智慧政务中的文本挖掘应用

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，大幅度加重了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负担。智慧政务服务已经是社会治理创新发展必然趋势，有助于极大地提高政府管理水平和施政效率。因此，运用自然语言处理和数据挖掘技术对留言及答复意见的研究具有重大意义。

对于问题 1，本文通过分词、去停用词等文本预处理，得到留言详情的词语列表，并根据列表计算得出相应的 TF-IDF 权值矩阵及对应的标签值(这里将七个一级标签用数字 0-6 标记)。并采用 python 中的 train_test_split 方法将数据分为训练集和测试集用于训练和测试 LinearSVC 模型，最后将分类结果用 F1-score 进行评价。

对于问题 2，本文在文本预处理后生成 TF-IDF 权值矩阵，将每一条留言当作一个簇，采用自底向上的方法(距离矩阵的计算才用的是余弦相似度和簇间平均距离结合的方法)进行层次聚类，生成热点问题留言明细表。之后分别对归完类的留言的点赞数、反对数、出现此类问题的频率、反应此类问题的用户数、此类问题的持续时间做量化分析，用层次分析法得到这五个特征的权值后，带入 TOPSIS 模型计算出分类问题的热度值，最终通过热度值进行排名，得到排名前五的热点问题。

对于问题 3，本文从信息量、相关性和可读性三个方面来考虑答复意见的合理性给出一套评价方案。信息量是先通过大致观察文本的特征，发现文本的信息量小时，内容大多是停用词或无意义的称呼；相关性是通过提取答复意见的关键字与对应留言问题进行比对判断是否相关；可读性是通过计算文本可读性分数来判断。一条合理的答复意见需要经过信息量、相关性和可读性的三次筛选，最终剩下的答复意见可以认为是合理的。

关键词：中文分词;TF-IDF；层次聚类；Linear SVC;TOPSIS;AHP

Application of text mining in intelligent government

Abstract: In recent years, with Wechat, micro-blog, mayor's mailbox, sunshine hotline and other online political platforms have gradually become an important channel for the government to understand public opinion, gather people's wisdom and gather people's morale. In addition, the amount of text data related to various social conditions and public opinions has been increasing, which has greatly increased the workload of the relevant departments, which mainly rely on manual work to divide up messages and sort out hot spots.The intelligent government service has been the inevitable trend of social governance innovation and development. Therefore, the use of natural language processing and data mining technology to comment and reply to the Study of opinion is of great significance.

For the first problem, this paper gets the detailed list of words by text preprocessing, such as word segmentation, word de-stop, etc. , the TF-IDF Weight Matrix and the corresponding tag values are calculated from the list. Here seven first-level tags are marked with the number 0-6. The data are divided into training sets and test sets for training and testing LINEAR SVC models by using the train method in Python. The classification results are evaluated by F1-score.

For the secoond problem, TF-IDF Weight Matrix is generated after text preprocessing, and each message is treated as a cluster, the bottom-up method is used to calculate the distance Matrix, and the cosine similarity and the average distance between clusters are used to cluster hierarchically to generate the hot topic list. Then they did a quantitative analysis of the likes and dislikes of the categorized comments, the frequency of such questions, the number of users who responded to such questions, and the duration of such questions, then, TOPSIS model is used to calculate the heat value of the classification problem, and finally the top five hot problems are obtained by ranking the heat value.

For the third problem, this paper considers the rationality of the response from three aspects: completeness, relevance and Interpretability, and gets a set of evaluation scheme. Integrality is first through a rough observation of the characteristics of the text, found that the text of small amount of information, most of the content is stop words or meaningless call; Relevance is measured by extracting the key words of the response and comparing them to the corresponding question left in the comment. Interpretability can be achieved through a readability response, where a reasonable response is screened three times for information volume, relevance, and readability, resulting in

a response that is considered reasonable.

Keywords: Chinese word segmentation;TF-IDF;Hierarchical clustering;Linear SVC;TOPSIS;AHP

目录

1. 挖掘目标

互联网正融入到社会生活的各个方面，全方面改造着传统行业，互联网与传统行业的融合能更好地满足人们个性化的需求，提供智能化的解决方案、产品和服务，进而催生出更多的新产品、新服务、新模式。如今，智慧政务已成为电子政务升级发展的突破口，是管理型政府走向服务型政府的必然产物，也是引导智慧城市建设的主干线。

本次建模目标是利用互联网公开来源的群众问政留言记录以及相关部门对部分群众留言的答复意见数据，其中包括结构化和非结构化的文本数据，对文本数据进行基本的预处理、中文分词、停用词过滤后，一方面根据一定的划分体系（三级标签体系），采用 SVM 中的 LinearSVC 模型进行文本分类(仅针对一级标签)。同时利用层次聚类算法，对留言内容进行聚类，将某一时间段内反映特定地点或特定人群问题的留言进行归类，选取了热度评价指标，采用层次分析法、TOPSIS 法，给出排名前 5 的热点问题；另一方面，给出答复意见的评价体系，从信息量、相关性、完整性、可解释性等方面对答复意见进行评价，分析政府对于群众留言答复意见的质量。

2. 分析方法与过程

2.1 总体流程

本文的总体架构及思路如下：

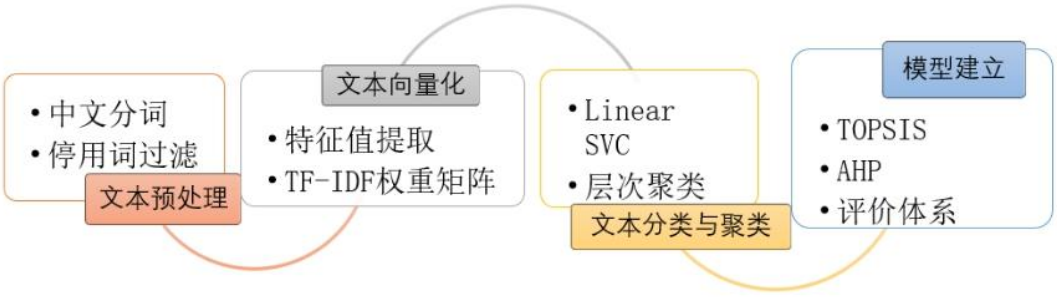


图 1　总体流程图

步骤一：数据预处理，对附件 2 非结构化文本中文文本分词、停用词过滤，以便后续分析，并用 F-Score 对分类方法进行评价；

步骤二：文本向量化，基于 TF-IDF 权重法提取关键词，构造词汇-文本矩阵；

步骤三：文本聚类，根据文本向量，计算文档间的欧式距离，再基于层次聚类算法对留言进行聚类，根据选取的热度评价指标，给出排名前 5 的热点问题；

步骤四：从完整性、相关性、完整性、可解释性等方面评价答复意见质量。

2.2 具体步骤

2.2.1 问题 1 分析方法与过程

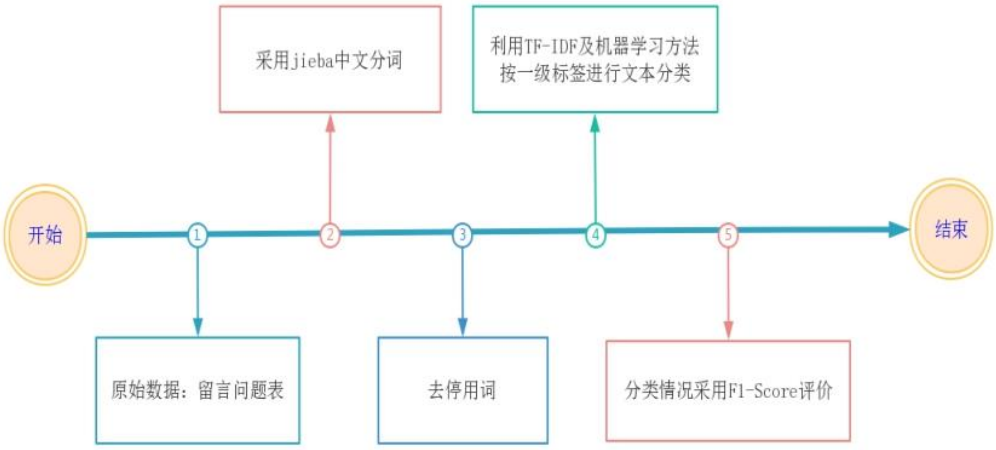


图 2 问题一流程图

2.2.1.1 数据预处理

我们把文本的预处理分为两个部分：中文分词和停用词过滤。

◆对留言数据的中文分词

词是句子语义的最小单位，句子需要切分由词构成的集合，为后续对句子进行分析处理和理解奠定基础。英文语句中词与词之间有明显的分隔符，中文语句中词之间没有明显的分隔标志，切分时需要根据语义，另外中文词在句子中的前后关系复杂，不同词在不同的语境中意义不同，因此中文分词难度较大[1]。

本文分词采用当前广泛使用的基于 Python 的中文分词工具 jieba，jieba 中文分词工具内置多个算法，支持多种模式进行分词，能有效解决未登陆词和歧义词，准确率高达 97%，召回率高达 90%。结巴分词是一种适合中文分词的方法，主要支持精确模式、全模式和搜索引擎模式 3 种分词模式，支持繁体分词、支持自定义词典以及 MIT 授权协议。

结巴分词的原理如下：

（1）是基于统计词典，先构造一个前缀词典；

（2）利用前缀词典对输入句子进行切分，得到所有的切分可能，根据切分位置构造有向无环图（DAG）；

（3）通过动态规划算法，计算得到最大概率路径，找出基于词频的最大切分组合；

（4）对于未登录词，采用了基于汉字成词能力的 HMM 模型，并运用 Viterbi 算法进行计算及词性标注；

（5）分别基于 TF-IDF 和 TextRank 模型抽取关键词。

以下是结巴分词技术路线图。

图 3　结巴分词技术路线图

对原文本进行 jieba 分词后部分分词结果如下图所示：

```
0       [A3，区，大道，西行，便，道，，，未管，所，路口，至，加油站，路段...
1       [位于，书院，路，主干道，的，在水一方，大厦，一楼，至，四楼，人为，拆...
2       [尊敬，的，领导，：，A1，区苑，小区，位于，A1，区，火炬，路，，...
3       [A1，区，A2，区华庭，小区，高层，为，二次，供水，，，楼顶，水箱,...
5       [我，在，2015，年，购买，了，盛世，耀凯，小区，17，栋，3，楼...
                                ...
9205    [ ，  ，  ，  ，  ，  ，  ，  ，  ，  ，  ，  ，我们，夫妻，都...
9206    [ ，  ，  ，  ，  ，  ，  ，  ，  ，  ，  ，  ，本人，2015,...
9207    [ ，  ，  ，  ，  ，  ，  ，  ，  ，  ，  ，  ，我们，是，再婚...
9208    [ ，  ，  ，  ，  ，  ，  ，  ，  ，  ，  ，K8，县惊现，奇葩，证明...
9209    [ ，  ，  ，  ，  ，  ，  ，  ，  ，  ，  ，  ，领导，你好，，...
```

图 4　部分分词结果

分词之后进行词频统计，绘制的词云图如下：

图 5　留言问题词云图

由图可以看出，“业主”、“小区”、“市”、“县”、“领导”、等词语出现频数较大，说明留言问题的范围主要是市级或者县级的业主给领导的留言内容；其次，“项目”、“规划”、“开发商”、“部门”、“房屋”、“居民”、“公司”、“物业”、“政府”等词语出现频率也比较大，说明居民的住房和物业以及公司情况等问题是较为主要的问题。总而言之，与人民群众生活相关的问题是群众所关注的主要问题。

◆停用词过滤

图 4 的分词结果是没有停用词过滤的结果，其中有大量标点及表达无意义的字词，较大程度影响了之后对留言情况的具体分析，因此接下来需要进行停用词过滤，将文本中没有意义或只有极小意义的词语去除掉。

具体原理如下：

停用词（stopword）是指文本中没有意义或只有极小意义的词语。通常在处理过程中要将他们从文本中剔除，这样做的目的是保留具有最大语境及意义的词语。如果基于单个标识聚合语料库，再去查看词语的频率，会发现停用词出现的频率是最高的。目前还没有统一或已穷尽的停用词列表，不同领域可能都会有各自独有的停用词表。

Zipf 定律指出，在文本中，一个单词出现的频率与它在频率表里的排名成反比，换句话说，在文本中出现非常频繁的词在语义上很可能是没有意义的，而稀有词占字典的主要部分，但出现的频率却不高。也就是说，停用词有两个特征：一是极其普遍、出现频率高；二是包含信息量低，对文本标识无意义。[2]因此，对文本进行停用词过滤就十分必要了。

本文筛选出如下停用词：会、月、日、相关、真的等。将筛选出的停用词加入停用词表，再利用停用词表过滤停用词，将分词结果与停用词表中的词语进行匹配，若匹配成功，则将其删除。

去除停用词后的部分结果示例如图 6：

```
0       [大道，西行，道，未管，路口，加油站，路段，人行道，包括，路灯，杆，圈...
1       [位于，书院，路，主干道，在水一方，大厦，一楼，四楼，人为，拆除，水，...
2       [A1，区苑，小区，位于，A1，火炬，路，小区，物业，市程明，物业管理,...
3       [A1，区华庭，小区，高层，二次，供水，楼顶，水箱，长年，不洗，自来水,...
5       [购买，盛世，耀凯，小区，栋，楼，楼，两层，共计，平方，足额，缴纳，...
                              ...
9205    [夫妻，农村户口，女，岁，岁，斤，治疗，两年，一级，脑瘫，女户，招郎,...
9206    [号，B，市中心，医院，做，无痛，人流，手术，手术，怀孕，症状，号，...
9207              [再婚，想，小孩，不知，我省，二胎，新，政策，先，怀孕，做]
9208    [K8，县惊现，奇葩，证明，西地省，K8，县人，想，生二孩，告知，要开,...
9209    [未，婚生子，2013，接受，处罚，小孩，上户，小孩，外地，上学，需，...
```

图 6　停用词过滤后分词结果

2.2.1.2 基于 TF-IDF 的关键词抽取模型

为了提取文本数据中的关键词，我们需要确定一个指标，来对文本中的词语进行排序。TF-IDF 是一种统计方法，用来评价词语对于文本或者语料库中的一个文件的重要性。根据 TF-IDF 数值的高低，可以排序得到词语与文本的相关程

度，用以确定留言的热点问题的范围。因此，本文采用 TF-IDF 的方法对关键词在各文档中的权重进行量化。

TF-IDF(term frequency-inverse document frequency) 即词频-逆文档频率，是基于统计学的计算词权重的方法，是特征向量化的一种常用方法，在信息检索、数据挖掘等领域应用非常广泛。该方法用于评估一个词在该文档中对于区分语料库中其他文档的重要程度，即如果单词出现在本文档中的次数越多，在其他文档中出现的次数越少，则表示该词语对于这篇文档具有越强的区分能力，其权重值就越大[3]。

TF 表示一个词在该篇文档中出现的频率，用于计算这个词描述文档内容的能力。其计算公式如下。

$$T_{i,j}=\frac{n_{i,j}}{\sum_k n_{k,j}}$$

式中，$n_{i,j}$表示在第 j 篇文档中该词出现的次数，$\sum_k n_{k,j}$ 表示对第 j 篇文档中出现的所有词的次数求和。

IDF 即逆文档频率主要是度量一个词语的普遍重要性，如果一篇文档的某个词出现在语料库中的大多数文档中，则说明该词不能够对文档进行区分，反之，则说明该词能够将该篇文档与语料库中的其他文档区分开来。某一词语的 IDF，是用语料库中所有文档的总数目除以含有该词的文档数目的商取对数。计算公式如下。

$$I_i=\log\frac{|D|}{|\{j:t_i\in d_i\}|}$$

其中，$|D|$表示语料库中所有文档的数目，$|\{j:t_i\in d_i\}|$表示语料库中包含词语 t_i 的文档数目。如果词语不在语料库中则会导致分母为 0，为了避免这种情况的发生，通常分母使用

$$|\{j:t_i\in d_t\}|+1$$

然后

$$W_{i,j}=T_{i,j}\times I_i$$

其中，Wi，j 表示所计算文本在语料中的 TF-IDF 权重，文档内的高频率词语以及该词语在整个语料库中的低文档频率能够产生较高的 TF-IDF 权重值。

2.2.1.3 基于 SVC 的文本分类模型

◆基本思想

SVM 理论的主要思想是在样本空间中建立一个最优超平面，将两类样本的隔离边缘最大化，是结构风险最小化（SRM）的近似实现。虽然 SVM 不利用问题所属的领域知识，但仍能够提供良好的泛化性能。SVM 的目标是在样本点所在的向量空间中，找到一个满足分类要求的最优分类超平面，这个超平面能把不同类的

样本分开，使其在满足分类精度的同时，两侧的空白区域（分类间隔）最大化。对于分类间隔最大化的原因，可以简单理解为最大化的分类间隔下，对于未知点的分类更准确，也即泛化能力更强。获得了这个最优超平面后，就可以使用它对于训练样本所属类别进行决策。

在分类问题中，存在着线性不可分的样本,是不能够直接求得决策曲面的。SVM 对低维空间中线性不可分的样本集进行映射，让样本获得在高维特征空间中线性可分的特性，并在这个高维特征空间中求得最优超平面，最终实现样本的分类。

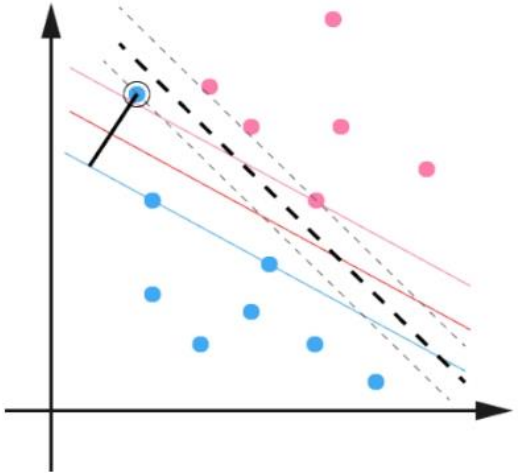


图 7　SVM 分类基本原理

◆具体步骤

对文本单词进行统计，统计出所有单词（去除重复的），然后将这些单词作为特征向量，将行数作为维度。

步骤一：分割数据集

步骤二：使用特征向量化库对文本进行特征向量转化（将文本转化成多维度的特征向量）

步骤三：初始化 SVC 模型，用分割好的训练数据，训练模型，使模型 get 到参数

步骤四：用训练好的模型，预测 X_test

2.2.1.4 F-score 对分类结果的评价

文本分类研究的一个重要环节就是评估分类模型的性能。对于结果的评测我们采用 F1-score 法进行评价。

F1-Score，是准确率和召回率的综合指标：

$$\Pr ecision(c_j)=\frac{TP(c_j)}{TP(c_j)+FP(c_j)}$$

$$\mathrm{Re}call(c_j)=\frac{TP(c_j)}{TP(c_j)+FN(c_j)}$$

$$F1-Score(c_j)=\frac{2\Pr ecision(c_j)\times \mathrm{Re}call(c_j)}{\Pr ecision(c_j)+\mathrm{Re}call(c_j)}$$

其中，TP(cj) 表示属于 cj 的样本且被正确分为 cj 类的样本数；FN(cj) 表示属于 cj 类的样本但没有被正确分为 cj 类的样本数；FP(cj) 表示不属于 cj 类的样本，但被分为 cj 类的样本数。

最初选用三种分类模型，分别为多项式朴素贝叶斯分类器(MultinomialNB)、SGD 分类器(SGDClassifier)和线性支持向量分类器(LinearSVC)来分类，用 F1-score 值来决定最终的分类模型，经过计算可以得到，(一次中)三种模型的 F1-score 值如下：

类别	F1-score 值
MultinomialNB	0.8305314505752082
SGDClassifier	0.831095300554181
LinearSVC	0.9034513554641376

表 1　三种模型 F1-score 值对比

经过多次计算比较，选择效果最好的 LinearSVC 模型来进行热点问题的一级标签分类。

通过对 F1-score 评价指标的分析可知，F1-score 值能达到 0.90 以上，说明 SVC 模型训练较好，分类结果较为理想。

2.2.2 问题 2 分析方法与过程

2.2.2.1 基于层次聚类算法的留言归类

◆思路分析

相似度是用来衡量文本间相似程度的一个标准。在文本聚类中，需要对文本信息进行相似度计算，将根据相似特性的信息进行归类。层次聚类算法分成凝聚的和分裂的两种，取决于层次分解是以自底向上(合并)还是以自顶向下(分裂)方式形成。本题中采用“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。如下图：

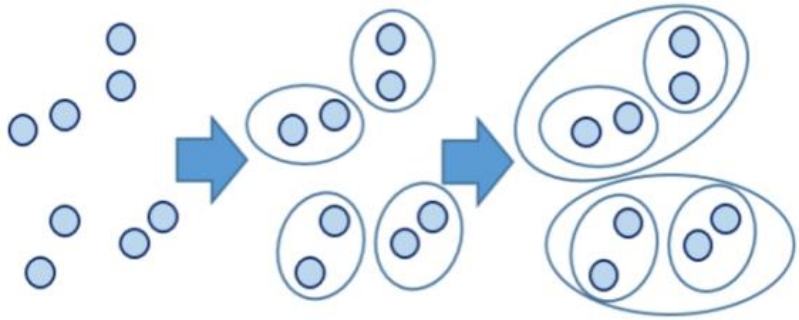


图 8　层次聚类(自底向上)算法示意图

采用余弦相似度作为相似性的评价指标，即该算法认为两个对象的距离越近，其相似度就越大。 用的余弦相似度来判断，其中余弦值的计算公式为：

$$\cos\theta = \frac{\sum_1^n (A_i \times B_i)}{\sqrt{\sum_1^n A_i^2} \times \sqrt{\sum_1^n B_i^2}}$$

其中 A、B 为两个向量，在本题中指的是文本权值向量。余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，这就叫"余弦相似性"。余弦值的范围在[-1,1]之间，值越趋近于 1，代表两个向量的方向越接近；越趋近于-1，他们的方向越相反；接近于 0，表示两个向量近乎于正交。一般情况下，相似度都是归一化到[0,1]区间内，因此余弦相似度表示

$$\cos ineSIM = 0.5\cos\theta + 0.5$$

关键在于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：

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

最大距离：

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

平均距离：

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

本题使用平均距离。

◆具体实现

层次聚类算法的具体描述如下：

输入：给定要聚类的 N 个对象以及 N*N 的距离矩阵(或者是相似性矩阵)

（1）将每个对象归为一类，共得到 N 类，每类仅包含一个对象。类与类之间的距离就是它们所包含的对象之间的距离；

（2）找到最接近的两个类并合并成一类，于是总的类数少了一个；

（3）重新计算所有类与所有旧类之间的距离；

（4）重复(2)(3)步，直到满足一定条件终止(本题中提供一个阈值)。

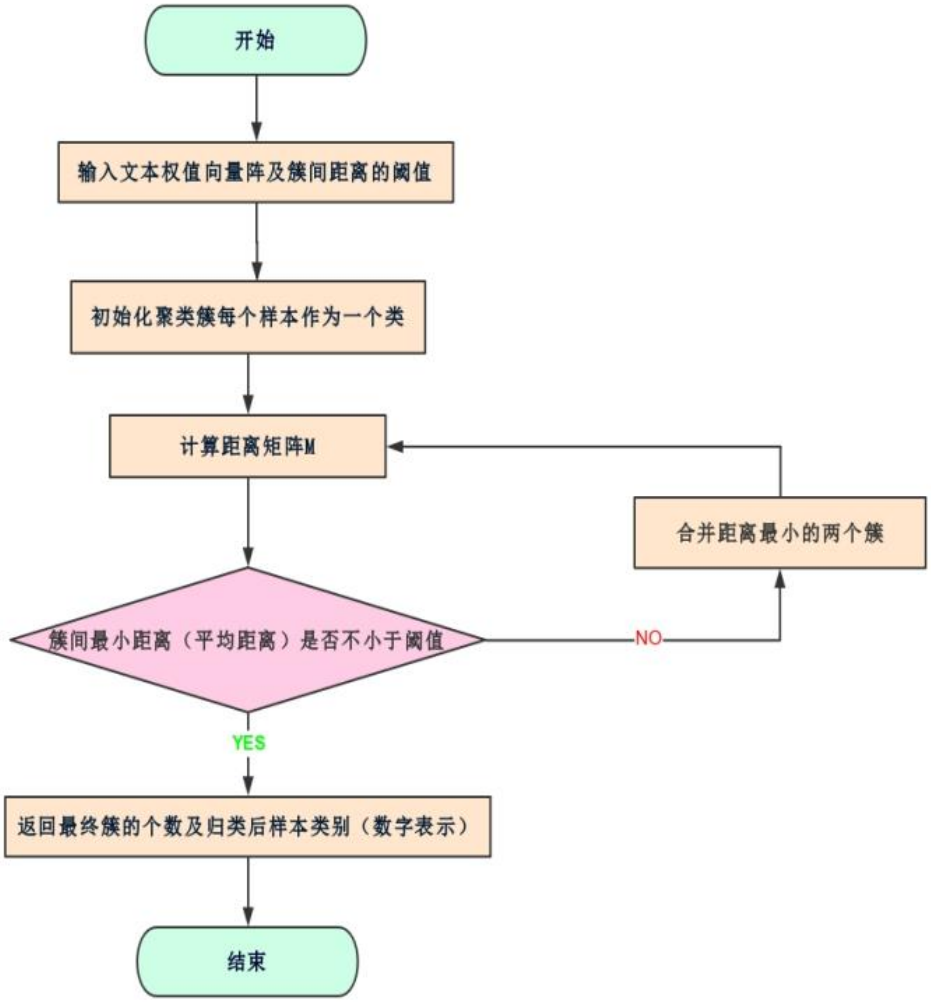


图 9　层次聚类算法流程图

◆层次聚类结果

最终聚类结果将 4326 条留言共分为 1456 个簇。

1151	195686	A00091998	A市滨	2019/	润万滨江天著在售的3/6/8/9
1151	218489	A0001040	请依	2019/	尊敬的A市住建委局长：您好
1151	264735	A0007508	要求A	2019/	申请事项：请求行使监督权
1151	287331	A909217	A2区	####	举报A2区丽发新城小区附近建
1152	195179	A0007788	A市阳	2019/	A市阳光医院2017年6月至7月
1152	197021	A0001065	A7县	2019/	在没有报警立案，也非案件调
1152	248547	A0009806	A7县	2019/	A7县楚龙街道楚郡社区三一街
1153	219069	A0001806	A市交	2019/	A市交通运输局10月23日关于
1153	248228	A0002863	只有A	2019/	尊敬的领导：您好！目前西地
1154	199836	A0006793	A3区	2019/	您好！本人购买的A3区学士街
1154	222275	A0001432	请问A	2019/	地铁3号线什么时候开通

图 10　最终聚类结果

部分聚类结果

5	360107	A0283523	A5区劳动东路魅力之城小区一楼的夜宵摊严重
5	360104	A012417	A市魅力之城商铺无排烟管道，小区内到处油烟
5	360102	A1234140	A5区劳动东路魅力之城小区底层餐馆油烟扰民
1052	248182	A0001580	A4区建筑工地施工
1052	248350	A0001073	A4区万国城3期附近一建筑工地不分昼夜

图 11　部分聚类结果

2.2.2.2 热点问题模型的建立

◆热点问题的定义

某一时段内群众集中反映的某一问题可称为热点问题，本文通过分析认为热点问题需要具备以下特征：

（1）出现频率高

（2）涉及的人群范围广

（3）持续时间较长

针对上述分析，用附件 3 所给的数据，本文用点赞数、反对数、出现此类问题的频率、反映该问题的用户数以及此类问题出现的持续时间长短来反应留言的热度。

◆AHP 层次分析法

（一）基本原理

层次分析法（简称 AHP）是将与评价目标有关的因素主要分解成三个层次，即目标层（Z）、准则（X）、指标层（Y），并且在这个三个主要层次的基础之上开展定性、定量分析的一种决策方法。层次分析法是可以通过模型的建立来确切反映出事态走向的一种科学的研究方法。这种分析法的最大特点就在于对复杂决策问题的本质、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。[4]

本文用层次分析法得出留言热度排名前 5 的问题，基本步骤如下：

(1)建立层次结构模型。在深入分析实际问题的基础上，将有关的各个因素按照不同属性自上而下地分解成若干层次，同一层的诸因素从属于上一层的因素或对上层因素有影响，同时又支配下一层的因素或受到下层因素的作用。最上层为目标层，通常只有 1 个因素，最下层通常为方案或对象层，中间可以有一个或几个层次，通常为准则或指标层。当准则过多时(譬如多于 9 个)应进一步分解出子准则层。

本文构造的热点问题层次结构模型如下：

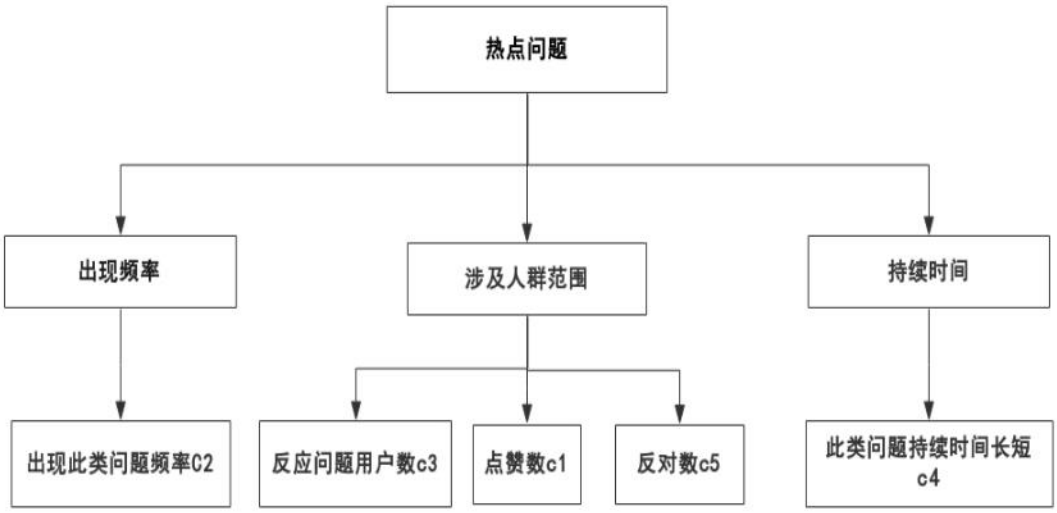


图 12　热点问题层次分析模型

(2)构造成对比较阵。从层次结构模型的第 2 层开始，对于从属于(或影响)上一层每个因素的同一层诸因素，用成对比较法和 1—9 比较尺度构造成对比较阵，直到最下层。

序号	重要性等级	C_{ij} 赋值
1	i, j 两元素同等重要	1
2	i 元素比 j 元素稍微重要	3
3	i 元素比 j 元素明显重要	5
4	i 元素比 j 元素强烈重要	7
5	i 元素比 j 元素极端重要	9
6	i 元素比 j 元素不重要	1/3
7	i 元素比 j 元素明显不重要	1/5
8	i 元素比 j 元素强烈不重要	1/7
9	i 元素比 j 元素极端不重要	1/9

表 2　比较判断矩阵标度及其含义

我们选取了点赞数 c1、出现此类问题的频率 c2、反映该问题用户数 c3、此类问题出现的持续时间长短 c4 以及反对数 c5 五个指标,得到成对比较矩阵如下：

$$\begin{pmatrix} 1 & \frac{1}{7} & \frac{1}{7} & \frac{1}{5} & 3 \\ 7 & 1 & 1 & 3 & 8 \\ 7 & 1 & 1 & 3 & 8 \\ 5 & \frac{1}{3} & \frac{1}{3} & 1 & 7 \\ \frac{1}{3} & \frac{1}{8} & \frac{1}{8} & \frac{1}{7} & 1 \end{pmatrix}$$

(3)计算权向量并做一致性检验。对于每一个成对比较阵计算最大特征根及对应特征向量，利用一致性指标、随机一致性指标和一致性比率做一致性检验。若检验通过，特征向量(归一化后)即为权向量：若不通过，需重新构造成对比较阵。

采用权重算术平均法确定各影响因素的权重，步骤如下：

（a）计算各个有效判断矩阵的权重，即计算判断矩阵的最大特征根及其特征向量。计算方法有很多，精度要求不同方法也不同。精度要求不高时采用和法和方根法，精度要求较高时，一般采用幂法计算权重。判断矩阵每一行元素的乘积M_i，如式所示：

$$M_i = \prod_{j=1}^{n} c_{ij} (i = 1,2,\ldots\ldots,n)$$

（b）计算M_i的 n 次方根$\overline{W}_i$，如式所示：

$$\overline{W_i} = \sqrt[n]{M_i}$$

（c）对向量正规化，如式所示：

$$W_i = \frac{\overline{W_i}}{\sum_{j=1}^{n} \overline{W_j}}$$

就计算出特征向量，即相应的权重系数。

阶数	1	2	3	4	5	6	7	8	9
RI 值	0.00	0.00	0.58	0.90	1.12	1.24	1.32	1.41	1.45

表 3　平均随机一致性指标

(4)计算组合权向量并做组合一致性检验。计算最下层对目标的组合权向量，并根据公式做组合一致性检验，若检验通过，则可按照组合权向量表示的结果进行决策，否则需要重新考虑模型或重新构造那些一致性比率较大的成对比较阵。

（二）具体实现

根据上述步骤，计算出权重以及一致性检验结果。如下表所示：

	权重	CR
点赞数	0.0594737	0.052
出现此类问题的频率	0.3636	
反应该问题的用户数	0.3636	
此类问题持续时间长短	0.179528	
反对数	0.0337991	

表 4　各指标权重

由上表可知，权向量为：

$$(0.0594737,0.3636,0.3636,0.179528,0.0337991)^T$$

一致性比率为 CR=0.052＜0.1，通过一致性检验

权向量 w 代表点赞数和反对数、出现此类问题的频率、反映该问题人群范围的大小以及此类问题出现的持续时间长短对于留言热度的权重。

◆基于 TOPSIS 的热点问题模型

（一）思路分析

TOPSIS 法 (Technique for Order Preference by Similarity to Ideal Solution)，又叫逼近于理想的排序法，是一种多目标决策下解决离散问题的有效方法。由于这种方法计算简便，并不受数据样本多少的影响，所以在经济决策问题中得到了广泛的运用。[5]TOPSIS 是一种多目标决策的离散技术，在许多的经济评价模型中体现了它的有效性。本文采用这种方法建立热度评价模型，得出排名前 5 的热点问题。

与一般的方法相比，它在以下几个方面表现出了明显的优势：

①实现了多目标综合评价，克服了评价标准单一的缺点。

②基于数据的评价方法,有效的减少了人的主观作用,增强了评价的可信度。

③定量的核算，易于在计算机系统中应用。

④整个模型(包括指标体系在内)是一个动态的评价过程。

（二）具体步骤：

假设有 n 类留言问题，则留言问题集为$X=(x_1,x_2,\cdots,x_n)$，热点问题评价体系共选取了五个指标，$C=(c_1,c_2,c_3,\cdots,c_m)$。$X_{ij}$表示$x_i$在第$c_j$指标下的指标值，$x_{ij}>0$，各指标的权重向量记为$w=(w_1,w_2,\cdots,w_m)$，$\sum_{j=1}^{m} w_j=1$

模型建立过程如下：

（1）由假设得原始数据矩阵为：$(X_1,X_2,\cdots,X_m)$

（2）构造标准化数据矩阵

$$Y=(y_{ij})_{nxm}$$

$$y_{ij}=\frac{x_{ij}}{\sqrt{\sum_{i=1}^{n}x_{ij}^2}}$$

（3）构造加权的数字化矩阵 $z=(z_{ij})_{n\times m}$，其中

$$z_{ij}=w_i y_{ij}$$
$$i=1,2,\cdots,n$$
$$j=1,2,\cdots,m$$

（4）构造理想解和负理想解，构造“热点问题特征”和“负热点问题特征”

$$x^{+}=(x_1^{+},x_2^{+},\cdots,x_m^{+})$$

$$x^{-}=(x_1^{-},x_2^{-},\cdots,x_m^{-})$$

其中：

效益型指标，即指标值越大越好的指标，如指标体系中的效益型指标，即指标值越大越好的指标，如指标体系中的 c1,c2,c3,c4。对于这一类指标

$$x_j^{+}=\max_i z_{ij},\quad x_j^{-}=\min_i z_{ij}$$

成本型指标，即指标值越小越优的指标，如指标体系中的 c5。对于这一类指标

$$x_j^{+}=\max_i z_{ij},\quad x_j^{-}=\min_i z_{ij}$$

（5）根据标准化数据矩阵求得每类热点问题与“理想的热点问题特征”和“负理想热点问题特征”的欧氏距离 S^{+}、S^{-}

$$S^{+}=\left\|z_i-x^{+}\right\|=\sqrt{\sum_{j=1}^{m}(z_{ij}-x_j^{+})^2}$$

$$S^{-}=\left\|z_i-x^{-}\right\|=\sqrt{\sum_{j=1}^{m}(z_{ij}-x_j^{-})^2}$$

（6）计算各热点问题与“理想热点问题”和“负理想热点问题”的贴近度

$$u_i=\frac{S_i^{-}}{(S_i^{+}-S_i^{-})}$$

（7）求得的u_i越大说明该热点问题的特征与“理想的热点问题特征更相近”，反之与“负理想热点问题特征更相近”越接近，从而根据u_i得到热点问题序列。

（三）热点问题排名

本文得到排名前五的热点问题如下：

热度指数	时间范围	地点人群	问题描述
0.9731	2019/3/26 至 2019/11/22	A 市 A5 区汇金路五矿万境业主	小区租房现象严重影响小区住户生活
0.8432	2019/1/27 至 2019/12/15	A 市金毛湾业主	学区划分没有之前承诺的金毛湾
0.5314	2019/2/21 至 2019/8/30	A 市 A4 区贷车案受害人	希望书记多关注贷车案，为广大受害人讨回公道
0.4977	2019/2/25 至 2019/9/11	A 市 A2 区暮云复绿业主	工程为什么一直无进展
0.3167	2019/4/24 至 2019/7/19	A 市 A6 区乌山镇居民	A 市外嫁女征地补偿政策不合理

表 5　热点问题表

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
	211218	A0006585	A 市火星镇熊孩子引燃“僵尸车”有感	2019/3/16 20:50:49	刚才看到熊孩子引燃“僵尸车”让人感慨万千啊。为什么单单每次“僵尸车”出事都是火星镇?	0	0
	288536	A0005258O	对 A 市地铁 3 号线的两点建议	2019/1/3 10:34:29	第一：建议 A 市地铁 3 号线合理安排、倒排工期、加快进度，力争在 2019 年	0	0

					10 月 1 日前基		
	2437 29	A 0003 7605	A 市半山壹号三期无限期延迟交房	2019/2/26 11:00:44	奋斗十年才买套房，也是为了小孩能如期上学。开发商无耻，多次约谈未果，没有给一个明确	0	0
	2483 03	A 0003 9772	A7 县金科代时中心小区物业公司违规收取停车费	2019/7/8 18:08:44	A7 县金科代时中心小区物业在没有物价局的审批和业主投票情况下，违规并私自收取地下停车场费用。	0	9
..	...	. ..	...	. ..	...	.. .	.. .
456	2316 95	A 0001 0078 1	关于 A7 县 R 区地坪提质改造管理，请求政府加速实施的请示	2019/3/5 10:00:35	关于 R 区地坪提质改造管理请求政府加速实施的请示县委县政府、街道党工委办事处、	0	0
456	2345 56	A 0009 561	A5 区地质中学附近拥堵问题很严重	2019/3/8 16:11:11	地质中学，位于 A 市 A5 区人民中路 72 号，在校师生近 5000	0	0

					人，周围		
456	241203	A00097317	A5 区中航城 3 期 9 栋开放走廊设计存在安全隐患	2019/5/7 19:26:13	您好！您百忙当中也不知道能不能看到这条消息，所以抱着所有一线可能给您留言。A5 区	0	0
456	249734	A00011640	A 市双龙警苑房屋漏水问题一直没解决	2019/10/10 9:29:22	本人为 A6 区双龙警苑一栋一单元 1703 顶楼业主，于 2010 年入住	0	0

表 6　热点问题留言明细表

2.2.3 问题 3 分析方法与过程

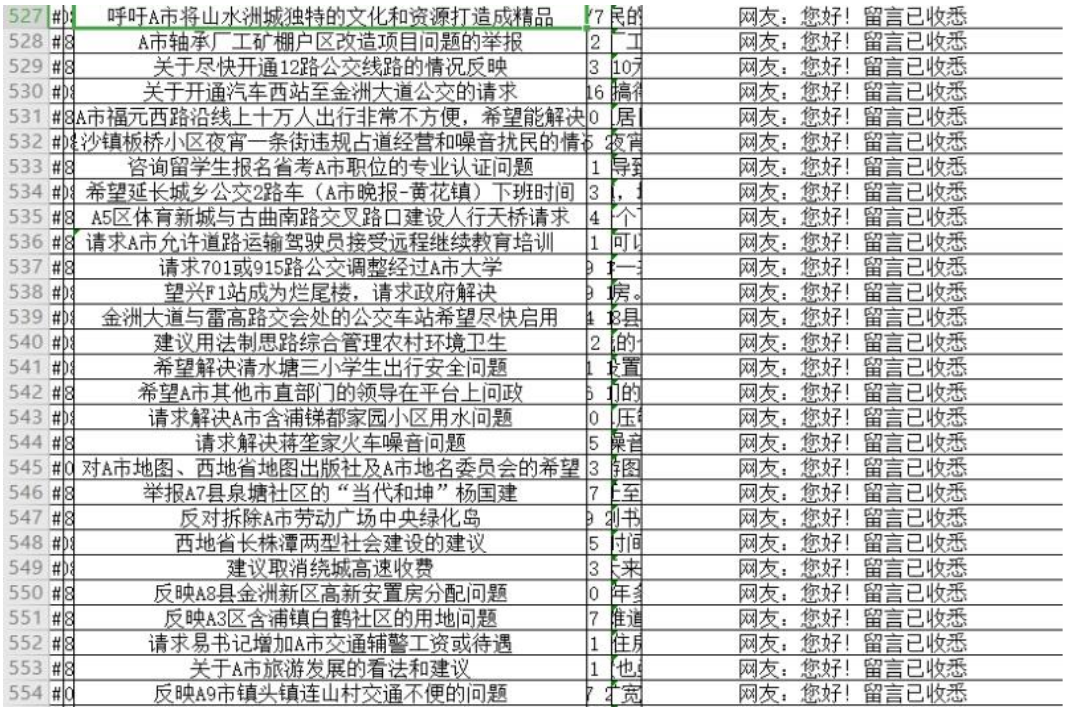
2.2.3.1 完整性的分析

本文通过信息量来反应完整性，经过观察可以看出，一些答复意见仅仅只写了一些无关紧要的信息，往往只有十几个字，与留言主题无关，也没有意义。用去停用词(此处为特殊停用词，如：“网友”、“你好”、日期等)来得到每条留言的词语列表，通过判断列表的长度(是否小于特定阈值)来判断文本的信息量是否合理。

最终输出结果为信息量不足(去停用词后词语列表小于长度阈值 3 的)的的答复意见序号：

```
[243, 517, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538,
539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 549, 550, 551, 552,
553, 554, 838, 1078, 1181, 1361, 1373, 1376, 1425, 1459, 1875, 1913, 2000, 2233, 2339, 2687]
```

部分结果如下图：



527	#0	呼吁A市将山水洲城独特的文化和资源打造成精品	7	民的	网友：您好！留言已收悉
528	#8	A市轴承厂工矿棚户区改造项目问题的举报	2	厂工	网友：您好！留言已收悉
529	#8	关于尽快开通12路公交线路的情况反映	3	10万	网友：您好！留言已收悉
530	#0	关于开通汽车西站至金洲大道公交的请求	16	搞得	网友：您好！留言已收悉
531	#8	A市福元西路沿线上十万人出行非常不方便，希望能解决	0	居[	网友：您好！留言已收悉
532	#0	沙镇板桥小区夜宵一条街违规占道经营和噪音扰民的情	6	夜宵	网友：您好！留言已收悉
533	#8	咨询留学生报名省考A市职位的专业认证问题	1	导致	网友：您好！留言已收悉
534	#0	希望延长城乡公交2路车（A市晚报-黄花镇）下班时间	3	[, [	网友：您好！留言已收悉
535	#8	A5区体育新城与古曲南路交叉路口建设人行天桥请求	4	一个[	网友：您好！留言已收悉
536	#8	请求A市允许道路运输驾驶员接受远程继续教育培训	1	可以	网友：您好！留言已收悉
537	#8	请求701或915路公交调整经过A市大学	9	1一[	网友：您好！留言已收悉
538	#0	望兴F1站成为烂尾楼，请求政府解决	9	1房。	网友：您好！留言已收悉
539	#0	金洲大道与雷高路交会处的公交车站希望尽快启用	4	8县	网友：您好！留言已收悉
540	#0	建议用法制思路综合管理农村环境卫生	2	的[	网友：您好！留言已收悉
541	#0	希望解决清水塘三小学生出行安全问题	1	设置	网友：您好！留言已收悉
542	#8	希望A市其他市直部门的领导在平台上问政	5	1的	网友：您好！留言已收悉
543	#0	请求解决A市含浦锑都家园小区用水问题	0	压[	网友：您好！留言已收悉
544	#8	请求解决蒋垄家火车噪音问题	5	噪音	网友：您好！留言已收悉
545	#0	对A市地图、西地省地图出版社及A市地名委员会的希望	3	特图	网友：您好！留言已收悉
546	#8	举报A7县泉塘社区的“当代和坤”杨国建	7	上至	网友：您好！留言已收悉
547	#8	反对拆除A市劳动广场中央绿化岛	9	2书	网友：您好！留言已收悉
548	#0	西地省长株潭两型社会建设的建议	5	时间	网友：您好！留言已收悉
549	#0	建议取消绕城高速收费	3	大来	网友：您好！留言已收悉
550	#8	反映A8县金洲新区高新安置房分配问题	0	年多	网友：您好！留言已收悉
551	#8	反映A3区含浦镇白鹤社区的用地问题	7	准道	网友：您好！留言已收悉
552	#8	请求易书记增加A市交通辅警工资或待遇	1	住房	网友：您好！留言已收悉
553	#8	关于A市旅游发展的看法和建议	1	也[	网友：您好！留言已收悉
554	#0	反映A9市镇头镇连山村交通不便的问题	7	2宽	网友：您好！留言已收悉

图 13　完整性的判断结果

显然，在上图中，答复意见均为“网友：您好！留言已收悉”，而没有具体的给出解决的方式，信息量很小，答复的十分不合理。

2.2.3.2 相关性的分析

计算每条答复意见的 TF-IDF 权值向量，并将每条意见的词语按照权值大小进行排序，选取前 6 个词语(认为是答复意见关键词)的权值组成一条意见的向量，将所有意见的向量排列成意见 TF-IDF 权值矩阵。将每条答复意见对应的留言主题也按照上述的方式提取出对应的留言主题 TF-IDF 权值矩阵。算出每条答复意见与留言主题之间的余弦相似度，借此来判断相关性。

下图是部分答复意见的关键词的提取过程(前面为词语，后面为对应的 TF-IDF 权值)：

```
----------writing all the tf-idf in the  2733
民政部门 0.7464611126958236
询问 0.6654290399681342
----------writing all the tf-idf in the  2734
残疾人 0.4431052536402531
残联 0.29485405070564574
十八 0.25237255308810735
----------writing all the tf-idf in the  2735
移民 0.639637534950053
湖北省 0.37422082356039077
水库 0.2834734319644894
```

图 14　部分答复意见 TF-IDF 权值

最终的结果输出在给定阈值条件下认为不相关的答复意见(输出意见序号)：

```
[243, 257, 1078, 1109, 1181, 1361, 1373, 1376, 1425, 1459, 1682, 1875, 1913, 2136, 2233, 2294, 2687, 2796]
```

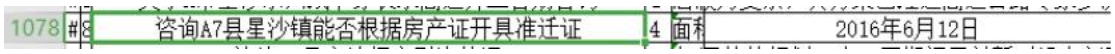
◆部分判断结果图如下：

1.

第 243 条：

留言主题为“咨询打狂犬疫苗报销比例是多少”，而答复意见为“已收悉”，没有合理回复。

2.



第 1078 条：

留言主题为“咨询 A7 县星沙镇能否根据房产证开具准迁证”，而答复意见为“2016 年 6 月 12 日”，显然毫不相关，答复意见不合理。

3.

第 2687 条：

留言主题为“坚决反对在 M5 市拖坪村办石场”，而答复意见为“你好！2019 年 6 月 13 日”，显然与主题不相关。

2.2.3.3 可解释性

本文用可读性来反应文本的可解释性，文本的可读性指的是文本的内容被人理解的难易程度。本文认为答复意见的可读性可以参照公式法来计算可读性级别或分数，经过查找，大致确认有下列几种方法能够实现中文文本可读性指数的计算：

陈世敏在总结了关于英语的可读性公式的研究后，认为不能照搬英语的可读性公式应用到中文中去。他改进了 Gunning Fog 公式，同时结合了 Flesch 公式和 Dale&Chall 公式，建立了如下的中文可读性公式[6]：

$$\text{RedabilityScore} = 0.8\text{cps} + \text{dcr}$$

公式中的自变量依次是平均句子字数和难字比例。

Yang 基于 85 篇中文繁体字选文，总共抽取了 39 个因素变量。他通过统计分析确定了最重要的几个因素，利用中学生对这些短文的阅读分数，建立了如下的可读性公式：

$$\text{Re}adability\text{Score} = 14.9596 + 39.07746 * dwr + 1.01156 * csn - 2.48 * stpc$$

上述公式中的 csn 为完整句子的数量，stpc 是平均字笔画数。Dur 指 5600 常用词表以外的词的比例。另外该公式是对由繁体中文字构成的中文文本的可读性度量，而现在中文的使用基本上都采用简体字。本文的工作都是基于中文简体字构成的文本。

孙汉银选取了 20 篇中文文章，每篇 250 字左右，然后对中学生进行完型测试。基于抽取的可读性因素和测试成绩，创建可读性公式如下：

$$ComprehensionScore = -7.00685 + 14.34587 * stpc - 2.13791 * dwr - 3.38799 * cps + 4.00731 * wps$$

公式中的自变量依次为平均字笔画数、难词比例、平均句子字数和平均句子单词数。

2.2.4 结果分析

2.2.4.1 问题 1 结果分析

F-score 对分类结果的评价

本文通过分词、去停用词等文本预处理，得到留言详情的词语列表，并根据列表计算得出相应的 TF-IDF 权值矩阵及对应的 labels。用 python 中的 train_test_split 方法将数据分为训练集和测试集用于训练和测试 LinearSVC 模型。最终的分类结果用 F1-score 来评价，F1-score 值在 90 以上，分类结果较好，模型训练较为成功。

但是不足之处在于单独判断分类结果，每个分类的 F1-score 均在 95 以上，可能是因为综合判断时，许多文本在预处理后有较大的重复区段，造成分类的失误。

2.2.4.2 问题 2 结果分析

本文在文本预处理后生成 TF-IDF 权值矩阵，将每一条留言当作一个簇，采用自底向上的凝聚层次聚类的方法。其中距离矩阵的计算才用的是余弦相似度和簇间平均距离结合的方法，最终生成热点问题留言明细表，就生成的表格人工检查来看，聚类效果不是很好

但是由于一些留言详情的长度较长，可能 TF-IDF 计算权值的方法更适合短文本，而且留言详情的长短不均衡，可能计算两条留言之间的距离时会产生问题，同时文本预处理不够精细，导致 TF-IDF 矩阵较为稀疏，聚类结果会产生一定偏差。

2.2.4.3 问题 3 结果分析

本文从完整性（信息量）、相关性和可读性三个方面来考虑答复意见的合理性，得出一套评价方案。经过信息量、相关性和可读性的筛查后，余下的答复意见我们认为是相对合理的。信息量的解决：通过观察，可以看出信息量较小的文本均只有“你好!”、“网友”、“已收悉”和日期等信息，不能解决问题，不是正确的答复方式；相关性的解决：分别计算答复意见中关键字(选取排名前六的词语)和留言主题相对应的 TF-IDF 权值矩阵，并计算它们之间的余弦相似度来判断答复意见是否与主题相关，由问题三的结果可以看出，能较好判断答复意见的质量。但是不足之处在于定距离的阈值没有硬性指标，不太好界定如何才算最优结果。

3. 结论

总结本次比赛，我们采用 jieba 分词，停用词过滤等，基于 TFIDF 权重法提取特征词，构造词汇-文本矩阵，然后通过层次聚类算法对热点问题进行了聚类；随后选取了热度评价指标，采用 AHP 得出各指标的权重，用 TOPSIS 得出排名前五的热点问题；最后，本文从信息量、相关性、完整性、可解释性等方面给出了

答复意见的评价方案。

本文所采用的分词算法数据处理速度较快，分词的准确率也较高。不足之处为在处理复合词时，会将较长的词语分开，造成重复，影响研究的准确性。基于 TF-IDF 算法的关键词抽取模型的主要优点为它能过滤掉常见的词语，保留重要的词语。在选出出现频率高的词的同时，也排除了重复出现的词语。但此模型对于用户语料库的要求较高，且无法识别新词，具有一定的局限性.而基于 SVM 的分类模型需要相当数量的训练集才能得到有效的分类器。这都对数据的准备提出了较高的要求。而且，我们最后得到的聚类结果准确度不是特别的好，与准确的结果有一定程度上的出入，这可能由于文本过长等产生的误差，反应了当今中文文本挖掘模型的不足，我们之后也会对中文文本挖掘做更加深入的研究。

4. 参考文献

[1]李春林,冯志骥.基于文本挖掘的新能源汽车用户评论研究[J].特区经济,2020(04):148-151.
[2]张波,黄晓芳.基于 TF-IDF 的卷积神经网络新闻文本分类优化[J].西南科技大学学报,2020,35(01):64-69.
[3]谭章禄,陈孝慈.基于文本挖掘的煤矿安全隐患管理研究[J].中国安全生产科学技术,2020,16(02):43-48.
[4]徐蕾,张科伟.基于文本挖掘的京东商品评论分析[J].内蒙古科技与经济,2020(03):41+43.
[5]冯梦莹,李红.文本卷积神经网络模型在短文本多分类中的应用[J].金融科技时代,2020(01):38-42.
[6]邱智学. 面向 B2C 电商平台的短文本挖掘研究[D].浙江工商大学,2020.

附录

附录一：AHP

```
    # -*- coding: utf-8 -*-
"""
Created on Tue Apr  7 16:48:35 2020

@author: HP
"""

import numpy as np
```

```
RI_dict = {1: 0, 2: 0, 3: 0.58, 4: 0.90, 5: 1.12, 6: 1.24, 7: 1.32, 8: 1.41, 9: 1.45}

def get_w(array):
    row = array.shape[0]   # 计算出阶数
    a_axis_0_sum = array.sum(axis=0)
    # print(a_axis_0_sum)
    b = array / a_axis_0_sum   # 新的矩阵 b
    # print(b)
    b_axis_0_sum = b.sum(axis=0)
    b_axis_1_sum = b.sum(axis=1)   # 每一行的特征向量
    # print(b_axis_1_sum)
    w = b_axis_1_sum / row   # 归一化处理(特征向量)
    nw = w * row
    AW = (w * array).sum(axis=1)
    # print(AW)
    max_max = sum(AW / (row * w))
    # print(max_max)
    CI = (max_max - row) / (row - 1)
    CR = CI / RI_dict[row]
    if CR < 0.1:
        print(round(CR, 3))
        print('满足一致性')
        # print(np.max(w))
        # print(sorted(w,reverse=True))
        # print(max_max)
        # print('特征向量:%s' % w)
        return w
    else:
        print(round(CR, 3))
        print('不满足一致性，请进行修改')

if __name__ == '__main__':
    # 由于地方问题，矩阵我就写成一行了
    e = np.array([[1    , 1/7,    1/7, 1/5    ,3],
                  [7    , 1    ,    1, 3    ,8],
                  [7, 1,    1, 3    ,8],
                  [5, 1/3,    1/3,1 ,7],
                  [1/3    , 1/8    ,    1/8,    1/7 ,1    ]])
    w = get_w(e)
```

附录二：TOPSIS

```
- coding: utf-8 -*-
"""
Created on Tue Apr  7 17:23:21 2020

@author: HP
"""
import numpy as np

X = np.array([[0,0,1,1,2],
              [2,1,1,1,2],
              [3,5,2,1,0]])

w = np.array([0.0482906,0.39359,0.39359,0.06453,0.1])

Y = np.linalg.norm(X, ord=2, axis=0, keepdims=True)

Z = X*Y*w

x_pos = np.array([max(Z[:,0]),
                  max(Z[:,1]),
                  max(Z[:,2]),
                  max(Z[:,3]),
                  min(Z[:,4])])

x_neg = np.array([min(Z[:,0]),
                  min(Z[:,1]),
                  min(Z[:,2]),
                  min(Z[:,3]),
                  max(Z[:,4])])
S_pos = []
S_neg = []
for i in range(Z.shape[0]):
    S_pi = np.linalg.norm(Z[i:]-x_pos, ord=2, keepdims=True).tolist()
    S_ni = np.linalg.norm(Z[i:]-x_neg, ord=2, keepdims=True).tolist()
    S_pos.append(S_pi)
    S_neg.append(S_ni)

S_pos = np.array(S_pos).reshape(Z.shape[0],1)
S_neg = np.array(S_neg).reshape(Z.shape[0],1)
```

U = S_neg/(S_pos-S_neg)