2020 年第五届“数维杯”大学生
数学建模竞赛论文

题　　目　关于舆情监测情感倾向的分析建模

摘　要

公共危机事件爆发时，相关信息在短时间内迅速传播，其中负面消息甚至影响到政府及公共单位的公信力和企业的形象及口碑。通过舆情的情感倾向预测，有助于企业及时对口碑进行维护。问题 1 是对于抓取的部分媒体或网民评论的数据，提供一个针对某一主题的舆情筛选方法；问题 2 是提供一个全新数据的抓取方法，包含具有深层次分析价值的数据；问题 3 是提供一种能够合理引导网民们情感倾向逐步转向对政府或企业有利的干预方法；问题 4 是提供一个充分考虑疫情传播时间、规模及网民情感倾向的舆情处理等级的划分方法。针对于问题 1 采用 LDA 主题模型的方法解决；针对于问题 2 采用 TF-IDF 模型的方法解决；针对于问题 3，在问题 1 和问题 2 的基础上提供相应的一种合理引导网民们情感倾向逐步转向对政府或企业有利的干预方法；针对于问题 4，采用聚类分析和已有数据集得结果进行分析解决。

对于问题 1，对于抓取的部分媒体或网民评论的数据，提供一个针对某一主题的舆情筛选方法这个问题，我们建立了狄利克雷分布（LDA）主题模型。首先，选取了部分所给数据，进行了数据的预处理，即对特殊符号进行处理，对词的长度进行预览；其次，使用 jieba 进行分词，优化主题使其显得更加集中；最后，通过 LDA 模型来对进行文本处理和分词后的文档进行分类，从而达到分类主题的目的。

对于问题 2，对于提供一个全新数据的抓取方法，其中尽量包含诸如发表时间、评论人数、关注人数及具体内容等具有深层次分析价值的数据这个问题，我们使用了四种不同的方法以适用不同情况，分别为建立 TF-IDF 模型，词性标注，正则表达式和自建序列标注平台。首先，先对文章的常用词进行分析，使用序列标注模型中的词性标注对于特定词的提取，使用正则表达式直接对特定内容进行搜索。建立序列标注平台对所需的特定内容进行标注，达成标注要求后，标注完的文本可以作为训练样本，再利用 bert 等框架对模型进行训练；最后，完成训练后的模型即可对任意输入的语料标注出所需的特定内容，达到抓取具有深层次分析价值的数据的目的。

对于问题 3，对于提供一种能够合理引导网民们情感倾向逐步转向对政府或企业有利的干预方法的问题来说，根据问题 1 和问题 2 所得出得结果，分别制定合理引导网民们情感倾向逐步转向对政府或企业有利的干预策略。

对于问题 4，对于提供一个充分考虑疫情传播时间、规模及网民情感倾向的舆情处理等级的划分方法这个问题，首先，我们搜集了情感倾向统计的相关数据，并对这些数据进行聚类分析；其次，根据聚类分析的结果，以及现有收集到的已经标注好的疫情传播时间和情感倾向数据；最后，找到充分考虑疫情传播时间、规模及网民情感倾向的舆情处理等级的划分方法。

关键词　LDA 主题模型；jieba 分词；TF-IDF 模型；正则表达式；聚类分析

目　　录

一、问题重述……（1）
二、问题分析……（2）
三、模型假设……（3）
四、定义与符号说明……（3）
五、模型的建立与求解……（4）
……
5.1 问题 1 的模型……（4）
5.2 问题 2 的模型……（7）
5.2 问题 3 的分析与结论……（10）
5.2 问题 4 的模型……（11）
……
六、模型的评价及优化……（15）
6.1 问题 1 的 LDA 模型……（15）
6.2 问题 2 的 TF-IDF 模型……（16）
6.3 问题 3 的 R 型聚类模型……（16）
参考文献……（17）
附录……（18）

一、问题重述

如今多媒体发展迅速，信息的传播渠道多种多样，而在公共危机事件爆发时，如拍石击水，相关信息在短时间内迅速传播，引起群众的广泛关注其中负面报道或者主观片面的一些失实评判常常在一定程度上激发人们普遍的危机感，甚至影响到政府及公共单位的公信力，影响到企业的形象及口碑。如果不及时采取正确的措施分析和应对，将对相关部门或者企业造成难以估计的后果。所以关注相关舆情对政府或者企业来说非常重要。同时，通过舆情的情感倾向预测，有助于企业能够了解媒体或网民对相关事件或者品牌的舆情情感倾向分布和情感倾向趋势，同时能快速识别负面情感倾向的文章或评论，及时对口碑进行维护。

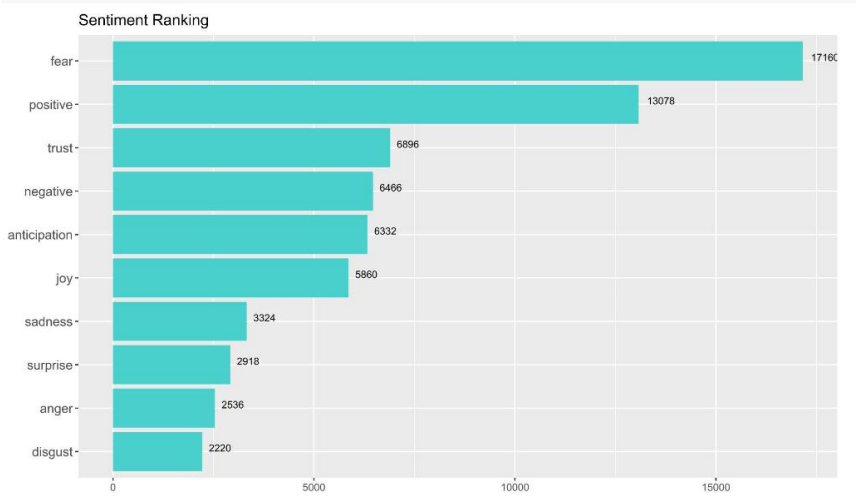

图 1-1 文本挖掘情感分析数据统计

问题 1 的研究意义在于，通过技术手段抓取的部分媒体或网民评论的数据，提供一个针对某一主题的舆情筛选方法，对于某一主题的舆情筛选在生活当中对于创业公司极其重要。舆情能够结合关键信息，第一时间获取预订的与企业产品密切相关的价值信息，针对企业的情况个性化定制数据源并对其进行系统的分类处理，筛选最具价值的信息进行分析总结[1]，是对于企业来说具有参考价值的。

问题 2 的研究意义在于，提供一个全新数据的抓取方法，其中尽量包含诸如发表时间、评论人数、关注人数及具体内容等具有深层次分析价值的数据，对于企业来说能够准确抓住情感倾向方向。数据挖掘的技术水平和获取信息的能力是各大舆情服务平台的核心竞争力，而信息源就尤为关键，信息源的广度与深度直接决定了数据的质量，抓取具有深层次分析价值的数据更是能提供更加准确的信息，有利于能快速识别负面情感倾向的文章或评论，及时对口碑进行维护。

问题 3 的研究意义在于，合理引导网民们情感倾向逐步转向对政府或企业有利的干预方法，有利于阻止另外一种形式继续传播舆情，及时制止传播。当网络中公众从舆论传播的接受者转变为传播者和评论者，其身份不再局限于传播媒介，甚至在事件的舆论中同时扮演着多种身份于一体的角色[2]，因此需要合理、有效的干预策略。

问题 4 的研究意义在于，充分考虑疫情传播时间、规模及网民情感倾向的舆情处理等级的划分方法，对于政府或企业而言对处于不同阶段的舆情需要进行干预的等级不同，等级的划分能够更好地制定干预方法。

二、问题分析

2.1 问题 1 的分析

问题 1 探究的舆情筛选方法，属于分类的数学问题。为解决此类问题，我们进行了数据预处理，以及对特殊符号进行处理、对词的长度进行预览，然后选用前部分文本进行分类，也有助于使用自建的停用词库，并找到针对某一主题的舆情筛选方法。

由于以上原因，我们首先使用 jieba 进行分词，并且通过优化，让主题显得更加集中，随着文本选取的扩大化，主题会逐渐增多。我们通过 LDA 模型来对进行文本处理和分词后的文档进行分类，从而达到分类主题的目的。

2.2 问题 2 的分析

问题 2 探究的全新数据的抓取方法，其中尽量包含诸如发表时间、评论人数、关注人数及具体内容等具有深层次分析价值的数据，属于文本数据挖掘的数学问题。为解决此类问题，我们对于无监督模型中关键词的提取，使用 TF-IDF 模型：用于反映一个词对于某篇文档的重要性，过滤掉常见的词语，保留重要的词语

由于以上原因，我们首先对文章的常用词进行分析，其次考虑到文章或评论的长度，对于特定词的提取，使用序列标注模型中的词性标注，另外考虑到文本本身的复杂性，如上下文等，还需要通过类似于命名实体识别模型 NER，来辅助去获得更精确的结果，达成标注要求后，标注完的文本可以作为训练样本，再利用 bert 等框架对模型进行训练，完成训练后的模型即可对任意输入的语料标注出所需的特定内容。最后，考虑到如果文本较为复杂，有上下文对结果造成影响的和对于具体内容等如对人名，时间，地名等有特殊要求的，可以自建序列标注平台进行标注，训练模型，达成对文本进行分析的结果。

2.3 问题 3 的分析

问题三探究的提供一种能够合理引导网民们情感倾向逐步转向对政府或企业有利的干预方法，是寻找引导、干预的策略问题，基于问题 1 和问题 2 的分析，以及不同的舆情对不同的人群存在着不同的价值，期间不同的人员在舆情传播过程中起到了不同的作用的情况来看，面向不同的人群需要有不同的干预方法。

2.4 问题 4 的分析

问题四探究的对于政府或企业而言对处于不同阶段的舆情需要进行干预的等级不同，提供一个充分考虑疫情传播时间、规模及网民情感倾向的舆情处理等级的划分方法，是一个聚类分析的数学问题，在整理统计的相关情感倾向数据表格基础之上进行聚类分析，并结合现有的收集到的已经标注好的疫情传播时间和情感倾向数据，找到合适的舆情处理等级的划分方法。

三、模型假设

1.假设题目所给的数据真实可靠；
2.选用前部分文本进行分类也能保证文本传递的讯息的完整性；
3.优化缩小主题数，仍保留主题的类别；
4.过滤词语的时候，重要的词语能较多地被保留；
5.利用 bert 等框架对模型进行训练后的模型准确性较高；
6.自建序列标注平台进行标注能满足特殊要求。

四、定义与符号说明

符号定义	符号说明
LDA	生成式模型
XX	需要建模的数据
YY	标签信息
W	词
M	词构成的词汇表矩阵
K	主题数量
V	词汇表大小
D	文档集合
T	主题集合
wi	文档中第 i 个单词
VOC	文档集合中每一个单词的大集合
Topic	主题

表 4-1 符号说明

对每个 D 中的文档 d，对应到不同 Topic 的概率θ_d<P_{t1},...,P_{tk}>，其中，P_{ti}表示 d 对应 T 中第 i 个 topic 的概率。计算方法是直观的，$P_{ti}=n_{ti}/n$，其中n_{ti}表示 d 中对应第 i 个 topic 的词的数目，n 是 d 中所有词的总数。

对每个 T 中的 topic，生成不同单词的概率 Φt<P_{w1},...,P_{wn}>，其中，P_{wi}表示 t 生成 VOC 中第 i 个单词的概率。计算方法同样很直观，$P_{wi}=N_{wi}/N$，其中N_{wi}表示对应到 topic 的 VOC 中第 i 个单词的数目，N 表示所有对应到 topic 的单词总数。

LDA 模型核心公式：$p(w|d)=p(w|t)*p(t|d)$ ①

Topic 作为中间层，可以通过当前的θ_d和Ψ_t给出了文档 D 中出现单词 W 的概率。其中 p(t|d)利用θ_d计算得到，p(w|t)利用Ψ_t计算得到。

实际上，利用当前的θ_d和Ψ_t，我们可以为一个文档中的一个单词计算它对应任意一个 Topic 时的 p(w|d)，然后根据这些结果来更新这个词应该对应的 topic。然后，如果这个更新改变了这个单词所对应的 Topic，就会反过来影响θ_d和Ψ_t。

五、模型的建立与求解

数据的预处理：

	正文
0	\n 近日，哪吒汽车旗下哪吒N01的两款新车型上市，补贴后的售价分...
1	2020-05-18 11:32:51 特斯拉 / 汽车 / 新车\n\n出品 | 搜狐汽车...
2	在当下这个时代,都市的快节奏生活和日常的琐碎藩篱,不仅没有掩盖年轻人身上的锋芒与锐利,甚至带...
3	为了带动新兴产业投资，提高投资强度和经济密度，上海围绕产业园区建设将出台落地一系列重磅政策。...
4	在当下这个时代,都市的快节奏生活和日常的琐碎藩篱,不仅没有掩盖年轻人身上的锋芒与锐利,甚至带...

图 5-1 数据展示

对特殊符号进行处理：

```
#text = re.sub('[!！]+', "!", text)
#text = re.sub('[??]+', "?", text)
#text = re.sub('[~~]+', "~~", text)
text = re.sub('[!！]+', " ", text)
text = re.sub('[??]+', " ", text)
text = re.sub('[~~]+', " ", text)
text = re.sub('[～～]+', " ", text)
text = re.sub("[a-zA-Z#$%&\'()*+,-./:;: <=>@, 。★、…【】《》“” ‘’ [\\]^_`{|}~]+"," ", text)
return re.sub("\s+", " ", text)
```

图 5-2 特殊符号处理

对词的长度进行预览：

```
df['word_length'] = df['正文'].astype(str).apply(lambda x: len(seg.cut(x)))
```

```
np.percentile(df['word_length'].tolist(),99)
```

```
5729.9800000000105
```

图 5-3 词的长度预览

5.1 问题 1 的模型建立与求解

5.1.1 LDA 模型的建立

LDA 模式是生成式模型，在这里，假设需要建模的数据为，标签信息为 YY。

判别式模型：对 YY 的产生过程进行描述，对特征信息本身不建模。判别式模型有利于构建分类器或者回归分析生成式模型需要对 XX 和 YY 同时建模，更适合做无监督

学习分析。

生成式模型：描述一个联合概率分布 P(X,Y)P(X,Y) 的分解过程，这个分解过程是虚拟的过程，真实的数据不是这么产生的，但是任何一个数据的产生过程可以在数学上等价为一个联合概率分布。

LDA 是一种矩阵分解技术，在向量空间中，任何语料（文档的集合）可以表示为文档（Document - Term，DT）矩阵。下面矩阵表达了一个语料库的组成：

.	W_1	W_2	...	W_m
D_1	0	2	...	3
D_2	1	4	...	0
...	...	...	...	...
D_n	1	1	...	0

其中，N个文档 $D_1, D_2, \ldots, D_n$的组成语料库，M 个词 $W_1, W_2, \ldots, W_m$ 组成词汇表。矩阵中的值表示了词 W_j 在文档 D_i 中出现的频率，同时，LDA 将这个矩阵转换为两个低维度的矩阵，M_1 和 M_2。

.	W_1	W_2	...	W_m
ϕ_1	0	2	...	3
ϕ_2	1	4	...	0
...	...	...	...	...
ϕ_k	1	1	...	0

上面显示了 M_2 矩阵的情况，它是一个 $K * V$ 维的 topic - term矩阵，K 指主题的数量，V 指词汇表的大小。M_2 中每一行都是一个 ϕ 分布，也就是主题 ϕ_k 在 m 个词上的多项式分布情况，可以通过学习得到。

LDA 文档生成流程如下：

LDA 假设文档是由多个主题的混合来产生的，每个文档的生成过程如下：

从全局的泊松分布参数为 β 的分布中生成一个文档的长度 N；

从全局的狄利克雷参数为 alpha 的分布中生成一个当前文档的θ；

对当前文档长度 N 的每一个字都有：

1、从θ为参数的多项式分布生成一个主题的下标 Z_n

2、从θ和 Z 共同为参数的多项式分布中，产生一个字 W_n

这些主题基于词的概率分布来产生词，给定文档数据集，LDA 可以学习出，是哪些主题产生了这些文档。对于文档生成过程，则有，首先对于文档 N 中的每一个字，都先从文档矩阵 M_1 中的 θ_i 中产生一个下标，告诉要从主题矩 M_2 中的哪一行 Ψ_m 生成当前的字[2]。

5.1.2 LDA 模型的求解

在进行了数据预处理过后，发现文本内容过大，因此选用前部分文本进行分类，有助于使用自建的停用词库：

```
texts=['近日，哪吒汽车旗下哪吒N01的两款新车型上市，补贴后的售价分别为8.96万元与9.96万元。在外观与内饰方面，两款新车均与现款车型保持了一致，仅在
'2020-05-18 11:32:51 特斯拉 / 汽车 / 新车'
'日前，我们从特斯拉官方了解到，由于美国加州Fremont工厂已逐步恢复生产，而产能、供应链恢复巅峰情况仍需时间，公司同步调整了Model 3与Model Y的交付

'根据官方信息显示，Model 3的交付时间被调整为4-7周、Model Y的交付时间则调整为8-12周。消化存量订单成为Fremont工厂目前的主要任务.'

'相比之下，中国特斯拉工厂 Model 3(参数|图片)交付时间为2-4周、 Model Y(参数|图片)的交付时间则预计为2021年，交付能力表现较好。随着中国疫情被控制

'回顾：受疫情影响，特斯拉的新车交付、财务情况均收到巨大影响。此前特斯拉CEO马斯克曾“威胁”加州政府，要求允许工厂立刻复工，否则企业将离开加州。

'为了带动新兴产业投资，提高投资强度和经济密度，上海围绕产业园区建设将出台落地一系列重磅政策。'

'5月15日，上海市政府新闻办发布了《关于加快特色产业园区建设 促进产业投资的若干政策措施》（以下简称《若干政策措施》）。'

'''文字丨「Bob鲍勃」图片丨「来自网络」今天是“国际不再恐同日”，普通人可能很少听说过这个节日，但全球许多国家的同性恋者们会在今天自豪地高举彩
"[车友头条-车友号-ams车评网] 不管您是图新鲜，还是因为摇不上号，或者由于其他原因，反正买电动车的人是越来越多，并且电动车的样式也越来越丰富了，

不知道您是否发现了，就是在介绍电动车的续航里程时，总会出现“NEDC”这四个字母，都说的是“NEDC续航里程”为XX公里。或许大多数朋友并不关心它是什么

NEDC全名叫做“New European Driving Cycle” ，翻译成中文就是“新欧洲驾驶周期”，大家也可以称它为“新标欧洲循环测试”。从名字也可以看出，NEDC是

有些朋友会说了，这个NEDC根本就不准，实际续航跟标称数字相差挺多的。其实这就要说到NEDC的测试方法了，它主要是包含4个市区循环和1个郊区循环。不过，

这下您就明白了吧，NEDC都是在台架上测出来的，肯定与实际道路驾驶相差较大，所以参考价值也就不那么高了。那么有没有其他测试标准呢？当然有了，比如WL

产品规格，货运公司，物流专线直达 回程车
```

图 5-4 使用的部分文本

加载停用词库如下：

```
flags = ('n', 'nr', 'ns', 'nt', 'eng', 'v', 'd')  # 词性
stopwords = ('根据', '图片','Moment','是','不','也')  # 停词
```

图 5-5 停用词库

使用 jieba 进行分词，对分词后的文本使用 LDA 主题模型如下：

```
(0, '0.007*"车型" + 0.007*"汽车" + 0.006*"客户" + 0.006*"产业" + 0.006*"共享"')
(1, '0.006*"车型" + 0.005*"NVIDIA" + 0.004*"汽车" + 0.004*"客户" + 0.004*"产业"')
(2, '0.007*"车型" + 0.005*"客户" + 0.005*"汽车" + 0.005*"NVIDIA" + 0.005*"都"')
(3, '0.007*"车型" + 0.005*"汽车" + 0.005*"AI" + 0.004*"产业" + 0.004*"NVIDIA"')
(4, '0.008*"汽车" + 0.008*"车型" + 0.007*"AI" + 0.006*"客户" + 0.006*"有"')
(5, '0.008*"汽车" + 0.007*"车型" + 0.006*"项目" + 0.005*"共享" + 0.005*"产业"')
(6, '0.007*"汽车" + 0.006*"NVIDIA" + 0.005*"产业" + 0.005*"AI" + 0.005*"车型"')
(7, '0.005*"车型" + 0.004*"产业" + 0.004*"汽车" + 0.004*"NVIDIA" + 0.004*"客户"')
(8, '0.006*"车型" + 0.005*"汽车" + 0.004*"客户" + 0.004*"发布" + 0.004*"AI"')
(9, '0.007*"车型" + 0.005*"客户" + 0.004*"NVIDIA" + 0.004*"汽车" + 0.004*"项目"')
(array([[1.00017861e-01, 1.00008182e-01, 1.00011684e-01, 1.00007705e-01,
         2.62177734e+03, 1.00024208e-01, 1.00014105e-01, 1.00003801e-01,
         1.00005589e-01, 1.00006223e-01]], dtype=float32), None)
```

图 5-6 LDA 模型应用后的 10 个主题

由于选取文本的原因，大部分主题都与车有关，但是可以大致分为‘AI 汽车’‘共享汽车’‘NVIDIA 汽车项目’等，主题但是过于重复，因此优化缩小主题数，可以使主题更加明确：

```
(0, '0.007*"车型" + 0.006*"汽车" + 0.006*"产业" + 0.006*"项目" + 0.005*"客户"')
(1, '0.006*"汽车" + 0.006*"NVIDIA" + 0.005*"有" + 0.005*"车型" + 0.005*"产业"')
(2, '0.008*"车型" + 0.007*"汽车" + 0.005*"客户" + 0.005*"AI" + 0.005*"发布"')
(array([[2.5367661e+03, 3.9890265e-01, 8.7140007e+01]], dtype=float32), None)
```

图 5-7 优化后的主题

5.1.3 结果

通过优化，主题显得更加集中，主要围绕‘汽车车型’，‘NVIDIA 汽车’‘AI 汽车’三个主题，随着文本选取的扩大化，主题会逐渐增多。因此通过 LDA 模型来对进行文本处理和分词后的文档进行分类，从而达到分类主题的目的。

5.2 问题 2 的模型建立与求解

5.2.1 TF-IDF 模型的建立

首先对于无监督模型中关键词的提取可以使用 TF-IDF 模型：用于反映一个词对于某篇文档的重要性，过滤掉常见的词语，保留重要的词语。

如果某个词在一篇文档中出现的频率高，则 TF 高；并且在其他文档中很少出现，则 IDF 高，TF-IDF 就是将二者相乘为 TF*IDF，这样这个词具有很好的类别区分能力。考虑到点赞数，评论数每篇文章或评论都会有，可能会具有较高的词频，因此先对文章的常用词进行分析：

```
freq = pandas.Series(' '.join(df['正文']).split()).value_counts()[:20]
freq
```

```
the                                      23535
https://yqms3.zhxgimg.com/download/img/  18389
0                                        16569
to                                       14393
in                                       12968
and                                      12721
of                                       12569
1                                        10754
,                                        10732
a                                        10532
5                                        10171
3                                         9429
-                                         8851
2                                         7958
6                                         7918
|                                         6533
。                                        6464
der                                       5817
4                                         5610
8                                         5597
dtype: int64
```

图 5-8 常用词分析

发现大部分常用词均为英文单词或数字，再考虑到文章或评论的长度：

```
count    100668.000000
mean         39.992401
std        1585.903677
min           1.000000
25%           1.000000
50%           3.000000
75%           9.000000
max      440227.000000
Name: word_count, dtype: float64
```

图 5-9 字数分析

发现平均字数为近 40 词，最长的一篇为将近 40 万词，最短的仅 1 个词，文本的内容相差较大，因此对于词频即 TF-IDF 模型的分析，并不能够有很好的效果。

5.2.2 词性标注

对于特定词的提取，如发表时间，评论人数，具体内容等，可以使用序列标注模型中的词性标注，即根据词性对词进行标注，如发表时间如 2020-3 即为 2020（数词/m）3（数词/m），只需要提取特定的词性组合，就可以识别出时间，比如 2020-3 的词性组合为 m，m。对于微博等发表具有特定的时间标注，可以较好的识别出来:

```
import jieba.posseg as pseg
words =pseg.cut("2020-05-18 11:32:51 特斯拉新汽车")
for w in words:
    print(w.word,w.flag)
```

```
2020 m
- x
05 m
- x
18 m
  x
11 m
: x
32 m
: x
51 m
  x
特斯拉 nrt
新 a
汽车 n
```

图 5-10 基于原文本中可能是发表时间的评论经行的词性分析

但是由于时间格式具有普遍性，如果没有特定的指明，经常会提取出词性均相同但并不是评论时间。如结果中展示的特斯拉发布的时间就可能与微博发布时间的格式相同。因此简单的词性标注有可能在在实际中表现不好。

5.2.3 正则表达式

正则表达式(regular expression)描述了一种字符串匹配的模式（pattern），可以用来检查一个串是否含有某种子串、将匹配的子串替换或者从某个串中取出符合某个条件的子串等。

如果仅仅通过特定词，如点赞人数，评论人数等，可以通过正则表达式来得到相应的结果，如点赞人数为 点赞人数+数字的形似，可以表现为 点赞人数+\d*\.?\d 的形式来提取具有特定字符串的信息。但是如果想要提取特定的文本内容等其他更深层次的信息需要对文章文本进行分析。

5.2.4 建立序列标注平台

考虑到文本本身的复杂性，如上下文等，还需要通过类似于命名实体识别模型 NER，来辅助去获得更精确的结果，但是 NER 模型的识别实体类型为人名、地名、组织机构名，但是我们往往也会有识别其它实体的需求，比如时间、点赞人数，具体内容等。如果是对所需内容进行识别的话，可以首先自建序列分类平台，自建的序列标注平台需要大量的人工对特定所需内容进行标注。即输入特定的文本，会对文本进行标注，其中特定的文本会被用 BIO 规范标注。

8	B-TIME
月	I-TIME
8	I-TIME
日	I-TIME
是	O
“	O
全	O
民	O
健	O
身	O
日	O
”	O
,	O

图 5-11 序列标注平台应达到的效果图

达成标注要求后，标注完的文本可以作为训练样本，再利用 bert 等框架对模型进行训练。完成训练后的模型即可对任意输入的语料标注出所需的特定内容。但是这种方法所需时间太大，需要对大量数据进行标注训练才能达到较好的效果。

5.2.5 结果

综上所述，如果对于有特定格式的，如发表日期，评论人数，可以使用词性标注或正则表达式进行搜索，如果文本较为复杂，有上下文对结果造成影响的和对于具体内容等如对人名，时间，地名等有特殊要求的，可以自建序列标注平台进行标注，训练模型，达成对文本进行分析的结果。

5.3 问题 3 的分析与结论

5.3.1 问题分析

问题三在于提供一种能够合理引导网民们情感倾向逐步转向对政府或企业有利的干预方法，是寻找引导、干预的策略问题，基于问题 1 和问题 2 的分析，以及不同的舆情对不同的人群存在着不同的价值，期间不同的人员在舆情传播过程中起到了不同的作用的情况来看，面向不同的人群需要有不同的干预方法。

5.3.2 结论

企业在舆情监测发现任何问题时需要尽快开展舆论的引导，通过各种方式或策略积极开展公众之间的沟通，再着手实施其他应对或处理策略，并以避免和解决危机为目标持续完善危机管理预案中舆情引导部分的内容[1]。

网络舆论情况，错综复杂，尤其是对于那些知名企业或者是知名品牌而言，当他们发生危机事件的时候，舆情的复杂程度将会进一步加剧。网络舆情情况难以捉摸，因为所有人都不知道群众的关注点会从哪一个角度看待问题，从源头把握有影响力的观点，是企业引导舆论情况的基本理念。

因此，在网络环境下，有这么一类人群他们所提出的观点会更容易成为影响力舆论点，他们对于舆论发展情况，更具备决定性导向力。

1、网络红人

网红是个特殊群体，他们是这个快速发展网络环境下的产物，他们对于自己粉丝情绪的引导力和观点的认可率，都是非常高的。

3、专业领域的专家

他们本身就代表着权威。他们用自己的专业知识和傲人的头衔，对广大吃瓜群众造成“成吨”的冲击力。

4、行业内公认具备影响力的人物

我国市场内各行各业，都存在一些领军人物，他们对于国内某行业的发展，起到了推动性的历史作用。因此，他们所说的一些观点或言论，更容易得到群众认可。

通过上述四类关键人群的列举和分析，我们不难他们都拥有的特性都会存在差异。但只要某人具备这些特性的一种或多种，都有成为“引导性”人群的潜质[3]，因此要对以上四类可能成为引导性的人物进行沟通向导，才能把控舆情的方向。

5.4 问题 4 的模型建立与求解

5.4.1 R 型聚类分析模型的建立

模型建立的数据的获取:

第四问要综合考虑网民的舆论倾向，规模和时间来划分处理等级。因此引用了 datafunation 疫情期间网民情绪识别竞赛的数据集。通过分析处理数据中规模，时间，情感倾向，进行聚类，再将其方法运用到本数据集中来达到实际要求:

	微博id	微博发布时间	发布人账号	微博中文内容	微博图片	微博视频	情感倾向
0	4456072029125500	01月01日 23:50	存曦1988	写在年末冬初孩子流感的第五天，我们仍然没有忘记热情拥抱这2020年的第一天。带着一丝迷信，早...	['https://ww2.sinaimg.cn/orj360/005VnA1zly1gah...	[]	0
1	4456074167480980	01月01日 23:58	LunaKrys	开年大模型...累到以为自己发烧了腰疼膝盖疼腿疼胳膊疼脖子疼#Luna的Krystallife#?	[]	[]	-1
2	4456054253264520	01月01日 22:39	小王爷学辩论o_O	﹏傣癯空笃褪俏业﹏，爹，发烧快好，毕竟美好的假期拿来养病不太好，假期还是要好好享受快乐，爹，...	['https://ww2.sinaimg.cn/thumb150/006ymYXKgy1g...	[]	1
3	4456061509126470	01月01日 23:08	岑鎏	新年的第一天感冒又发烧的也太衰了但是我要想着明天一定会好的?	['https://ww2.sinaimg.cn/orj360/005FL9LZgy1gah...	[]	1
4	4455979322528190	01月01日 17:42	changlwj	问：我们意念里有坏的想法了，天神就会给记下来，那如果有好的想法也会被记下来吗？答：那当然了。...	[]	[]	1

图 5-12 数据集预览

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 7 columns):
微博id       100000 non-null int64
微博发布时间     100000 non-null object
发布人账号      100000 non-null object
微博中文内容     99646 non-null object
微博图片       100000 non-null object
微博视频       100000 non-null object
情感倾向       99919 non-null object
dtypes: int64(1), object(6)
memory usage: 75.0 MB
```

图 5-13 数据集基本信息

通过分析数据集，发现所需的规模即发布数量，发表时间，情感倾向均能得到。因此对数据集进行进一步分析。

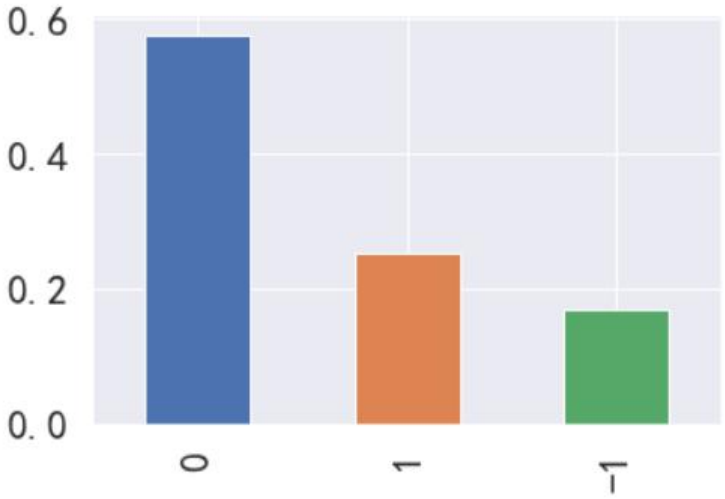

图 5-14 情感倾向的分布图

发现本数据集的情感倾向分为三类，0 为中立态度，1 代表正面态度，-1 代表负面态度，情感标签的分布较为符合实际。因此对发布数量（规模），发布时间，情感态度进行整合。

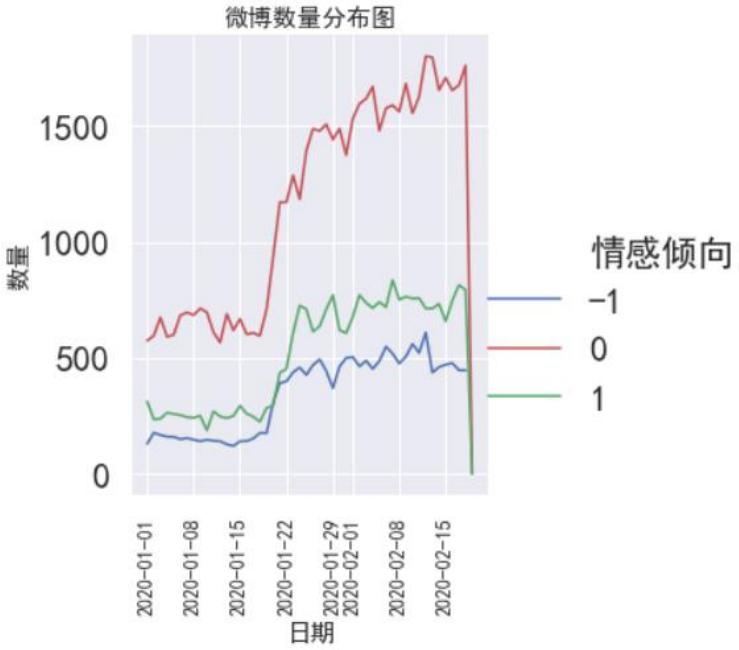

图 5-15 基于微博数量，时间，情感分析的概览

因此，根据疫情传播时间从 2020 年 1 月 1 日起到 2020 年 2 月 15 日为止分别对正面、中立、负面的情感倾向下的微博数量进行聚类分析。

5.4.2 R 型聚类分析模型的求解

聚类分析的结果即反应对于等级的划分：

（1）对情感倾向为正面的微博数量进行分析，结果如下：

```
第1类的有1  2  3  4  5  6  7  8  9  10  11  12  13  14  15  16  17  18  19  20  21  22
第2类的有23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46
```

图 5-16 聚类结果

聚类树结果如下：

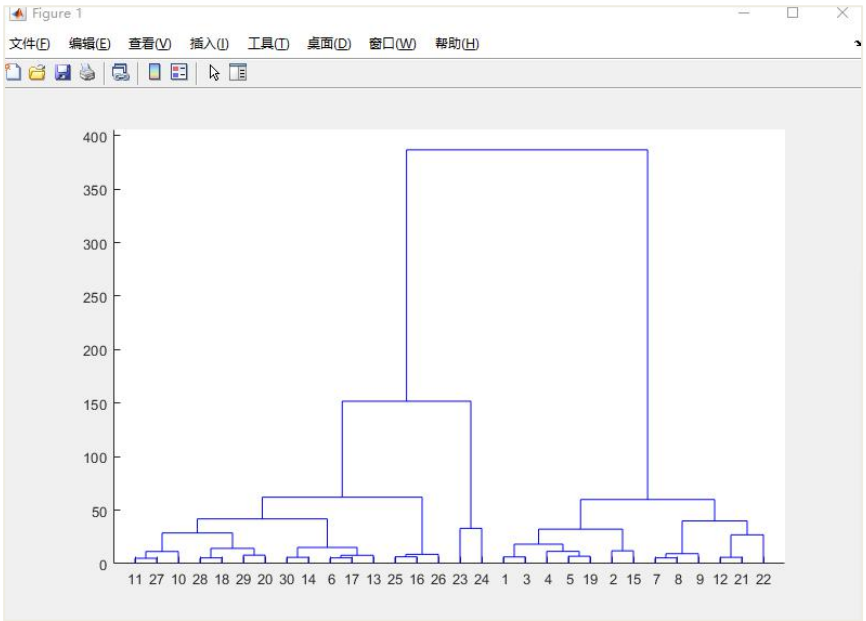

图 5-17 聚类树结果

由图可知，可划分为两大类。划分为第一类的为从 2020 年 1 月 1 日到 2020 年 1 月 22 日为止;第二类为从 2020 年 1 月 23 日到 2020 年 2 月 15 日为止。

（2）对情感倾向为中立的微博数量进行分析，结果如下：

```
第1类的有1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
第2类的有21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46
```

图 5-18 聚类结果

聚类树结果如下：

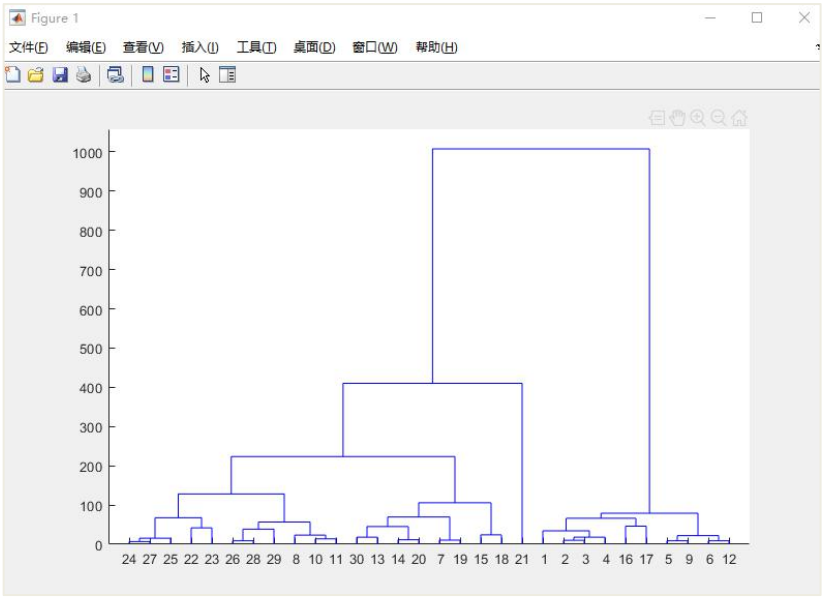

图 5-19 聚类树结果

由图可知，可划分为两大类。划分为第一类的为从 2020 年 1 月 1 日到 2020 年 1 月 20 日为止;第二类为从 2020 年 1 月 21 日到 2020 年 2 月 15 日为止。

（3）对情感倾向为负面的微博数量进行分析，结果如下：

```
第1类的有1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19
第2类的有20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41  42  43  44  45  46
```

图 5-20 聚类结果

聚类树结果如下：

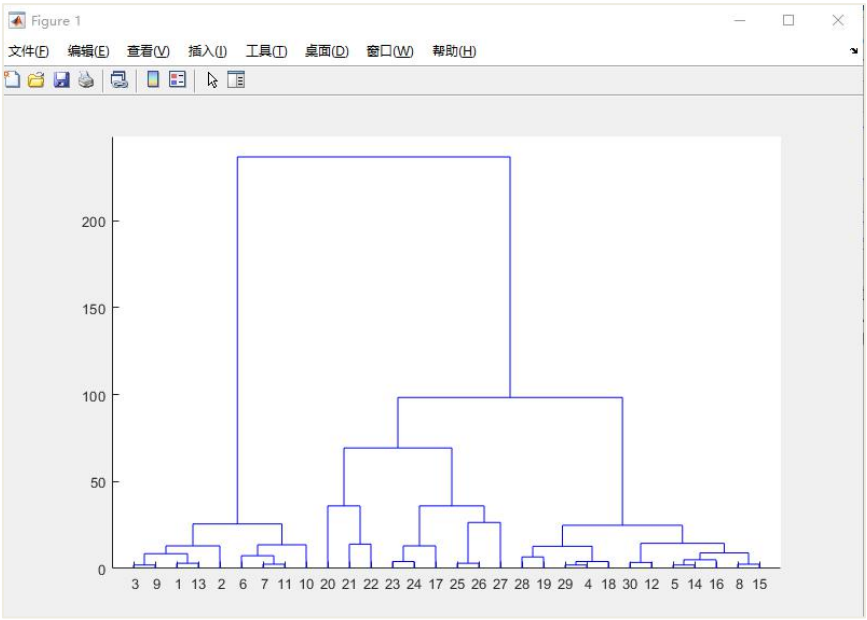

图 5-21 聚类树结果

5.4.3 结果

由求解过程可知，可划分为两大类。划分为第一类的为从 2020 年 1 月 1 日到 2020 年 1 月 19 日为止;第二类为从 2020 年 1 月 20 日到 2020 年 2 月 15 日为止。据此，我们将划分为两类，可以发现随着疫情的传播，舆论的规模和情感分析是随之变化的，因此对于 1 月 19 号之前，舆论的管控等级可以为低级，进入 2 号，随着发布规模的剧增，网民的不满情绪也是激增，因此需要加大管控力度，可以划为高级的层次。如果迁移运用到题目所给的数据集，也需要提前提取出日期等关键词，并以日期统计发布数量，并预先对文本给出标签，之后再通过聚类的方法，得到分类结果，由此调整监控等级。

六、模型的评价及优化

6.1 问题一的 LDA 模型

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。运用 LDA 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。过程如下:

1.对每一篇文档，从主题分布中抽取一个主题;

2.从上述被抽到的主题所对应的单词分布中抽取一个单词;

3.重复上述过程直至遍历文档中的每一个单词;

其优点为: 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。能够较好的通过分析关键词词

频的方式来得到主题，通过停用词库能很好的降低一些没有特征的词的权重，因此关键词在一篇评论中出现越多，词频越高，越能反映出主题。

其缺点为: 词袋方法没有考虑词与词之间的顺序，难于区分同类型的评论，即这些词在同类型的评论重复出现，但并不能通过停用词库解决，因此会降低部分与主题相关词的频率，造成部分主题相似但无法区分。

6.2 问题二的 TF-IDF 模型

TF-IDF 是一种统计方法,用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降。TF-IDF 加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。除了 TF-IDF 以外，因特网上的搜索引擎还会使用基于链接分析的评级方法，以确定文件在搜寻结果中出现的顺序。

TF-IDF 的优点为: 实现简单，相对容易理解。可对频率进行分析,认为文本频率小的单词就越重要，文本频率大的单词就越无用。通过四种模型分别分析每种模型在提取关键词的优劣，并给出每种模型的适用方法，挖掘更深层次的信息。

TF-IDF 的缺点为: 严重依赖语料库,需要选取质量较高且和所处理文本相符的语料库进行训练, 不能反应词的位置信息，在对关键词进行提取的时候，词的位置信息，例如文本的标题、文本的首句和尾句等含有较重要的信息，应该赋予较高的权重。

6.3 问题四的 R 型聚类模型

在实际工作中，为了避免漏掉某些重要因素，往往在一开始选取指标的时候尽可能考虑所有的相关因素，而这样做的结果，则是变量过多，变量间的相关度较高，给统计分析与建模带来极大不便，因此人们希望能够研究变量间的相似关系，按照变量的相似关系把他们聚合成若干类，进而找出影响系统的主要因素，引入了 R 型聚类方法。

其优点为：既可以了解个别变量的模型关系，还可以了解各个变量的模型关系。聚类分析模型通过外部数据进行划分，较为符合实际。

其缺点为：针对于过多数据的分类标准，结果无法直观的得到细致的分类，只能得到粗略分类，从而聚类树和聚类结果无法统一。聚类分析模型受制于数据集的局限性，无法运用 25 号以后的数据，因此只能针对前期的舆论严重性进行划分。

参考文献

[1] 林萍，黄卫东. 基于 LDA 模型的网络舆情事件话题演化分析[J]. 情报杂志. 2013. 23(12). 26-30

[2] 陈晓美，高铖，关心惠. 网络舆情观点提取的 LDA 主题模型方法[J]. 图书情报工作. 2015. 59(21). 21-26

[3] 方雪琴. 信息公开与媒体理性——试论危机传播中的舆论引导策略[J]. 中州学刊. 2004. 12(6). 183-185

[4] 胡欣杰，路雨楠，路川. 基于聚类分析的网络舆情倾向性分析研究[J]. 兵器装备工程学报. 2019. 12(5). 115-118

[5] 梁杰，丁嘉瑞，禹常隆. Python 语言及其应用[M]. 北京. 人民邮电出版社. 2015. 122-128

[6] 龚纯，王正林. MATLAB 语言常用算法程序集[M]. 北京. 电子工业出版社. 2008. 135-145

附录

（一）问题 1 的 Python 源代码

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('seaborn')
sns.set(font_scale=2)

plt.rcParams['font.sans-serif']=['SimHei'] #用来正常显示中文标签
plt.rcParams['axes.unicode_minus']=False #用来正常显示负号
f = open(r'G:\附件 1.csv',encoding='utf-8"')
df= pd.read_csv(f,delimiter='\t',sep='\t',error_bad_lines=False)

df.head()
df['char_length'] = df['正文'].astype(str).apply(len)
np.percentile(df['char_length'].tolist(),75)

from gensim import corpora, models
import jieba.posseg as jp, jieba
texts=['部分数据
。'
flags = ('n', 'nr', 'ns', 'nt', 'eng', 'v', 'd')  # 词性
stopwords = ('根据', '图片','Moment','是','不','也','到')  # 停词
# 分词
words_ls = []
for text in texts:
    words = [w.word for w in jp.cut(text) if w.flag in flags and w.word not in
stopwords]
    words_ls.append(words)
# 构造词典
dictionary = corpora.Dictionary(words_ls)
# 基于词典，使【词】→【稀疏向量】，并将向量放入列表，形成【稀疏向量集】
corpus = [dictionary.doc2bow(words) for words in words_ls]
# lda 模型，num_topics 设置主题的个数
lda    =    models.ldamodel.LdaModel(corpus=corpus,    id2word=dictionary,
num_topics=3)
# 打印所有主题，每个主题显示 5 个词
for topic in lda.print_topics(num_words=5):
    print(topic)
# 主题推断‘
print(lda.inference(corpus))
```

（二）问题 2 的 Python 源代码

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
plt.style.use('seaborn')
sns.set(font_scale=2)

plt.rcParams['font.sans-serif']=['SimHei'] #用来正常显示中文标签
plt.rcParams['axes.unicode_minus']=False #用来正常显示负号
f = open(r'G:\附件 1.csv',encoding='utf-8"')
df= pd.read_csv(f,delimiter='\t',sep='\t',error_bad_lines=False)
df.head()

import jieba

import jieba.analyse

import jieba.posseg

def dosegment_all(sentence):

    ,,,

    带词性标注，对句子进行分词，不排除停词等

    :param sentence:输入字符

    :return:

    ,,,
    sentence_seged = jieba.posseg.cut(sentence.strip())
    outstr = ''
    for x in sentence_seged:
        outstr+="{}/{},".format(x.word,x.flag)
        #上面的 for 循环可以用 python 递推式构造生成器完成
        #   outstr   =   ",".join([("%s/%s"   %(x.word,x.flag))   for   x   in
sentence_seged])
        return outstr

import jieba.posseg as pseg
words =pseg.cut("2020-05-18 11:32:51 特斯拉新汽车")
```

```
for w in words:
print(w.word,w.flag)

import pandas
#平均字数为 39 个词，最少的仅有一个词，做多的一篇有 40 万词
#预览常见词
freq = pandas.Series(' '.join(df['正文']).split()).value_counts()[:20]
freq
df.word_count.describe()
```

（三）问题 3 的 matlab 聚类分析子程序

```
a=[x];   %日期矩阵
d=1-abs(a); % 进行数据变换,将相关系数转换为距离
y=linkage(d,'average'); % 按类平均法聚类
j=dendrogram(y); % 画聚类图
L=cluster(y,'maxclust',2) % 把样本点分为 2 类
for i=1:2
    b=find(L==i); % 求第 i 类的对象
    b=reshape(b,1,length(b)); % 变成行向量
    fprintf('第%d 类的有%s\n',i,int2str(b)); % 显示分类结果
end
```