基于网络问政留言的数据挖掘分析

摘要

本题目根据互联网公开来源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进行数据挖掘。利用自然语言处理和文本挖掘的方法对群众留言进行分类，查找热点问题并做出评价。

问题一：本题对于群众留言内容进行文本挖掘，首先利用 jieba 分词对留言文本进行数据预处理，依据信息增益方法筛选特征词，基于伯努利朴素贝叶斯、多项式朴素贝叶斯和逻辑回归算法建立三种不同的模型，最初拟定 4000 个特征词作为训练样本进行模型训练，根据特征量确定特征在七类一级标签中的所属类别。然后对测试样本数据进行模型测试，利用选择十折交叉验证的方法对验证样本进行验证，利用混淆矩阵分析模型分类结果。验证后用 F-Score 对分类结果进行初步评价，选择一种模型并进行优化。最后得出，依据伯努利朴素贝叶斯算法建立的模型得到的结果最优，特征量选择再 5100 词时，最好评价结果F_1为 0.88。

问题二：此问根据留言内容中的关键词、留言时间、点赞数等相关信息，处理留言主题中的文本进行热点问题挖掘。对留言文本简单数据预处理，利用信息增益算法选择词频在 4~500 之间的特征词进行 DBSCAN 粗聚类。首先，通过计算两个特征向量之间的余弦相似度计算文本相似度，并计算每个点的领域和密度，粗聚类结果的半径为 0.7，每个类别中最少个数为 4。接着对初步聚类簇进行再次聚类，每个类别中最少个数为 3。当每个类别中只存在一个特征数据时，结束聚类。根据聚类结果进行建立热度评价指标计算热度指数，得到热度指数最高的五个问题，热度最高的问题热度指数为 21.30。

问题三：此问题要从相关性、完整性的角度对答复意见进行评价，答复意见和留言详情两个文档的相似度越高，其相关性越强，完整度也越高。具体可以用到 doc2bow 方法、TF-IDF 模型和 token2id 方法来计算文本的相似度。

最后，在模型建立过程中用到了不同模型对比，以及多次改进模型，进行了模型的总结并分析了模型的优缺点，使用的主要软件是 python。

关键词：信息增益，伯努利贝叶斯，十折交叉验证，基于密度的聚类算法

目录

一、挖掘背景和挖掘目标

（一）挖掘背景

近年来，我国的城镇化建设取得了巨大成功，越来越多的人口向城镇流动。专家预测，在2050年我国城镇化率将达到75%，将有112500万人口居住在城镇。这将给城市发展带来很多新的困难和挑战。随着物联网、云计算等电子通信技术的高速发展，建设智慧城市，为国家解决城市发展的难题已成为一种新的发展趋向。而电子政务是是智慧城市发展的重要组成部分，是解决民生问题的重要保障。

在信息化时代，政府提供公共服务、促进经济社会发展的职能需要大数据的支持。现如今，很多民众通过在各种网络问政平台留言来反映生活中遇到的问题。在群众和政府进行互动时，会沉淀大量的数据文本，这些数据与社会经济、民众生活息息相关，具有数量庞大、涉及面广、动态实时、可用性强等特点[1]。也正是因为这些特点，面对海量的数据，人工整理分类的效率十分有限。

利用大数据技术进行数据挖掘和分析，将能够使政府提高治理效能，作出科学决策。凭借大数据，城乡建设、环境保护、交通运输、教育问题等方面均能得到改善，政府部门服务水平得以加强，基于大数据分析得到的政府决策能够更好地满足群众的需求。同时，若对数据加以运用和共享，可促进各级政府各部门之间的分工合作，减少重复的数据采集，降低采集成本。因此，促进大数据和电子政务的融合，大力推动智慧政务系统创新型发展显得尤为重要。

（一）　挖掘目标

本次建模的主要目标是按照一定的划分体系对网络问政平台的群众留言进行分类处理，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，挖掘出热点问题，以及制订评价方案对相关部门的答复进行评价，具体如下：

（1）本题主要目标是进行群众留言分类，利用信息增益的方法对附件 2 总体数据进行分词处理，选取出特征词样本。比较朴素贝叶斯中伯努利贝叶斯、多项式贝叶斯以及逻辑回归算法模型，运用模型样本训练模型，并对模型分类结果进行评价，并找出最优模型。

（2）本体主要目标是热点问题挖掘，从附件 3 留言主题中选取词频较高的特征词，重复利用 DBSCAN 算法进行聚类，初步找到热点问题。构造热度评价指标，对热点问题进行热度评价，得到热度指数排名高的前五个问题。

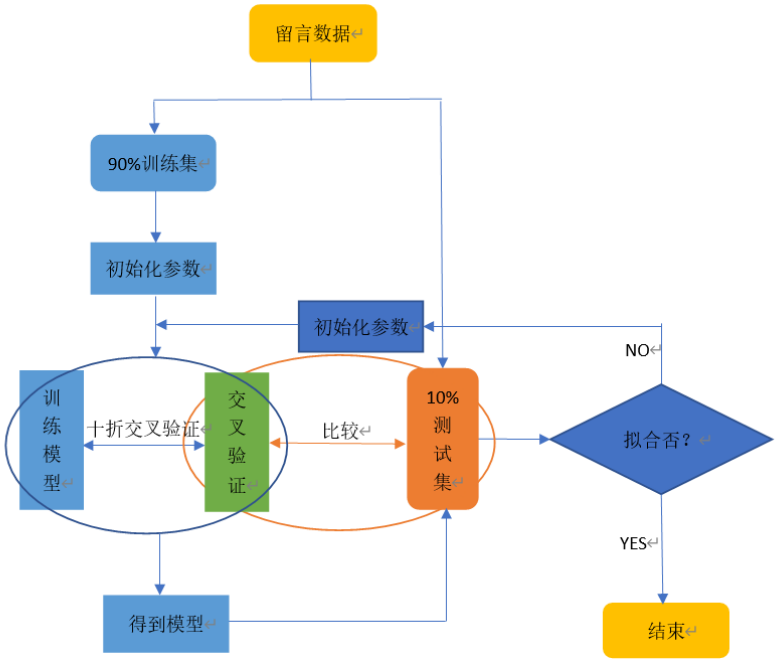
（3）针对附件 4 相关部门对留言的答复意见，从答复的相关性、完整性、可解释性等角度对答复意见的质量给出一套评价方案。

二、问题一

（一）问题分析

根据所给出的附件二数据，要建立模型对留言内容的一级标签进行分类。首先要对留言内容进行自然语言处理，将留言详情的文档内容分割为句子，句子分割为单词，并标注单词词性。拆分留言详情内容后，在拆分出的特征词中删去低频特征词，利用信息增益算法计算剩余特征词熵值，进行排序，在留下的高频特征词中选取一部分熵值高的特征词作为训练样本进行模型训练和模型测验和模型验证。

基于伯努利朴素贝叶斯、多项式朴素贝叶斯和逻辑回归算法建立三种不同的模型，利用选取训练样本数据中一部分进行模型训练。其中，模型中的七类一级标签为：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。根据特征量确定其所属类别，另外测试样本数据作为模型测试数据，模型建立以后，利用验证样本数据进行交叉验证，利用混淆矩阵分析模型分类结果。验证后用 F-Score 对模型进行初步的评价，选择最优的模型作为该题模型，在评价结果的基础上对模型进行改进。思路流程图如下：



图表 1：问题一分析流程图

（二）数据预处理

（1）文本分词

首先对民众留言内容进行分类，用 python 脚本读取群众留言文件并对读取的留言详情进行数据预处理。将原有文本进行处理，删去文本中无用的数据，例如空格、标点符号、数字等。第一步去除文本中无用的符号，让文本只保留文字部分；之后对文本进行 jieba 分词操作，将句子分割为若干单词，并标注词性；在此基础上设置了停用词表，去除掉指定停用词，以此提炼出与分类模型相关的特征词。

（2）文本特征处理

对留言内容进行分词之后，利用自然语言处理中的 TF 策略统计每个词在文本中出现的次数，将文本向量化。将词频小于 4 的词语删去后得到 50000 左右的特征词，接下来用信息增益的方法对这些特征进行选择。

信息增益是一种比较有效的特征选择方法，要在对特征的重要程度量化之后再进行选择。在信息增益中，重要性的衡量标准就是看特征词能够为分类系统带来多少信息，带来的信息越多，该特征词越重要。

信息增益对于一个特征（例如“学生”,用 t 表示）来说，就是系统中有 t 时的信息量和系统中没有 t 时的信息量的差值，也是 t 为系统带来的信息量，即增益。

当系统中存在 t 时，熵的表达式

$$H(C) = -\sum_{i=1}^{n} P(C_i)\log_2 P(C_i)\text{。}$$

其中，C 为类别，n 为类别总数，赛题中 n=7

当系统中不存在 t 时，熵的表达式

$$H(C|T) = P(t)H(C|t) + p(\bar{t})H(C|\bar{t})\text{。}$$

其中

$$\begin{cases} H(C|t) = -\sum_{i=1}^{n} P(C_i|t)\log_2 P(C_i|t) \\ H(C|\bar{t}) = -\sum_{i=1}^{n} P(C_i|\bar{t})\log_2 P(C_i|\bar{t}) \end{cases}\text{。}$$

则信息增益

$$IG(T) = H(C) - H(C|T)。$$

利用信息增益对特征词进行选择后，约剩下 4000 个词。下图为部分特征词的信息增益：

```
In [17]: sorted(featureAndweight.items(),key = lambda item:item[1] ,reverse=True)

Out[17]: [('学校', 0.13474048302005448),
          ('学生', 0.105292691163698),
          ('教育', 0.1037946214777381),
          ('污染', 0.09545365988542076),
          ('教师', 0.09269742350517873),
          ('老师', 0.08364213787349839),
          ('教育局', 0.0813537892913847),
          ('社保', 0.07807426637469472),
          ('家长', 0.06764865271532772),
          ('工资', 0.06361636561163464),
          ('排放', 0.05412373004005877),
          ('医生', 0.05376569408943799),
          ('居民', 0.05115588102902224),
          ('劳动', 0.05070737179394391),
          ('环保', 0.04917072869044503),
          ('医院', 0.048858694013449),
          ('职工', 0.04693254634865385),
          ('业主', 0.04643562428350423),
```

图表 2 信息增益部分结果图

上表得出了各个特征量信息熵，其信息熵值也一定程度上反映了该特征量的重要程度，因此在之后的模型建立中将信息熵值作为特征矩阵加权至模型中，以提升模型的训练结果[2]。

（三）模型建立

1、 模型理论分析

对于此问题我们初步的想法想到采用朴素贝叶斯算法或者逻辑回归算法建立模型，为了得到具体选择哪种算法应用的更好，我们选择建立三种不同的模型并比较，最终得出较优的模型，三种模型应用的主要算法分别为：伯努利贝叶斯，多项式贝叶斯，逻辑回归算法[4]。

以下为不同算法的思路：

朴素贝叶斯分类算法是贝叶斯分类中最简单的，是一种基于概率论的分类方法，有三种常用模型：高斯、多项式、伯努利。其模型过程基本思路相似。

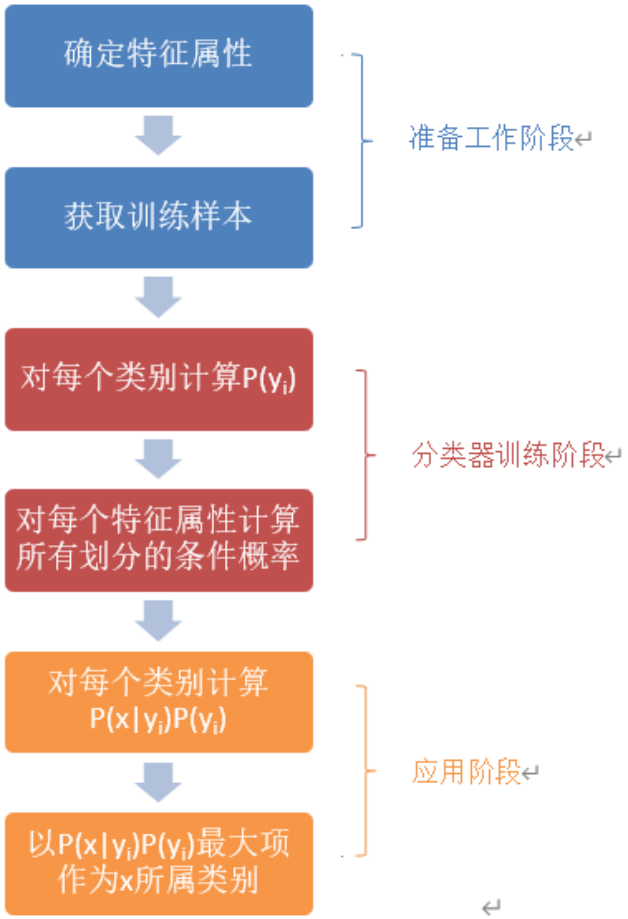
首先是准备工作阶段，主要是为朴素贝叶斯分类做必要准备，主要工作是根据确定数据的特征属性，并对每个特征属性进行适当排序和划分，然后对一部分待分类项进行分类，形成训练样本集合。这一阶段的输入是所有留言内容，输出

是特征属性和训练样本。分类器的准确性很大程度上由特征属性、特征属性划分及训练样本质量决定。

其次是分类器训练阶段，该阶段就是构造分类器，主要工作是计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，并将结果记录。其输入是特征属性和训练样本，输出是分类器。这一阶段是机械性阶段，根据公式可以由程序自动计算完成。

最后是应用阶段，该阶段是使用分类器对待分类项进行分类，其输入是分类器和待分类项，输出是待分类项与类别的映射关系。这一阶段也是机械性阶段，由程序完成。

分析流程图如下：



图表 3 朴素贝叶斯流程图

假设给定训练数据集(X,Y)，每个样本 X 包含 n 维特征，即$x = (x_1, x_2 \dots x_n)$，类别 Y 有 k 种，即$y = (y_1, y_2, \dots y_k)$。模型中 X 为 4000 个特征词，Y 为 7 种类别。从概率论角度来看，对于题中给出的待分类项 X，求解在 X 出现的条件下各个类

别出现的概率$P(y_1|x),P(y_2|x),\ldots,P(y_k|x)$，可以认为 X 属于概率最大的类别。

由贝叶斯公式和全概率公式得

$$P(y_k|x)=\frac{P(x|y_k)P(y_k)}{\sum_k P(x|y_k)P(y_k)}。$$

其中，$P(y_k)$是先验概率，根据训练集可计算出来。另外，需要对条件概率作出特征条件独立性假设

$$P(x|y_k)=P(x_1,x_2,\ldots,x_n|y_k)=\prod_{i=1}^{n}P(x_i|y_k)。$$

则朴素贝叶斯分类器可表示为：

$$f(x)=argmax_{y_k}P(y_k|x)=argmax_{y_k}\frac{P(y_k)\prod_{i=1}^{n}P(x_i|y_k)}{\sum_k P(y_k)\prod_{i=1}^{n}P(x_i|y_k)}。$$

2、建立三种模型

1） 伯努利模型

当特征词x_i为在留言文档中出现时，$P(x_i|y_k)=P(x_i=1|y_k)$；

当特征词x_i为在留言文档中不存在时，$P(x_i|y_k)=P(x_i=1|y_k)$；

2） 多项式模型

多项式模型在计算先验概率$P(y_k)$和条件概率$P(x_i|y_k)$时，会做一些平滑处理，如果不做平滑，当某一维特征的值x_i没在训练样本中出现过时，会导致$P(x_i|y_k)=0$。公式为：

$$P(y_k)=\frac{N_{y_k}+\alpha}{N+k\alpha}。$$

N 是总的样本个数，k 是总的类别个数，N_{y_k}是类别为y_k的样本个数，α是平滑值。

$$P(x_i|y_k)=\frac{N_{y_k,x_i}+\alpha}{N_{y_k}+n\alpha}。$$

N_{y_k}是类别为y_k的样本个数，n 是特征的维数，N_{y_k,x_i}是类别为y_k的样本中，第 i 维特征的值是x_i的样本个数，α是平滑值。

3）逻辑回归模型

除了建立上述两个模型，还想到用逻辑回归算法用来建立分类模型，逻辑回归模型属于广义线性模型，逻辑回归函数在线性函数的基础上进行函数转换，即 Y=g(Z),Z=X θ ,通过线性回归的方式将原有值域映射到[0，1]区间中，取值大

于临界值为一类，小于则为另一类，从而实现分类，其中，g 的函数形式一般如下：

$$g(z) = \frac{1}{1 + e^{-z}}$$

逻辑回归的假设函数为：

$$h_\theta(X) = g(X\theta) = \frac{1}{1 + e^{-X\theta}}$$

其中为 X 样本输入，$h_\theta(X)$为模型输出， θ 为要求解的模型参数。设一值为临界值，当输出大于 0.5 时，y 为 1；当输出小于 0.5 时，y 为 0。模型输出值在区间内取值，因此可从概率角度进行解释：越接近于 0，则分类为的概率越高；越接近于 1，则分类为的概率越高；越接近于临界值 0.5，则无法判断，分类准确率会下降。应用以上几种理论训练模型。

（四）模型初步验证与评价

1、验证过程

对初步建立的朴素贝叶斯模型进行有效性的验证，对此模型采用交叉验证的方法，其典型模式是 k 折交叉验证。k 折交叉验证方法将样本集分成 k 份，每次选择其中 1 份用来测试模型的性能，剩下的 k-1 份用来训练模型。交叉验证重复 k 次，并将 k 次的平均交叉验证结果作为最终对模型精度的估计[3]。

本问采用十折交叉验证方法，在初步建立模型选用特征量为 4000 个，将其分为 10 份，其中一份作为测试样本，剩余九份训练模型，交叉验证重复十次，即保证此测试数据是训练过程中未使用过的，十折交叉验证结果后，建立测试结果的混淆矩阵，以便后续进行评价，混淆矩阵的矩阵分布如下：

混淆矩阵		真实值	
		对	错
预测值	对	TP	FP
	错	FN	TN

图表 4：混淆矩阵结果分布图

其中，混淆矩阵中分类结果为预测值，原特征值所属类别为真实值。

下面使用 F-Score 对模型分类结果进行初步评价：

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

其中，P_i为第 i 类的查重率，即预测正确占全部预测正确的比例：

$$P_i = \frac{TP}{TP + FP}。$$

R_i为第 i 类的查全率，即预测正确占实际正确个数的比例：

$$R_i = \frac{TP}{TP + FN}。$$

其中 TP 为正确分为该类的样本数，FP 为将错误样本判断为正确样本的样本数量，FN 为将正确样本判断为错误样本的样本数量。

2、三种模型验证结果

对上述三种算法模型的测试结果分别进行评价，首先对混淆矩阵结果进归类，计算其查重率和查全率，计算得出各一级标签分类结果的查重率和查全率结果如下：

伯努利朴素贝叶斯结果：

```
print(eval_BernoulliNB)
```

```
[0.84335305 0.85563974 0.83197623 0.82654736 0.82694736 0.8172793
 0.80954699 0.83531274 0.83834699 0.84911931]
(       Label  Precision    Recall        F1  Support
0       城乡建设   0.820580  0.771712  0.795396      403
1       环境保护   0.878453  0.868852  0.873626      183
2       交通运输   0.482587  0.757812  0.589666      128
3       教育文体   0.903333  0.860317  0.881301      315
4   劳动和社会保障   0.872000  0.867374  0.869681      377
5       商贸旅游   0.791489  0.726562  0.757637      256
6       卫生计生   0.859649  0.816667  0.837607      180
nan       总体   0.801156  0.809900  0.800702     1842,)
```

图表 5：伯努利贝叶斯交叉验证结果图

由上表数据进行十折交叉验证后用 F-Score 评价得出结果为：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} = 0.833。$$

多项式朴素贝叶斯结果：

```
print(eval_MultinomialNB )
```

```
准确率：0.8545059717698155
交叉验证：[0.85675676 0.87702703 0.85385656 0.86197564 0.86856369 0.85597826
 0.84489796 0.87874659 0.84741144 0.8744884 ]
测试集f1：0.8545059717698155
(       Label  Precision    Recall        F1  Support
0       城乡建设   0.827423  0.868486  0.847458      403
1       环境保护   0.880597  0.967213  0.921875      183
2       交通运输   0.807692  0.656250  0.724138      128
3       教育文体   0.883077  0.911111  0.896875      315
4    劳动和社会保障   0.849148  0.925729  0.885787      377
5       商贸旅游   0.836449  0.699219  0.761702      256
6       卫生计生   0.902439  0.822222  0.860465      180
nan       总体   0.855261  0.835747  0.842614     1842,)
```

图表 6：多项式贝叶斯交叉验证结果图

由上表数据进行十折交叉验证后用 F-Score 评价得出结果为：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} = 0.85。$$

逻辑回归结果：

```
交叉验证：[0.79403973 0.81823624 0.81397347 0.80341165 0.80910933 0.79850681
 0.78331139 0.81773042 0.76576295 0.80859596]
测试集f1：0.8161105338635316
(       Label  Precision    Recall        F1  Support
0       城乡建设   0.743478  0.848635  0.792584      403
1       环境保护   0.877907  0.825137  0.850704      183
2       交通运输   0.851064  0.625000  0.720721      128
3       教育文体   0.879630  0.904762  0.892019      315
4    劳动和社会保障   0.847826  0.931034  0.887484      377
5       商贸旅游   0.824324  0.714844  0.765690      256
6       卫生计生   0.865385  0.750000  0.803571      180
nan       总体   0.841373  0.799916  0.816111     1842,)
```

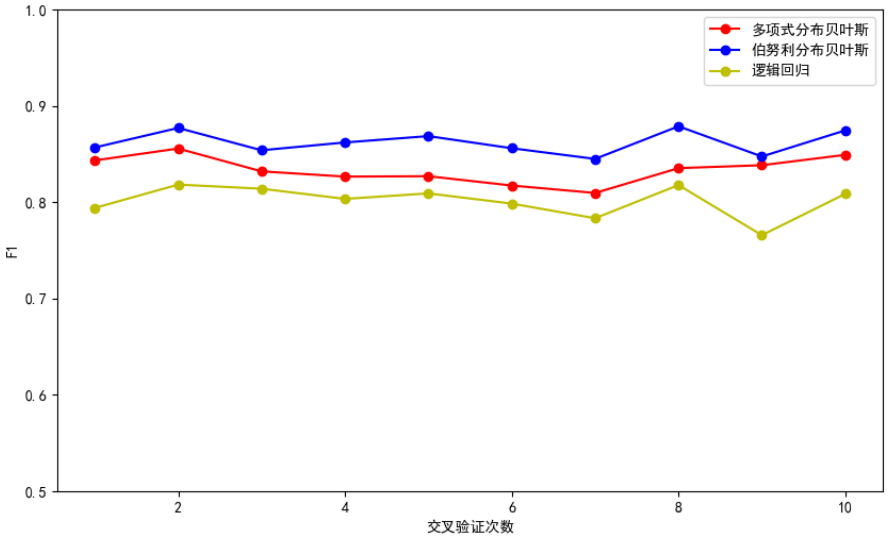
图表 7：逻辑回归交叉验证结果图

通过上述数值计算得出，初步建立的模型 F-Score 的值为：

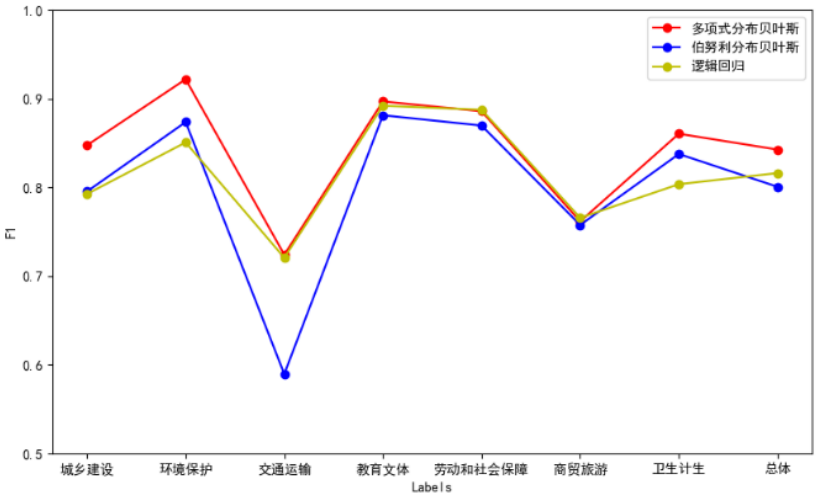
$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} = 0.82。$$

3、三种模型结果比较

对上述三种模型的评价结果进行可视化处理，对每一个模型的十折交叉验证结果进行比较，同时对一级标签的分类结果进行评价，他们的十折交叉验证和 F-Score 结果图分布如下：



图表 8：三种模型十折交叉验证结果图



图表 9 ： 三种模型 F-Score 结果图

根据上图可以比较直观的看出，多项式分布贝叶斯算法模型的评价结果最高，同时，伯努利贝叶斯算法和逻辑回归算法的 F-Score 评价结果相对来讲较好，整体处于 0.8-0.9 的区间范围内。在交叉验证结果中，可以看出伯努利贝叶斯整体的十折交叉验证结果较好。

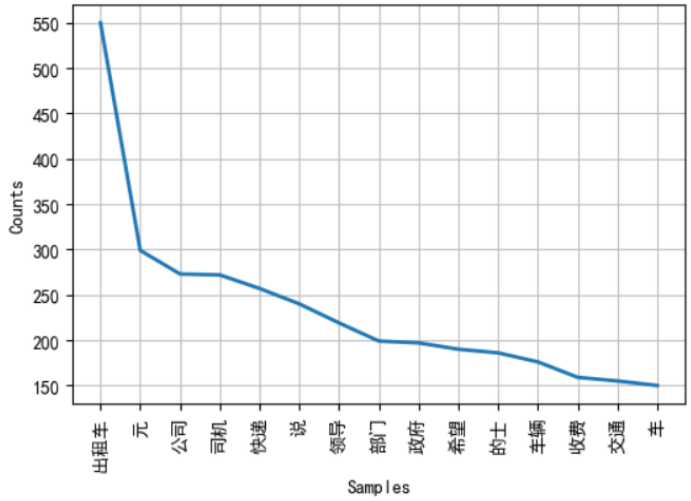
综上我们认为伯努利贝叶斯模型较好，并对初步建立的伯努利贝叶斯模型进行下一步的改进。

（五）模型改进

1、对交通运输类的改进

在伯努利贝叶斯模型中，通过直观的各个一级标签与 F-Score 图像可以看出，

交通运输类的评价值明显偏低，这表示在此模型中交通运输类的分类结果比较不准，因此在模型中对交通运输类进行改进，起初选择在原有特征词样本的基础上添加交通运输类的样本数量，首先统计交通运输类的词频如下：



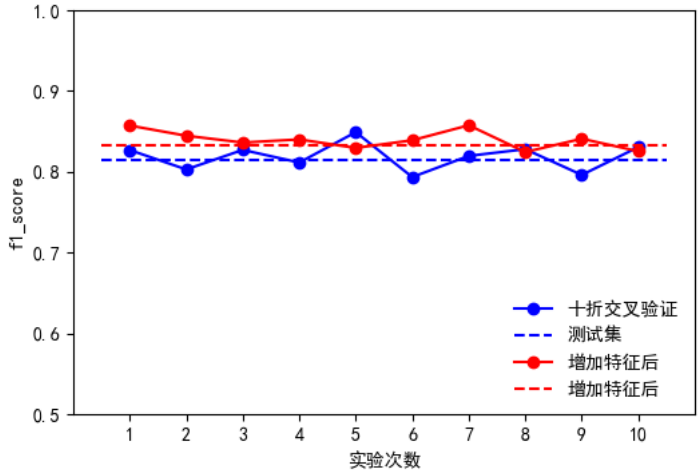
图表 10：交通运输类词频图

根据信息增益的方法，选择熵值高的前 800 个特征词添加为新样本，与原有的 4000 个样本相加，去掉重合样本后，样本总数变为 4602 个，再次训练模型后，进行交叉验证，改进后的混淆矩阵分类结果为：

```
交叉验证：[0.85719276 0.84461116 0.83616867 0.84004665 0.82951958 0.83900076
 0.85756202 0.82441264 0.84087375 0.82537309]
测试集f1：0.8327064847139568
[[343   5  52   7   6  13   1]
 [ 10 167   2   1   0   3   0]
 [ 11   1  98   1   4   4   0]
 [  7   0   7 272  18  16   3]
 [  3   0   9  11 328   0  13]
 [ 16   3  22   4   2 199   1]
 [  1   0   6   1  15   8 148]]
```

图表 11：改进伯努利贝叶斯混淆矩阵结果图

由上表得出改进交通运输类的词频后，评价模型的 F-Score 值为 0.833，与原来相比并没有太大的差异。因此观察对比每次交叉验证，得到的验证结果的 F-Score 值与改进前的值比较图如下：

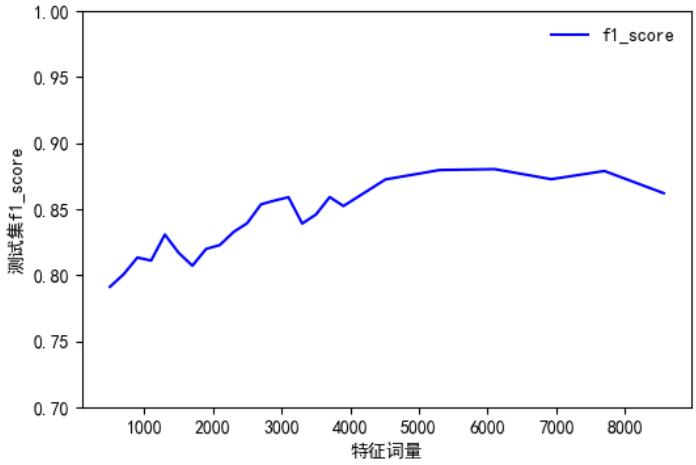


图表 12：优化前后 F-Score 结果对比图

从上图中可以看出，增加特征后的验证结果为红色折线，改进前的验证结果为蓝色折线，改进后的 F1-Score 值有了显著的提升。

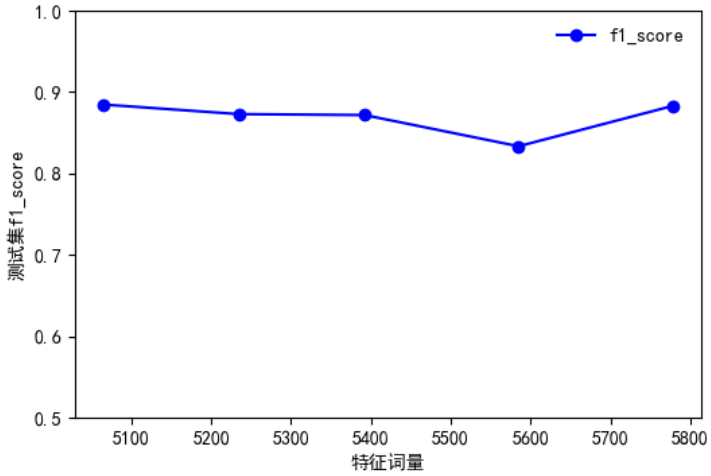
2、从整体改进

由于在初步建立模型时，模型训练样本容量即特征词量为人为规定，由此得出的 F-Score 值在该样本容量条件下的训练模型中得到了较好结果，但是特征词的数目会影响到模型进而影响模型评价结果，在之前的模型中特征词量并非为 F-Score 值最高值时的特征词量。因此，要观察模型训练特选取征词量对模型评价结果的影响，影响结果如下图：



图表 13：评价结果随特征词量变化图

由上图可以较为直观的看出，当特征词量选择在 5000 左右时，F-Score 的值较高，即模型较为优秀。因此，我们改变模型初始的特征词量对模型进行改进，选取特征词量在 5000 左右时，观察模型的评价结果如下图：、



图表 14：不同特征词量的评价结果图

上表中，特征词数量与其具体 F1-Score 值对应关系如下：

特征词量	F1-Score 值
5065	0.88
5235	0.87
5392	0.87
5585	0.83
5778	0.88

图表 15：不同特征词量的凭借结果具体数值表

从数据中可以看出在这个区间范围内评价结果整体较高但是略有浮动，这代表了模型也存在一定的差异性和随机性，因此选择特征词量在 5100 左右，此时，模型评价结果 F1-Score 值为 0.88。

三、问题二

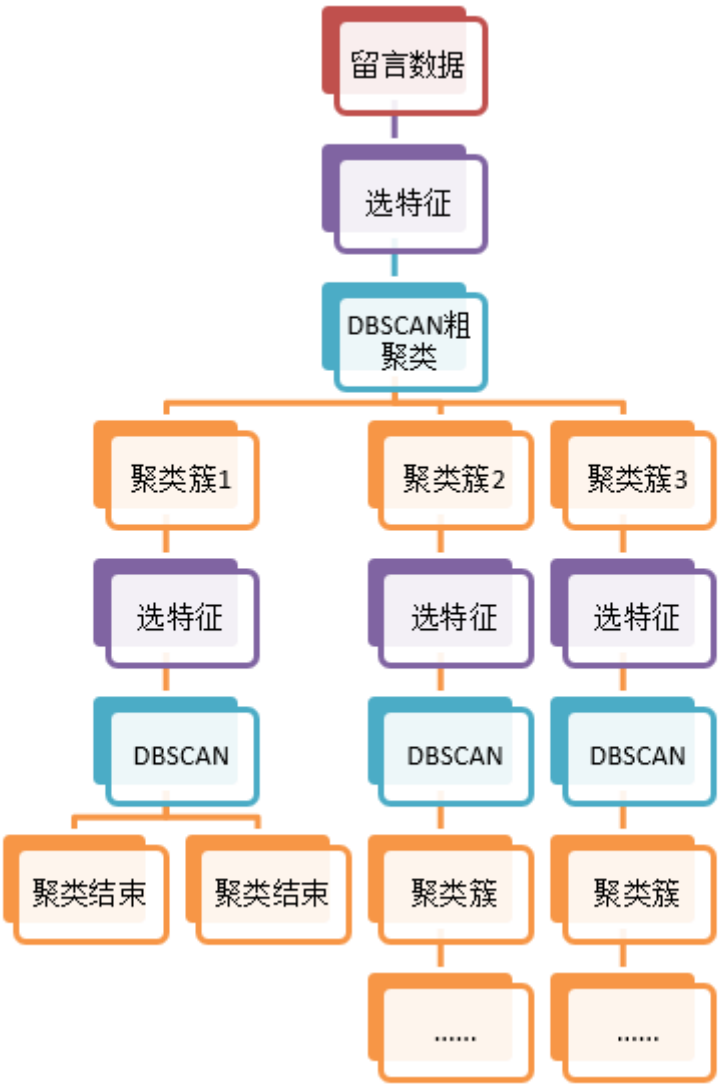
（一） 问题分析

该问题要对热点问题进行挖掘，问题之所以成为热点问题，因其在短时间内的反应比较大，提问的时间较为集中，地点较为高频，点赞人数较高，因此对热点问题的挖掘也从以上几个相关信息入手。

由于留言主题中一般拥有留言详情中的关键词，所以我们选择处理留言主题中的文本。将留言文本简单进行数据预处理，分词处理和统计词频之后，选择词频在一定范围内的特征词进行 DBSCAN 粗聚类。粗聚类之后，得到不同密度的聚类簇。由于粗聚类选取的半径大，聚类不精确，接着对得到的聚类簇进行再次聚

类。假设聚类簇 1 的留言数量为 n，选取频次大于 n/2-1 的词语作为第二次聚类的特征词。往后的重复聚类直至收敛。根据聚类结果得到的问题进行热点问题热度计算，设定热度评价指标，并评价得出的热度结果[5]。

主要思路的流程图如下：



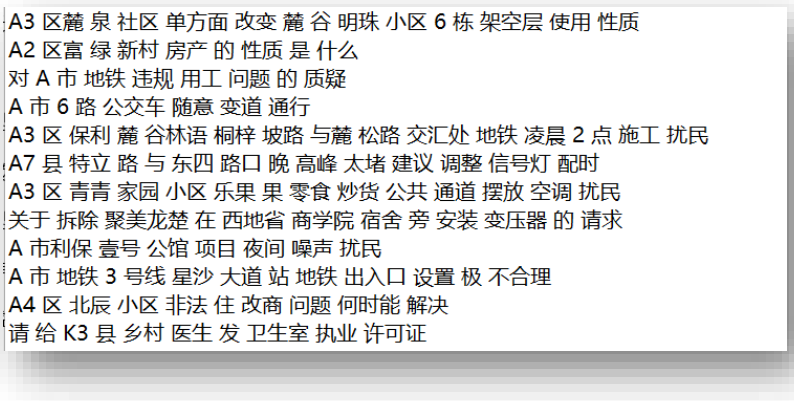
图表 16：聚类方法流程图

（二） 具体步骤

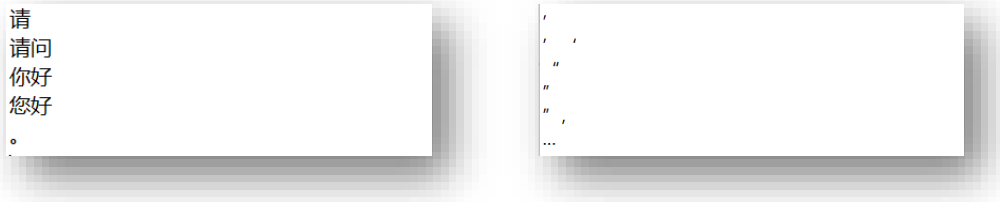
1、数据预处理

我们选择对附件 3 留言主题中的文本进行预处理，方法同问题一相似。首先，对文本进行分词处理（结果如下图），并统计每个单词的词频，标注词性，选择词频在 4~500 之间的词语；然后，我们设置了停用词表（如图），删除掉一部分符号、数字、无用的代词等，使文本保留关键的人名地名等词语；最后利用信息增益算法，将词语按照 IG 值进行降序排列，筛选出较为重要的一部分词作

为特征词。接下来，利用 DBSCAN 算法对得到的特征词进行粗聚类。



图表 17：分词部分结果图



图表 18：部分去停用词图

2、基于密度的聚类算法

DBSCAN 算法的核心思想就是先发现密度较高的点，然后把相近的高密度点逐步都连成一片，进而生成各种聚类簇。

首先通过计算两个特征向量之间的余弦相似度来计算文本相似度，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；将相似度带入到 DBSCAN 的距离公式中，计算每个点的邻域和密度值。粗聚类时，对每个数据点为圆心，以 0.7 为半径画个圈，然后我们在 0-1 之间选取一个密度阈值，设置每个类别中最少个数为 4。根据邻域的大小和密度阈值，判断一个点是核心点、边界点或者异常点，并将异常点删除。如果核心点之间的距离小于密度阈值，就将两个核心点连接在一起，这样就形成了若干组聚类簇。将边界点分配到距离它最近的核心点范围内，形成最终的聚类结果。

粗聚类时，我们选取的半径为 0.7，相对来说数值有些大，聚类不够精确。因此，需要对得到的聚类簇进行再次聚类。假设聚类簇 1 的留言数量为 n，选取频次大于 n/2-1 的词语作为第二次聚类的特征词。根据聚类的次数，计算出每个

数据点的半径，使每个类别中最少个数为 3，重复上面第一次聚类的步骤，从而得到其领域和密度。根据情况判断是否进行下一次聚类，当每个类别中只存在一个特征数据时，结束聚类。此外也可以自己设置聚类的次数来结束聚类。若要继续聚类，选取特征词的方法和计算半径的公式都按照第二次聚类来算。最终聚类收敛得到的聚类簇一共有 73 个。

3、热度评价指标

定义热度评价指标主要由留言的出现次数，留言点赞数，留言时间这几个参量决定，因此，我们对以上几个参量进行权重等划分，对聚类后得到的热点问题进行热度评价指标评价，用 Q 代表指标评价结果，其公式为：

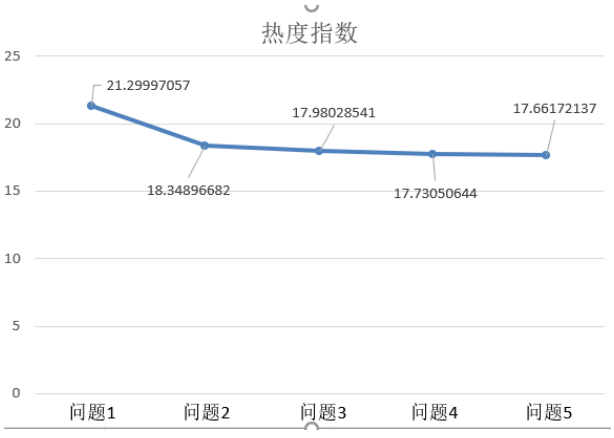
$$Q = \frac{c}{K}\sum_{i=1}^{n}\frac{\sqrt{posi}+a}{\sqrt{negi}+b}。$$

其中，pos 代表该留言的点赞数，a 为每篇文章的因子，用于平衡点赞数过高导致热度不均匀的问题，经过尝试最终取得此处 a=3，代表 9 个点赞数与一条留言重要性相当。neg 为该条留言的反对赞数，b 用于防止分母为 0，此处 b=1，c 用于控制评价指标的整体量，取值 c=20，K 为热度指数中的时间指数，其公式为：

$$K = \sqrt{\sigma_t + d}。$$

时间指数中，σ_t为该条留言时间在此聚类簇中所有时间的标准差，d 为防止分母为 0，因此 d 取值为 1。

根据上述热度评价指标得出了热点问题的热度指数，其中前五个热点问题的热度指数的折线图如下：



图表 19：前五热点问题热度指数图

前五个问题的热度指数分别为：21.30、18.35、17.98、17.73、17.66 根据以上评价公式计算得出的前五名热点问题以及热点问题明细见附件。

四、问题三

（一）方案

要对答复意见进行评价，我们认为需要计算答复意见和留言详情两个文本的相似度。两个文档的相似度越高，其相关性越强，完整度也越高。计算步骤如下：首先，读取答复意见文档，对文档进行 jieba 分词，并整理分词后的文档，使其成列表格式，以便于后续方便计算。随后，计算每个词语的频率，对低频率的词进行过滤，将词语向量化。再通过语料库建立词典。接下来，加载要对比的文档——留言详情。将留言详情文档通过 doc2bow 转化为稀疏向量，对其进一步处理，得到新的语料库。将新的语料库通过 TF-IDF 模型处理，得到 TF-IDF 值。通过 token2id 得到特征数，也就是字典里面键的个数。计算稀疏矩阵的相似度，建立一个索引，根据索引得到最终的相似度。

在 TF-IDF 模型中 TF 指词频，表示一段文字中单词出现的频繁程度。IDF 指逆文档频率，表示在所有文本中单词出现的不频繁程度。

第一步，计算词频：

$$\text{词频（TF）}=\text{某个词在文章中出现的次数}$$

考虑到文章的长短字数不同，为了便于不同文章的比较，可以将“词频”标准化。

$$\text{词频（TF）}=\frac{\text{某个词在文章中出现的次数}}{\text{文章的总次数}}$$

第二步，计算拟文档频率：

需要建立一个语料库（corpus），用来模拟语言的使用环境。

$$\text{逆文档频率（IDF）}=\log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}+1}\right)$$

第三步，计算 TF-IDF 值：

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

可以得到，TF-IDF 与一个词在文档中的出现次数成正比，与该词在整个语言中的出现次数成反比。若一个词在文章中的 TF-IDF 值越大，那么意味着一般

来说这个词在这篇文章的重要性越高。

（二）预测结果

上述给出的方案中，通过 TF-IDF 将两条内容进行稀疏矩阵测算，计算相似度测算来估计答复意见的结果，从相关性角度看，相似度越高特征量重合度越高，答复的相关性越高，以下面的回复为例子，留言详情与回复内容的特征词相关性高，两条内容的相似度很高。

易书记：　　您好！黄兴小学问题塑胶跑道在家长们一月来联名不断上访的努力下，由教育局出面督促施工单位下对学校问题塑胶（大小操场）于2日铲除完了，家长和孩子们很开心；然而，开心不出半日烦恼又生；因拆除露出水混面的小操场，有家长发现紧挨教室的小操场上开始铺沥青了！沥青不透水，气味重，并含荧光物质，其中含致癌物质3．4苯并芘，高温处理时随烟气一起挥发出来。沥青烟气是青黄色的气体，其中含焦油细雾微粒。沥青烟和粉尘可经呼吸道和污染皮肤而引起中毒，发生皮炎、视力模糊、眼结膜炎、胸闷、腹病、心悸、头痛等症状。经科学试验证明，沥青和沥青烟中所含的3，4苯并芘是引起皮肤癌、肺癌、胃癌和食道癌的主要原因之一。　在受沥青污染的空气中生活，易致免疫力下降。第二日就是孩子们上课了，此做法对孩子们身体肯定有影响的，而且极不负责！为何不采用80%家长们提出的环保建议：校园里多栽绿色植被；建议使用水泥，沙坑，草坪，沙质跑道（环保，维护也容易）；请书记出面吧，给予全校六七百名师生们健康更好选择，谢谢！

网友“UU0081308”　　您好，您的留言已收悉。现将有关情况回复如下：　　针对家长提出的使用水泥、沙质、草皮铺设操场的建议，黄兴小学之前进行过认真的研究与论证。如果使用水泥操场，施工期和维护期都比较长，影响学校的正常教学秩序，并且水泥操场因表面较硬且遇水较滑，其安全性低于改性沥青操场，学生在操场上活动容易摔跤且容易受伤。如果使用草皮铺设操场，维护的成本太高、难度太大。因为小学的操场是学生上体育课和课余活动的重要场地，不同于专业的比赛球场有专人进行维护和保养，学生的踩踏很容易对草坪造成根本性的破坏。如果使用沙质或土质的操场，会出现“天晴一脸灰，下雨一身泥”的情况，也会给学生和老师带来不便。因此，经过多方考量，黄兴小学决定铺设改性沥青，因为改性沥青广泛应用于城市道路、生活小区、校园及公共广场等地面的铺设，安全性较高，且易于维护，是较理想的建筑材料。　　针对家长提出的校园内多栽种绿色植物的建议，黄兴小学在今年暑期的提质改造中，已对校园进行了最大限度的绿化，并购买了许多盆栽绿色植物放置于校园内，以美化校园的环境。　　感谢您对我区教育工作的关心、理解与支持！中共A2区委办公室2013年12月4日

图表 20：示例恢复特征词

五、总结

（一）问题一模型对比

问题一中用到了伯努利贝叶斯模型，多项式贝叶斯模型，逻辑回归模型，分析比较三种模型结果可以看出模型评价指标结果中多项式贝叶斯模型的评价指标较高，但是在十折交叉验证结果中，伯努利贝叶斯模型的各项分类结果指标较为优异，在伯努利贝叶斯模型中，没有参数变化，但是其余两个模型中可以设置参数的量，使分类结果更偏向一方。

（二）模型优缺点

1、问题一

优点：问题一模型中的模型评价指标较高，模型准确性较好。在训练模型的过程中，通过信息增益的方法得出特征量及熵增值，将特征量的熵增作为矩阵权重用于模型中，能够增加模型的精确度。

缺点：问题一最终选择了伯努利贝叶斯模型，伯努利贝叶斯的算法中，不能拟定参数使分类偏向一方，只能靠模型训练结果，在模型训练结果中可以看出交通运输类的分类结果准确率较低，即评价指标低，且在特征量的选取上模型结果有一定随机性。

2、问题二

优点：问题二的基于密度的聚类算法中，通过粗聚类再聚类直至之收敛的方式，能够准确的查找出密度高的热点问题，建立的热度评价指标中，综合了留言频率、点赞数量、留言时间的集中度综合总结出了一个公式进行评价。

缺点：在评价热点问题的指标中，其中的参量值均是自己拟定的，没有准确的测算各相关量之间的关系，如果有更多的理论支撑准确度可能更好。

六、参考文献

[1]李军,乔立民,王加强,高杰.智慧政务框架下大数据共享的实现与应用研究[J].电子政务,2019(02):34-44.

[2]李海瑞. 基于信息增益和信息熵的特征词权重计算研究[D].重庆大学,2012.

[3] 秦彩杰, 管强. 一种基于 F-Score 的特征选择方法[J]. 宜宾学院学报, 2018, 018(006):4-8.

[4]杨杰明. 文本分类中文本表示模型和特征选择算法研究[D].吉林大学,2013.

[5]高茂庭. 文本聚类分析若干问题研究[D].天津大学,2007.