基于“智慧政务”的群众留言和政府回应的分析和挖掘

摘 要

近年来，随着微信、微博、市长信箱、阳光热线等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。

对于问题 1，首先统计各个类别的留言短信，通过图形化方式可知数据间存在不均衡现象，利用百度翻译对已有的数据回译进行数据增强。然后对数据做去重去空、降噪处理、去停用词，再利用 $jieba$ 中文分词对短信进行分词。利用 $TF-IDF$ 算法得到每条短信描述的 $TF-IDF$ 权重向量，再通过卡方检验，找出每个分类中关联度最强的两个词和两个词语对。然后通过比较算法准确度，选择线性支持向量机算法作为分类模型算法。最后构建 $LinearSVC$ 模型完成对群众留言的分类。

对于问题 2，首先基于附件 3 的留言主题，使用基于$k-means$算法改进的层次聚类算法，建立层次聚类模型，对数据完成聚类分类。对上一步的数据进行处理，以达到预期效果。然后是定义热点评价指标，使用$StackOverflow$问答排名算法，对于上面生成的数据进行统计，主要统计每一类问题的留言总数、点赞总数、反对总数、问题持续时间、问题最后一次留言的时间、留言得分，然后利用$StackOverflow$问答排名算法来计算出留言热度指数。最后，使用LDA模型对热度排名前五的热点问题通过建立 LDA 主题模型提取各类的特征词总结出热点问题的问题描述，并用$jieba$库来对数据进行词性标注。对于地名的识别，利用正则表达式来实现。

对于问题 3，根据赛题数据对答复意见的回应力进行评价。从相关性，完整性，可解释性，及时性等方面采用层次分析法 AHP 进行综合性评价，建立完善的综合评价指标体系，给出一套评价方案。

关键词：$LinearSVC$，层次聚类算法，$TF-IDF$ 算法，LDA 主题模型，$StackOverflow$ 问答排名算法 ，层次分析法 AHP

Abstract

In recent years, with the online questioning platforms such as WeChat, Weibo, M-ayor 's Mailbox, and Sunshine Hotline, etc., it has gradually become an important c-hannel for the government to understand public opinion, gather people 's wisdom, a-nd gather public opinion. The work of the relevant departments, which mainly relied on manual work to divide the message and organize hotspots, brought great challenges. At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of a smart govern-ment system based on natural language processing technology has become a new tren-d in the development of social governance innovation, which has a great impact on im-proving the management level and efficiency of government Promote the role.

For question 1, first count the message messages of various categories, through t-he graphical way we can see that there is an imbalance between the data, use Baidu tr-anslation to enhance the existing data back-translation. Then, the data is deduplicated, de-noised, noise-reduced, and stop words removed, and then the Chinese word segm-entation is used to segment the SMS. Use the TF-IDF algorithm to obtain the TF-ID-F weight vector described by each message, and then pass the chi-square test to find t-he two most relevant words and two word pairs in each category. Then by comparin-g the accuracy of the algorithm, the linear support vector machine algorithm is selecte-d as the classification model algorithm. Finally, the LinearSVC model is constructe-d to complete the classification of the mass messages.

For question 2, first, based on the message subject of Annex 3, a hierarchi-cal clustering algorithm based on the improved k-means algorithm is used to e stablish a hierarchical clustering model to complete the data classification. Pro-cess the data from the previous step to achieve the desired effect. Then, defin-e the hotspot evaluation indicators, use the StackOverflow question and answer ranking algorithm, and count the data generated above, mainly counting the to-tal number of comments, likes, total oppositions, question duration, the last ti-me the question was left, and the message Score, and then use the StackOverfl-ow question and answer ranking algorithm to calculate the message popularity

index. Finally, the LDA model is used to summarize the top five hot issues b y creating LDA topic models to extract various types of feature words to sum m-arize hot topic problem descriptions, and the jieba library is used to tag part s of speech. For the recognition of place names, the use of regular expressions t-o achieve.

For question 3, evaluate the responsiveness of the answers based on the question d-ata. From the aspects of relevance, completeness, interpretability, and timeliness, AH -P is used to conduct comprehensive evaluation, establish a comprehensive comprehe-nsive evaluation index system, and give a set of evaluation plans.

Keywords: LinearSVC ,hierarchical clustering algorithm, TF-IDF algorithm ,LDA to -pic model ,StackOverflow question and answer ranking algorithm, AHP

目 录

图 录

表 录

0.引言

0.1 问题重述

（一）针对附件 2 中的网络问政平台的群众留言数据，建立关于留言内容的一级标签分类模型，并通过 $F-Score$ 对分类方法进行评价。

（二）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，给出前五的热点问题，并保存为文件“热点问题表.xls”。将相对应的留言信息保存为“热点问题留言明细表.xls”。

（三）针对附件 4 相关部门对留言的答复意见，从相关性、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。

0.2 本文主要工作和创新点

本文基于题目中所给的分析角度和学术领域内已有的研究工作基础上，主成分分析方法、机器学习算法与层次分析模型，建立了一套较为完整的解决方案。

（一）基于建立留言内容的一级标签分类模型的角度。本文首先对数据进行统计，利用百度翻译 API 对不均衡数据进行增强，其次对预处理后的数据使用 $TF-IDF$ 算法将数据转为向量，然后使用四种不同类型的机器学习算法，来比较它们多分类数据的结果，选择其中效果最优的线性支持向量机分类器，再利用混淆矩阵来显示预测结果与实际标签之间的差异性，最终进行有针对性调整，使得 $F1$ 值达到 0.98。

（二）基于留言归类与定义热度评价指标的角度，首先基于附件 3 的留言的留言主题进行聚类处理，使用基于 $k-means$ 算法改进的层次聚类算法，建立层次聚类模型，利用计算文本间的余弦距离，对数据完成聚类。在定义热度评价指标上，我们在 $StackOverflow$ 问答排名算法的基础上进行改进，产生了合理的热点评价指标体系。最后使用 LDA 模型来进行主题提取。经过以上各步可以得出 “热点问题表.xls”与“热点问题留言明细表.xls”。

（三）针对答复意见质量的评价方案角度。根据赛题数据对答复意见的回应力进行评价。我们通过分析数据决定从相关性，完整性，可解释性，及时性方面来制定方案，后采用层次分析法 AHP 进行综合性评价，以得到较为合理的评价方案。

1.挖掘目标

本次建模目标是根据发布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，利用 $jieba$ 中文分词工具对留言信息进行分词，支持线性向量机算法，层次聚类算法，LDA 主题模型，$StackOverflow$ 问答排名算法和层次分析法达到以下目标：

利用文本分词和文本分类的方法对非结构化的数据进行文本挖掘，根据分类结果，对留言进行分类，以便后续将群众留言分派至相应的职能部门。

根据网络留言数据，分析出某一时段内群众反映的排名前五的热点问题及相应的留言信息，有助于相关部门进行针对性地解决，提升服务效率。

根据相关部门对留言的答复意见，从答复的相关性，完整性，可解释性和及时性给相关部门的答复意见提供真实可靠的评价。

2.分析方法与过程

2.1 问题 1 分析方法与过程

总体流程图

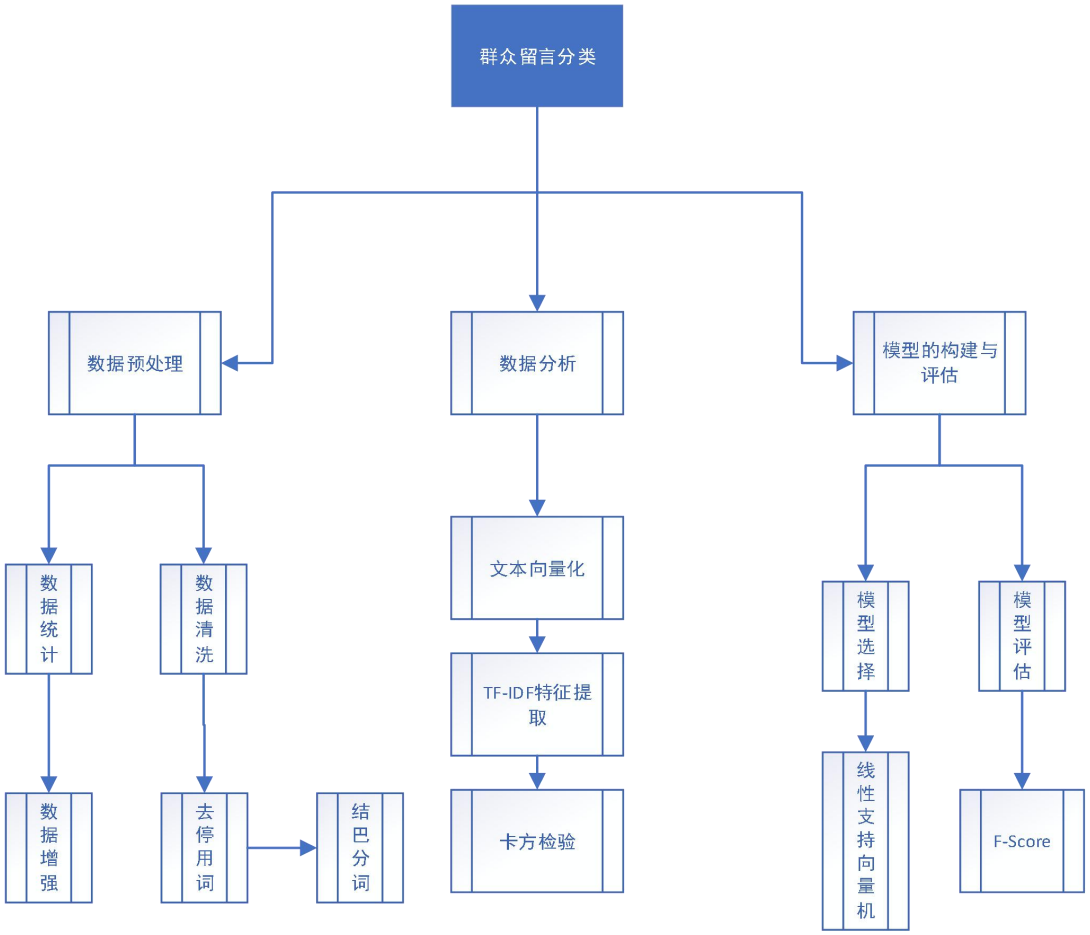

图 1　问题 1 流程图

本题主要包括如下步骤：

步骤一：数据预处理，在题目所给的数据中，数据杂乱重复，所以使用去重处理，清洗空格，统计数据。删除文本中的标点符号，特殊符号，还要删除一些无意义的停用词（$stopwords$），在此基础上进行中文分词。统计数据，因数据存在不均衡现象，我们利用百度翻译对“附件 1.xlsx”中的三级分类和“附件 2.xlsx”中的不均衡留言数据进行了数据增强。

步骤二：数据分析，在对留言短信分词后，需要把这些词语转换为向量，以供挖掘分析使用。然后采用 $TF-IDF$ 算法，找出每条短信描述的关键词，把留言短信转换为权重向量。采用卡方检验来检验数据的拟合度和关联度。

步骤三：模型选择及评价，这里我们比较四种机器学习模型的准确度，综合比较得出线性支持向量机（$Linear\ Support\ Vector\ Machine$）准确率最高。$Classification\_report$ 函数用于显示分类指标的文本报告。

2.1.1 数据预处理

2.1.1.1 统计各个类别的数据量

统计各个类别的留言短信，发现各个类别的数据量并不一致，且分布不均匀，接下来我们用图形化的方式查看各个类别的分布。因数据存在不均衡现象，所以我们利用百度翻译对部分数据进行数据增强处理。

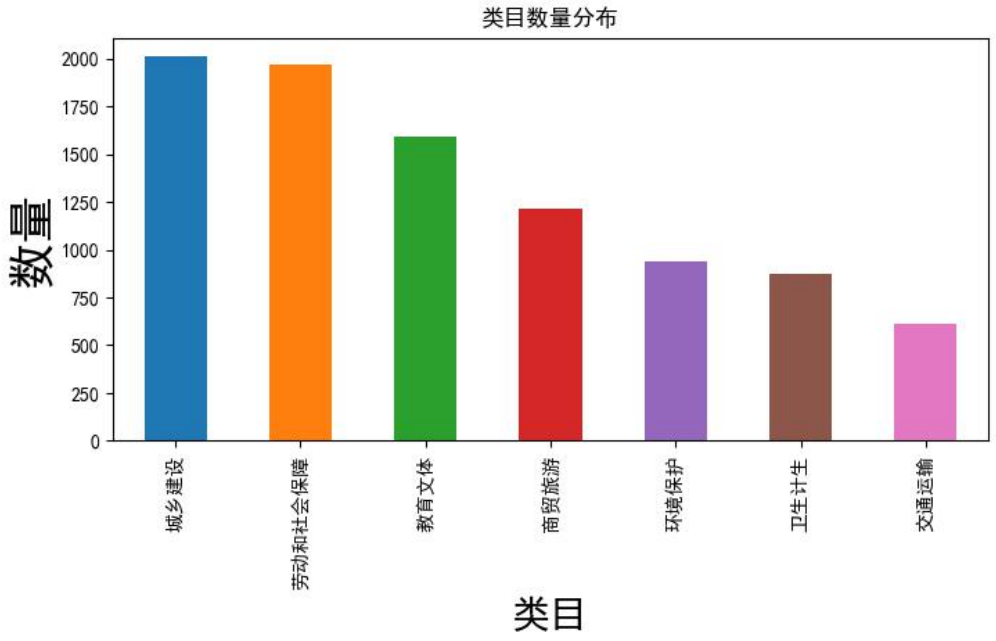

图 2　数据分布

2.1.1.2 数据增强

由上一步数据量统计可知，附件 2 中的留言短信表中数据分布不均匀，所以采用回译的方法来将数据增强，首先利用网络爬虫技术来连接百度翻译的接口。其次编写回译部分，将目标语言先翻译成五种其他语言，后将他们翻译回中文，并将数据存储在字典结构中。最后，将已回译的数据整理装入 excel 表格中。进行增强的数据是附件 1 中的三级分类和附件二中的不均衡数据，来作为回译的对象。将附件 1 中的三级分类进行增强的原因：对于分类器来说三级分类就相当于附件 2 中留言详情与对应标签的特征。所以大量的物理增强了三级分类，使得分类器更容易捕获标签对应的特征。

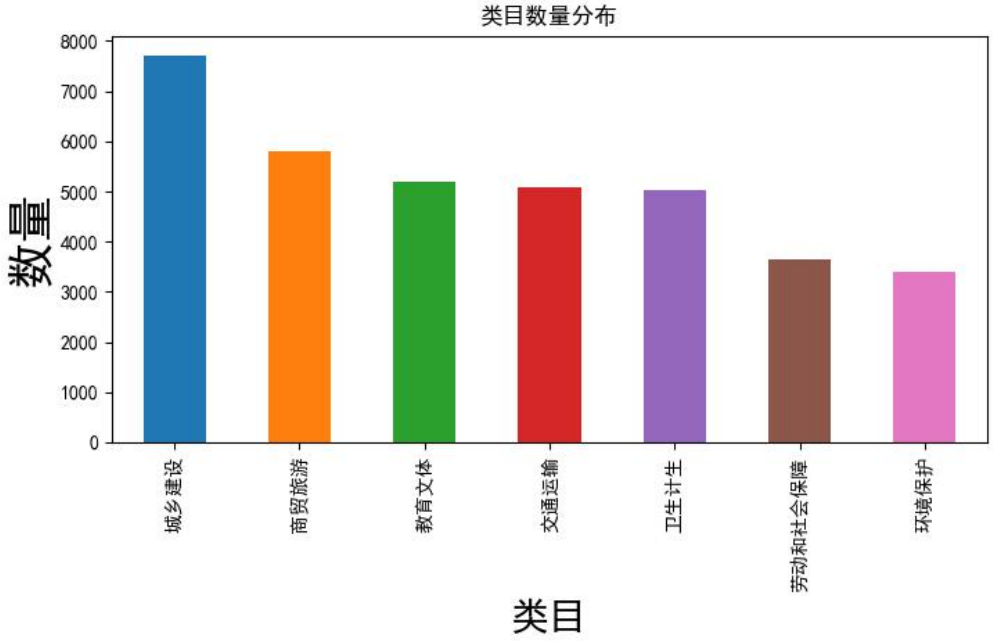

图 3　数据增强后数据分布

2.1.1.3 留言短信的清洗与中文分词

1.数据清洗

在题目给出的数据中，存在很多重复的留言数据。例如城乡建设中出现了很多相同的留言短信。清洗空格，统计数据。删除文本中的标点符号，特殊符号，还要删除一些无意义的停用词（$stopwords$），因为这些词和符号对系统的分析预测文本的内容没有任何帮助，反而会增加计算的复杂度和增加系统开销，所以在使用这些文本数据之前必须将它们清理干净。通过定义 $remove\_punctuation$ 函数对文本进行处理。

2.中文分词

在对留言短信进行挖掘分析之前，先要把非结构化的文本信息转换为计算机能够识别的结构化信息。在附件 2 留言短信表中，以中文文本的方式给出了数据。为了便于转换，先要对这些职位描述信息进行中文分词。

这里采用 $python$ 的中文分词包 $jieba$ 进行分词。$jieba$ 采用了基于前缀词典实现的高效词图，进行扫描，生成句子中汉字所有可能成词情况所构成的有向无环图(DAG),同时采用了动态规划查找最大概率路径，找出基于词频的最大切合分

组，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使得能更好的实现中文分词效果。

在分词的同时,采用了 $TF-IDF$ 算法,抽取每个职位描述中的前 5 个关键词,这里采用 $jieba$ 自带的语义库。对数据增强之后的数据进行去除停用词和结巴分词等操作，后将中文分词和回译的数据保存在作品附件 1 和作品附件 2 中。

2.1.2 数据分析

2.1.2.1 TF-IDF 算法

在对处理之后的留言数据进行分词后，需要把这些词语转换为向量，以供挖掘分析使用。

这里采用 $TF-IDF$ 算法，把留言短信转换为权重向量。$TF-IDF$ 算法的具体原理如下：

第一步，计算词频，即 TF 权重（$Term\ Frequency$）。

$$词频(TF) = 某个词在文本中出现的次数 \tag{1}$$

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：

$$词频(\boldsymbol{TP}) = \frac{某个词在文本中出现的次数}{文本的总次数} \tag{2}$$

或

$$词频(\boldsymbol{TF}) = \frac{某个词在文本中出现的次数}{该文本出现次数最多多次出现的次数} \tag{3}$$

第二步，计算 IDF 权重，即逆文档频率（$Inverse\ Document\ Frequency$），需要建立一个语料库（$corpus$），用来模拟语言的使用环境。$IDF$ 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$逆文档频率(\boldsymbol{IDF}) = \boldsymbol{log}(\frac{语料库对文本总数}{包含该词的文本数+\boldsymbol{1}}) \tag{4}$$

第三步，计算 $TF-IDF$ 值（$Term\ Frequency\ Document\ Frequency$）。

$$TF-IDF = 词频(TF) \times 逆文档频率(IDF) \tag{5}$$

实际分析得出 $TF-IDF$ 值与一个词在留言短信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留言短信表中文本的关键词。

2.1.2.2 生成 TF-IDF 向量

生成 $TF-IDF$ 向量的具体步骤如下：

使用 $TF-IDF$ 算法，找出每条留言描述的前 5 个关键词；

对每条留言描述提取的 5 个关键词，合并成一个集合，计算每条留言的留言详情对于这个集合中词的词频，如果没有则记为 0；

生成各个留言详情的 $TF-IDF$ 权重向量，计算公式如下：

$$TF-IDF = 词频(TF) \times 逆文档频率(IDF) \quad (6)$$

2.1.2.3 卡方检验

卡方检验就是统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，如果卡方值越大，二者偏差程度越大；反之，二者偏差程度越小；若两个值完全相等时，卡方值就为 0，表明理论值完全符合。即用来检验数据的拟合度和关联度。在这里我们使用 $sklearn$ 中的 $chi2$ 方法，来检测标签与数据的相关性。经过卡方（$chi2$）检验后，找出了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地反映出分类的主题。

```
# '交通运输':
  . Most correlated unigrams:
       . 快递
       . 出租车
  . Most correlated bigrams:
       . 赵子 议论
       . 出租车 司机
# '劳动和社会保障':
  . Most correlated unigrams:
       . 职工
       . 社保
  . Most correlated bigrams:
       . 劳动 关系
       . 退休 人员
# '卫生计生':
  . Most correlated unigrams:
       . 医生
       . 医院
  . Most correlated bigrams:
       . 社会 抚养费
       . 乡村 医生
```

图 4　卡方检验示例

2.1.3 模型构建及评估

2.1.3.1 模型的选择和比较

在对于模型的选择过程中，通过比较以下四种不同的机器学习模型,并评估它们的准确率。

1.*Logistic Regression*(逻辑回归)

基本原理：

1)找一个合适的预测函数（*Andrew Ng* 的公开课中称为 *hypothesis*），一般表示为 h 函数，该函数就是我们需要找的分类函数，它用来预测输入数据的判断结果。这个过程是非常关键的，需要对数据有一定的了解或分析，知道或者猜测预测函数的“大概”形式，比如是线性函数还是非线性函数。

2)构造一个 $cost$ 函数（损失函数），该函数表示预测的输出(h)与训练数据类别(y)之间的偏差，可以是二者之间的差$(h-y)$或者是其他的形式。综合考虑所有训练数据的“损失”，将 $cost$ 求和或者求平均，记为$J(\theta)$函数，表示所有训练数据预测值与实际类别的偏差。

3)显然，$J(\theta)$函数的值越小表示预测函数越准确（即 h 函数越准确），所以这一步需要做的是找到$J(\theta)$函数的最小值。找函数的最小值有不同的方法，$Logistic\ Regression$ 实现时用的是梯度下降法$(Gradient\ Descent)$。

$(Multinomial)\ Naive\ Bayes$(多项式朴素贝叶斯)

基本原理：

在多项式模型中， 设某文档 $d=(t_1,t_2,\cdots,t_k)$，t_k是该文档中出现过的单词，允许重复，

则先验概率：

$$P(c)=\frac{类c\ 下单词总数}{整个训练样本的单词总数} \tag{7}$$

类条件概率：

$$P(t_k|c)=\frac{类c\ 下单词t_k\ 在各个文档中出现过的次数之和+1}{类c\ 下单词总数+|V|} \tag{8}$$

V 是训练样本的单词表（即抽取单词，单词出现多次，只算一个），

$|V|$则表示训练样本包含多少种单词。在这里，$m=|V|,P=\frac{1}{|V|}$。

$P(t_k|c)$可以看作是单词t_k在证明 d 属于类 c 上提供了多大的证据，而 $P(c)$则可以认为是类别 c 在整体上占多大比例(有多大可能性)。

2.$Linear\ Support\ Vector\ Machine$(线性支持向量机)

基本原理：

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。

从结果的角度，$LinearSVC$ 和使用 SVC 且 $kernel$ 传入 $linear$，结果是一致的。但是由于 $LinearSVC$ 只能计算线性核，而 SVC 可以计算任意核，所以，他

们的底层计算方式不一样，这使得同样使用线性核的 SVC，用 $LinearSVC$ 的计算速度，要比用 SVC 且 $kernel$ 传入 $linear$ 参数，快很多。所以，整体而言，如果决定使用线性 SVM，就使用 $LinearSVC$。

因为处理的数据较多而且是多分类问题，为了节省时间提高效率，选择 $LinearSVC$ 代替 SVC 或 $NuSVC$。SVC 用于二分类问题很方便，但是用于多分类问题的开销较大，而 $LinearSVC$ 既可以用于二分类问题，也可以用于多分类问题，而且效率较高。

3.$Random\ Forest$(随机森林)

基本原理：

随机森林是一种特殊的 $bagging$ 方法，它将决策树用作 $bagging$ 中的模型。首先，用 $bootstrap$ 方法生成 m 个训练集，然后，对于每个训练集，构造一颗决策树，在节点找特征进行分裂的时候，并不是对所有特征找到能使得指标（如信息增益）最大的，而是在特征中随机抽取一部分特征，在抽到的特征中间找到最优解，应用于节点，进行分裂。随机森林的方法由于有了 $bagging$，也就是集成的思想在，实际上相当于对于样本和特征都进行了采样（如果把训练数据看成矩阵，就像实际中常见的那样，那么就是一个行和列都进行采样的过程），所以可以避免过拟合。

4.综合分析选择分类器

利用赛题数据测试 4 种模型，通过 $seaborn$ 函数将其测试结果呈现在箱式图上，在如图箱体图中，可以看出随机森林分类器的准确率是最低的，因为随机森林属于集成分类器(有若干个子分类器组合而成)，一般来说集成分类器不适合处理高维数据(如文本数据),因为文本数据有太多的特征值,使得集成分类器难以应付,另外三个分类器的平均准确率都在 90%以上。综合比较后线性支持向量机的准确率最高。因此选择线性支持向量机模型。

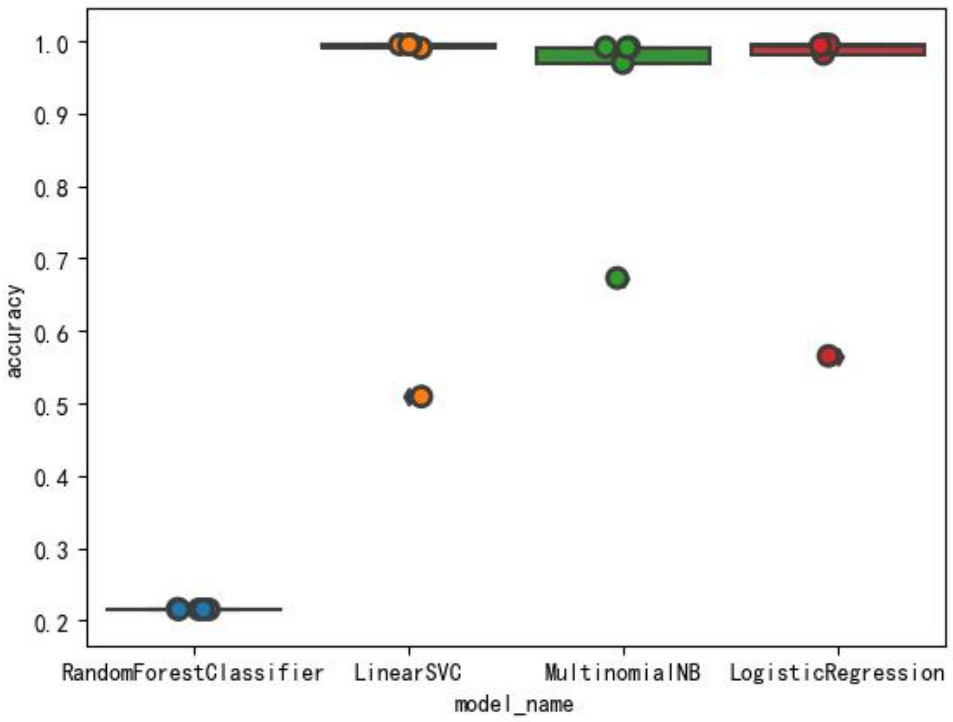

图 5　箱式图

2.1.3.2 线性支持向量机

支持向量机分类器($Support\ Vector\ Classifier$)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。

从结果的角度，$LinearSVC$ 和使用 SVC 且 $kernel$ 传入 $linear$，结果是一致的。但是由于 $LinearSVC$ 只能计算线性核，而 SVC 可以计算任意核，所以，他们的底层计算方式不一样，这使得同样使用线性核的 SVC，用 $LinearSVC$ 的计算速度，要比用 SVC 且 $kernel$ 传入 $linear$ 参数，快很多。所以，整体而言，如果决定使用线性 SVM，就使用 $LinearSVC$。

因为处理的数据较多而且是多分类问题，为了节省时间提高效率，选择 $LinearSVC$ 代替 SVC 或 $NuSVC$。SVC 用于二分类问题很方便，但是用于多分类

问题的开销较大，而 $LinearSVC$ 既可以用于二分类问题，也可以用于多分类问题，而且效率较高。

2.1.3.3 多分类模型的评估

多分类模型一般不使用准确率($accuracy$)来评估模型的质量,因为 $accuracy$ 不能反映出每一个分类的准确性,因为当训练数据不平衡(有的类数据很多,有的类数据很少)时，$accuracy$ 不能反映出模型的实际预测精度,这时候我们就需要借助于 $F1$ 分数、ROC 等指标来评估模型。在 $sklearn.metrics$ 中我们调用 $classification\_report$ 函数，其主要用于显示主要分类指标的文本报告．在报告中显示每个类的精确度，召回率，$F1$ 值等信息。

针对准确率最高的 $LinearSVC$ 模型，我们将查看混淆矩阵，并显示预测标签和实际标签之间的差异。

2.2 问题 2 分析方法与过程

2.2.1 文本聚类分类

聚类分析是一种无监督机器学习（训练样本的标记信息是未知的）算法，它的目标是将相似的对象归到同一个簇中，将各类相似的对象归到各个簇中。如果要使用聚类分析算法对一堆文本分类，关键要解决这几个问题：

（1）如何衡量两个对象是否相似

（2）算法的性能怎么度量

（3）如何确定分类的个数或聚类结束的条件

（4）选择算法

2.2.1.1 聚类需求

留言数据存在数据重复，信息不规范，错别字等现象，所以需要先对数据进行聚类分类，将相同或相似的数据划到一起，再通过重复多次筛查数据，最终形成规范的留言数据。

2.2.1.2 理论基础

1.相似度计算

相对于欧式距离，余弦相似度更适合计算文本的相似度，所以此处使用余弦相似度处理相似度的计算。

2.余弦相似度简介：

余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体差异的大小。相比欧氏距离度量，余弦相似度更加注重两个向量在方向上的差异，而非距离或长度上的差异。余弦值的计算公式如下：

$$cos\,\theta = \frac{\sum_1^n (A_l \times B_l)}{\sqrt{\Sigma_1^n A_i^2} \times \sqrt{\Sigma_1^n B_i^2}} \tag{9}$$

首先将文本转换为权值向量，通过计算两个向量的夹角余弦值，就可以评估他们的相似度。余弦值的范围在[-1,1]之间，值越趋近于 1，代表两个向量方向越接近；越趋近于-1，代表他们的方向越相反。为了方便聚类分析，我们将余弦值做归一化处理，将其转换到[0,1]之间，并且值越小距离越近。

3.性能度量

在选择聚类算法之前，首先来了解什么样的聚类结果是比较好的。希望同一个簇内的样本尽可能相似，不同簇的样本尽可能不同，也就是说聚类结果的“簇内相似度”高且“簇间相似度”低。

考虑聚类结果的簇划分 $C = \{C_1, C_2, \cdots, C_k\}$，定义：

$$avg(C) = \frac{2}{(|C|(|C|-1))} \sum_{1 \le i \le j \le |c|} dist\,(x_i, x_j) \tag{10}$$

$$diamC = \max_{1 \le i \le j \le |C|} dist(x_i, x_j) \tag{11}$$

$$d_{min}(C_i, C_j) = \min_{x_i \in C_i, x_j \in C_j} dist(x_i, x_j) \tag{12}$$

$$d_{cen}(C_i, C_j) = dist(\mu_i, \mu_j) \tag{13}$$

其中，μ代表簇 C 的中心点；$avg(C)$代表簇内样本的平均距离；$diamC$ 代表簇内样本间的最远距离；$d_{min}(C_i, C_j)$对应于簇C_i和簇C_j最近样本间的距离；$d_{cen}(C_i, C_j)$对应于簇C_i和C_j中心点间的距离。基于以上公式可导出下面两个常用的聚类性能度量内部指标：

DB 指数（$Davies - Bouldin\,Index$，简称 DBI）：

$$DBI = \frac{1}{k}\sum_{i=1}^{k} max_{j\neq i}\left(\frac{avg(C_i)+avg(C_j)}{d_{cen}(C_i,C_j)}\right) \quad (14)$$

$Dumn$ 指数（$Dumn\ Index$，简称 DI）：

$$DI = \min_{1\leq i\leq k}\left\{\min_{j\neq i}\left(\frac{d_{min}(C_i,C_j)}{max_{1\leq l\leq k}diam(C_l)}\right)\right\} \quad (15)$$

DB 指数的计算方法是任意两个簇内，样本的平均距离之和除以两个簇的中心点距离，并取最大值，DBI 的值越小，意味着簇内距离越小，同时簇间的距离越大；$Dumn$ 指数的计算方法是任意两个簇，最近样本间的距离除以簇内样本的最远距离的最大值，并取最小值，DI 的值越大，意味着簇间距离大而簇内距离小。因此，DBI 的值越小，同时 DI 的值越大，意味着聚类的效果越好。

2.2.1.3 聚类过程

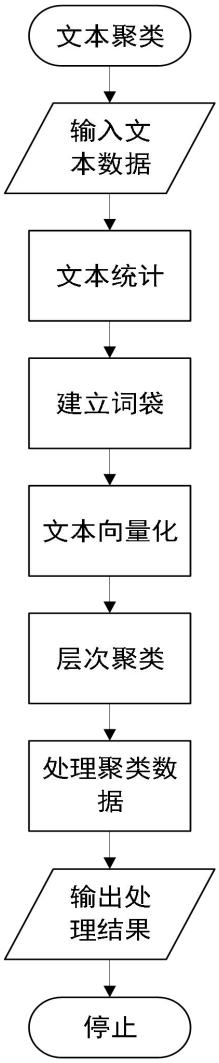

图 6　文本聚类流程图

具备相似度计算方法和性能度量这两个理论基础，下面就开始对文本的主题进行分类。

1.分词

要对中文文本做聚类分类，首先要对文本做分词处理，例如“小区临街门面油烟直排扰民”，我们希望将其切分为“小区 临街 门面 油烟 直排 扰民”。$Python$ 提供中文切词工具 $jieba$，它可以将中文长文本划分为若干个单词。

2.构建词袋模型

文本被切分成单词后，需要进一步转换成向量。先将所有文本中的词汇构建成一个词条列表，其中不含重复的词条。然后对每个文本，构建一个向量，向量的维度与词条列表的维度相同，向量的值是词条列表中每个词条在该文本中出现的次数，这种模型叫作词袋模型。例如，“经济学院强制学生实习”和“经济学院组织学生打工”两个文本切词后，结果是“经济 学院 强制 学生 实习”和“经济 学院 组织 学生 打工”。它们构成的词条列表是[经济， 学院， 强制， 学生， 实习， 组织， 打工]，对应的词袋模型分别是[1，1，1，1，1，0，0]，[1，1，0，1，0，1，1]。

定义 $creatVocabList$ 函数实现创建不重复的词条列表；定义 $bagOfWordsVec$ 函数实现将文本转化为词袋模型。

3.权值转换

$TF-IDF$ 是一种统计方法，用来评估一个词条对于一个文件集中一份文件的重要程度。$TF-IDF$ 的主要思想是：如果某个词在一篇文章中出现的频率 TF 高，并且在其他文件中很少出现，则认为此词条具有很好的类别区分能力，适合用来分类。将词袋向量转换为 $TF-IDF$ 权值向量，更有利于判断两个文本的相似性。

TF(词频 $term\ frequency$)：

$$tf_{i,j}=\frac{n_{i,k}}{\Sigma_k n_{k,j}} \qquad (16)$$

分子是词条t_i在文件d_j中出现的次数，分母是文件d_j中所有词条出现的次数之和。

IDF(逆向文件频率 $inverse\ document\ frequency$)：

$$idf_i = log\frac{|D|}{|\{j:t_i \in d_j\}|} \quad (17)$$

对数内的分子是文件总数，分母是包含词条的文件数，如果该词不存在，就会导致分母为零，因此一般使用 $1 + |\{j:t_i \in d_j\}|$作为分母。

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (18)$$

定义 $wordsCount$ 函数计算所有文本包含的总词数；

用 $wordInFileCount$ 函数计算包含某词的文本数；

用 $calTFIDF$ 函数计算权值。

4.计算余弦相似度

前文已经介绍过。定义 gen_sim 函数实现。

5.聚类算法（层次聚类算法）

由于初始质心的随机性对 $k-means$ 的结果影响很大，数据很可能收敛到局部最小值，并且会产生空簇，所以为了得到更好的聚类效果，对传统 $k-means$ 算法做出了改进。

如果文本所含的词汇量较多，并且已知分类的个数 k，可以选择二分 $k-$*均值聚类算法*。而层次聚类算法根据样本距离分类，并且可以以样本距离作为分类结束的条件，比较适合 k 位置的情况。层次聚类算法根据样本数据来分类，并且可以以样本距离作为分类结束的条件，比较适合 k 未知的情况。为此我们利用层次聚类算法对数据进行分类。

层次聚类试图在不同的层次对数据集进行划分，可以采用“自底向上”的聚类策略，也可以采用“自顶向下”的分拆策略。一般采用“自底向上”的策略，它的思路是先将数据集中的每个样本看作一个初始聚类簇，然后找出两个聚类最近的两个簇进行合并，不断重复该步骤，直到达到预设的聚类个数或某种条件。关键是如何计算两个簇之间的距离，每个簇都是一个集合，因此需要计算集合的某种距离即可。例如，给定簇C_i和C_j，可通过以下 3 种方式计算距离：

最小距离：

$$d_{min}(C_i,C_j) = \min_{x_i \in C_i, x_j \in C_j} dist(x,z) \quad (19)$$

最大距离：

$$d_{max} = (C_i,C_j) = \max_{x_i \in C_i, x_j \in C_j} dist(x,z) \quad (20)$$

平均距离：

$$d_{avg}(C_i,C_j) = \frac{1}{|C_i||C_j|}\sum_{x\in C_i}\sum_{z\in C_j} dist(x,z) \quad (21)$$

最小距离由两个簇的最近样本决定，最大距离由两个簇的最远样本决定，平均距离由两个簇的所有样本决定。

接下来要考虑如何确定一个合适的聚类个数或某种结束条件，具体思路是：

（1）选定一部分测试样本，对其进行层次聚类分析。

（2）计算性能度量指标 DBI 和 DI 的变化趋势，结合人工校验，得到一个合适的聚类个数和对应的距离阈值。

（3）将此距离阈值作为聚类结束的条件，对所有样本做聚类分析。此时无需再计算 DBI 和 DI 值，计算效率可以大幅提升。

基于上述考虑，结合层次聚类法中的分裂思想对 $k-means$ 算法的部分流程做出改进。分裂的层次聚类算法思想是，首先将所有的数据当作一个整体，划分到一个簇中，然后根据情况使用不同的计算准则例如欧式距离、余弦公式等，将该簇渐渐细化分为越来越小的簇。直到簇的数量到达预先指定的数目或者簇之间的 距离大于某个阈值，聚类结束。层次聚类的流程示意图如图所示:

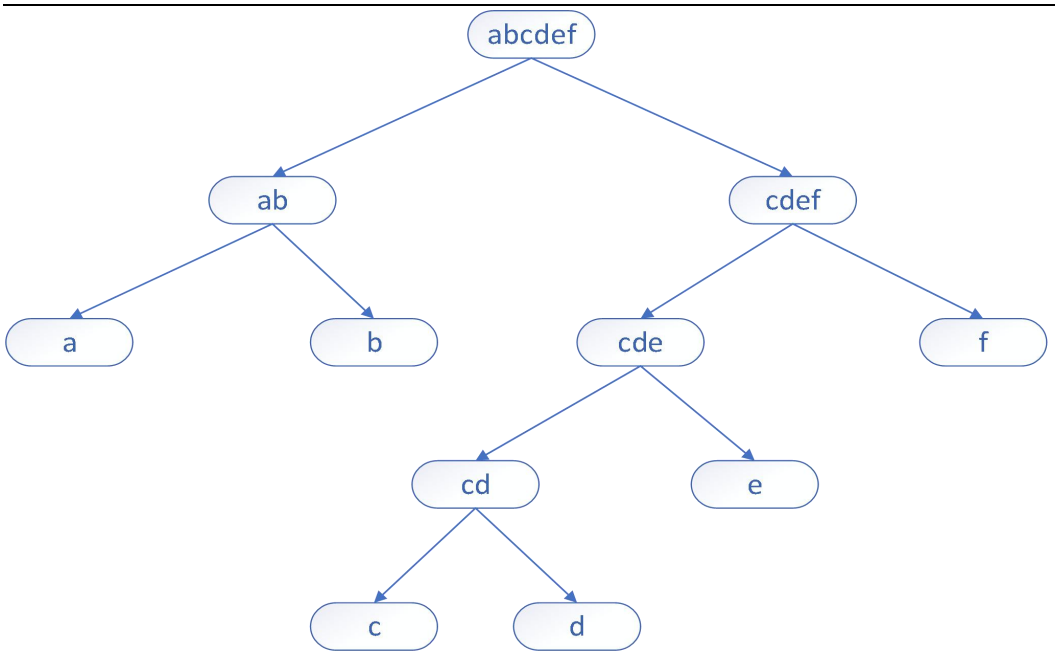

图 7　层次聚类流程图

在程序中：

定义 $distMin$ 函数实现计算两个簇的最小距离，定义 $distmax$ 函数实现计算两个簇的最大距离；定义 $distAvg$ 函数实现计算两个簇的平均距离；定义 $fidMin$ 函数实现找到距离最近的两个簇；定义 $DBIvalue$ 函数实现 DBI 指数的计算；定义 $DIvalue$ 函数实现 DI 指数的计算；定义 $HCluster$ 实现层次聚类算法。

聚类模型的构建：

通过建立层次聚类模型 $HCluster$，对于数据中的留言主题进行聚类处理。利用计算簇间的余弦距离，使簇间距离最小的两个簇归并，直至两个簇间距离小于设定簇间距离 $dist$, 通过设定适当参数，使聚类分类效果尽量完善，以此对数据完成进行聚类，后将聚类后的数据存入作品附件 3 文件中。

然后对于聚类后的数据，通过统计各类数量，对每一类的数据量排序，设定“问题 ID”为分类标签，取数量排名前 80 类数据存入作品附件 4。

2.2.2 热度评价指标

2.2.2.1 问题热度的影响因素

问题热度反映了问题本身的重要程度，也反映了公众的参与度，这是一个综合性的指标，因此影响问题热度的因素主要有以下几个方面：

1.群众关注度

发表某问题的数量是群众关注度的重要体现。相关问题的数量越多，表明问题越被群众重视，这也必然会引起相应部门与社会的广泛的关注，问题的热度也就会增高。

2.话题普遍性

话题的核心事件是否与公众中的每个人相关，这在很大程度上决定了话题的热度。若一个核心事件只是在某个领域有很大影响力，那么这个话题的传播范围仅限于该领域，然而如果话题涉及的范围没有局限性，那么热度可想而知。在信息传播不发达的时代，人们只能语言交流传播，但现在可以通过文章的浏览次数、评论数来评估话题的传播情况。这些指标反映了公众的知情度和参与度，越多人了解和参与其中，就表明话题越火热。

3.话题时间性

一般来说，话题都是对某个时间所发生的某件事情的即时报道和讨论。随着时间慢慢过去，人们对于某个话题的关注也会下降，从而被新的话题所取代。时间久远的话题理所当然要比新的话题受到的关注要少，所以最新产生的话题永远拥有更多的关注和更高的热度。

2.2.2.2 在文章聚类中话题热度排序的研究与实现

1.数据处理

在对数据进行热点问题挖掘之前，首先要对我们已分类并处理的 作品附件 4 再进行加工处理。

对前 80 类的数据以“问题 ID”、“热度指数”、“时间范围”、“地点”、“问题描述”、“反对总数”、“点赞总数”、“问题具体数目”和“持续时间”为目录创建 $DataFrame$。“时间范围”就是某个问题最晚出现的时间与最早出现时间的差值。如下表：

表 1　数据处理

	问题ID	热度指数	时间范围	地点/人群	问题描述	反对总数	点赞总数	问题具体数目	持续时间
0	1	NaN	2019-11-02至2020-01-26	NaN	NaN	2	50	53	85
1	2	NaN	2019-07-02至2019-09-01	NaN	NaN	1	25	48	61
2	3	NaN	2018-11-15至2019-12-02	NaN	NaN	4	45	27	382

再对以上数据进行去重复处理，得到作品附件5。

对作品附件4进行数据清洗，因为我们的数据里面有特殊的词汇如“A市”，“$K3$县”，“魅力之城小区”，“L市”，“梅溪湖”，“$A5$区”，“劳动东路”，“夜宵摊”等，于是我们通过添加新词的方式添加到$jieba$中进行中文分词等操作。通过构建LDA模型可得到该类的主题词，通过逐一计算每类留言的每条数据对于该类主题词的推断值作为该条留言数据的留言得分。如下图：

```
tf-idf版主题词
(0, '0.009*"实习" + 0.009*"过年" + 0.008*"校方" + 0.008*"做" + 0.008*"家长" + 0.008*"孩子" + 0.007*"安排" + 0.007*"临近" + 0.007*"500" +
0.007*"点"')
['A市涉外经济学院组织学生外出打工合理吗']
主题0推断值86.00
['A市经济学院强制学生实习']
主题0推断值37.00
['A市经济学院寒假过年期间组织学生去工厂工作']
主题0推断值66.00
['西地省涉外经济学院变相强制学生社会实践']
主
```

图 8　推断值

得出每类留言基于该类主题词的推断之后，我们在处理的结果获得对应的推断值。结合上述处理的 作品附件 5 ，我们求出每类的推断值之和、推断值均值、推断之方差。如下图：

表 2　数据量化

	问题ID	热度指数	时间范围	地点	问题描述	反对总数	点赞总数	问题具体数目	持续时间	距离现在的时间	推断值和	推断值均值	推断值方差
0	1	NaN	2019-11-02至2020-01-26	NaN	NaN	2	50	53	85	100	1489	28.09	127.2
1	2	NaN	2019-07-02至2019-09-01	NaN	NaN	1	25	48	61	247	2204.9	45.94	541.8
2	3	NaN	2018-11-15至2019-12-02	NaN	NaN	4	45	27	382	155	1408.9	52.18	1780
3	4	NaN	2019-01-08至2019-10-14	NaN	NaN	1	42	19	279	204	808.97	42.58	1923
4	5	NaN	2019-07-21至2019-12-04	NaN	NaN	9	11	17	136	153	857.96	50.47	1008

详情可见作品附件 6。

2.$StackOverflow$ 排名算法

采用 $StackOverflow$ 排名算法，计算各类留言数据的热度，并根据热度指标排序，选出热度排名前五的热点问题。

$StacOverflow$ 排名算法详情：

$$热度指标=\frac{(log_{10}Qviews)\times 4+\frac{Qanswers\times Qscore}{5}+sum(Ascorces)}{((Qage+1)-(\frac{Qage-Qupdated}{2}))^{1.5}} \quad (22)$$

（1）$Qviews$（问题的浏览次数）

某个问题的浏览次数越多，就代表越受关注，得分也就越高。这里使用了以 10 为底的对数，用意是当访问量越来越大，它对得分的影响将不断变小。有效浏览量可以建立一个停留时间的阀值去衡量，浏览得越多则越热门。

（2）$Qscore$（问题得分）和 $Qanswers$（留言数量）

首先，$Qscore$(问题得分) = 赞成票数量 − 反对票数量 。如果某个问题越受到好评，排名自然应该越靠前。$StackOverflow$ 允许用户投反对票，所以这里可以使用绝对投票数，即正面票数量 − 负面票数量。绝对数越高问题越热门。$Qanswers$ 表示回答的数量，代表有多少人参与这个问题。这个值越大，得分将成倍放大。

（3）$Ascores$（留言得分）

这一项的得分越高，就代表留言更加能反映这个问题。简单加总的设计还不够全面。这里有一个问题。首先，对于反应的不同问题，留言的数量与切合问题的程度是不同，但是，简单加总会导致，1 个得分为 100 的留言与 100 个得分为 1 的留言，总得分相同。

（4）$Qage$（距离问题发表的时间）和 $Qupdated$（距离最后留言的时间）

改写一下，可以看得更清楚：$Qage$ 和 $Qupdated$ 的单位都是秒。如果一个问题的存在时间越久，或者距离上一次回答的时间越久，$Qage$ 和 $Qupdated$ 的值就相应增大。

2.2.2.3 热度影响因素量化

话题热度在平时沟通的过程中是一个很模糊、难以界定的概念，为了能对话题热度进行比较排序，必须将影响话题热度的因素逐一量化，使之成为能够参与数值计算的数据。综合考虑以上所列的因素，本文使用的文本数据全部从政府相关部门网站上下载，具有可靠性和专业性。此外，本文采用问题的浏览次数、问题得分、留言数量、留言得分、距离问题发表时间、距离最后留言的时间六个因素作为热度排序依据，并将这些因素量化。

（1）$Qviews$（问题的浏览次数）

利用每类留言数据点赞数与反对数数值之和作为该类热点问题的浏览次数。

（2）$Qscore$（问题得分）

$Qscore$（问题得分）可用每类留言数据点赞数与反对数数值之差数值量化。

（3）$Qanswers$（留言数量）

通过聚类分类后统计每类热点问题的频数代表每类留言的数量。

（4）$Ascores$（留言得分）

通过构建 LDA 模型可得到该类的主题词，通过逐一计算每类留言的每条数据对于该类主题词的推断值作为该条留言数据的留言得分，类内的总推断值即为该类热点问题的留言得分。

（5）$Qage$（距离问题发表的时间）和 $Qupdated$（距离最后留言的时间）

计算该类留言的所有留言数据中的最早与最晚留言时间，再计算当前时间与最早留言时间之差作为 $Qage$，计算当前时间与最晚留言时间之差作为 $Qupdated$。

2.2.2.4 问题热度排序实现流程

对上一步已量化的热度影响因素通过上述 $StackOverflow$ 排名算法 进行计算，再利用计算得到的各类热度指标来对问题的热度进行排序。处理后的数据存入“热点问题表.xls”中的热度指标列中。

2.2.3 主题分析

2.2.3.1 主题模型

主题模型($Topic\ Model$)是用来在一系列文档中发现抽象主题的一种统计模型。直观来讲，如果一篇文章有一个中心思想，那么一定存在一些特定词语会出现的比较频繁。

如果现在一条留言是在讲噪音扰民的，那么“噪音”和“商户扰民”等词语出现的频率会更高一些；如果现在一篇文章是在描述强制学生实习的，那么“学院学生”和“企业实习”等词语出现的频率会更高一些；

主题模型就是一种自动分析每个文档，统计文档中的词语，根据统计的信息判断当前文档包含哪些主题以及各个主题所占比例各为多少。

2.2.3.2 LDA 简介

LDA 主题模型是一种三层贝叶斯模型，三层分别为：文档层、主题层和词层。

该模型基于如下假设：

1）整个文档集合中存在 k 个互相独立的主题；

2）每一个主题是词上的多项分布；

3）每一个文档由 k 个主题随机混合组成；

4）每一个文档是 k 个主题上的多项分布；

5）每一个文档的主题概率分布的先验分布是 $Dirichlet$ 分布；

6）每一个主题中词的概率分布的先验分布是 $Dirichlet$ 分布。

LDA 主题模型的参数：

α：表示 $document-topic$ 密度，α越高，文档包含的主题更多，反之包含的主题更少。

β：表示 $topic-word$ 密度，β越高，主题包含的单词更多，反之包含的单词更少。

主题数量：主题数量从语料中抽取得到，使用 $Kullback\ Leibler\ Divergence\ Score$ 可以获取最好的主题数量。

主题词数：组成一个主题所需要的词的数量。这些词的数量通常根据需求得到，如果说需求是抽取特征或者关键词，那么主题词数比较少，如果是抽取概念或者论点，那么主题词数比较多。

迭代次数：使得 LDA 算法收敛的最大迭代次数

文档的生成过程如下：

1）对于文档集合 M，从参数为 β 的 $Dirichlet$ 分布中采样 $topic$ 生成 $word$ 的分布参数 φ ；

2）对于每个 M 中的文档 m，从参数为 α 的 $Dirichlet$ 分布中采样 doc 对 $topic$ 的分布参数 θ；

3）对于文档 m 中的第 n 个词语 W_mn，先按照 θ 分布采样文档 m 的一个隐含的主题 Z_m ，再按照 φ 分布采样主题 Z_m 的一个词语 W_mn。

LDA 是一种非监督机器学习技术，可以用来识别大规模文档集（$document\ collection$）或语料库（$corpus$）中潜藏的主题信息。它采用了词袋（$bag\ of\ words$）的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

2.2.3.3 构建 LDA 模型提取主题

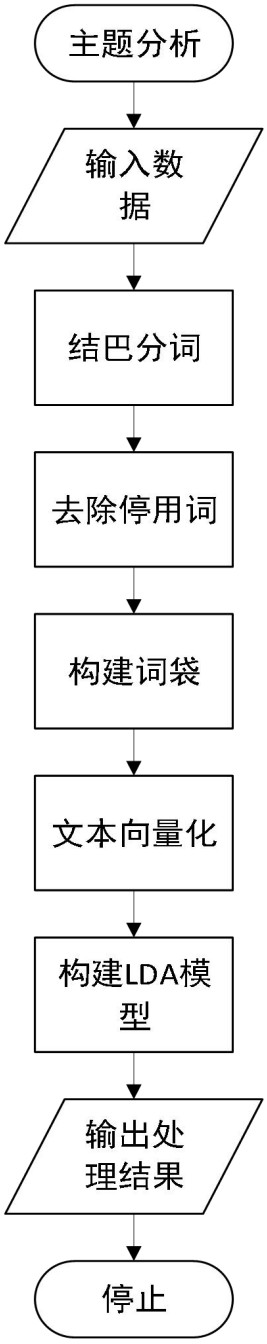

图 9　主题分析流程图

1.数据预处理

对留言主题和留言详情去除空格，使用 $drop_duplicates()$方法去重，删除字母，数字，汉字以外的所有符号，得到去噪清洗过的数据。然后导入停用词表去除停用词以及利用 $jieba$ 中文分词对短信进行分词。

2.过滤低词频 token

定义 $txts_filter_once()$函数，计算清洗处理过的数据的词频，过滤掉只出现一次的 $token$，返回过滤后的数据，可发现数据量大幅减小。

3.词性标注

本题使用 $jieba$ 库对分词结果标注相应词性。

词性标注（$part-of-speech\ tagging$），又称为词类标注或者简称标注，是指为分词结果中的每个单词标注一个正确的词性的程序，也即确定每个词是名词、动词、形容词或者其他词性的过程。

词性标注基本可以参考分词，在中文中，大多数词语只有一个词性，或者出现频次最高的词性远远高于第二位的词性。选取最高频词性，就能实现 80%准确率的中文词性标注程序。

4.稀疏向量集

由处理过的留言主题和留言详情生成语料词典后使用 $TfidfModel$ 将文本表示为向量，并使用 $gensim$ 生成稀疏向量集。

5.构建模型

构建 LDA 主题模型，合理设置参数，并选择适当的词数输出主题词。

6.主题模型推断

$lda.inference$ 简介：

利用 $topic$ 个数以及 $word$ 个数，初始化γ以及φ。严格实现这个过程，工程做了优化，对φ取了对数 $log\varphi$，这样降低计算复杂度，同时利用 log $\_sum$ 接口，计算$\log(\varphi_1)\log(\varphi_2)\cdots\log(\varphi_k)$计算出 $\log(\varphi_1+\varphi_2+\cdots+\varphi_k)$，这样利用$(log\varphi_1)-\log(\varphi_1+\varphi_2+\cdots+\varphi_k)$即可完成归一化。

利用 $lda.inference$ 方法推断每个语料库中的主题类别对于每个主题的推断值，可明确得到某数据对于每一个主题的推断值，从而验证提取效果。效果如下：

['A 市伊景园滨河苑开发商违法捆绑销售无产权车位']

主题 0 推断值 0.20

主题 1 推断值 0.19

主题 2 推断值 0.19

主题 3 推断值 9.22

主题 4 推断值 0.17

['伊景园滨河苑项目绑定车位出售是否合法合规']

主题 0 推断值 0.20

主题 1 推断值 1.60

主题 2 推断值 3.67

主题 3 推断值 2.33

主题 4 推断值 0.17

把热度排序后的前五类热点问题建立模型，得出每一类的问题描述。结果保存在“热点问题留言明细表.xls”中。

2.2.4 地点识别

在程序中使用正则表达式匹配留言中的地名，以提取前五类热点问题的每条留言数据中的特定地点，为热点问题的归类补充完整，结果存入“热点问题表.xls”中的地点/人群列。

2.3 问题 3 分析方法与过程

2.3.1 答复质量影响因素

1.相关性：指的是相关部门的答复意见是否群众的问题相关，是否答非所问。

2.完整性：指的是否满足某种固定的格式，从目前的数据中看出，大多数答复意见都有一个固定的开头和固定的结尾格式，而且大多数答复字数在一定的范围内，这些都可以用来检验完整性。

3.可解释性：指的是答复意见中的内容的相关解释，对问题的答复是否有具体的解释或者是一些理论支撑，例如引经据典和引进法律等。

4.及时性：指的是相关部门对群众的问题答复是否及时，这也是一个重要的指标。它可以从表中的留言时间和答复时间中反映出来。

2.3.2 影响因素的量化

1.相关性

采用的方法是首先对附件 4 中的留言详情与答复意见两列的数据进行特征词提取，然后再利用第二问中的提到的余弦相似度来计算两者之间的距离，距离越小，两者相似度越大，也就越相关。

(1)特征词提取

这里我们直接使用的是 $jieba.analyse.extract_tage()$函数来对数据进行特征词提取。下面是对该函数的简介：

$$Keywords = jieba.analyse.extract\_tags(content,topk = 5,withWeight = True,allowPOS = ()) \quad (23)$$

第一个参数：待提取关键词的文本；

第二个参数：返回关键词的数量，重要性从高到低排序；

第三个参数：是否同时返回每个关键词的权重；

第四个参数：词性过滤，为空表示不过滤，若提供则仅返回符合词性要求的关键词。

(2)计算相似度

对上一步提取的数据处理，其处理步骤如下：

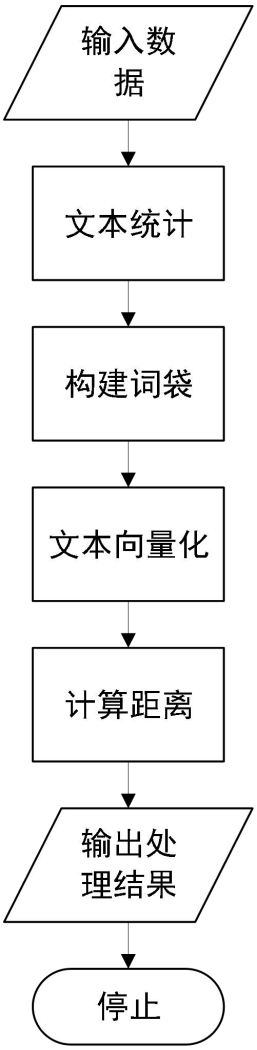

图 10　相关性计算流程图

经过以上处理我们便得到了群众留言与对应答复的相似度，再利用统计学的方法计算出一个阀值，若大于阀值，则相关性差，反之，则相关性良好。

2.完整性

这里分为两步，一是答复格式的确定和答复意见字数的统计，二是利用正则表达式来判断答复意见是否满足格式。

(1)格式的确定

这里我们先对部分答复意见进行粗略的格式提取。后对答复意见进行文本处理，找到其中的高频词组，用来使答复的格式更加精炼。具体步骤如下：

数据清洗：

利用 *pandas* 库来处理表格，去除空行，首先使用 *jieba* 库对数据进行分词处理。后加载停用词表，对数据进行去除停用词操作，可根据数据情况进行扩充停用词表。

绘制词云：

利用前面定义的 *txts_filter_once*()函数，计算清洗处理过的数据的词频，过滤掉出现次数较少的数据，返回过滤后的数据，可发现数据量大幅减小。后利用 *wordcloud* 包实现生成中文词云，最后就是提取高频词组。

图 11　词云

(2)按格式进行正则匹配

直接利用正则表达式来表示答复格式，利用 *re* 库来判断答复意见是否满足格式，并且判断答复意见的字数是否满足要求。如果答复意见满足上面的判定，就说明它具有完整性。

3.可解释性

直接使用正则表达式进行匹配，如果含有某些法律或条例，就说明该答复意见具有可解释性。

4. 及时性

通过计算各个数据对应的答复时间和留言时间的时间差作为答复质量及时性的评判依据。通过统计学的方法计算出一个阀值，时间差小于阀值，则满足及时性，反之，即不满足。

2.3.3 层次分析法 AHP

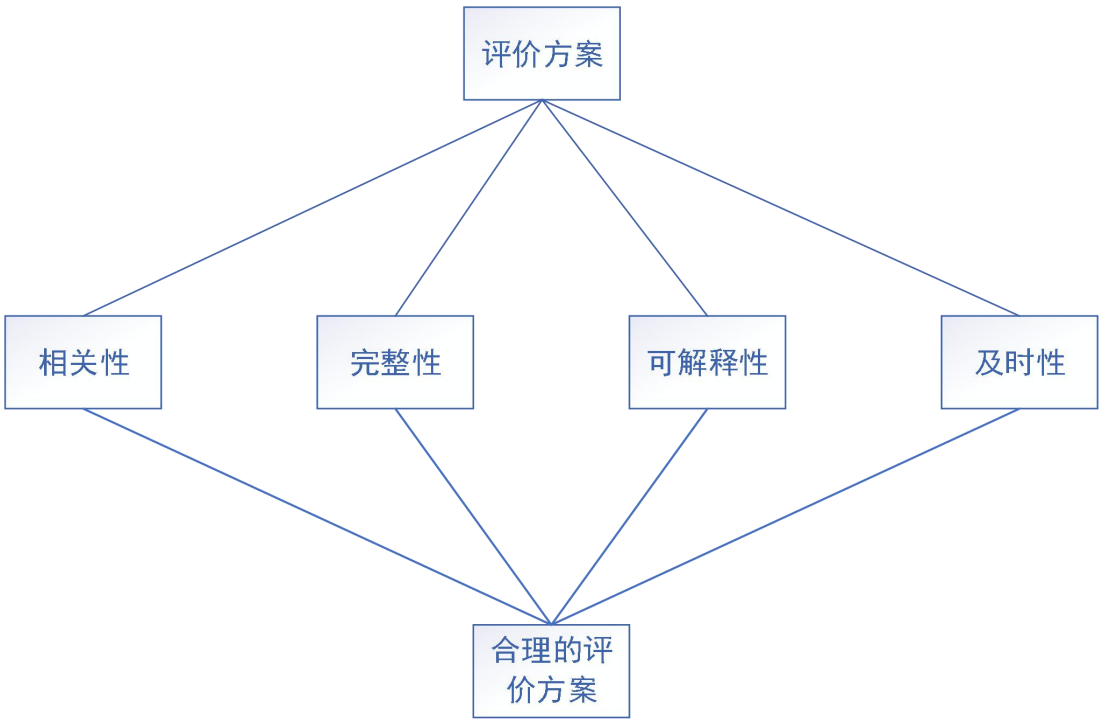

图 12　层次分析法

层次分析法，简称 *AHP*，是指将与决策总是有关的元素分解成目标、准则、方案等层次，在此基础之上进行定性和定量分析的决策方法。该方法是美国运筹学家匹茨堡大学教授萨蒂于 20 世纪 70 年代初，在为美国国防部研究“根据各个工业部门对国家福利的贡献大小而进行电力分配”课题时，应用网络系统理论和多目标综合评价方法，提出的一种层次权重决策分析方法。

层次分析法的基本步骤：

1.建立层次结构模型

一般分为三层，最上面为目标层，最下面为方案层，中间是准则层或指标层。

若上层的每个因素都支配者下一层的所有因素，或被下一层所有因素影响，称为完全层次结构，否则称为不完全层次结构。

2.构造成对比较矩阵

设某一层有 n 个因素，$X=\{x_1,x_2,\cdots,x_n\}$。要比较该层的每一个因素对上一层的某个因素的影响程度，确定在该层中相对于某一准则所占的比重。

假设上一层有 m 个因素，该层有 n 个因素，那么对于该层我们需要构建 m 个 $n*n$ 的成对比较矩阵。

用表示第 i 个因素相对于第 j 个因素的比较结果，比较时取 1~9 尺度。

A 则称为成对比较矩阵。

3.层次总排序及其一致性检验

计算某一层次所有因素对于最高层（总目标）相对重要性的权值，称为层次总排序。

这一过程是从最高层次到最低层次依次进行的。

算法总结：

应用领域：经济计划个管理，能源政策和分配，人才选拔和评价，生产决策，交通运输，科研选题，产业结构，教育，医疗，环境，军事等。

处理问题类型：决策、评价、分析、预测等。

建立层次分析结构模型是关键一步，要有主要决策层参与。

构造成对比较矩阵是数量依据，应由经验丰富、判断力强的专家给出。

2.3.4 评价方案

根据层次分析法分配的权重综合比较，可得出一套评价方案：

根据“附加 4”中所给部分群众留言及相关部门对于它们的答复意见，为了构建答复意见质量综合评价体系，对提升政府的管理水平和施政效率具有极大的推动作用。

一、评价方式

利用层次分析法 AHP 从相关性，完整性，可解释性，及时性等方面进行综合性评价，对其重要性分配相应权重，对于有关部门对群众留言的重视程度和完善程度进行探究。

二、评价内容

1.相关性

相关性是指相关部门的答复意见是否群众的问题相关，是否答非所问。

2.完整性

完整性是指否满足某种固定的格式，从目前的数据中看出，大多数答复意见都有一个固定的开头和固定的结尾格式，而且大多数答复字数在一定的范围内，这些都可以用来检验完整性。

3.可解释性

可解释性是指答复意见中的内容的相关解释，对问题的答复是否有具体的解释或者是一些理论支撑，例如引经据典和引进法律等。

4.及时性

及时性是指相关部门对群众的问题答复是否及时，这也是一个重要的指标。它可以从表中的留言时间和答复时间中反映出来。

三、评价过程

1.数据处理与统计

由上文所提及的数据处理方法和统计，可得到所有给定数据的总体相关性占91.726%，满足范式的比例为50.3169%，满足可解释性的比例为66.4%，答复意见字数均值约为361字，时间差均值为19.8217天。

2.层次分析法确立权重

数据统计之后我们需要判断相关性、完整性、可解释性和及时性的相对重要程度。

首先确定判断矩阵，再通过层次分析法AHP确定各个影响因素指标分配的权重，即相关性占0.25，完整性占0.25，可解释性占0.25，及时性占0.25。

3.评价答复质量

综合以上所述，可以得到各个影响因素的数据和所分配的权重。

由此可得，相关部门在答复意见中绝大部分能够充分做到答符所问，答复意见与群众留言紧密相关，能够就群众的问题提出解决方案或者策略；但其中有少部分答非所问，滥竽充数；基本满足答复范式，答复意见字数也较高，完整性也可以相应得到体现；在答复意见中对于相关法律法规或者引经据典占比较大，大部分数据具备对答复意见的理论支持；整体数据中答复意见发布时间与留言时间时间差约为20天，除去答复时间与发布时间之间的时间差，可得到相关部门对

于群众留言的答复意见相对及时，能够对群众的问题及时给予反馈和解答。

3.结果分析

3.1 问题 1 结果分析

3.1.1 准确度分析

Accuracy= 0.9785809906291834

$$分类正确率=\frac{人决定和系统决定相同的个数}{总督决定个数}=\frac{Y(+)+N(-)}{KN} \quad (24)$$

分类正确率越大越好。缺点：平均审视每一个决定：错误的种类可能不同，不能平均审视。可看出构建的线性支持向量机分类模型准确度很高，达到绝佳的分类效果。

3.1.2 F1 值分析

多分类模型一般不使用准确率($accuracy$)来评估模型的质量,因为 $accuracy$ 不能反应出每一个分类的准确性,因为当训练数据不平衡(有的类数据很多,有的类数据很少)时，$accuracy$ 不能反映出模型的实际预测精度,这时候我们就需要借助于 $F1$ 分数、ROC 等指标来评估模型。

精确度和查全率：

FN：假负

FP：假正

TN：真负

TP：真正

精确度：

$$P=\frac{TF}{(TP+FP)} \quad (25)$$

测量当系统说对时，有多少正确率。

查全率：

$$R=\frac{TP}{(TP+FN)} \quad (26)$$

测量是否所有的文档都有该有的分类。

对于 $Precision$ 和 $Recall$，虽然从计算公式来看，并没有什么必然相关性关系，但是，在大规模数据集合中，这 2 个指标往往是相互制约的。理想情况下做到两个指标都高当然最好，但一般情况下，$Precision$ 高，$Recall$ 就低；$Precision$ 低，$Recall$ 就高。

需要综合权衡这 2 个指标，这就引出了一个新的指标 $F-score$。这是综合考虑 $Precision$ 和 $Recall$ 的调和值。

表 3　问题 1 结果分析

	Precision	Recall	F1-score	support
交通运输	0.99	0.98	0.98	1270
商贸旅游	0.98	0.98	0.98	1450
环境保护	0.99	0.98	0.98	850
卫生计生	0.99	0.98	0.99	1254
城乡建设	0.96	0.97	0.97	1930
教育文体	0.99	0.99	0.99	1300
劳动和社会保障	0.96	0.98	0.97	910
Avg/total	0.98	0.98	0.98	8964

从结果中可看出每个类以及各类平均的精确度，召回率，$F1$ 值等信息。在利用百度翻译经过数据增强之后，弥补了数据量不均衡的缺点，通过均衡各类数据的比例，从而高效提升 $F1$ 值。从结果可以看到，$F-Score$ 值达到 0.98，得出构建的线性支持向量机分类器分类效果拔尖。

3.1.3 混淆矩阵与判错示例

混淆矩阵的主对角线表示预测正确的数量,除主对角线外其余都是预测错误的数量。

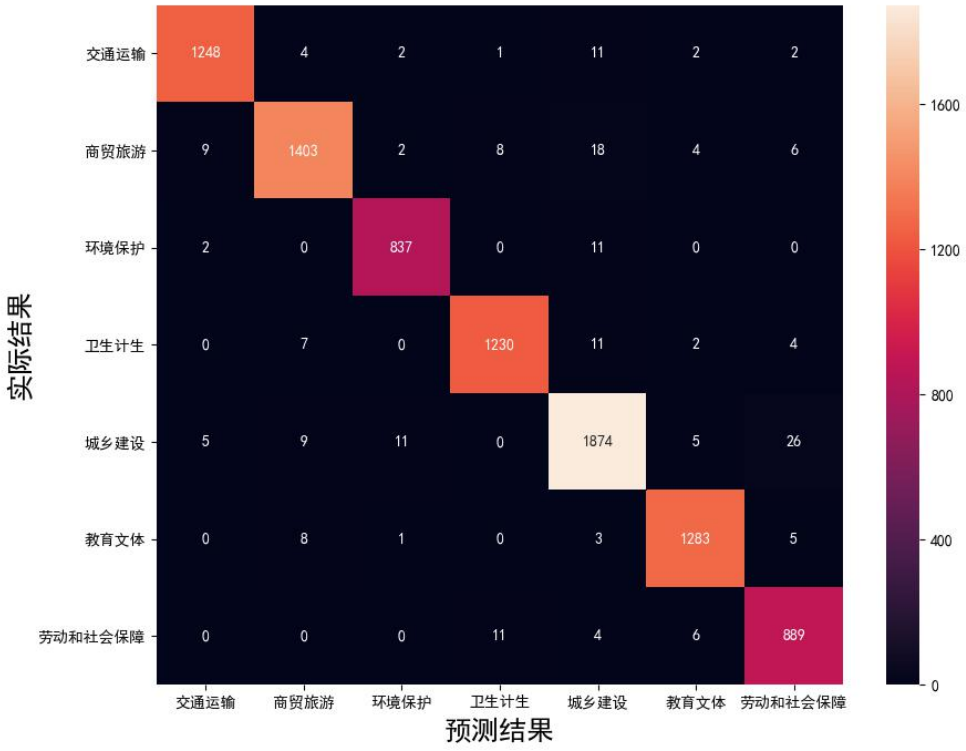

图 13　混淆矩阵

从上面的混淆矩阵可以看出"交通运输"和"卫生计生"准确率较高。"城乡建设"预测的数量较多。

可通过查看一些预测失误的例子,来改善我们的分类器。

例如：城乡建设 预测为 交通运输 ：13 例.

交通运输 预测为 城乡建设 ：17 例.

环境保护 预测为 城乡建设 ：14 例.

通过观察判断失误的示例和比例对模型的参数进行调试与修改，以期达到更好的效果。

3.2 问题 2 结果分析

表 4　热点问题表

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	1	8.733612875	2019-11-02至2020-01-26	A市A2区丽发新城小区	搅拌站灰尘噪音污染严重
2	2	6.695558398	2019-05-05至2019-09-19	A市A5区汇金路五矿万境K9县	房屋存在一系列问题
3	3	5.123023243	2018-11-15至2019-12-02	A市人才	租房购房补贴问题
4	4	3.161019957	2019-07-02至2019-09-01	A市武广新城伊景园滨河苑	强制捆绑销售车位
5	5	3.103456481	2017-06-08至2019-11-27	A 市经济学院学生	学校强制学生实习

表 5　热点问题留言明细表示例

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
1	188809	A909139	A市万家丽南路丽	2019/11/19	A市万家丽南路丽发新城居民	0	1
...							
1	287458	A909241	环境污染	2019/12/2014:06:	丽发新城小区旁建了一个大型搅	0	0
2	208069	A00094436	A5区五矿万境K9县	2019/5/513:52:50	本人是A5区洞井街道汇金路五矿	0	2
...							
2	275491	A00061339	A市五矿万境K9县负一楼面	2019/9/109:10:22	关于五矿万境·K9县负一楼面积缩水的反馈。	0	0
3	189180	A00010651	A市人才购房补贴	2019/6/18	胡书记，您好。我于19年三月在	0	3
...							
3	289408	A0012413	在A市人才app上申请购房	2018/11/1516:07:12	我叫朱琦梦，是2017年12月落户，并于2018年	0	0
4	188801	A909180	滨河苑针对广铁职工购房的	2019/8/10:00:00	尊敬的张市长，您好！我叫李建义，来自湖北仙	0	0
...							
4	289950	A00044759	A市伊景园滨河院	2019/7/77:28:06	提问A市政府就广州局集团公司	0	0
5	195917	A909119	A市涉外经济学院组织学生	2019/11/0510:31:38	一名中职院校的学生,学校组织我们学生在外边	0	1
...							
5	360113	A3352352	A市经济学院强制学生外出	2018/5/178:32:04	A市经济学院强制16届电子商务跟企业物流专业	3	0

从最终输出的“热点问题表.xls”与“热点问题留言明细表.xls”中可以看出，经过文本聚类分类、主题分析和热度排序后，可以得到热度排名前五的特定时段、特定地点发生的热点问题，即 A 市暮云街道丽发新城社区搅拌站灰尘噪音污染严重、A 市伊景园滨河苑捆绑销售车位、A 市 A5 区汇金路五矿万境 K9 县存在

一系列问题、A 市人才购房补贴政策和 A 市经济学院强制学生实习及其对应的留言信息。及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。

3.3 问题 3 结果分析

我们根据群众留言和政府回应的数据，提取出相关性、完整性、可解释性和及时性等因素，综合评价答复质量，给出了一个可靠完善的评价方案。

由指标的全局权重可以看出，相关性、完整性、可解释性和及时性这四个指标是评价指标体系中占权重最大的四个指标，这也与我们的主观感觉相符。根据指标的综合权重结果可以看出，我们给出的这个综合评价指标体系，综合评价了这四个方向的差异与融合，这样就可以得到我们的答复意见质量评价指标体系。高富峰指出，现代社会信息快速流动，要求相关部门必须提高输入信息的广泛性、判断信息的准确性、处理信息的有效性、输出信息的可靠性，来对发生或可能产生的变化做出及时有效的反应。

由此可得相关部门要更加重视对民众的回应力的效率与质量，进而满足社会公众的需求。共同建立和谐可持续发展的文明社会。

4.结论

对网络问政平台的群众留言进行分析研究，把群众和社会在平常生活中所遇到的难题与担忧进行分类汇总，对相关部门能高效率的针对性处理问题具有重大意义，但也是文本分析中的一个难题。传统的文本经验处理已不能满足新时代数据量庞大的群众留言信息。本文通过线性支持向量机分类模型，对留言短信进行高准确率的分类。解决了人工分类的难题。

对群众留言短信进行分类以后还要对其中相应的热点问题进行发现挖掘。由分析结果可知，A 市暮云街道丽发新城社区搅拌站灰尘噪音污染严重、A 市伊景园滨河苑捆绑销售车位、A 市 A5 区汇金路五矿万境 K9 县存在一系列问题、A 市人才购房补贴政策和 A 市经济学院强制学生实习，为五大热点问题。通过发现热点问题为相关部门能及时有效的处理大众的问题。

文本挖掘可根据公众意见的文本数据，通过挖掘相关部门的答复意见文本数据，根据本文设立的一套评价方案，有效地判断相关部门对群众反映问题的感知程度与执行情况。统计可得，大部分部门能够及时的较完整的解决群众问题，小部分回应观念淡薄。加强相关部门回应力，强化回应力平台建设是我国进入深化行政体系改革的重要内容。其中主要是转变思想，强化“以人为本“的理念。完善体制建设，切实保障公众相关权益。

5.参考文献

[1]-派神-，使用 python 和 sklearn 的中文文本多分类实战开发，[EB/OL]. https://blog.csdn.net/weixin_42608414/article/details/88046380,2019-03-02

[2] 王小小小草，主题模型 LDA 实践与应用 [EB/OL]. https://blog.csdn.net/sinat_33761963/article/details/53945581,2016-12-30

[3] 黄大侠 aa, 层次分析法 AHP ,[EB/OL],https://blog.csdn.net/weixin_40683253/article/details/81222318,2018-07-26

[4]wamg 潇潇,模糊数学模型（一）：隶属函数、模糊集合的表示方法、模糊关系、模糊矩阵，[EB/OL]https://blog.csdn.net/qq_29831163/article/details/89892822.2019-05-06

[5]Em_dark，用户投票的排名算法 [EB/OL]https://blog.csdn.net/Em_dark/article/details/68063236,2020-04-13

[6]Carl-Xie，LDA(Latent Dirichlet Allocation)主题模型，[EB/OL].https://blog.csdn.net/aws3217150/article/details/53840029，2016-12-24

[7]weixin_44941795 ， DataFrame 对时间序列的操作，[EB/OL]https://blog.csdn.net/weixin_44941795/article/details/10086260,2019-09-15

[8]张国锋，在文本聚类中话题热度排序的研究与实现，东华大学，2019

[9]李少温，基于网络问政平台大数据挖掘的公众参与和政府回应问题研究，2019

[10]杨凯，基层政府回应力评价指标体系的建立与应用，2016

[11]基基伟，Python+gensim-文本相似度分析，[EB/OL].https://blog.csdn.net/Yellow_python/article/details/81021142，2018-07-12

[12]zhuhengv,基于用户投票的排名算法（三）：StackOverflow, [EB/OL],https://blog.csdn.net/zhuhengv/article/details/50475925,2016-01-07