基于机器学习算法的政务信息挖掘应用

——让网络问政“问”出好效果

摘要

随着各类网络问政平台的兴起，如何从不断攀升的各类社情民意相关的文本数据中，归类、整合，并提取出有效信息，成为了“智慧政务”中要解决的关键问题。本文利用机器学习中的方法有效且准确地实现了文本分类，通过两类模型的建立，提出了多指标热度评价体系，进而对留言热度进行排名，本文还从多角度挖掘文本信息，建立了完整的对留言答复意见的评价方案。

针对问题一，建立基于有监督学习的一级分类模型。首先对所有的文本进行预处理，包括无关信息剔除，中文分词，删除停用词处理，然后取每一类的三分之二文本作为训练集，三分之一的文本作为测试集，构建主题和留言分开的词向量矩阵。利用**改进的 TF-IDF 算法**对特征向量作进一步地特征处理和降维处理，最终获得 6144×52711 的训练集矩阵和 3066×52711 的测试集矩阵。在已知分类的前提下，本题使用**基于特征加权的改进贝叶斯算法**，经检验，其 F-score 得分为 0.806。

对于问题二，本题建立了舆情事件预警模型和热门关注提取模型进行热点问题挖掘，其中使用了**基于余弦相似度的谱聚类算法**，通过动态调整簇中心数目，经过多次调式得到较为稳定的**轮廓系数**为 0.7 的最大的两个蔟，这两个蔟可认为是集中爆发事件，相关部门急需处理；另一个模型则提取了点赞数和反对数加起来最多的前三类话题；最后定义了四个热度评价指标(相关留言篇数、点赞数、反对数、问题的持续时间)对提取的这五类话题进行热度评价以及排名。

对于问题三，从答复意见的相关性，及时性，完成度，直观性，可理解性多个角度进行评价，通过构建层次的树状目标结构体系来计算综合得分。其中利用**余弦相似度算法**实现答复意见与留言之间的相关性评价，计算了答复与留言的时间差，并进行了可视化。

关键词：**改进的 TF-IDF 算法；特征加权的改进贝叶斯算法；余弦相似度；谱聚类算法；轮廓系数**

目录

1. 问题的提出与分析

1.1. 问题背景

随着互联网的快速发展和普及，网络问政平台已成为党和政府了解民意、汇聚民智、凝聚民气，科学决策的重要渠道，也是公众行使知情权、参与权、表达权和监督权的重要途径，而随着各类社情民意相关的文本数据量不断攀升，若仅依靠以往的人工进行留言划分和热点整理，则相关部门的工作量将会非常庞大。因此，在大数据、云计算、人工智能等技术高速发展的时代，将基于自然语言处理技术的智慧政务系统运用到社会治理方法中，已经是社会治理创新发展的新趋势，对提升政府的管理水平和施政效率具有极大的推动作用，本文将利用所提供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，运用自然语言处理和文本挖掘的方法，建立起关于留言内容的一级标签分类模型；定义合理的热度评价指标，评判出的排名前 5 的热点问题；从多角度对留言的答复意见的质量给出一套合理、可行的评价方案。

1.2. 问题重述

针对问题一，为了解决依靠人工经验对留言进行分类所存在的工作量大、效率低，且差错率高等问题，本问将利用所给数据，建立出立关于留言内容的一级标签分类模型，并使用 F-Score 对分类方法进行评价。

针对问题二，及时发现热点问题，有助于相关部门有针对性地处理，提升服务效率，本问将利用附件 3 所给的留言进行归类，定义合理的热度评价指标，并且评定出排名前 5 的热点问题。

针对问题三，利用附件 4 提供的相关部门对留言的答复意见，从相关性、完整性、可解释性等多个角度，对答复意见的质量建立出一套合理、可行的评价方案。

1.3. 问题分析

1.3.1. 问题一

观察问题一，群众留言的信息量巨大，为了方便后续对群众留言内容以及标签进行分类处理，故而采用了有监督学习的方法，建立基于特征加权朴素的贝叶斯模型。我们首先需要对所有的数据进行文本预处理，将已有的文本数据转化为词向量矩阵，后对预处理过的文本进行特征处理，运用改进后的 TF-IDF 算法对文本数据进行特征选择并加权，从而选择出群众留言中的与相关问题分类标签，观察他们的关联性，接着利用改进的朴素贝叶斯算法，又称朴算贝叶斯算法，对其进行文本分类，在得出分类结果之后，我们仍需通过 F-Score 对分类方法进行评价，观察模型是否合理，从而解决问题一。

1.3.2. 问题二

观察问题二，附件三提供的数据相比附件二来说，缺少已有的标签分类且数据量更为庞大，采用人工标注的方法显然欠佳，因此本文采用无监督学习的方法对附件三的群众留言进行分类，对于短时间内收到的多条类似反馈，本文使用了一种基于改进相似度计算的文本聚类方法，建立出舆情事件预警模型。但上述模型忽略了具有篇数较少，但是点赞数高、持续时间长等特点的热点事件，为了解决此问题，本文建立了热门关注事件提取模型。将两个模型结合起来，通过资料查阅以及分析，本文定义出多个合理的热度评价指标，并且赋予每个指标一定的权重，建立出多指标热度评价体系，共同评价出排名前 5 的热点问题。

1.3.3. 问题三

为了解决问题三，我们选择建立任务型对话系统的回复质量评价方案，利用附件 4 提供的相关部门对留言的答复意见，从相关性、完整性、可解释性等多个方面的具体指标，对任务型对话系统的回复质量进行定性定量的分析评价，具体指标中我们选择了与该问题的已有回复相似性和回复的时间差来判断该评价方案是否合

理、可行。本题利用余弦相似度算法来实现了答复与留言的相关性，计算了答复与留言的时间差，并进行了可视化最后通过构建层次的树状目标结构体系来进行评分。

2. 符号说明

符号	说明
t	特定的某个特征词
C	特指某一个类的群众留言
m	类 C 的群众留言文本内包含特征词 t 的留言数
o	类 C 的群众留言文本内不包含特征词 t 的留言数
k	非类 C 的群众留言文本包含特征词 t 的留言数
q	非类 C 的群众留言文本不包含特征词 t 的留言数
w	权重
TF	特征词 t 在类 C 的群众留言文本中出现的次数
A	类 C 群众留言的一组标签
B	一条群众留言中的所以词语的集合
b_i	代表了一条群众留言中的某个词语
P_i	第 i 类的查准率
R_i	第 i 类的查全率
W	文本相似度矩阵
S	样本的相似度矩阵
M	无向赋权图的邻接矩阵
D	n×n 的度矩阵
L	拉普拉斯矩阵
n、k_1、i、j、N	正整数
$\mathbf{f}$	特征向量
F	特征向量 f 构成的特征矩阵

d_i	第 i 个文本的特征向量
d_j	第 j 个文本的特征向量
T	阈值
G	无向赋权图
V	留言数据集中点的集合
E	留言数据集中点与点间连成的边的集合
w_{ij}	点 v_i 和点 v_j 之间的权重
v_i	第 i 个点
D_i	点 v_i 对应的度
K	轮廓系数
w_i	各项指标的权重
V_r	群众留言文本的向量表示
$V_{r'}$	相关部门对留言的答复意见文本的向量表示
SR	有参考度量

3. 模型建立与问题解答

3.1. 问题一

3.1.1. 模型的建立与求解

为了实现对群众留言这一块文本进行系统性分类，我们采用的是朴算贝叶斯算法。此题中我们使用到了有监督学习的概念，即从给定的训练数据集中学习到一个模型参数，当新的数据给出时可以根据这个参数去预测结果，利用机器学习的思想建立一级标签分类系统模型——特征加权朴素贝叶斯模型。

本题的解题步骤分为四个部分，如下图 1。

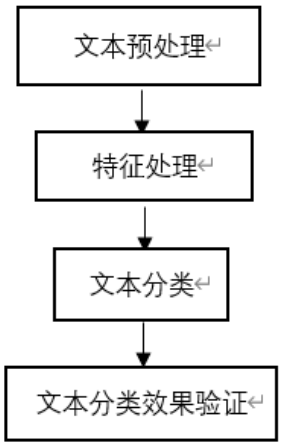

图 1 问题一解题思路

1、文本预处理

本题我们构造了特征加权朴素贝叶斯模型。在采用朴算贝叶斯算法计算前，我们需要对文本进行文本预处理，具体操作步骤可见图 2。在中文语料库的文本预处理主要包括去除文本标记、删除重复语词、去除停用词、中文分词等操作。首先在进行的是对文本的处理工作，通过 Python 软件对群众留言的文本信息进行阅历，将文本加载到内存中，便于计算机识别，而后进行文本预处理，我们在这里所用到是结巴分词工具开发包，用其来进行中文分词操作。在本题中，我们对其进行了删除文本标记、中文分词并标记词性和删除停用词等处理。

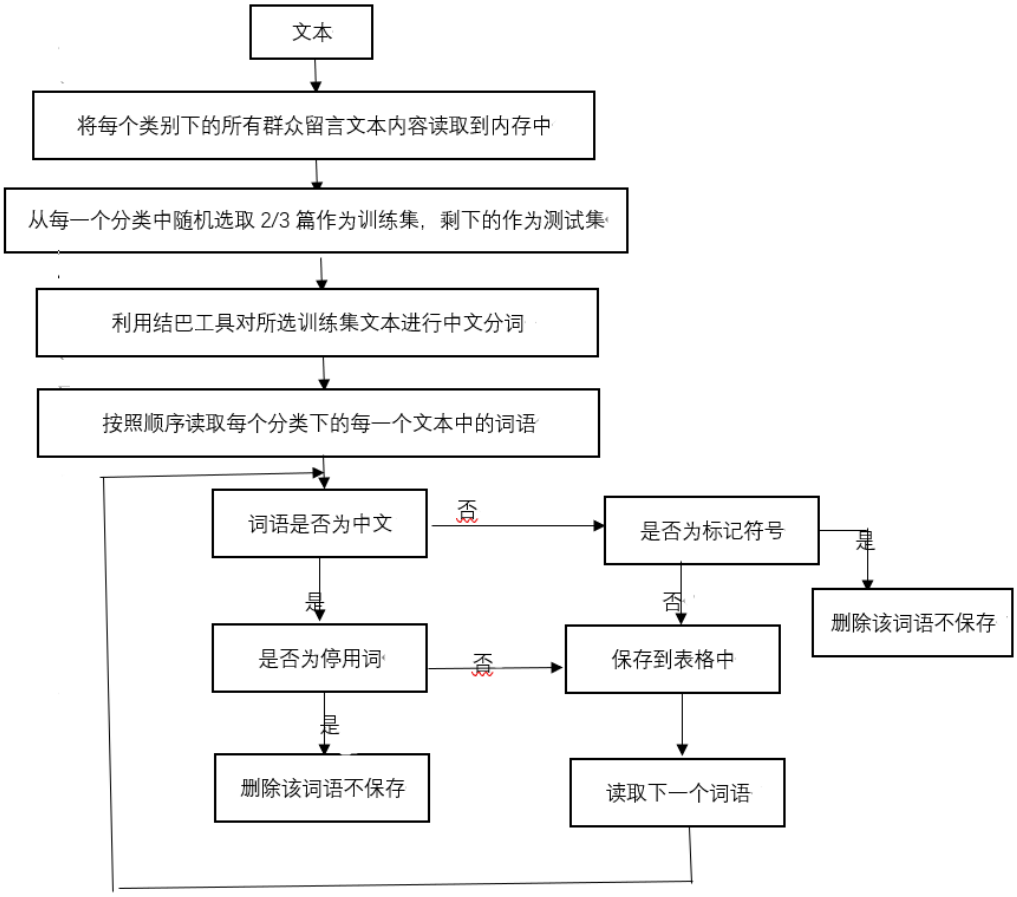

图 2 文本预处理过程图

2、特征处理

在进行了文本预处理后，我们进入到了第二个步骤，对处理了的群众留言文本进行特征选取并加权的处理，此过程运用了改进的 TF-IDF 算法。在特征选择的过程中，要求着特征词要为名词、动词、动名词以及命名实体，其中名词最为重要，需要重点提高其 TF-IDF 权重。特征选择就是用于高维特征空间的降维处理。

根据实操后，可以得知，群众留言在经过了文本预处理后，将变成由一系列词语所组成的集合。每一个词语代表了该文本的即为了选择出能够代表某个文本的特征项，这时，需要构造特征词的向量矩阵，但它的维度很高，计算率很低，为了提升其计算率，我们需要做进一步的特征处理，选择出能够明确代表该本文的特征项，并计算出每一个特征项的 TF-IDF 权重值。TF-IDF 权重值即表示了一个词语在该类的文本中出现的次数越多，并在其它类的文本中出现的次数越少，说明该词语能够代表这个文本的核心内容的一个数据。当计算出了每一个特征项的 TF-IDF 值，我们根据所给题目数据，利用附件一的三级分类作为关键词对特征进一步的加

权，进行权重修改，后运用词频-逆文档频率（TF-IDF）和等级因子结合方式的新特征权重的计算方法来确定特征权重，提高其权重。在经过特征处理之后，大大有利于数据的精确度的提升，减少了内存的开销和繁杂的工作量。

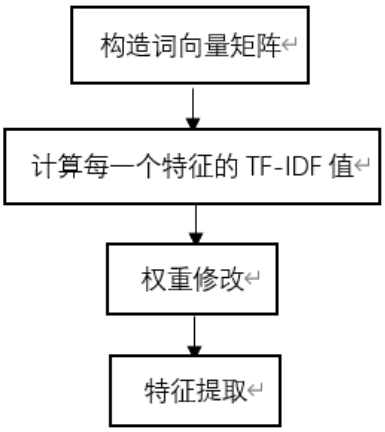

图 3　特征处理过程

在特征处理过程中，所使用到的改进的 TF-IDF 算法，建立模型如下

$$IDF' = \log\left(\frac{\frac{m}{m+o}}{\frac{m}{m+o}+\frac{k}{k+q}} \times N\right) \quad (1)$$

在该式子中$\frac{m}{m+o}$表示了包含特征词 t 的留言在类 C 的群众留言集中的比例，$\frac{k}{k+q}$表示了包含特征词 t 的留言在非类 C 的群众留言集中的比例。

$$w(\text{权重}) = TF(\text{词频}) \times IDF' \quad (2)$$

在确定特征权重后，需要做特征提取即在每一篇文章中提取 TF-IDF 值比较高的 30 个关键词，从而降低特征词向量矩阵的维度，为后续采用朴算贝叶斯算法给出相关数据。

3、文本分类

经过了上述两个步骤，下一步朴算贝叶斯算法进行文本分类处理，

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (3)$$

根据公式（3），我们需要根据本题做一些假设，用 P(A)来表示的为原题中所给出一级分类，如城乡建设等。用 B 来表示一条群众留言中的每一个词语。其中

A=$\{a_1, a_2, \ldots, a_k\}$即代表一组类标签，B（$b_1, b_2, \ldots, b_n$）为一个向量，其中一条群众留言是由 n 个词语组成的。

p(A)为是类 C 群众留言的先验概率。P(B|A)为当这条群众留言为这一类时，含有的 B 的概率，而 B 因为是确定的一条群众留言，所以群众留言中的词语也是确定的。只需要看这一组类标签的词语的概率。

$$P(B|A)=P(b_1, b_2, \ldots, b_n| A) \quad (4)$$

公式（4）所说明的是在类 C 群众留言中 A 的出现跟 B 相同的概率是多少。但是利用公式（4），过于精细，概率就会变得很小，所以判断起来也更加困难，因此可以做如下扩展：

$P(b_1, b_2, \ldots, b_n| A)$ 转化为$P(b_1 |A)P(b_2 | A)\ldots P(b_n | A)$

即有如下公式：

$$P(B|A) = P(b_1 |A)P(b_2 | A)\ldots P(b_n | A) \quad (5)$$

通过一系列计算后，我们只需要求出每个b_i在类 C 群众留言中出现的频率，即可得到 P(A|B)。通过以上步骤我们就可以完成对一条群众留言的分类工作，让机器学习了该分类模型的流程后，对剩下的全部群众留言进行自动分类。

4、文本分类效果验证

完成关于留言内容的一级标签分类步骤，后使用 F-Score 对分类方法进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i + R_i} \quad (6)$$

其中为P_i第 i 类的查准率，R_i为第 i 类的查全率。

3.1.2. 结论及其分析

问题一建立了有监督学习的一级分类模型，让机器学习了我们所建立的算法后，得出了以下数据图，如图 4 测试集分类后的文本可视化散点图。

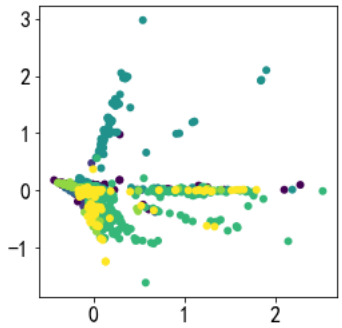

图 4　测试集分类后的文本可视化散点图

上图不同颜色的数据点代表了不同的分类。从图中可以直观的看出题目以及标签的分类情况。

本题最关键的步骤为使用 F-Score 对分类方法进行评价，运用公式（6），借助计算机，通过机器运算，我们得到了以下各个分类标签的 F—Score 值，如下图 5　每一类的 F—Score 值折线图图 5　每一类的 F—Score 值

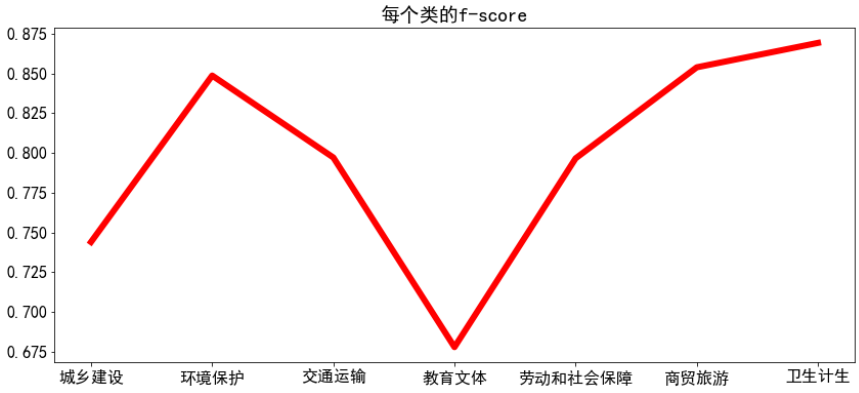

图 5　每一类的 F—Score 值折线图

F-score 值，F1 值为算数平均数除以几何平均数，且越大越好，将查准率和查全率的上述公式带入会发现，当 F1 值小时，True Positive 相对增加，而 false 相对减少，即查准率和查全率都相对增加，即 F1 对查准率和查全率都进行了加权。

本文运用到的特征加权朴素贝叶斯算法与传统的贝叶斯算法的比较如图 6，从而我们可以看出改进后的朴素贝叶斯算法的准确率有很明显的提高。

	查准率	查全率	F-score
传统的贝叶斯算法	0.521	0.504	0.49
改进后的贝叶斯算法	0.809	0.806	0.806

图 6

根据最后的计算，我们可以得到有监督学习的一级分类模型对全部分类的 F-

Score 值为 0.806，故可以得知所建立的分类模型相对合理。

3.2. 问题二

3.2.1. 舆情事件预警模型的建立

1、建模思路

要将某一时段内反映特定地点或特定人群的留言进行分类，可以利用文本聚类算法，将留言中文本相似度较高的聚为一类，并选出所含条目最多的类进行舆情事件预警。针对文本在聚类或分类时，为解决文本数据高维、稀疏而导致相似度值低的问题，常用的方法有降维或增维，但考虑到维度增加会导致文本特征的冗余度提升，本文采用特征提取的方式达到降维的效果，为了保留原文本信息的准确性，在特征提取时，会通过 TF-IDF 权重计算，保留更多有意义的特征词。

2、文本聚类

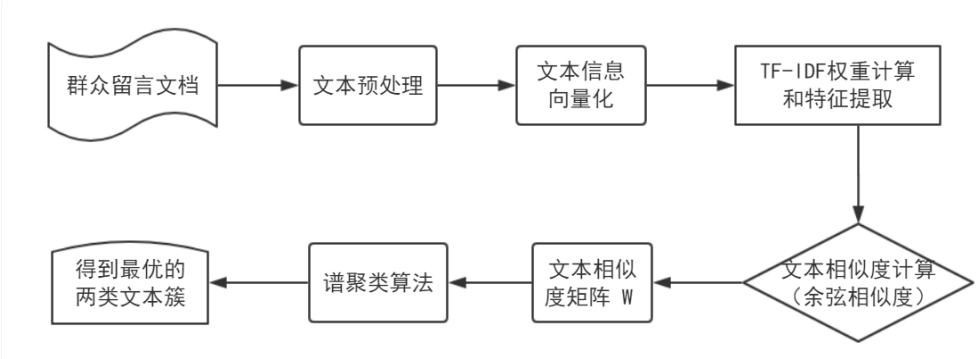

图 7 文本聚类流程图

① 文本预处理：其包含三个步骤，删除文本标记、进行中文分词并标记词性、删除停顿词；

② 文本信息向量化：通过词频计算，从而实现文本信息进行向量化；

③ TF-IDF 权重计算和特征提取：通过计算出每个特征的 IF-TDF 值，量化出每一个词条在文本中的重要程度，然后去掉一些携带信息相对较少的常用词汇，通过权重修改，保留重点词汇的特征词，达到特征提取以及降维的目的；

④ 计算文本相似度计算：文本相似度计算是文本聚类过程中的重要步骤，本文采用余弦函数来计算两个文本之间的相似度，构建出文本相似度矩阵；

⑤ NJW 谱聚类算法：

1) 根据输入的相似度矩阵的生成方式，构建样本的相似度矩阵 S

2) 根据样本的相似矩阵 S 构建邻接矩阵 M，以及构建度矩阵 D

3) 根据矩阵邻接矩阵 M、D 计算出拉普拉斯矩阵 L

4) 构建标准化后的拉普拉斯矩阵 $D^{\frac{1}{2}}LD^{\frac{-1}{2}}$

5) 计算矩阵 $D^{\frac{1}{2}}LD^{\frac{-1}{2}}$ 最小的 k_1 个特征值，并得到他们各自对应的特征向量 f，

6) 将各自对应的特征向量 f 组成的矩阵按行标准化，最终组成 $\mathrm{n}\times\mathrm{k}_1$ 维的特征矩阵 F，

7) 对 F 中的每一行作为一个 k_1 维的样本，共有 n 个样本，然后进行 K-means 聚类

8) 若特征矩阵 F 中第 i 行属于 j 簇，则将第 i 行的文本也划分的第 j 簇中，进而得到众多文本簇。

⑥ 使用轮廓系数来评价文本聚类的效果

3.2.2. 舆情事件预警模型的求解

1、TF-IDF 权重计算和特征提取

利用 TF-IDF 算法，得到每个特征词的 IF-TDF 值，量化出每一个词条在文本中的重要程度并利用文本预处理时，对词性的进行的标注，本文通过权重修改来达到特征提取的目的，词性与权重设计如下：

词性：	名词 n	动词 v	动名词 vn	地名 ns	其他词
权重：	1.25	1.1	1	1.05	0

即提取名词、动词、动名词、地名的这四类重点词汇，以达到特征提取以及降低维度的效果，这样做的好处是，可以减少内存的开销，提高计算效率并且提高后续进行文本聚类的准确率。

2、计算文本相似度计算

文本聚类是一种无监督自动分类方法，它根据文本之间的相似程度将文本划分到不同的簇，属于同一簇的文本数据具有较高的相似度，而不同簇中的文本数据相似度较低，所以文本相似度计算是文本聚类过程中的重要步骤。

本文采用余弦函数来计算两个文本之间的相似度，文本d_i与d_j之间的余弦相似度计算公式为：

$$\mathrm{sim}(d_i,d_j)=\cos dis(V_{d_i},V_{d_j})=\frac{V_{d_i}\cdot V_{d_j}}{|V_{d_i}|\times|V_{d_j}|} \tag{7}$$

利用此公式，可计算出任意两个文本间的相似度，得到文本的余弦相似度矩阵 W，如下图 8 所示；

	0	1	2	3	4	5	6	7	8	9	10	11
0	1	0.00135798	0	0	0.00303771	0	0	0.000943523	0	0.00249072	0	0
1	0.00135798	1	0.00134447	0	0.00105711	0.00509817	0.000642794	0.00112099	0	0	0.00300566	0
2	0	0.00134447	1	0.00298444	0.000405501	0.00287414	0.00819412	0	0	0.00162514	0	0
3	0	0	0.00298444	1	0.00327947	0.00313175	0.00284058	0	0	0.00548849	0	0.00382705
4	0.00303771	0.00105711	0.000405501	0.00327947	1	0.080628	0.00443709	0.000247027	0	0.158432	0.00117342	0.0497899
5	0	0.00509817	0.00287414	0.00313175	0.080628	1	0.269823	0	0	0.00310772	0.0019618	0.0305421
6	0	0.000642794	0.00819412	0.00284058	0.00443709	0.269823	1	0.000532421	0.00077314	0.0042655	0	0.00563244
7	0.000943523	0.00112099	0	0	0.000247027	0	0.000532421	1	0	0.120545	0	0
8	0	0	0	0	0	0	0.00077314	0	1	0	0	0
9	0.00249072	0	0.00162514	0.00548849	0.158432	0.00310772	0.0042655	0.120545	0	1	0	0.0490637
10	0	0.00300566	0	0	0.00117342	0.0019618	0	0	0	0	1	0
11	0	0	0	0.00382705	0.0497899	0.0305421	0.00563244	0	0	0.0490637	0	1
12	0	0.00294402	0.00150185	0.00402446	0.00378133	0.00210586	0.00455831	0.000592047	0.00243065	0.00937749	0.00123219	0
13	0.000942111	0.0010138	0	0.00176566	0.0856572	0.00145281	0.00482055	0.0025639	0	0.0863385	0	0.0448352
14	0	0	0.0033077	0.00691917	0.00183376	0.0039865	0.00237404	0.213809	0.00784147	0.198006	0.000736061	0.00286394
15	0.00121766	0	0.00264089	0.00679125	0.0118804	0.0683975	0.0149529	0.000591493	0	0.00689717	0	0.0594621
16	0	0	0	0.00359092	0.00579892	0	0.00253162	0	0	0	0.00295637	0.00185848

图 8 余弦相似度矩阵

余弦相似度矩阵 W 的形式可归结为：

$$W=\begin{bmatrix} s_{11} & \cdots & s_{1N} \\ \vdots & \vdots & \vdots \\ s_{i1} & s_{ij} & s_{iN} \\ \vdots & \vdots & \vdots \\ s_{N1} & \vdots & s_{NN} \end{bmatrix} \tag{8}$$

由上图可知，W 是一个对称矩阵，其中$S_{ij}=sim(d_i,d_j)$，且$S_{ij}>0$。S_{ij}越大，表明文档内容的相似性越大，并且此类文档内容更容易分到同一个文本簇中。

在具有 N 篇文档的数据集中，文本之间的余弦相似度以$N\times N$的矩阵即 W 表示，扫描矩阵 W 的每一行，统计与文档d_i的相似度$sim(d_i,d_j)>T$的文档集合，即

矩阵 W 第 i 行元素大于阈值 T 的列元素集，得到与文档d_i的相似度大于阈值 T 的文本集合。

3、谱聚类算法求解如下：

由于谱聚类是基于图论的知识，因此将留言的特征集转化为一个无向赋权图G，我们一般用点的集合V和边的集合E来描述。即为$G(V,E)$。其中V即为我们数据集里面所有的点$(v_1,v_2,...v_n)$。对于V中的任意两个点，可以有边连接，也可以没有边连接。我们定义权重w_{ij}为点v_i和点v_j之间的权重。由于我们是无向图，所以$w_{ij}=w_{ji}$。

首先输入的相似度矩阵的生成方式，构建出样本的相似度矩阵 S，再根据样本的相似矩阵 S，构建出无向赋权图的邻接矩阵 M，以及利用每个点度的定义：

对于有变连接的两个点v_i和v_j，$w_{ij}>0$，对于没有边连接的两个点v_i和v_j，$w_{ij}=0$。对于图中的任意一个点v_i，它的度D_i定义为和它相连的所有边的权重之和，即

$$D_i=\sum_{j=1}^{n}w_{ij} \tag{9}$$

利用每个点度的定义，我们可以得到一个$n\times n$的度矩阵D,它是一个对角矩阵，只有主对角线有值，对应第i行的第i个点的度数，定义如下：

$$D=\begin{pmatrix} d_1 & \cdots & \cdots \\ \cdots & d_2 & \cdots \\ \vdots & \vdots & \ddots \\ \cdots & \cdots & d_n \\ & & \end{pmatrix} \tag{10}$$

利用所有点之间的权重值，我们可以得到图的邻接矩阵 M,它也是一个$n\times n$的矩阵，第i行的第 j 个值对应我们的权重w_{ij}。

根据上面的邻接矩阵 M 以及度矩阵 D，计算出拉普拉斯矩阵 L，由公式可知拉普拉斯矩阵

$$L=D-M \tag{11}$$

其中 D 为上文中的度矩阵，它是一个对角矩阵，W 是上文中的邻接矩阵

将上面得到的拉普拉斯矩阵进行标准化后，可得到拉普拉斯矩阵 $D^{\frac{1}{2}}LD^{\frac{-1}{2}}$，计算矩阵 $D^{\frac{1}{2}}LD^{\frac{-1}{2}}$ 最小的前 k 个特征值，并得到他们各自对应的特征向量 f，接着将各自对应的特征向量 f 组成的矩阵按行标准化，最终组成 $n\times k$ 维的特征矩阵 F，对 F 中的每一行作为一个 k 维的样本，共有 n 个样本，然后进行 K-means 聚类，若特征矩阵 F 中第 i 行属于 j 簇，则将第 i 行的文本也划分的第 j 簇中，进而得到众多文本簇。

3.2.3. 舆情事件预警模型的结果与分析

由于未知数据样本点的正确类别信息的情况下，本文采用轮廓系数作为文本聚类算法的评价指标，假定样本点 m，设 a 是与 m 同类的其他样本的平均距离，b 是与 m 距离最近的其他类簇中样本的平均距离，则轮廓系数 K 为：

$$K=\frac{b-a}{\max(a,b)} \tag{12}$$

所有样本的轮廓系数平均值就是整个样本数据集的轮廓系数，它的取值范围是 $[-1,1]$，同类数据样本点距离越近同时不同类数据样本点距离越远，则分数越高，分值越高则聚类效果约好。

又由于不知道选取多少个簇中心进行文本聚类，所以选了 8~60 个聚类中心进行动态聚类，并计算每一次聚类的轮廓系数平均值来评价聚类效果，但由于文本噪音点太多，导致每一次聚类的轮廓系数平均值总是在 0.2 附近，如下图 9 每一次聚类的轮廓系数平均值折线图所示。

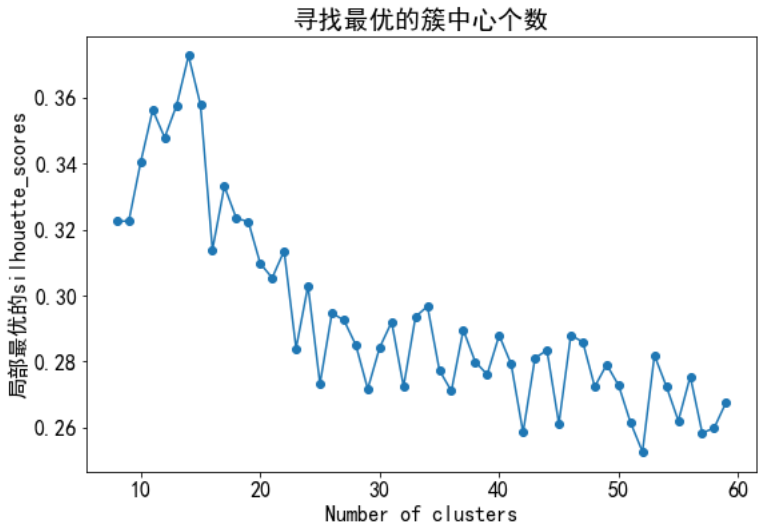

图 9 每一次聚类的轮廓系数平均值折线图

对此本文提出了改进方案：利用 python 的 metrics 库计算出每个文本的轮廓系数，再计算出每一类簇的轮廓系数的平均值，并取其最大值作为整体聚类的轮廓系数。经过笔者多次迭代运行，发现最优的聚类的轮廓系数保持在 0.7 附近，并且有两个平均轮廓系数最大的簇趋于稳定，这两类就是集中爆发事件的最大团。如下图 10 的可视化。

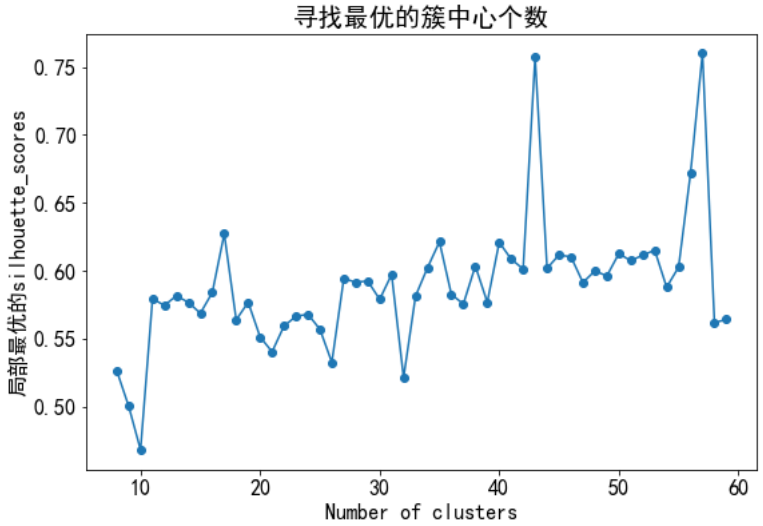

图 10 选定不同蔟中心个数的聚类图

经上面得到最优的簇中心个数 43，并选定该簇中心个数进行一次谱聚类，然后计算得到了每个蔟的轮廓系数平均值，如下图 11 所示。

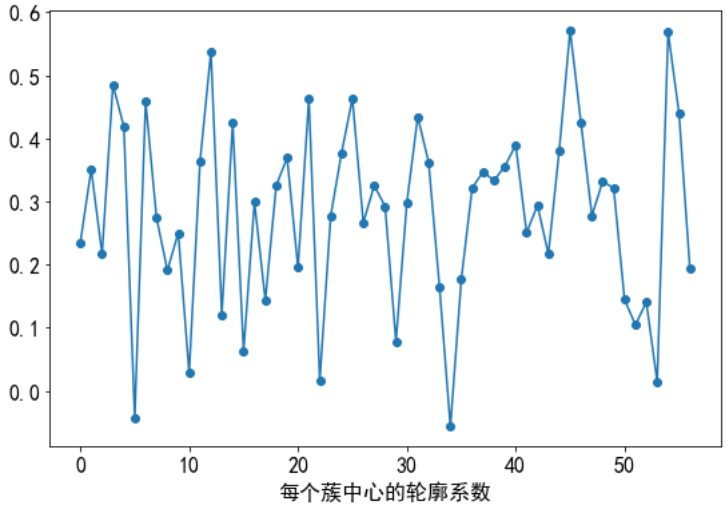

图 11 每个簇中心的轮廓系数折线图

聚类的可视化如下，不同的颜色代表不同的簇（即不同的话题）。

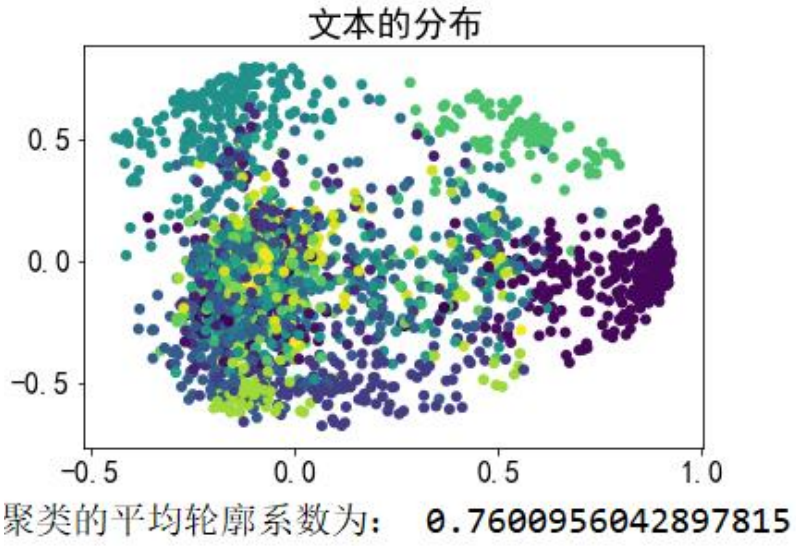

图 12 文本分布散点图

3.2.4. 热门关注事件提取模型的建立

1、建模思路

对于留言篇数较少，但是单条留言点赞数高的这类热点事件，利用上面的模型进行文本聚类，没办法得到轮廓系数可靠的文本簇，为了解决此问题，本文建立了热门关注事件提取模型，由点赞数与反对数共同反映留言的受关注度，找到较高的前三名，利用文本相似度矩阵，在附件三的留言信息中找出与这三篇留言相似的其他留言。

2、具体步骤如下

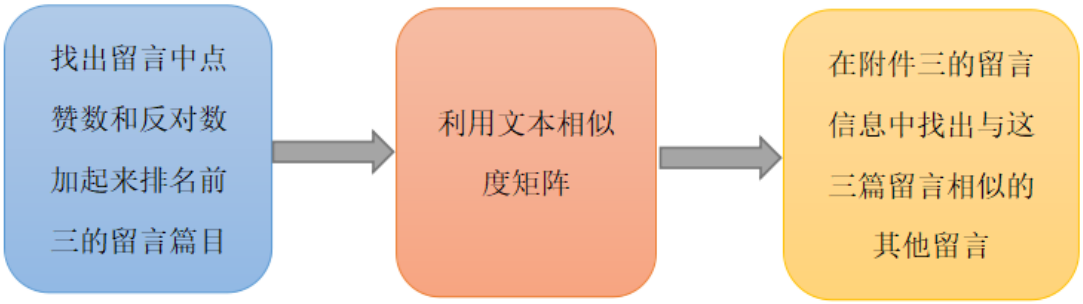

图 13　热门关注事件提取模型具体步骤图

3.2.5.　热门关注事件提取模型的求解

首先对附件 3 的点赞数进行降序处理，利用 python 的 xlrd 库提取点赞数最高的前三篇文章，然后计算与其余文章的余弦相似度，通过人工观察决定设置 0.361 作为阈值，大于 0.361 即为是与这三类话题的相似留言。

3.2.6.　热门关注事件提取模型的结果与分析

提取出来的点赞数最高的前三篇文章的留言编号分别为：208636、223297、220711，并且找到了与其相似的其余留言文本，如下图 14 所示。

A	B	C	D	E	F	G	H
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
3	208636	A0007717	A市A5区汇金路五矿万境K9县	2019/8/19 11:34:04	我是A市A5区汇金	0	2097
3	220716	A0008904	A市中意一路与汇金路之间路	2019/1/23 10:32:59	A市A5区万芙南路	0	0
3	265503	A0008413	A市、区的基础教育科怎么都	2019/5/15 18:15:40	想咨询下小孩子	0	2
4	222919	A0001719	A市辰北三角洲幼儿园入园难	2019/2/15 10:24:04	幼儿园要摇号，	0	1
4	223297	A0008752	反映A市金毛湾配套入学的问	2019/4/11 21:02:44	书记先生：您好	5	1762
4	253314	A0008995	反映A5区万科魅力之城小孩	2019/4/17 23:59:08	尊敬的市委书记：	0	0
4	277139	A0002401	A3区卓越浅水湾南门对面金	2019/9/25 0:17:51	梅溪湖卓越浅水	0	0
5	194343	A0001061	承办A市58车贷案警官应跟进	2019/3/1 22:12:30	胡书记：您好！	0	733
5	220711	A0003168	请书记关注A市A4区58车贷案	2019/2/21 18:45:14	尊敬的胡书记：	0	821

图 14

3.2.7.　多指标热度评价体系

对于反应篇数较多的这类热点问题，我们已经通过对第一个模型的求解，找到了留言文本中的对于同一类问题反映篇数最多的两类热点问题，而通过第二个模型

的求解，我们可以找到篇数较少，但是群众对于留言的关注度比较高的热点问题，并且利用相似度矩阵找到与此留言相似的文本，两者结合，可得到五类群众反映比较强烈的热点问题，并且汇总出“热点问题留言详细表”。

但是这五类热点问题该如何进行排名，为了解决此问题，我们将建立出如下多指标热度评价体系进行评价与分析。

3.2.8. 多指标热度评价体系的求解过程

1、对各项指标进行赋权

指标权重确定的方法多种多样，本文利用主观赋权法，通过资料查找以及对各指标按其重要程度进行比较，从而确定出如下各项热度评价指标的权重值

如下表 1 热度指标体系。

表 1 热度指标体系

	一级指标	二级指标	权重
留言热度综合评价指标体系	传受众特征热度影响力	指标 1：点赞数	3
		指标 2：反对数	3
	内容特征热度影响力	指标 3：留言篇数	5
		指标 4：持续时间	2

2、综合评价排序

$$\text{score}=\sum_{i=1}^{4} w_i v_i' \quad （w_i 为第 i 个指标的权重） \tag{13}$$

通过此公式算得总分之后，采用 max-min 归一化的方法，对所有总分进行 0-1 的归一化，公式如下：

$$V'=\frac{V-\min}{\max-\min}(\text{new}-\max-\text{new}-\min)+\text{new}-\min \tag{14}$$

由此可将留言热度值映射成分数（0-10 分）

$$\text{留言热度值}=V'\times 10 \tag{15}$$

最后再对所有的热度值从大到小进行排序，选出排在前五的热门留言。

3.2.9. 多指标热度评价体系的结果与分析

通过上述建立的多指标评价体系，可以求得五类热点问题中，各项评价指标特征的得分情况，如下图 15 所示。

	问题的持续时间	问题的留言篇数	问题的点赞数
问题1	3.69565	3	0
问题2	5	2.67073	0.0214634
问题3	3.93478	0	2
问题4	4.65217	0.0731707	1.67902
问题5	0	0	1.47024

图 15 五类问题的各项评价指标的得分情况

将上图中各项指标的得分，按照其各自对应的权重值进行再次整合计算可得到五类问题的总体得分情况，如下图 16 所示。

```
问题0的评价指数为：  6.695652173913043
问题1的评价指数为：  7.892485612496575
问题2的评价指数为：  5.934782608695652
问题3的评价指数为：  6.398614033624459
问题4的评价指数为：  1.4702439024390244
```

图 16 五类问题的总体得分情况

即在群众留言中，排名第一的是问题 2，排名第二的是问题 1，排名第三的是问题 3，排名第四的是问题 2，排名第五的是问题 5。

对“热点问题留言详细表”进行人工整合，可以得到如下图 17 所示的“热点问题表”

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
2	1	6.69	2019/7/28至	A市伊景园·滨河苑业主	广州铁路集
1	2	7.69	2019/1/9 至	A市、搭乘公交车的居民	公交车线路
4	3	5.93	2019/8/19	A市A5区汇金路五矿万境	小区群租房
3	4	6.4	2019/4/11	A市金毛湾业主	梅溪湖的设
5	5	1.47	2019/2/21至	A市车贷受害者	A市车贷案

图 17 热点问题表

3.3. 问题三

3.3.1. 模型的建立与求解

通过前两问解答与计算，我们已经建立了特征加权朴素贝叶斯模型对群众留言文本进行一定的分类，而后我们选择建立任务型对话系统的回复质量评价模型对系统回复质量作出一定的评价，由于任务型对话系统的目标任务非常明确，完成过程完整可观测，所以我们需要选择几个具体指标对于任务型对话系统的回复质量进行定性定量分析评价。利用附件 4 提供的相关部门对留言的答复意见，通过线性回归来对群众留言数据集进行相关处理，求出一个权重指标，并认为这个指标可以反映出用户的满意度。除此之外，还可以根据附件内容选取相关指标。

从完整性角度分析，需要通过衡量分类标准中定义的特征词覆盖群众留言中词语的程度来量化完整性，接着从相关性角度继续进行分析，我们可以选择到文本相似性这一指标，运用到任务型对话系统的回复质量评价模型中，可以判断出相关部门对群众留言的回复是否合理，方法可行。

为了得到文本相似度这一指标，我们引入 RUBER 这一概念，即一种有参考和无参考相结合的度量。其中本题我们运用到了有参考的度量，参考了附件 4 中相关部门对留言的答复意见，使用人工回复作为参考对机器回复的质量进行判断。

而后我们运用到了余弦距离来衡量文本相似度。需要得到群众留言的向量表达和相关部门留言答复的向量表达，算出它们的余弦距离，余弦距离越大，代表距离越近，两者之间越相似。

计算文本相似度，我们首先需要获得群众留言以及留言回复的向量表示。假设两个文本中有均 N 个词语，通过对话模型得到一个维度为 d 的向量，因此文本可以表示为一个 d×N 的矩阵，第 i 列表示句子中第 i 个词语对应的词向量。对每行分别取 max-pooling 和 min-pooling，得到两组 d 维的向量 Xmax 和 Xmin。二者拼接得到的维度为 2d 的向量即为该群众留言或留言回复的向量表示。

接着，计算其余弦距离。分别按照上述方法得到群众留言以及留言回复的向量表示，分别设为V_r和$V_{r'}$，余弦距离的计算公式如下：

$$S_R(\mathrm{r},\mathrm{r}') = \cos(V_r, V_{r'}) = \frac{V_r{}^T V_{r'}}{\|V_r\| \cdot \|V_{r'}\|} \quad (16)$$

其中 SR 表示有参考度量。

因此我们可以计算出两者之间文本相似度，从而获得任务型对话系统的回复质量评价模型的一个指标。

继续观察附件 4，我们可以找到相关部门回复留言的时间，通过分析，我们可以知道，当回复时间越短时，群众的满意度会更高，因此我们接着选择了回复时间差作为模型的第二个指标。

得到了两个指标后，我们可以开始运用任务型对话系统的回复质量评价模型，其流程如下图 18。

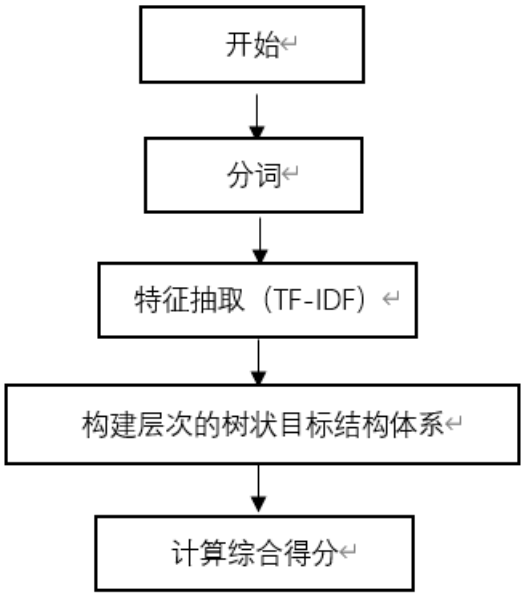

图 18 任务型对话系统回复质量评价流程图

对于评价任务完成度的预测，本题中使用的分类模型为特征加权朴素贝叶斯模型，在对这个模型进行评价时，仍然采用 F-Score 评价模型计算其准确率与查全率的调和平均值。

3.3.2. 结论及其分析

在计算相似度这一指标时运用公式（16），借助机器运算，我们可以得到如下

结果图。

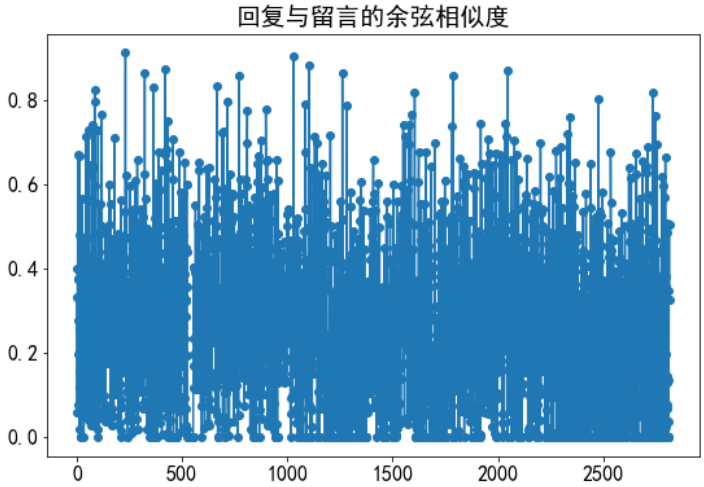

图 19 回复与留言的余弦相似度

同时，利用附件 4 的相关数据，我们也可以得到群众留言的时间与相关部门对留言回复的时间差。

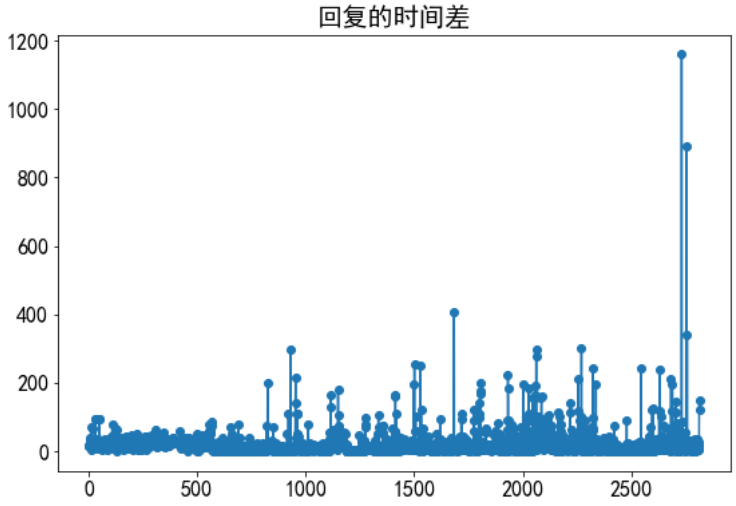

图 20 回复的时间差

除了以上两个指标之外，对于本题我们还提出了相关性，及时性，完成度，直观性，可理解性多指标评价方案，并通过构建层次的树状目标结构体系来计算综合得分。

接下来将介绍几个指标的测量标准：

1、 完成度指的是回复群众留言内容时，要将问题完整回答并且都回答到点子上。

2、 直观性即相关部门在回复群众留言时应该做到将回复文本长度尽量缩短，

且段落划分要清晰，这样直观性会更好。

3、 回复留言时，不应该出现方言、网络用语、专业名词等，不同文化程度的用户都可以读懂，则可理解性好。这一指标往往要求了有关部门在回复时需要统一用语。

4. 模型评价

4.1. 模型优点

4.1.1. 问题一的优点

1.特征选择和加权采用了改进的 TFIDF 算法，能更好的筛选出更有代表意义的词语，并且本文采用了双文本向量矩阵(即主题和留言赋予不同的权重），主题设置更高的权重，使主题更具有代表性。

2.采用特征加权朴素贝叶斯分类模型。本文将 TF-IDF 特征权重合并到贝叶斯公式的条件概率中，其次，将 TF-IDF 值决定的等级因子特征权重导入贝叶斯公式中，增强了词语的重要性，大大削弱了其特征独立性假设的影响，进一步提高分类结果。分类取得了较好的效果，准确率由 0.5 提高到了 0.8。

3.改进的贝叶斯算法相比传统方法增强了词语的重要性，大大削弱了其特征独立性假设的影响，大大提升了分类的精度。

4.1.2. 问题二的优点

本文使用的谱聚类算法对数据分布的适应性强，计算量也小，而且易懂，实现起来不复杂，对于处理这种稀疏的文本的聚类很有效。由于使用了降维，因此在处理高维数据聚类时的复杂度比传统聚类算法好。

谱聚类算法使用的是文本之间的相似度矩阵，一定程度上解决了一意多词的问题，提高了聚类的效果。

5. 模型的优化以及改进

5.1.1. 问题一的改进

文本表示和特征权重的进一步研究。由于分类性能很大程度上受到输入表示的影响，本文构造的特征权重修改未能很好地对命名实体赋予更高的权重。因为使用的是 jieba 分词工具，不能对命名实体进行较好的分词。可以使用更多的特征选择、特征提取算法来提高特征加权朴素贝叶斯方法的性能。

考虑语义和语法。大多数文本特征项之间存在着语义和语法的联系，本算法仅在一定程度上削弱朴素贝叶斯的特征独立性假设的影响，但是没有考虑语义和语法，而且对未出现过的词语都会过滤掉，受限于对训练集的选取。可以加入语义分析（LSA)来提高算法的准确度。

5.1.2. 问题二的改进

在探索集中爆发事件中，使用的谱聚类方法很难预先确定聚类数，所以每一次迭代的最优簇中心个数都不尽相同，聚类的评价指标不稳定。

笔者只建立了两个热点挖掘的模型，可能忽略了一些潜在的热点问题。可以再加个情感分析模型，把负面情绪很大的问题找出来，即使解决，防止恶化。

聚类效果依赖于相似矩阵，不同的相似矩阵得到的最终聚类效果可能很不同。

6. 参考文献

[1] 宋晓敏.基于改进贝叶斯算法的中文信息分类研究[C].北京邮电大学,2019.

[2] 郑霖;徐德华.基于改进 TFIDF 算法的文本分类研究[J].计算机与现代化,2014,No.229,10-13+18.

[3] 贾威.基于武汉城市留言板的舆情热点监控研究[C].华中师范大学,2019.

[4] https://blog.csdn.net/Joseph__Lagrange/java/article/details/90813885

[5] 李征.李斌.一种基于改进相似度计算的文本聚类方法[J].河南大学学报,2018,Vol.48 No.4.

[6] https://www.cnblogs.com/pinard/p/6221564.html

[7] 魏银华.基于 Python 的古汉语文本聚类应用研究[C].大连理工大学,2018.

[8] 刘秋艳;吴新年.多要素评价中指标权重的确定方法评述[J].知识管理论坛,2017,v.2;No.12,48-58.

[9] 张杨子.面向对话系统回复质量的自动评价研究[C].哈尔滨工业大学,2018.

[10] https://www.tinymind.net.cn/articles/27464d3973fb08

[11] 袁红;张莹.问答社区中询问回答的质量评价——基于百度知道与知乎的比较研究[J].数字图书馆论坛,2014,No.124,45-51.