基于文本分类算法及热点问题指标、答复质量评价的探讨

摘要

本文主要对政府收集到的各类社情民意文本数据进行分类汇总处理，通过分析社情民意及相关部门对留言的答复意见的文本数据，结合当前的大数据技术，通过 python 软件建立模型，来实现文本数据处理。

针对问题一，本文构建了多项式朴素贝叶斯模型。首先对数据进行预处理，利用 python 软件导入数据，对留言内容一级标签的分类进行量值化分别为 1，2…,7，对群众留言文本数据（附件 2 中给出）进行去重，利用 python 中的 jieba 库对群众留言数据进行分词、去停用词；其次进行数据分析，利用 IF-IDF 算法对样本数据进行赋值，得到样本的 TF-IDF 文本权值向量；最后建立多项式朴素贝叶斯模型对文本权值向量进行分类，并利用 F-Score 对分类模型进行准确度评估与评价。

针对问题二，我们按照问题一的步骤，利用 python 软件对附件 3 给出的群众留言文本数据进行预处理（读取、分词及去停用词），对样本数据进行处理后通过 TF-IDF 算法获取文本权值向量，并进行 K-means 聚类分析，将数据合理归类，以此反映相似问题的频率。引入问题热度 H，将问题热度分为冷门问题、一般问题和热门问题。根据问题热度 H 与问题的反映频率 K 成正相关关系、H 与民众本身关心的问题外的问题总数（N-K）成负相关关系，建立热度评价指标模型通过计算得出这三类问题之间的问题热度临界值，并以此为依据得出排名前 5 的热点问题及热点问题留言明细表。

针对问题三，利用附件 4 的答复意见进行情感分析，我们按照问题一的步骤，利用 python 软件对附件 4 给出的答复意见文本数据进行预处理（读取、分词及去停用词），同时计算情感分数，对否定词和程度副词进行计算，定义分值计算函数，得到所有答复意见的情感得分，分析情感得分后定义得分大于 W(ave)表示该答复较正面，反之较负面，建立 LDA 模型提取答复意见的主题，与附件 4 的留言主题做余弦相似度计算，以余弦相似度计算出来的数值来说明答复的相关性及完整性。利用情感分数和余弦相似度数值建立答复意见质量评价指标。对答复意见进行定量计算后，对应质量评价指标便可得到该答复意见的质量。

关键词：多项式朴素贝叶斯模型、IF-IDF 算法、热度评价指标模型、LDA 主题模型、答复意见质量评价

Discussion on text classification algorithm, hot question index and reply quality evaluation

abstract

In this paper, the government collected all kinds of social information and public opinion text data are classified, collected and processed. By analyzing the social information and public opinion and the text data of the response of the relevant departments to the message, combined with the current big data technology, the model is established through Python software to realize the text data processing.

In order to solve the first problem, this paper constructs a polynomial naive Bayesian model. First of all, preprocess the data, import the data with Python software, and quantize the first level labels of the message content into 1, 2 , 7. De duplicate the mass message text data (given in Appendix 2), use the Jieba database in Python to segment and de stop the mass message data; secondly, analyze the data, use if-idf algorithm to assign the sample data, and get the TF-IDF text weight vector of the sample; finally, establish the polynomial naive Bayesian model to classify the text weight vector, and use it to F-score evaluates the accuracy of the classification model.

To solve the second problem, we use Python software to preprocess the text data of the mass message given in Annex 3 (reading, segmentation and deactivation words). After processing the sample data, we use TF-IDF algorithm to obtain the text weight vector, and carry out K-means clustering analysis to reasonably classify the data, so as to reflect the frequency of similar problems. By introducing the problem heat h, the problem heat is divided into cold door problem, general problem and hot problem. According to the positive correlation between the problem heat degree h and the reflection frequency K of the problem, and the negative correlation between H and the total number of problems (n-k) other than the problems concerned by the people themselves, a heat evaluation index model is established to calculate the critical value of the problem heat degree among the three types of problems, and based on this, the top 5 hot issues and hot issues message list are obtained.

In view of the third question, we use the response opinions of Annex 4 to carry out emotional analysis. According to the steps of question 1, we use Python software to preprocess the text data of the response opinions given in Annex 4 (reading, segmentation and de deactivation words), calculate emotional scores, calculate the negative words and degree adverbs, define the score calculation function, and get all the emotional scores of the response opinions Score: after analyzing the emotional score, if the definition score is greater than w (AVE), it means that the reply is more positive; otherwise, it is more negative. LDA model is established to extract the subject of the reply. Cosine similarity calculation is done with the message subject in Annex 4, and the value calculated by cosine similarity is used to explain the relevance and integrity of the reply. Using emotion score and cosine similarity to establish the evaluation index of response quality. After quantitative calculation of the reply, the quality of the reply can be obtained by corresponding quality evaluation index.

Keywords: polynomial naive Bayesian model, if-idf algorithm, heat evaluation index model, LDA subject model, response quality evaluation

目录

1 问题重述

1.1 问题背景

近年来，政府主要通过微信、微博、市长信箱等网络问政平台了解民意，故有关各种社情民意的文本数据日益增多，人工进行留言分类和热点整理的工作难度极大。同时，随着大数据、云计算等技术的发展，基于自然语言处理技术建立的智慧政务系统成为了社会治理发展的新趋势，极大的提升了政府的管理水平和施政效率。

附件中包括了群众问政留言记录，以及相关部门对部分群众留言的答复。本文主要利用自然语言处理和文本挖掘来解决以下问题。

1.2 要解决的问题

问题一：本题主要是对群众留言进行分类。工作人员首先会按照附件 1 给出的划分体系，对网络问政平台的群众留言进行分类，然后将分类好的留言发送至相应的部门进行处理。由于目前大部分电子政务需要人工处理，工作量大、效率低且出错率高的问题一直存在。本题需根据附件 2 给出的数据，建立关于群众留言的一级分类模型，并对分类方法进行评价。

问题二：本题主要对热点问题进行挖掘。热点问题是多数群众在某一时间段内反映的某一问题。热点问题的及时发现，可以让政府更有针对性地进行处理，从而提高服务效率。本题需根据附件 3 中的数据，在对某一时间段内反映地点或人群问题的留言进行分类的基础下，定义出合理的热度评价指标，给出评价结果。将排名前 5 的热点问题按表 1 的格式保存至文件“热点问题表.xls”，将给出的相应热点问题对应的详细留言信息按表 2 的格式保存至文件“热点问题留言明细表.xls”。

问题三：本题主要是对政府的答复意见进行评价。附件 4 中给出了相关部门对群众留言的答复意见，需要从答复的相关性、完整性及可解释性等角度进行分析，得出一套答复意见质量的评估方案，并尝试测试实现。

2 问题分析

随着互联网技术的发展，我们生活的这个世界出现了大量的文本信息，利用大数据技术高效快速地将这些信息进行分类汇总至关重要。本文通过各类社情民意的文本数据，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，并挖掘出热点问题。

2.1 问题一的分析

针对本题，我们首先利用 python 软件对附录 1、2 中的数据进行读取，并对附录 1 中的一级标签进行量化值，对附录 2 中数据进行数据去重，以便后续的数据处理。利用 python 中的 jieba 库对附录 2 中的数据进行分词处理、去停用词，利用 IF-IDF 算法对样本数据进行赋值，得到样本的 TF-IDF 文本权值向量。建立多项式朴素贝叶斯模型对样本数据进行分类，最后利用 F-Score 对分类模型进行

准确度评估。

2.2 问题二的分析

针对本题，我们按照问题一的步骤，对附件 3 给出的群众留言文本数据进行预处理（读取、分词及去停用词），对样本数据进行处理后通过 TF-IDF 算法获取文本权值向量，并通过 python 软件进行 K-means 聚类分析，将数据合理归类，即可反映相似问题的频率。引入问题热度 H，将问题热度分为冷门问题、一般问题和热门问题，建立热度评价指标模型通过计算得出这三类问题之间的问题热度临界值，并以此为依据得出排名前 5 的热点问题及热点问题留言明细表。

2.3 问题三的分析

针对本题，利用附件三的答复意见进行情感分析，我们按照问题一的步骤，利用 python 软件对附件 4 给出的答复意见文本数据进行预处理（读取、分词及去停用词），同时计算情感分数，对否定词和程度副词进行计算，定义分值计算函数，得到所有答复意见的情感得分，得分大于 0 表示该答复较正面，反之较反面，建立 LDA 模型提取答复意见的主题，与附件 4 的留言主题做余弦相似度计算，以余弦相似度计算出来的数值来说明答复的相关性及完整性。对答复意见进行定量计算后，对应质量评价指标便可得到该答复意见的质量，故要通过算法求出余弦相似度和情感得分，利用情感分数和余弦相似度数值建立答复意见质量评价指标。

3 符号说明

符号	解释与说明
N	总样本数
K	总类别数
α	平滑值
Y_K	类别
n	特征的维数
N_{YK}	类别 Y_K 的样本个数
$N_{Y_K,\ x_i}$	Y_K 中，第 i 为特征值是 X_i 的样本个数
P_i	第 i 类查准率
R_i	第 i 查全率
H	问题热度
K	问题反映频数
λ	比例系数
$\cos\theta$	余弦相似度
W	情感分数
W(ave)	情感分数平均值

4 模型的建立与求解

4.1 问题一：多项式朴素贝叶斯模型的建立与求解

4.1.1 模型思想

在这个信息爆炸的时代，非结构化的数据正在急剧增加，如文本型数据，而面对这些海量的数据，如果单单只凭人工来处理，就显得极其消耗人力和时间，而且存在较大的误差，于是，基于计算机算法实现自动分类建立的系统就显得极其重要，通过文本分类算法可以节省大量的时间和人力。因此，下面介绍本文用于文本分类的算法。

一、下面给出整个文本分类的框架图：

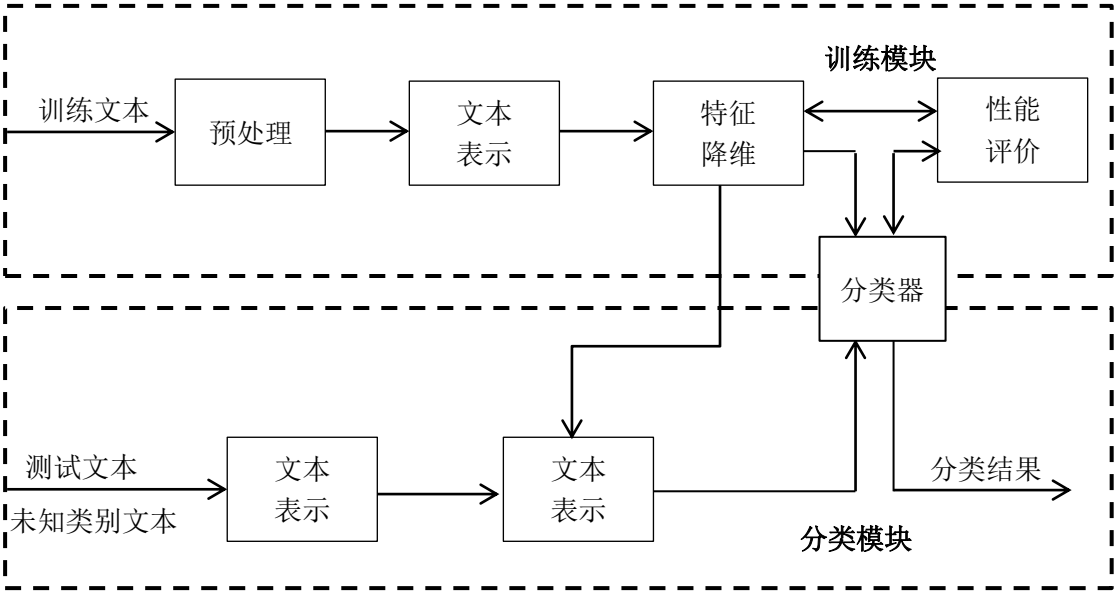

图 1 文本分类框架图

二、下面给出整个算法的思想图：

图 2 算法思想图

整个流程分为三个过程：数据预处理、数据分析、模型建立与求解。

(1)数据预处理包括了文本数据的去重、利用 python 中的 jieba 库进行分词、去停用词。

(2)数据分析包括对处理好的数据进行向量表示，利用 TF-IDF 算法对样本数据进行赋值，获取样本的 TF-IDF 权值向量。

(3)模型建立与求解中建立了多项式朴素贝叶斯分类器对样本获取到的 TF-IDF 权值向量进行分类、利用 F-Score 对分类模型进行测试与评价。

4.1.2 模型建立

一、朴素贝叶斯算法

对于文本的分类，我们建立了一级标签分类模型即朴素贝叶斯模型。朴素贝叶斯算法是基于贝叶斯定理与特征条件独立假设的分类方法[1]，利用了贝叶斯定理首先求出联合概率分布，再求出条件概率分布。这里的朴素贝叶斯是指在计算似然估计时假定了条件独立。基本原理可以用下面公式给出：

$$P(Y|X)=\frac{P(Y)P(X|Y)}{P(X)} \qquad \text{（公式 1）}$$

其中

$P(Y|X)=P(X_1,X_2,\cdots X_n|Y)=P(X_1|Y)P(X_2|Y)\cdots P(X_n|Y)$

$$P(X)=\sum_{i=1}^{n}P(Y_i)P(X|Y_i)$$

$P(Y|X)$叫做后验概率，$P(Y)$叫做先验概率，$P(X|Y)$叫做似然概率，$P(X)$叫做证据。

下面我们采用了朴素贝叶斯模型中的多项式朴素贝叶斯模型(MultinomialNB)。

二、多项式朴素贝叶斯模型(MultinomialNB)

当特征是离散的时候，使用多项式模型。符合本文数据类型的特点，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，具体公式为：

$$P(Y_K)=\frac{N_{Y_k}+\alpha}{N+K\alpha} \qquad \text{（公式 2）}$$

N 是总样本数，K 是总类别数，N_{YK}是类别 Y_K的样本个数，α是平滑值。

$$P(X_i|Y_K)=\frac{N_{Y_k,\ X_i}+\alpha}{N_{Y_K}+n\alpha} \qquad \text{（公式 3）}$$

N_{Y_K}是类别 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 为特征值是 Xi 的样本个数。

当α=1 时，称作 Laplace 平滑，当 0<α<1 时，称作 Lidstone 平滑，α=0 时不做平滑。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 $P(X_i|Y_K)=0$，从而导致后验概率为 0。加上平滑就可以克服这个问题。

对于预测出来的数据用 F-score 对分类方法进行评价。

$$F_1=\frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_iR_i} \qquad (公式\ 4)$$

P_i 为第 i 类查准率，R_i 为第 i 查全率。

4.1.3 模型的求解

一、求解前对数据预处理

(1)量值化

首先对数据经行预处理前，为了方便操作，本文对一级分类中的七大类标签进行量化值处理，如表 1 所示：

城乡建设	环境保护	交通运输	教育文体	劳动和社会保障	商贸旅游	卫生计生
1	2	3	4	5	6	7

表 1 标签量化值

(2)读取数据

利用 python 读取文本较短的留言主题(附件 2 中的留言主题)和文本标签，赋予标题分别为 massage、label，如图 3 所示读取结果：（代码参考作品附件附录 1）

Out[57]:

	message	label
0	A市西湖建筑集团占道施工有安全隐患	1
1	A市在水一方大厦人为烂尾多年，安全隐患严重	1
2	投诉A市A1区苑物业违规收停车费	1
3	A1区蔡锷南路A2区华庭楼顶水箱长年不洗	1
4	A1区A2区华庭自来水好大一股霉味	1
...	...	...
9205	两孩子一个是一级脑瘫，能再生育吗?	7
9206	B市中心医院医生不负责任，做无痛人流手术后结果还是活胚芽	7
9207	西地省二胎产假新政策何时出台?	7
9208	K8县惊现奇葩证明!	7
9209	请问J4县卫计委社会抚养费到底该交多少钱?	7

9210 rows × 2 columns

图 3 附件 2 读取结果

(3)数据去重

我们容易理解，由于反映问题的频率不止一次，因此数据中可能存在着重复的反映问题的数据，对此我们进行了去重的操作方便后面数据处理的操作，去重结果如图 4：（代码参考作品附件附录 1）

```
Out[58]: 1257                    L市2017年的房屋补贴什么时候发
         92                      建议A市行道树枝修剪至2.2米高
         1876                    H4县澧源镇八斗溪违章建筑十分猖獗
         1595                    投诉A市金毛地产精装修质量问题大！
         599      E市E10县金石镇山立花园小区，住户入住了6年没有办理房产证
                                  ...
         9097                    请救救K市瘫痪居民的双腿
         9139                K3县金沙湾业主坚决反对宁永健康体检营业
         8757                    超生二胎后妇女必须结扎吗？
         8354                希望给予村医与乡镇卫生院职工同待遇
         9209              请问J4县卫计委社会抚养费到底该交多少钱？
         Name: message, Length: 8905, dtype: object
```

图 4 数据去重

(4)分词

从原来的 9210 条数据（图 3）变成 8905 条数据（图 4），我们不难看出，清洗掉了部分重复的数据，接下来就是对数据经行分词，这里我们利用 python 中的 jieba 库进行分词，jiaba 采用了基于前缀词典实现的高效词图扫描，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。但我们观察数据可知，数据中存在着一些专有名词而 python 自带的 jieba 库里不能识别，为了解决这个问题我们导入新的专有名词，如附件二中的伊景园、滨河苑、丽发新城、魅力、妇幼保健院、A 市、A2 区、A3 区、A6 区、A7 县等以便减小计算误差。分词结果如图 5 所示：（代码参考作品附件附录 1）

```
Out[59]: 1136                  [请管，一下，K2区，跃进，小区，楼顶，搭建，违章建筑]
         1713                      [C4市，东方，巴黎，小区，规划局，严重，违规]
         1499                          [M5，市，六亩，塘棚，改办，乱收费]
         798          [J2区，石榴，湾，游园，河边，护栏，锈蚀，损毁，，，有，安全隐患]
         1297                  [L1区，薰衣草，家园，小区，房子，还，没有，房产证]
                                          ...
         8974         [A市，妇幼保健院，医生，护士，服务，差，，，不，开放，地下，车库]
         8521                    [L9县，无证，行医，和，借证，行医，现象，严重]
         8384               [M2县，一，非法，诊所，（，四方，诊所，），一针，致命]
         8803                [M3县，西河，镇，计生委，乱，征收，二胎，社会，抚养费]
         8555    [D市，农民，患者，举债，30，万，治疗，肾结石，竟，造成，无尿，，,...
         Name: message, Length: 8905, dtype: object
```

图 5 分词

(5)去停用词

对于分完词的数据，还不能作为数据的筛选，因为里面会出现一些常见的停用词，如“啊”、“在”、“的”之类。这些词也可称作虚词，包含副词、冠词、代词等，在文档中使用十分广泛，但却难以对文档分类提供帮助，它们在文本中并没有实际的意义，因此，在研究文本分类等数据挖掘问题时，经常会将它们预先剔除，既可以减少存储空间、降低计算成本，又可以防止它们干扰分类器的性能，我们导入已经准备好的停用词表对上面的数据进行去停用词操作。得到图 6：

（代码参考作品附件附录 1）

```
Out[60]: 852                                    [J11，健身，爱好者，活动场所]
         1095    [K2区，桥头，市场，那有，一栋，房子，一楼，沿街，门面，违规，搭建，棚子]
         158                              [A市，泉塘，街道，未来，蜂巢，下雨，漏水]
         1572                 [M12县，古阳镇，农村，危房改造，规避，招标，违法，分包]
         1230                                  [K市，购房，补贴，兑现]
                                          ...
         8802                    [请问，2015年，二孩，开放，新，政策，具备，条件]
         8551                      [D市，南华大学，附一，医院，导致，上半身，瘫痪]
         8667                          [A7县，超生，天文数字，罚款，减少]
         8348                               [M2县，黑诊所，取缔，难]
         8718                        [浅谈，卫生院，垄断，基药，采购，弊端]
         Name: message, Length: 8905, dtype: object
```

图 6 去停用词

此时得到的数据就是比较干净的数据，为后面算法提供了较干净的数据准备，在建立模型之前，还需要将文本型的数据转化为数字型的数据以便计算机解读和计算，为此，下面介绍了一种算法，TF-IDF 算法，通过该算法可获取文本数据的 TF-IDF 权值向量。下面介绍算法。

二、TF-IDF 算法获取文本权值向量

TF-IDF（term frequency－inverse document frequency，词频-逆向文件频率)是一种用于信息检索(information retrieval)与文本挖掘(text mining)的常用加权技术。TF-IDF 是一种统计方法，用于评估一个词对一个文件集或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，但同时会随着它在预料库中出现的频率成反比下降。

TF 是词频(Term Frequency)，词频（TF）表示词条（关键字）在文本中出现的频率。这个数字通常会被归一化(一般是词频除以文章总词数)，以防止它偏向长的文件。公式如下：

$$\mathrm{tf}_{ij}=\frac{n_{i,j}}{\sum_k n_{k,j}} \qquad \text{（公式 5）}$$

可得：

$$TF_{\omega}=\frac{\text{在某一类中词条}\omega\text{出现的次数}}{\text{该类中所有的词条数目}} \qquad \text{（公式 6）}$$

其中 $n_{i,j}$ 是该词在文件 d_j 中出现的次数，分母则是文件 d_j 中所有词汇出现的次数总和。

IDF 是逆向文件频率(Inverse Document Frequency)；

逆向文件频率 (IDF)：某一特定词语的 IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。如果包含词条 t 的文档越少，IDF 越大，则说明词条具有很好的类别区分能力。公式如下：

$$\mathrm{idf}_i=\log\frac{|D|}{|\{j:t_i\in d_j\}|} \qquad \text{（公式 7）}$$

其中，|D|是语料库中的文件总数。|{j:t_i∈dj}|表示包含词语 t_i 的文件数目

（即 $n_{i,j}\neq 0$ 的文件数目）。如果该词语不在语料库中，就会导致分母为零，因此一般情况下使用 $1+|\{j:t_i\in d_j\}|$，即可得下面公式：

$$IDF=\log\left(\frac{\text{预料库的文档总数}}{\text{包含词条}\omega\text{的文档数}+1}\right) \qquad \text{（公式 8）}$$

分母+1 是为了保证分母不等于 0。

TF-IDF 实质上就是 TF*IDF，某一特定文件内的高词语频率，以及该词语在整个文件集合中的低文件频率，可以产生出高权重的 TF-IDF。因此，TF-IDF 倾向于过滤掉常见的词语，保留重要的词语。

TF-IDF=TF*IDF （公式 9）

实际分析得出 TF-IDF 值与一个词再文本中出现的次数正正比，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本本的关键词。

通过该算法，可以得到关键词的 TF-IDF 权值向量，从而得到一个稀疏矩阵，如下图所示：（代码参考作品附件附录 2）

```
Out[7]: array([[0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.],
               ...,
               [0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.],
               [0., 0., 0., ..., 0., 0., 0.]])
```

图 7 关键词 TF-IDF 权值向量

由上图（图 8）可得到预处理过后的文本的 TF-IDF 的权值向量，矩阵里面的数字并非全都为 0，上图得到的是一个稀疏矩阵，由于关键词较多，所以每一行非 0 值的个数较少较稀疏。得到了文本关键字的 TF-IDF 后，便可用上面建立的多项式朴素贝叶斯模型对文本进行分类。

三、求解结果

对预处理后的数据进行分组，共有 8095 条数据，分为训练组和测试组，比例为 4：1。训练组有 7124 条，测试组有 1781 条：如表 2 所示：

留言主题数据（训练）	留言主题数据测试	标签（训练）	标签（测试）
7124	7124	1781	1781

表 2 预处理后数据分组

最终得到解如图 8 所示：（代码参考作品附件附录 2）

```
D:\anaconda\lib\site-packages\ipykernel_launcher.py:26: ParserWarning: Fa
upport regex separators (separators > 1 char and different from '\s+' are
gine='python'.
```

准确度： 0.7804604154969118

	precision	recall	f1-score	support
1	0.68	0.92	0.78	402
2	0.95	0.69	0.80	177
3	0.98	0.48	0.64	115
4	0.91	0.83	0.87	313
5	0.69	0.94	0.80	369
6	0.86	0.65	0.74	221
7	0.96	0.50	0.66	184
accuracy			0.78	1781
macro avg	0.86	0.72	0.76	1781
weighted avg	0.82	0.78	0.77	1781

图 8 评价精确度

模型准确率为 0.78，各类别的召回率和 F-Score 如表 3 所示：

一级标签	召回率	F-score
1	0.92	0.78
2	0.69	0.8
3	0.48	0.64
4	0.83	0.87
5	0.94	0.8
6	0.65	0.74
7	0.5	0.66

表 3 各类别召回率和 F-Score

从模型的评分看来准确率达到了 78%，而各类的 F-Score 评分也在都在 60 以上，召回率从表 3 也可看出模型的实用性。

4.2 问题二：热度评价指标模型的建立与求解

4.2.1 模型思想

对于第二问，我们知道某一时段内群众集中反映的某一问题可称为热点问题，因此评价此问题是否热门，跟它的反映数目有直接关系，如若反映的问题频率最高，显然该问题是民众所关注的，因此可称为热门问题，但对于热门问题的是与否也成为了民众所关注的，为此下面给出了合理的热度评价指标，以便于评价该问题的热度，便于被民众所关注到。

要建立热度评价指标模型，我们需要对海量数据进行统计，对相似的文本进行合理的归类，下面给出利用聚类分析进行文本聚类的框架图：

图 9 文本聚类分析框架图

我们借助第一题的思路，对数据进行预处理，包括数据的读取、文本的 jieba 分词，以及去停用词的处理，对样本数据进行处理后通过 TF-IDF 算法获取文本权值向量，通过计算机进行聚类分析，将数据合理归类，即可反映相似问题的频率，便于模型评价该问题是否为热门问题。

4.2.2 模型建立

热度评价指标模型

我们不难理解，问题的热度与问题的反映频率成正相关关系。随着问题反映频率的增加，该问题的热度也就越高，而当民众本身关心的问题外的问题总数减少时，更能突出民众所关心的问题。假设问题热度为 H，问题反映频数为 K，问题总数为 N，则有 $H \propto K$，而 $H \propto \frac{1}{N-K}$，N-K 表示民众本身关心的问题外的问题总数。

因此给出热度指标公式：

$$H_i = \lambda \frac{K_i}{N - K_i} \qquad \text{（公式 10）}$$

其中，H_i 表示第 i 个问题的热度，K_i 表示第 i 问题反映频数，λ 为比例系数，N 表示问题总数（i=1，2，3，4….）

H_i 的值越大，则说明该问题的热度越高，然而若只知道某个问题的 H 值，也很难得到评价第 i 的问题是否为热门的标准。为了解决这个问题，我们给出了评价 H_i 值是否为热门的指标，原理如下：

我们不难理解 K_i 取最大值，H_i 最大，因此可直接说明该问题为热门话题，设问题中第 j 类问题的频数最多为 K(max)、第 g 类问题的频数第二多为 K(max-1)、第 f 类问题的频数最少为 K(min)，由于 K（max）与 K（max-1）偶尔会差距过大，因此此过程不考虑 K（max），将 K=K（max-1）带入公式 10 可得到第 g 类问题的热度为 H_g，将 K=K（min）带入公式 10 可得到第 f 类问题的热度为 H_f。则做差为：

$$\Delta = H_g - H_f \qquad \text{（公式 11）}$$

我们将上述的 Δ 三等分，依次为

$$[0, \frac{H_g - H_f}{3}),\ [\frac{H_g - H_f}{3}, \frac{2(H_g - H_f)}{3}),\ [\frac{2(H_g - H_f)}{3},\ H_g - H_f]。$$

我们可知当 H_i 等于 0 或者接近 0 时，该问题的热度基本为 0，而 $H_j>H_g>H_g-H_f$，且上面上个区间的数值依次增大，对于热度来说该问题也就越热门。根据这个关系，将上面的数值分别设为冷门问题、一般问题（介于冷门问题和热门问题之间）、热门话问的评价标准，如下图所示：

热度	冷门问题	一般问题	热门问题
H_i	$[0,\frac{H_g-H_f}{3})$	$[\frac{H_g-H_f}{3},\frac{2(H_g-H_f)}{3})$	$[\frac{2(H_g-H_f)}{3}, H_g-H_f]$

表 4 热度指标表

因此我们给出了热度指标评价模型可计算热度，给出了评价该话题是否为热门话题的指标。便可求出排名前 5 的热点问题。由于 H_j、$H_g>\Delta$，因此默认总数最多和第二多的问题为热门问题。

4.2.3 模型的求解

一、K-means 聚类算法

对于模型的求解，我们需要获得每类问题的总数便可带入公式求解，对应指标评价是否为热门话题。为了从海量数据中获取相似文本的总数，我们采用了 K-means 聚类算法对预处理好的数据进行聚类。下面给出 K-means 聚类算法原理。

K-means 是基于给定的聚类目标函数，算法采用迭代更新的方法，每一次的迭代过程都是向目标函数减小的方向进行，最终聚类结果使得目标函数去到极小值，达到较好的分类过程。原始的 K-means 算法首先随机选取 k 个点作为初始聚类中心，然后计算各个数据对象到各个聚类中心的距离，把数据对象归到离它最近的那个聚类中心所在的类；调整后的新类计算新的聚类中心，如果相邻两次的聚类中心没有再发生任何变化；说明数据对象调整结束，聚类准则函数已经收敛。在每次迭代中都要考察每个样本的分类是否正确，若不正确，就要调整。在全部数据调整完后，再修改聚类中心，进入下一次迭代。如果在下一次迭代算法中，所有的数据对象被正确分类，则不会有调整，聚类中心也不会有任何变化，这标志着聚类准则函数已经收敛，算法结束。如下所示：

(1)给定大小为 n 的数据集，令 A=1，选取 k 个初始聚类中心 Zj(A),j=1,2,3，…，k，A 代表不同迭代轮数的聚类中心；

(2)计算每个样本数据对象与聚合中心的距离 D(xi,Zj(A))，i=1,2,3,…，并分类；

(3)令 A=A+1，计算新的聚类中心；

(4)判断：若|f(A+1)-f(A)|<theta(f 收敛)或者对象无类别变化，则算法结束，否则，A=A+1,返回(2)步。

本文中我们采用该算法的原理对文本进行聚类。

二、求解结果

通过 python 编写代码获得分类结果如图 10 所示（代码参考作品附件附录 4）：

```
    cluster_menmbers_list.append(menmbers_list)
# print cluster_menmbers_list

#聚类结果
for i in range(len(cluster_menmbers_list)):
    print ("============================")
    for j in range(len(cluster_menmbers_list[i])):
        a = cluster_menmbers_list[i][j]
        print (corpus[a])
    print(j+1)
```

```
============================
A市 楼盘 供暖 一事
A市 电工证 一事
K1区 学校食堂
A市 居住 凭证 牌 一事
A市 学历 落户 稳定 就业 凭证 指
A市 买房 居住 创业 条件
6
============================
A市 圣毅园 项目 出路
A市 经开区 泉星 二小 田径场 课余时间 居民 开放
项目 资金 A市 创业
A市 经开区 城北 污水 厂内 项目 消防栓 布局 不合理
A市 经开区 泉星 公园 项目 规划 需 优化
```

图 10 运行结果

由于运行的结果种类较多，因此我们将结果导入到文件中，下面给出文件中的部分数据：

```
1(1) - 记事本
文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)
A市 房产 不动产 权证 未满 四年 交易
小孩 转入 A市 就读 小学
19
============================
A市 涉外经济 学院 组织 学生 外出 打工
强制 封闭 煤球 厂 国家 政策
中南林 科大 强制 拆除 学生 空调 回购 价格 不合理
A市 商贸 旅游 学院 白田 宿舍 摆摊
A市 学院 拖欠 农民工 工资 三年 校领导 答应 付款 钱 踢皮球
A市 涉外经济 学院 强制 学生 实习
A市 学生 中考 提前 科 成绩 C
A市 商贸 旅游 职业 技术 学院 强制 学生 实习
A市 涉外经济 学院 寒假 过年 期间 组织 学生 工厂 工作
A市 电建 星 湖湾 强制 业主 收房
A市 经济 学院 寒假 过年 期间 组织 学生 工厂 工作
A市 经济 学院 组织 学生 外出 打工
A市 经济 学院 强制 学生 实习
A市 经济 学院 强制 学生 外出 实习
A市 经济 学院 体育 学院 变相 强制 实习
15
============================
承办 A市 58 车 贷案 警官 应 跟进 关注 留言
严惩 A市 58 车贷 特大 集资 诈骗案 保护伞
过问 A市 58 车贷 案件 进展 情况
恳请 A市 经侦 公正 办理 58 车贷 案件 受害人 公道
A市 聚利网 诈骗 派出所 立案 已超 月 依然 进展
A市 58 车贷 案件 臭名远扬
A市 58 车贷 老板 跑 路 美国 经侦 拖延 办案
A市 58 车贷 恶性 退出 立案 近半年 发 案情 通报
A市 58 车贷 恶性 退出 案件 发布 案情 进展 通报
9
============================
A市 地铁 号线 何时能 开工 建设
A市 加快 招商引资 有何
加快 A市 合规 网约车 治理
A市 加快 智能 制造业 招商引资 步伐
A市 加快 体育 强省 国家 战略 步伐
加快 A市 文化 体育 强省 步伐
A市 加快 流动 摊位 试点
A市 加快 招商引资 步伐 有何
A市 加快 公共场所 禁烟 有何
A市 加快 控制 物价上涨 过快 步伐
9
============================
A市 805 路车 改道 走 木莲 中路
A市 805 路 改道 走 木莲 中路
A市 机关 事务局 市民 留 一条 人行道
A市 地铁 五号线 围挡 区域 占用 人行道
A市 中路 主干道 旁 废品 回收站 影响 交通
A市 805 路 公交车 改道 走 木莲 中路
A市 木莲 路 新 姚路 交界处 红绿灯
A市 木莲 中路 百米 人行道 路 狭窄
A市 倾城 小区 规划 人行道 导致 行人 常年 走 马路
A市 万家 丽 北路 楚龙路 十字路口 人行道 围挡 影响 师生 出行
A市 805 路 公交车 改道 走 木莲 中路
11
============================
A市 五矿 万境 K9县 消防安全 隐患
A市 五矿 万境 K9县 房子 墙壁 开裂
A市 五矿 万境 K9县 交房 诸多
A市 五矿 万境 K9县 房屋 质量
A市 五矿 万境 K9县 负 一楼 面积 缩水
A市 五矿 万境 水岸 三期 违规 建设 垃圾站
```

图 11 部分文件数据

将各类数据分类好放入文件里，通过数据清洗我们可得到类别中数目由多到少的问题数目，如表 5 所示：（总数 N=4327）

问题数目	问题描述
21 条	A 市人才购房补贴政策
16 条	A 市公积金贷款问题
13 条	A 市 58 车贷诈骗案
12 条	加快 A 市中心城市建设
10 条	学校强制学生去定点企业实习
……	……
1 条	……

表 5 问题数目表

取比例系数 λ=1000，可得到：

$$\Delta=H_g-H_f=1000\times\left(\frac{16}{4327-1}-\frac{1}{4327-1}\right)\approx 3.4367$$

则计算可得到：

热度	冷门问题	一般问题	热门问题
H_i	[0 , 1.1557)	[1.1557 , 2.3113)	[2.3113 , 3.4367]

表 6 热度指标表

我们对问题数目量前五以及随机抽取其他数目的问题进行计算，依次代入公式 10 可得到结果如表 7：

问题数目	问题描述	H_i	评价结果
21 条	A 市人才购房补贴政策		热门问题
16 条	A 市公积金贷款问题		热门问题
13 条	A 市 58 车贷诈骗案	3.005	热门问题
12 条	加快 A 市中心城市建设	2.774	热门问题
10 条	学校强制学生去定点企业实习	2.317	热门问题
……	……	……	……
6 条	A 市五矿万境 K9 县房屋出现质量问题	1.3870	一般问题
……	……	……	……
2 条	A 市公交公司的清扫问题	0.462	冷门问题
1 条	……	……	……

表 7 问题评价结果

通过该表便得出排名前 5 的热点问题，结合附件 3 中的数据制作并保存为文件“热点问题表.xls”。并给出相应热点问题对应的留言信息，保存为“热点问题留言明细表.xls”。

4.3 问题三：答复意见质量方案的建立

4.3.1 模型思想

对于答复意见质量的评价，本文考虑了两个方面的因素，一是答复意见的主题与留言主题相关性如何，二是答复意见的情感得分如何，故该评价模型的建立需要用到文本相似度的计算和对文本进行情感分析得到情感得分。下面给出流程图：

图 12 流程图

通过上面的流程图的思想来建立答复意见质量评价方案。

4.3.2 模型建立

一、答复意见质量评价方案

基于我们的建模思想，考虑到答复意见的相关性、完整性和可解释性，参考的参数由文本相似度和情感得分结合得到，相似度利用了余弦相似度计算算法，求答复意见的文本主题与附件 4 提供的留言主题进行相似度的计算，求出的数值在 0 和 1 之间为正值，而对一般的评价得到情感得分的范围在一个有正有负的区间内，在本文里的答复与以往的评价不同，对于问题进行答复要求态度和语气诚恳作为前提，因此很少见到消极的答复内容，计算结果只有几个是负数，故本文得到了情感得分后剔除几个特殊的负数，保证所有的情感得分大于 0。取 W(ave)为所有情感得分的平均值，不妨设余弦相似度的值分为三个区间，分为：[0 ， 0.5），0.5 以及(0.5 ， 1]。

不妨设：

余弦相似度 $\cos\theta \in (0, 0.5)$ 时，文本的相关性较差，记为 - 号

余弦相似度 $\cos\theta = 0.5$ 时，文本的相关性位于差和强之间，记为 0

余弦相似度 $\cos\theta \in (0.5, 1)$ 时，文本的相关性较强，记为 + 号

情感得分 W< W(ave)时，答复意见较负面，即为 - 号
情感得分 W= W(ave)时，答复意见位于正面和负面之间，记为 0
情感得分 W> W(ave)时，答复意见较正面，记为 + 号
有了上面的定义，建立答复意见质量评价方案：

$$(\cos\theta,\ W) \in \begin{cases} (+,+)，\text{质量好} \\ (-,-)，\text{质量差} \\ (+,-)\text{或}(-,+)\text{或}(0,0)，\text{质量一般} \\ (0,+)\text{或}(+,0)，\text{质量较好} \\ (0,-)\text{或}(-,0)，\text{质量较差} \end{cases}$$

上面模型表示，若二者取+号，则该答复意见质量好，若两者取-号，则该答

复意见质量差，若两者异号或者都为 0，则表示不好也不坏，答复意见质量一般，若一个为 0，一个取+号，则该答复意见质量较好，若一个为 0，一个取-号，则该答复意见质量较差。通过该方案，计算答复意见的文本情感得分及余弦相似度的数值范围，便可定量的评价答复意见的质量。

要达到上面的评价结果，首先选需要对答复意见进行分析提取文本主题，因此构建 LDA 主题模型来求解，下面介绍 LDA 主题模型的建立。

二、LDA 主题模型

LDA 是一种文档主题生成模型，也可称为一个三层贝叶斯概率模型，包含词、主题和文档三层结构。LDA 是一种非监督机器学习技术，可以采用识别大规模文档集合或预料库中潜藏的主题信息。它采用了词袋的方法，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。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。

对于语料库中的每篇文档，LDA 定义了如下生成过程：

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。

语料库中的每一篇文档与 T（通过反复试验等方法事先给定）个主题的一个多项分布相对应，将该多项分布记为。每个主题又与词汇表中的 V 个单词的一个多项分布相对应，将这个多项分布记为 beta。

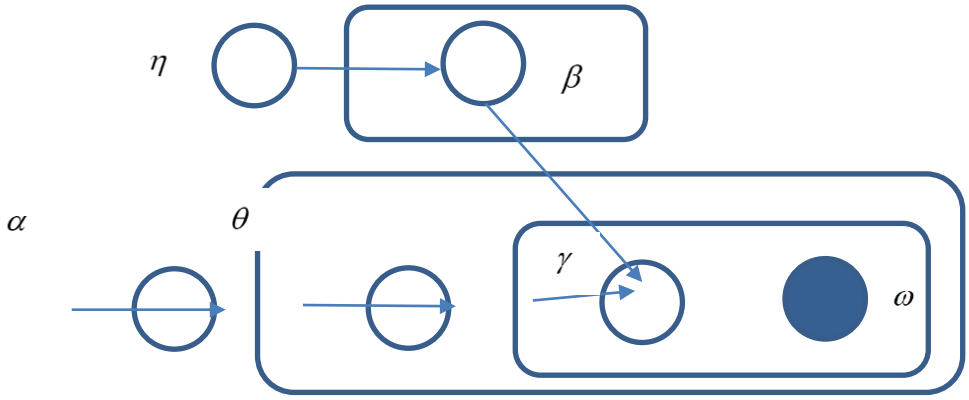

图 13 LDA 模型图表示

通过 LDA 主题模型的求解，便可得到答复意见的文本主题。

4.3.3 模型的求解

首先进行数据的预处理，通过第一问和第二问数据预处理的流程，很容易我们就能将附件 4 的答复意见进行预处理，包括文本去重、分词、去停用词得到干净的数据，(代码参考附件 5)然后计算情感词分数、程度副词的计算、否定词的计算之后，自定义分值计算函数，对所有的答复意见进行评分，便得到了文本数据的情感分数 W。如下图所示：(代码参考作品附件附录 4)

0	41.408096
1	25.427688
2	47.279369
3	49.147420
4	17.884715
5	22.664629
6	25.808803
7	98.530742
8	85.728831
9	22.560480
10	51.469673
11	51.493609
12	18.321933
13	7.023969
14	14.980929
15	68.179969
16	8.091125
17	24.997050
18	0.578455
19	20.672543
20	23.285822
21	58.441177
22	1.268424
23	21.841984
24	9.557643

2789	19.210415
2790	16.532666
2791	4.861719
2792	22.748240
2793	12.922087
2794	3.781577
2795	69.539961
2796	20.241308
2797	15.313357
2798	187.391543
2799	32.969130
2800	-0.661624
2803	18.989304
2804	8.277025
2805	27.124863
2806	58.004979
2807	44.106414
2810	35.100403
2811	1.859402
2812	0.000000
2813	70.209683
2814	62.955817
2815	22.386069

Name: 答复意见, Length: 2746, dtype: float64

图 14 部分情感分数图

对剔除特殊数据后的所有数据取平均值，通过计算可得到 W(ave)=38.1695

对于上面处理好的数据，建立 LDA 主题模型，读取答复意见的所有的数据，可以导出每条文本数据的主题。与附件 4 给出了留言主题做文本相似度的计算，这里我们采用余弦相似度计算算法求解。下面给出导出文档主题的部分结果：(代码参考附录 6)

```
#负面评论
pos_com = pd.read_csv('neg_com.csv', header=None, index_col=0)    # 读取预处理好后的数据

pos_com.columns = ['comment']  # 更改列名称
mid = list(pos_com['comment'].str.split(' '))    # 将评论文本分词（基于空格）
dictionary = Dictionary(mid)                     # 生成词典
bow = [dictionary.doc2bow(comment) for comment in mid]    # 将文档转成数值型预料库
pos_model = LdaModel(corpus=bow, id2word=dictionary, num_topics=3)    # 构建LDA主题模型
print(pos_model.print_topic(0))    # 打印主题
print(pos_model.print_topic(1))
print(pos_model.print_topic(2))
```

```
0.058*"  " + 0.009*"  " + 0.008*"鉴定" + 0.005*"患者" + 0.004*"学校" + 0.004*"转市" + 0.004*"驾驶证" + 0.004
*"我院" + 0.004*"业主" + 0.003*"机构"
0.027*"  " + 0.026*"  " + 0.012*"学校" + 0.006*"转学" + 0.005*"学生" + 0.005*"夜间" + 0.005*"转入" + 0.005*"施
工" + 0.005*"医保" + 0.004*"申请"
0.039*"  " + 0.036*"  " + 0.011*"医院" + 0.010*"报销" + 0.006*"学校" + 0.006*"医保" + 0.005*"感谢您" + 0.004
*"已转" + 0.004*"我局" + 0.004*"办理"
```

图 15 随机打印模型提取的文档主题

余弦相似度计算算法

余弦相似度是用空间向量中的两个向量夹角的余弦值大小，来衡量其个体间差异。相似度的值越小，则向量间相似程度越小，反之越大则相似程度越大。当余弦值接近于 1 时，表明向量夹角接近于 0，即两个向量越相似。

下面给出向量空间余弦相似度的详细推理过程。

如图所示，三角形 a，b 边的夹角的余弦定理公式为：

$$\cos(\theta)=\frac{a^2+b^2-c^2}{2ab} \quad （公式 12）$$

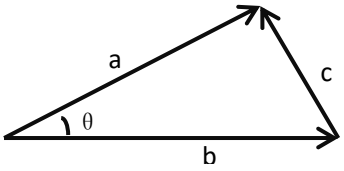

图 14 三角形

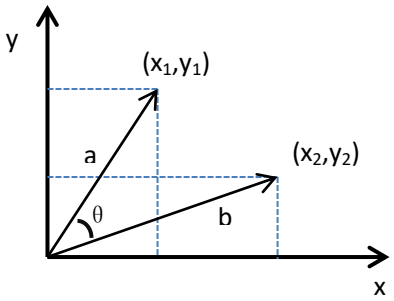

图 15 向量夹角

假设向量 a=(x_1,y_1)，向量 b=(x_2,y_2)，以此可将余弦公式改写成以下形式：

$$\cos\theta=\frac{a\bullet b}{\|a\|\times\|b\|}$$

$$=\frac{(x_1,y_1)\bullet(x_2,y_2)}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}}$$

$$=\frac{x_1x_2+y_1y_2}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}} \quad （公式 13）$$

通过拓展，有类比原理可得当 a，b 为 n 维向量的夹角余弦值为：

$$\cos\theta=\frac{\sum_{i=1}^{n}(x_i\times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}\times\sqrt{\sum_{i=1}^{n}(y_i)^2}} \quad （公式 14）$$

用余弦计算文本相似度原理：首先按照问题一的方法，对文本进行分词，去停用词及计算关键词词频，通过 IF-IDF 算法得出文本的词频向量，此时，问题就转变成了对向量相似程度的计算。将向量代入公式 3 即可求出文本的相似度 $\cos\theta$。如图下所示下面给出图 15 打印出三个主题与附件给出的留言主题做相似度计算：(代码参考作品文件附录 7)

```
vec1,vec2=get_word_vector(s1,s2)
dist1=cos_dist(vec1,vec2)
print(dist1)
```

```
[1. 0. 1. 1. 2. 0. 1. 0. 0. 0. 1. 1. 1. 1.]
[0. 1. 0. 0. 2. 1. 0. 1. 1. 1. 0. 0. 0. 0.]
0.3849001794597505
```

图 16 部分计算结果

随机抽取几个例子做测试并给出答复意见质量结果，如下表 8：
W(ave)=38.1695

情感分数	附件 4 给出的三条留言主题	（cosθ，W）	质量结果显示
41.4081	投诉 F9 市桃林供电所负责人为难小微企业	(0.3849，41.4081)	（-，+）质量一般
25.4277	反映小孩转学到 C 市的问题	（0.4003，25.4277）	（-，-）质量差
47.2794	咨询居民医保地区医院治疗报销问题	（0.2103，47.2794)	（+，+）质量好

表 8 测试结果显示

该模型可以实现对答复意见的评价结果，如上图 16 所示。

5 模型的评价与推广

5.1 模型的优点

本文利用了多项式朴素贝叶斯模型进行了文本的分类，它源于古典数学理论，有稳定的分类效率。能较好处理文本的多分类任务，适合增量式训练，尤其是数据量超出内存时，我们可以一批批的取增量训练，有利于提高精度。该算法比较简单，对缺失的数据不太敏感，很适合本文研究的文本分类问题。能给出较高的精度，可见模型的实用性。

本文还根据了 K-means 聚类的结果进行分析建立了热度指标评价模型，算法和模型容易理解、聚类效果良好、模型的计算量少、计算时间短以及速度快。给出了定性的热门评价指标，能定性简便的评价这类问题是否为热点问题。适用于数据量较多，类别较多的数据进行热度评价。

本文还采用 LDA 主题模型对文本提取了文本主题，可以把模糊不清的思想转化为直观的具有良好结构的模型，特别适用于多变量，关系复杂而结构不清晰的系统分析中，如本文的数据。可用于方案的排序，对文本进行分析可提取文本主题。提供了余弦相似度计算的算法给出了主题之间的相关性程度，能较好的解决了本文中所建立的答复意见质量评价方案中的计算问题。而建立的答复意见质量评价方案没有考虑数值的大小，只考虑范围，大大提高了批价的速率。

5.2 模型的局限性

利用了多项式朴素贝叶斯模型进行求解，理论上，它与其他分类方法相比具有最小的误差率。但实际上并非如此，正因为朴素贝叶斯模型假设属性之间相互独立，这个假设在实际应用中往往时不成立的，在属性个数比较多或者属性之间相关性较大时，分类的效果不佳，由本文中对量值化为 3 的类别可看出，召回率较低。二属性相关性较小时，朴素贝叶斯性能最为良好。对于这一点，具有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进。由于我们时通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。

本文用到的 K-means 聚类算法对异常数据敏感，需要提前确定 k 值，这就造成了聚类时存在着一定的误差。本文建立的热度指标模型没有考虑问题在点赞数和反对数，因为点赞数和反对数的多少也会影响问题是否成为热点问题。

采用的 LDA 主题模型的主观性强，在一定程度上系统各要素之间的关系依赖于人们的经验，客观性不足，计算结果存在着理解上的差异，争议性大。建立的答复意见质量评价方案评价过于简单，没有考虑数值的大小只考虑数值的正负性。

6 参考文献

[1] 李航.统计学习方法.北京：清华大学出版社，2012

[2] 于游，付钰，吴晓平.中文文本分类方法综述[J].网络与信息安全学报，2019，5(5):1-8

[3] 汪岿,刘柏嵩.文本分类研究综述[J].数据通信，2019(3):37-47

[4] 隗中杰.文本分类中 TF-IDF 权重计算方法的改进[J].软件导服，2018，17(12):39-42