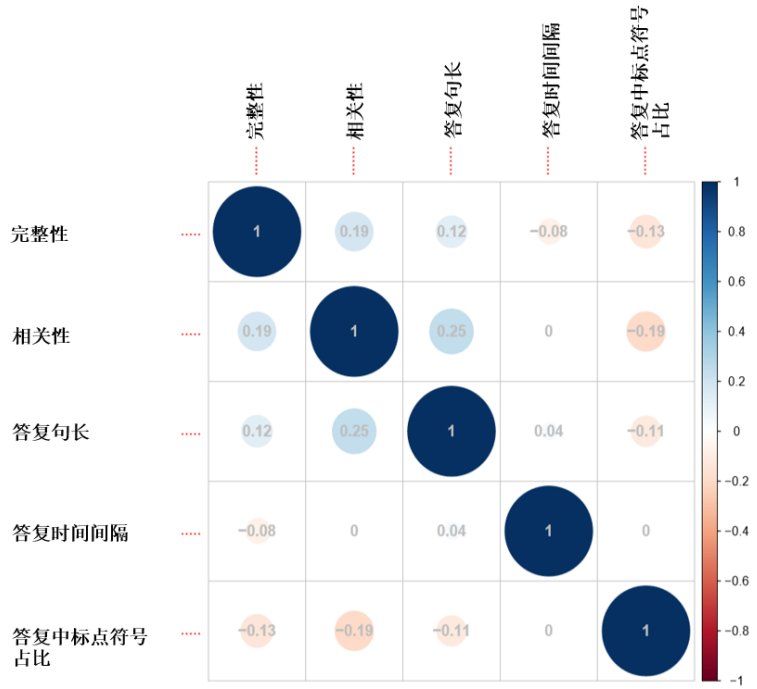


图 18 各指标相关系数

由图 18 可知，随着颜色逐渐变深相关性程度变强。主对角线元素表示指标自身的相关性，恒为 1；其他元素颜色较浅，最高达 0.25，属于弱相关，指标选用合理。

6.3.2 指标权重的确认

层次分析法是一种主观性较强的方法，其特点是在对复杂的决策问题的本质、影响因素及其内在关系等进行深入分析的基础上，利用较少的定量信息使决策的思维过程数学化，从而为多目标、多准则或无结构特性的复杂决策问题提供简便的决策方法。尤其适合于对决策结果难于直接量化的场合。但该方法中各指标权重的确定都是基于专家评分或模糊数学理论，为解决这一不足，引入主成分分析中载荷矩阵系数占比来作为指标权重的一个估计值。

主成分分析是将多个指标化为少数几个综合指标的方法，这些综合指标通常表示为原始变量的线性组合，变量在线性组合中的系数即可客观反映其对综合指标的贡献程度，对其归一化处理后用于做指标权重的参数估计显得较为合理。但应注意，主成分中有特征值累计贡献率这一指标，只有当累计贡献率达到要求的精度时，才可以选用，并由此确定模型的参数。

若上述各指标对应的特征值分别为λ_1，$\lambda_2 ... \lambda_6$，特征值对应的特征向量为$u_1, u_2 ... u_6$，则需对特征向量进行规格化处理，即

$$a_{ij} = u_{ij}\sqrt{\lambda_{ij}} \tag{13}$$

则得到因子载荷矩阵为：

$$A = (a_{ij}) = \begin{bmatrix} u_{11}\sqrt{\lambda_{11}}, u_{12}\sqrt{\lambda_{12}} ... u_{1n}\sqrt{\lambda_{1n}} \\ u_{21}\sqrt{\lambda_{21}}, u_{22}\sqrt{\lambda_{22}} ... u_{2n}\sqrt{\lambda_{2n}} \\ ..... \\ u_{n1}\sqrt{\lambda_{n1}}, u_{n2}\sqrt{\lambda_{n2}} ... u_{nn}\sqrt{\lambda_{nn}} \end{bmatrix} \tag{14}$$

计算可得C1-C6指标的因子载荷矩阵为：

$$A = \begin{bmatrix} 1,2,3,4,5 \\ 1,2,3,4,5 \\ 1,2,3,4,5 \\ 1,2,3,4,5 \\ 1,2,3,4,5 \end{bmatrix} = \begin{bmatrix} 1.000 & 0.029 & -0.005 & -0.003 & -0.078 \\ 0.029 & 1.000 & -0.096 & 0.237 & 0.105 \\ -0.005 & -0.096 & 1.000 & -0.193 & -0.134 \\ -0.003 & 0.237 & -0.193 & 1.000 & 0.185 \\ -0.078 & 0.105 & -0.134 & 0.185 & 1.000 \end{bmatrix}$$

主成分共筛选出 5 个综合指标，通过特征值的累计贡献率可知，前四项为 85.4%，效果较差，故用 5 个指标的系数平均值占总平均值的比例得出权重指标。

表 8 各指标的权重系数

目标层		准则层		指标层	
类别	权重值	类别	权重值	类别	权重值
政府回复留言评价指标体系A	1	及时性B1	0.186383	答复时间间隔 C1	0.186383
		相关性B2	0.161758	答复与留言的关度 C2	0.161758
		完整性B3	0.447038	规范性用词 C3	0.214514
				答复文本长度C4	0.232524
		可解释性B4	0.204822	标点符号使用的合理性C5	0.204822

6.3.3 计算留言总得分

（1）修正异常值

当指标中存在异常值时，会对结果造成不利的影响。故在数据处理之前应该先剔除异常值，如编号为“159285” 的留言提出到回复间隔为 1161 天，归一化对总体结果影响较大，但考虑到留言的真实情况，不应直接剔除，故依据样本数据分布予以修改为 171。

（2）指标数据标准化

在计算出各项指标的权重之后，可将指标量化，得到在区间（0,1）上的取值，称这一过程为数据的归一化或标准化。值得注意的是，各指标因其含义不同，其最优取值也不同，分“成本型”、“效益型”和“中间型”。顾名思义，成本越小越好，符合成本型的数据标准化满足取值越小，效果最优；效益值越大越好，符合效益型的数据标准化满足取值越大，效果最优；符合中间型的数据标准化满足当数据趋于中间某一取值时，效果最优。

五项指标中 C1“答复时间间隔”、C2“答复与留言相关度”为“成本型”指标，其标准化公式为：

$$Z_{ij}=\frac{y_i^{\max}-y_{ij}}{y_i^{\max}-y_i^{\min}} \tag{15}$$

其中 y_{ij} 表示第 i 项指标的第 j 个取值，$y_i^{\max}$ 表示指标取值的最大值，$y_i^{\min}$ 表示指标取值的最小值。如答复时间间隔最长的为 171 天，最短的为 0 天，故以第一条留言为例，其答复时间间隔为 15 天，则归一化公式为：

$$Z_{11}=\frac{y_i^{\max}-y_{ij}}{y_i^{\max}-y_i^{\min}}=\frac{171-15}{171-0}=0.9117647$$

同理，得出其它答复时间间隔的归一化数据。

指标 C3“规范性用词”和 C4 答复文本长度为“效益型”指标，其标准化公式为：

$$Z_{ij}=\frac{y_{ij}-y_i^{\min}}{y_i^{\max}-y_i^{\min}} \tag{16}$$

指标含义同成本型，这里不以赘述。

指标 C5“标点符号使用的合理性”为文本中标点符号的占比情况。可知当标点使用较少时，会造成句子冗余，词意难以理解；当标点符号使用较多时，可能存在仅仅是为了增加回复句长，敷衍回复的现象。故这一指标定义为“成本型”和“效益型”都显得不合理，选用 “中间型”解决这一问题。

“中间型”归一化是指求解出标点符号占比中间值，计算每条留言标点占比情况与中间值的距离，距离越小说明标点符号使用情况越合理，计算过程如下：

首先计算出标点符号占比情况的中位数为 0.093846388，则每条留言与这一中间值的距离为

$$y_{ij}'=|y_{ij}-0.093846388| \tag{17}$$

因为存在比中间值取值小和取值大两种的情况，故采用差的绝对值表示距离。后转化为“成本型”指标，即 y_{ij}' 取值越小，表示其占比约靠近中间位置，使用情况也就越合理，则归一化指标为：

$$Z_{ij}=\frac{(y_i')^{\max}-(y_{ij}')}{(y_{ij}')^{\max}-(y_i')^{\min}} \tag{18}$$

其中 y_{ij}' 表示第 i 项指标的第 j 个取值，$(y_i')^{\max}$ 表示指标取值的最大值，$(y_i')^{\min}$ 表示指标取值的最小值。在标点符号占比情况中，距离中间取值最远的情况为 0.25076936，

最近的情况为 0.00001，则以第一条样本为例，其标点符号占文本总长度的 0.0814977，则归一化处理后为：

$$Z_{11}=\frac{(y_i')^{\max}-(y_{ij}')}{(y_{ij}')^{\max}-(y_i')^{\min}}=\frac{0.25076936-0.0814977}{0.25076936-0.000001}=0.950831$$

6.3.4 计算各条留言的总得分

综合得分计算公式为：

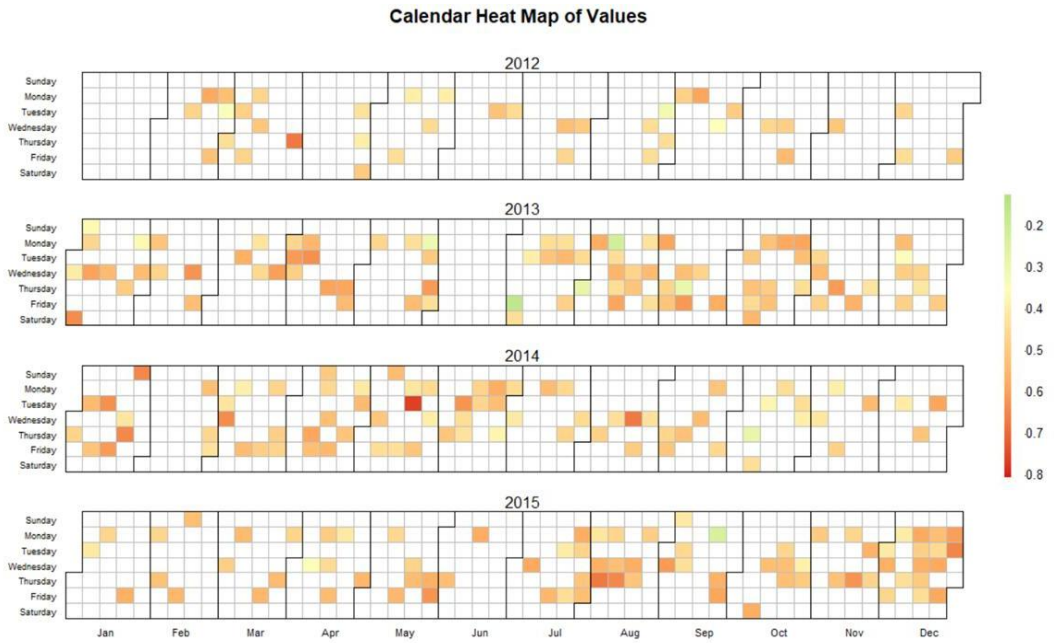
$$D=\sum_{i=1}^{5}Z_{ij}\times Q_i \tag{19}$$

其中 Z_{ij} 表示第 i 项指标的第 j 个取值，Q_i 表示第 i 个指标的权重值。则第一条留言的综合得分为

$$D=\sum_{i=1}^{5}Z_{ij}\times Q_i=\sum_{i=1}^{5}Z_{i1}\times Q_i=0.186383\times 0.98708+0.161758\times 0.087489$$

$$+0.214514\times 0.950831+0.232324\times 0.245353+0.204822\times 1=0.663966$$

由归一化后的数据取值不难看出，总得分情况越趋近于 1，说明留言回复的情况越好。为了解政府部门总体的留言回复情况，绘制留言综合得分随时间变化的日历图，如下所示：



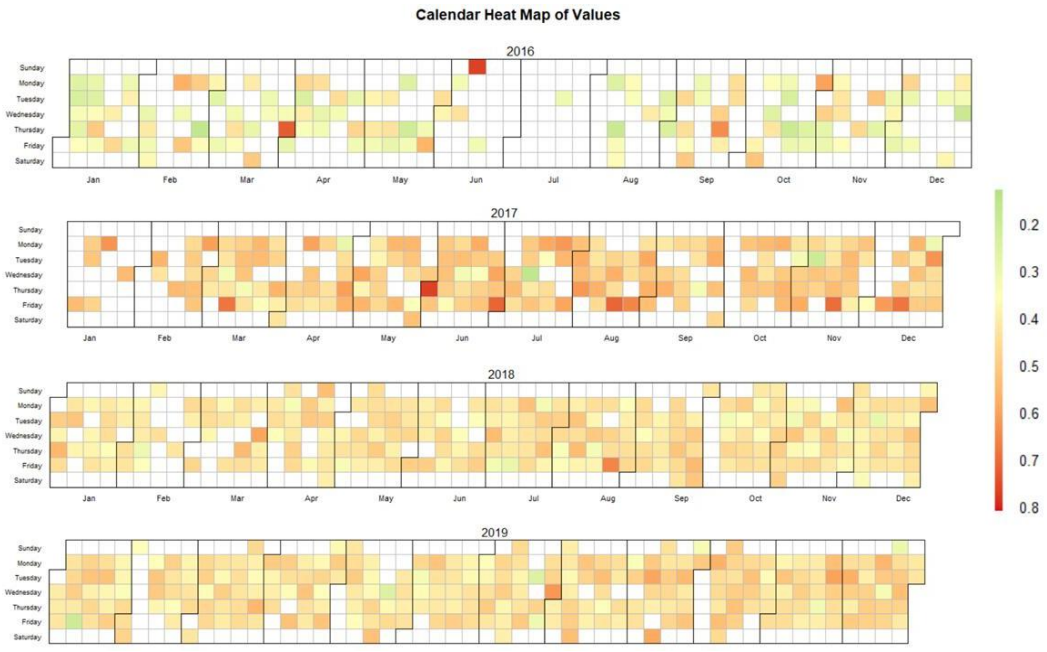


图 19 综合得分日历图

从图 19 可以看出，随着颜色加深，留言回复得分情况逐渐增加，且主要回复时间集中在周一到周五的工作日内。整体趋势中综合得分情况均在 0.5—0.7，在平均水平之上但仍存在改进的空间。2017 年整体回复情况较好，2018 年、2019 年次之。如编号为“11924”到“12458”，对应留言时间为 2014 年 3 月—2014 年 10 月的留言，回复信息仅为“网友：您好！留言已收悉”，没有给出具体的解决办法，对应得分均在 0.4—0.5，评分较低。故希望政府相关部门能在保证回复及时的同时，规范给出解决问题的具体措施，以更好地服务群众。

为了解相关部门回复的优缺点及具体的改进方案，本文分析了每年的综合得分情况并且进行了各指标的对比，具体内容如下图所示：

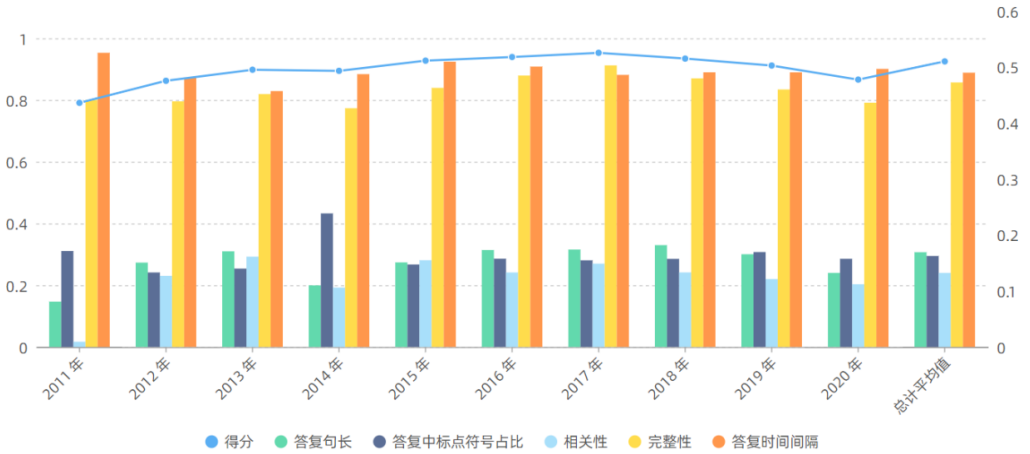


图 20 各指标对比图

上图中直方图以左边纵轴表示，折线图以右边纵轴表示。通过对比可知，2011 年综合得分较低的原因为回复的相关性和完整性较差，即答复句长较短，可能有些答复存在敷衍的情况。因 2020 年数据较少，故忽略。可明显看出从 2011 年到 2019 年，回复的综合得分由 0.436 生至 0.502，具有明显的提升，但在 2019 年，回复的完整性呈下降趋势，即回复中忽略了礼貌用词和日期等规范性用词，应加以注意。

7. 灵敏度检验

在问题二中筛选热点问题时，为统一数量级，定义了热度值，即将 1 次留言定义为 50 个热度值、1 个关注度定义为 1 个热度值。故等同于将留言数线性变换扩大了 50 倍，从而筛选出了热点问题。如若线性扩大 30、40 或 60、70 倍，热点问题会因此改变吗? 下文在 6.2.4 的相关理论的基础上，通过绘制线性扩大的不同倍数，探究热点问题排名变化情况：

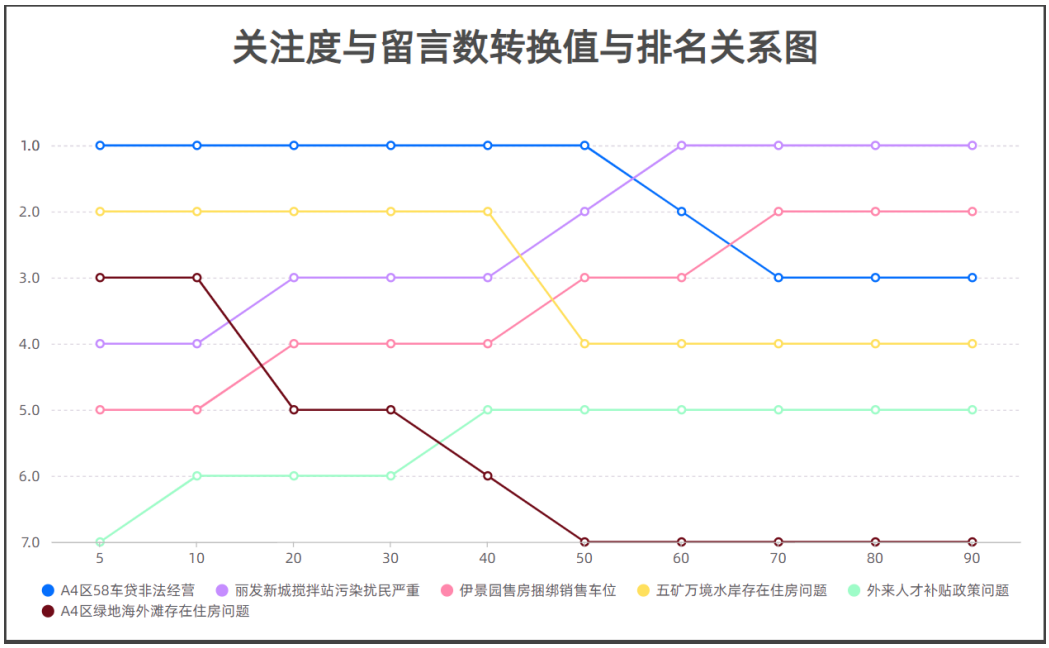


图 21 热点问题排名变化图

由上分析可知，当留言总数线性扩大 5 到 30 倍时，“A4 区绿地海外滩存在住房问题”的热度值高于“外来人才补贴政策问题”，居热点问题排名前 5；当留言总数线性扩大 40 倍以上时，“外来人才补贴政策问题”的热度值高于“A4 区绿地海外滩存在住房问题”，居热点问题排名前 5，且在 70 倍以上时，前 5 的热点问题排序不变，进一步证明了热度指标的合理性和模型的稳定性。

8. 模型优缺点

8.1 模型优点

1、对留言信息进行了 “去空”等预处理，筛选出了能够反映真实信息的特征值，排除了因字符串引起的操作失误，简化了计算。

2、相比较传统的朴素贝叶斯、多分类 logistics 回归分类，支持向量机算法不涉及较大概率测度和大数定律等故计算量小，实现速度快，最终决策的函数只涉及少量的支持向量，与样本自身的空间维数无关，解决了随机森林低维变量局限性。因此在处理留言问题时不仅运算速度快，而且精度较高。

3、在确定热点问题时，首先对地名进行筛选，避免同一地区的不同问题被聚为一类；其次针对某一特定地区分类，选出热点问题。这样可以减小计算量，提高分类的精度。

4、对层次分析法中指标权重的确定进行了统计学量化，结合主成分分析中因子载荷矩阵占比，进行了权重的参数估计，使得层次分析更具客观性。

8.2 模型缺点

1、支持向量机算法是一种现代化算法，在多分类算法理论研究中还不是很成熟，本文运用有向无环图对二分类算法进行了拓展，但仍存在准确度不足的情况，少数留言分类存在误判的情况，如关于学校、教师的社会保障类留言会误判成教育类，商务旅游留言因涉及地名信息会误判成城乡建设类，可后期加以人工检查。

2、在评价留言回复时，选用指标仍有挖掘的空间，如在附件 4 里留言信息结尾处有留言信息书写确定时间，这与留言公布时间存在着一个时间差，本文将这一时间差视为政府内部机制运行的合理时间，在之后的研究中可就这一问题展开深入讨论。

9. 参考文献

[1] 邓朝省，陈 莹．基于局部 SIFF 特征点的双阈值配准算法[J]．计算机工程与应用，2014，50(2)：189-193．

[2] 陈海涛, 徐嘉豪. 基于熵权模糊综合评价模型的河南省水资源承载力评价[J]. 人民珠江，2020，(1):48-53，116.

[3] 董金茂, 崔一民. 基于层次分析法的森林健康状况评价研究[J]. 林业调查规划，2020， (1):15-18.

[4] 王勇, 黄思奇, 刘永,等. 基于 K-means 聚类方法的物流多配送中心选址优化研究[J]. 公路交通科技，2020， (1):141-148.

[5] 严明，郑昌兴．Python 环境下的文本分词与词云制作[J]．现代计算机，2018,(34):86-89.