基于机器学习算法的“智慧政务”中的文本挖掘应用

摘 要

随着社会的发展，网络问政逐渐成为政府了解民意的渠道手段。但由于与民意相关的文本数量过于复杂，与民意相关的文本数据量不断攀升，而人工处理效率低错误率较高，构建智慧文本处理系统就有了相当重要的意义。本文利用机器学习里的多种算法，结合 Python 等工具，建立了一套比较完善的智慧文本处理系统，较好的解决了所给问题，同时具有一定的应用价值。

针对问题一，需要根据附件二的数据对一级标签进行分类。首先将附件二的留言内容进行数据处理，经过分词、清洗等操作后，按照一定比例构造训练集与测试集。接着对 KNN 分类器、朴素贝叶斯分类、支撑向量机算法进行研究分析，采取分类效果最好的 LinearSVC 算法，并将测试集的结果输出。

针对问题二，问题二要求筛选出热度前五的问题。首先留言主题作为文本分类的数据，经过去除脏数据、分词清洗之后，计算出文本相关度。采用 K-means 算法将留言分类并且计算热度，对热点事件的时间范围采取逐步压缩范围的方法实现范围的最终确定，最后将得到的热点问题写入文档。通过判断聚类是否合理不断调整 K 值。

针对问题三，问题三要求对附件四的留言回复从相关性，完整性，可解释性三个方面进行评价。本题中我们分别定义了衡量这三个性质的因子，通过对数字映射后的特征做余弦相似度匹配表示相关性，回复与留言字符长度的比值表示完整性，留言分词与特征词集的重复个数表示可解释性，经过数据标准化处理后加权得到最终评价结果，较好的对留言回复进行了评价。

关键词: 支持向量机算法、结巴分词、K-means 聚类、文本挖掘

ABSTRACT

With the development of society, Internet politics has gradually become a channel for the government to understand public opinion. However, because the number of texts related to public opinion is too complex, the amount of text data related to public opinion keeps rising, and the manual processing efficiency is low and the error rate is high, it is of great significance to build a smart text processing system. This paper uses various algorithms in machine learning, combined with Python and other tools, to establish a set of relatively perfect intelligent text processing system, which solves the given problems well and has certain application value.

In response to question 1, the primary labels need to be classified according to the data in Annex 2. First of all, the message content in Annex 2 is processed with data. After word segmentation, cleaning and other operations, the training set and test set are constructed according to a certain proportion. Then, KNN classifier, Naive Bayesian classification and support vector machine algorithm are studied and analyzed. LinearSVC algorithm with the best classification effect is adopted, and the results of the test set are output.

In response to question 2, question 2 requires screening out the top five hot issues. Firstly, the topic of the message is used as the data of text classification. After removing dirty data and word segmentation cleaning, the text correlation is calculated. The K-means algorithm is used to classify messages and calculate the heat. The time range of hot events is gradually compressed to realize the final determination of the range. Finally, the hot issues are written into the document. By judging whether clustering is reasonable or not, the K value is continuously adjusted.

In response to question 3, question 3 requires the comments in annex 4 to be evaluated in terms of relevance, completeness and interpretability. In this topic, we respectively define the factors to measure these three properties. Cosine similarity matching is used to represent the correlation of the features after digital mapping, the ratio of reply and message character length is used to represent the integrity, message word segmentation and the repeated number of feature word sets are used to represent the interpretability, and the final evaluation result is obtained by weighting after data standardization processing, which better evaluates the message reply.

Keywords: Support Vector Machine Algorithm 、jieba Word Segmentation、 K-means Clustering、 Text Mining

目录

一、 问题重述

1.1 问题背景

近年来，随着科技水平的不断发展，政府部门也越来越多的使用微信、微博、市长信箱、阳光热线等网络问政平台来了解民意、汇聚民智、凝聚民气。因而导致了各类与社情民意相关的文本数据量的不断攀升，这给以往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带来了极大挑战。与此同时，随着大数据、云计算、人工智能等技术的发展，建立基于自然语言处理技术的智慧政务系统已经是社会治理创新发展的新趋势，这对提升政府的管理水平和施政效率具有极大的推动作用。

1.2 需要解决的问题

1、群众留言分类

在处理网络问政平台的群众留言时，工作人员首先按照一定的划分体系（参考附件 1 提 供的内容分类三级标签体系）对留言进行分类，以便后续将群众留言分派至相应的职能部门处理。目前，大部分电子政务系统还是依靠人工根据经验处理，存在工作量大、效率低，且 差错率高等问题。请根据附件 2 给出的数据，建立关于留言内容的一级标签分类模型。 通常使用 F-Score 对分类方法进行评价：

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

2、热点问题挖掘

某一时段内群众集中反映的某一问题可称为热点问题，如“XXX 小区多位业主多次反映 入夏以来小区楼下烧烤店深夜经营导致噪音和油烟扰民”。及时发现热点问题，有助于相关 部门进行有针对性地处理，提升服务效率。请根据附件 3 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，定义合理的热度评价指标，并给出评价结果，按表 1 的格式给出 排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按表 2 的格式给出相应热点问题 对应的留言信息，并保存为“热点问题留言明细表.xls”。

3、答复意见的评价

针对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，并建立一套评价的系统

二、 分类器的评估与选择

分类是数据挖掘的一种非常重要的方法。分类的概念是在已有数据的基础上学会一个分类函数或构造出一个分类模型（即我们通常所说的分类器(Classifier)）。该函数或模型能够把数据库中的数据纪录映射到给定类别中的某一个，从而可以应用于数据预测。总之，分类器是数据挖掘中对样本进行分类的方法的统称，包含决策树、逻辑回归、朴素贝叶斯、神经网络等算法。

2.1 支持向量机

支持向量机（SVM：Support Vector Machine）是机器学习中常见的一种分类算法，一种二分类模型，它的目的是寻找一个超平面来对样本进行分割，分割的原则是间隔最大化，最终转化为一个凸二次规划问题来求解。支持向量机要做的就是要在这些可以选择的直线中选择一条最好的直线来作为分类的直线。线性模型有线性决策边界（交叉的超平面），而非线性核模型（多项式或者高斯 RBF）的弹性非线性决策边界的形状由核种类和参数决定。

下面分别计算 LinearSVC,SVC 的准确度。

2.1.1 LinearSVC 分类器

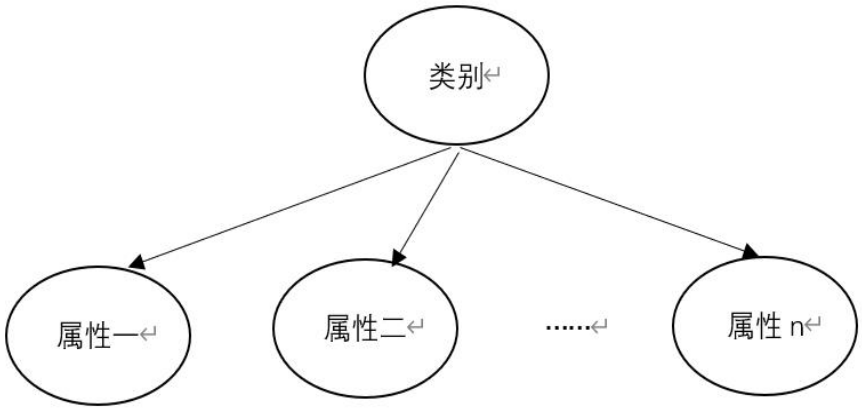
标号	实验次数	平均查准率	平均查全率	平均 F
1	3	86.06%	87.31%	86.68
2	3	86.2%	87.98%	87.08
3	3	86.77%	87.81%	87.29

2.1.2 SVC 分类器

标号	实验次数	平均查准率	平均查全率	平均 F
1	3	86.01%	87.15%	86.58
2	3	86.32%	87.69%	86.45
3	3	85.94%	87.56%	86.23

2.2 贝叶斯分类器

贝叶斯分类器是各种分类器中分类错误概率最小或者在预先给定代价的情况下平均风险最小的分类器。它的设计方法是一种最基本的统计分类方法。其分类原理是通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。研究较多的贝叶斯分类器主要有四种，分别是朴素贝叶斯、TAN、BAN 和 GBN。



对一些规模较小的数据集，朴素贝叶斯分类器和 TAN 分类器的效果比较好，并且当数据集属性间的关联性较弱时，朴素贝叶斯分类器的分类效果要优于 TAN 分类器。本题附件二的数据较少且数据属性间的关联性较弱，故只选择朴素贝叶斯分类器，分类效果如下

2.2.1 朴素贝叶斯分类器

标号	实验次数	平均查准率	平均查全率	平均 F
1	3	76.54%	75.91%	76.22
2	3	73.76%	73.5%	73.63
3	3	74.88%	74.14%	74.51

2.3 KNN 分类器

KNN 分类器，即 K 最近邻算法。它的原理十分简单:存在一个训练样本集合，该集合中每行数据包含多个特征和分类标签，输入没有标签但有多个特征的新数据，将新数据的每个特征与样本中每条数据对应的特征进行比较，然后提取出样本中与新数据最相似的 K 条数据，统计该 K 条数据中各类标签出现的次数，那么出现次数最多的标签即为新数据的分类标签。

使用 KNN 分类器得到的结果如下图所示

2.3.1KNN 分类器

标号	实验次数	平均查准率	平均查全率	平均 F
1	3	48.64%	56.65%	52.34
2	3	48.91%	55.15%	51.84
3	3	48.5%	54.77%	51.44

2.3.1KNN 分类器不足及改进

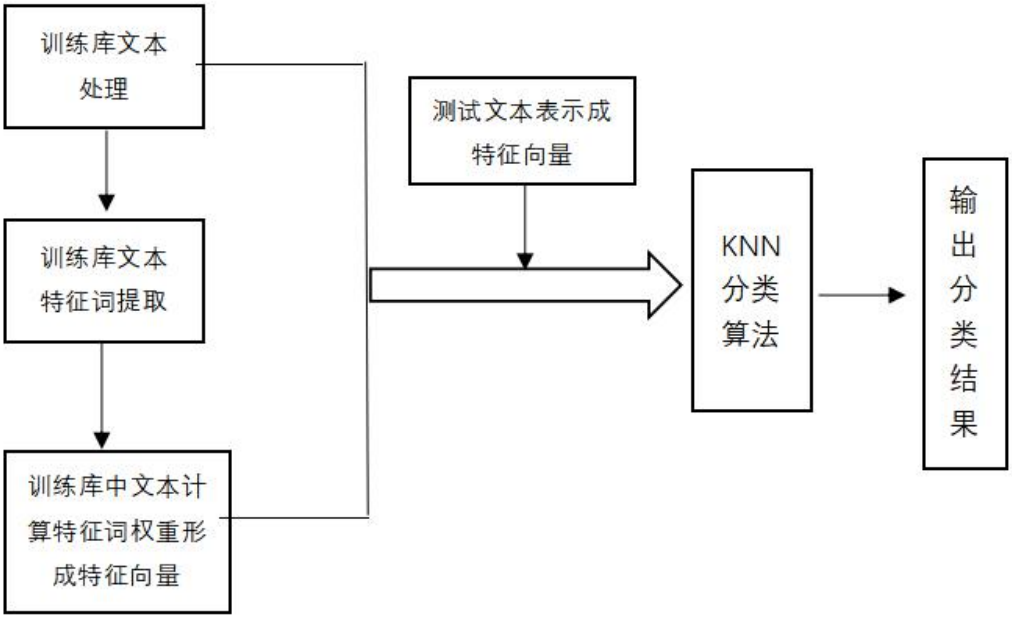
1、添加某些类别的样本容量非常大，而其它类样本容量非常小，即已知的样本数量不均衡。有可能当输入一个和小容量类同样的新样本时，该样本的 K 个近邻中，大容量类的样本占多数，从而导致误分类。

针对此种情况能够采用加权的方法，即和该样本距离小的近邻所相应的权值越大，

将权值纳入分类的参考根据。

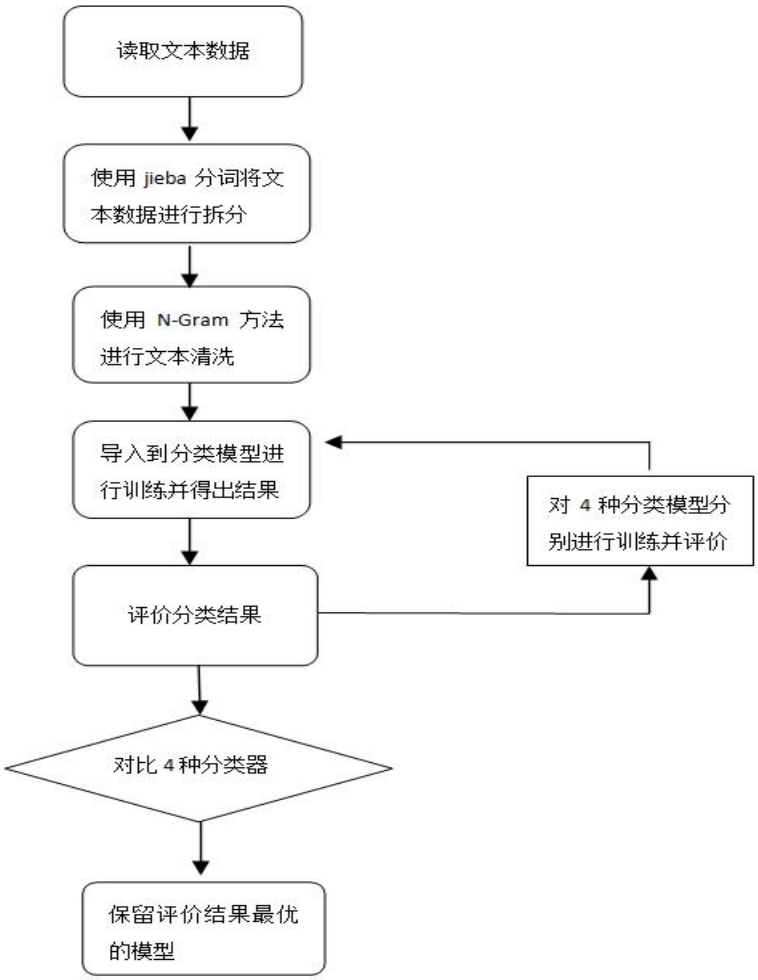
2、分类时须要先计算待分类样本和全体已知样本的距离。才干求得所需的 K 近邻点，计算量较大，尤其是样本数量较多时。

针对这样的情况能够事先对已知样本点进行剪辑。去除对分类作用不大的样本，这一处理步骤仅适用于样本容量较大的情况，假设在原始样本数量较少时采用这样的处理。反而会添加误分类的概率。



三、问题一的分析与求解

3.1 问题一算法流程图



3.2 数据预处理

通过对原始文本的数据提取可以观察到，原始数据对于我们所要采用的机器学习的分类算法来说是“脏数据”，这是因为机器学习的分类算法不能直接对一整段文字进行处理，而必须要先经过分割词，剔除无用词等步骤。因此我们采用结巴分词，以及 N-Gram 方法对数据进行清洗，具体操作如下:

3.2.1 jieba 中文分词

中文分词是对于中文文本处理的一个基础步骤，相对于英文，中文在一句话中没有词的界限，因此分词就显得尤为重要。分词，即将一段文字分割成一个个的词组或者是词。在 Python 语言中分词工具很多，包括盘古分词、Yaha 分词、jieba 分词、清华 THULAC 等，我们这里选用对中文分词效果优异的 jieba 分词

来进行文本处理。其处理效果如下所示：

原始文本:	A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆
分词文本:	'A3','区','大道','西行','便','道',', ','未管','所','路口','至','加油站','路段',', ','人行道','包括','路灯','杆'。

3.2.2N-Gram 清洗数据

N-Gram（有时也被为 N 元模型）是在自然语言处理中一个极其关键的概念，通常应用在 NLP 中，人们基于一定的语料库（在本题中，我们基于的是用于过滤无效符号或者词组的停用词库），可以利用 N-Gram 来预计或者评估一个句子是否合理

3.2.2.1 停用词库

停用词是指在信息检索中，为了节省存储空间和提高搜索效率，在处理自然语言数据（或文本）之前或之后会自动过滤掉某些字或词，这些字或词即被称为 Stop Words（停用词）。这些停用词都是人工输入、非自动化生成的，生成后的停用词会形成一个停用词库，这些词即一段文字中无用的词，这些词通常有以下的特征：

（1）不同类型的符号，例如一些常见的标点符号，转折符，引号，括号等，这些符号本身并不具有什么特殊的意义。相反，它们的存在可能会对机器学习的分类算法产生不利的影响，所以我们要在清洗数据的时候把这部分符号去除掉

（2）常见的中文文本停用词，例如“啊”，“哎”这类的语气词以及“强烈”这类的形容词，另外一些连接词，转折词也是在停用词库之内的，这些词往往只起到修饰作用，对理解文本的作用不大

（3）另有一些在基本停用词库中不包含的词组，但是这部分词在所给出的文本数据中也是对分类没有实际作用的，这部分词我们可以采取人工添加的方式进行补充

3.2.2.2 清洗文本举例

原始文本:	A3 区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆
分词文本:	'A3','区','大道','西行','便','道',', ','未管','所','路口','至','加油站','路段',', ','人行道','包括','路灯','杆'
清洗文本:	'大道','西行','道','未管','路口','加油站','路段','人行道','包括','路灯','杆'

通过对比可以看到形如，“A3”，“区”，“，”，“至”这样的词或者是符号都已经被清洗掉了，留下的是一些对判断文本内容更为重要的关键词

3.3 文本特征

3.2.1 划分训练集与测试集

将清洗后的文本中的 85%作为训练集，15%作为测试集，训练集用来对模型调参，训练集用来检验模型的精确度。

3.2.2 构造文本特征

在不考虑分词与文本中上下文的关系情况下，我们只是关心每个出现过的词

的权重。我们将分好的训练集和测试集单独进行如下步骤。将去除停用词后那些不重复的特征词单独作为一列特征词汇集合成为词表使用 CountVectorizer 类把所有训练文本中的特征词转换为词频矩阵

```
vocabulary={'00': 0, '0000': 1, '00000000': 2, '03': 3,
            '030': 4, '05': 5, '06': 6, '0638': 7, '064': 8,
            '07': 9, '0731': 10, '08': 11, '09': 12, '10': 13,
            '100': 14, '1000': 15, '100000': 16, '1000w': 17,
            '1003788': 18, '1005': 19, '100mg': 20, '100ml': 21,
            '100w': 22, '101': 23, '102': 24, '104': 25,
            '10404': 26, '105': 27, '106': 28, '107': 29, ...})
```

再通过此类下的 fit_transform 函数统计每个特征单词在文本中出现的频率作为该文本基于此特征词的特征。我们可以通过词频矩阵看到结果

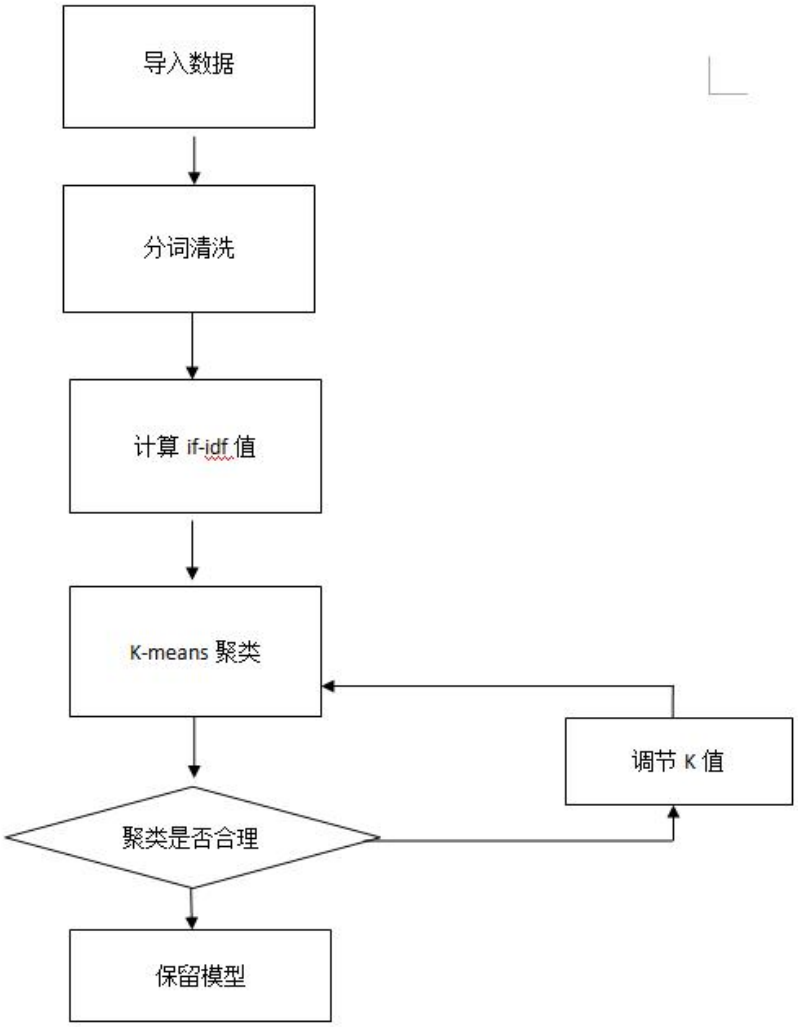
词频矩阵			
(0, 715)	1	(0, 1701)	1
(0, 1712)	1	(0, 3102)	1
...	...	...	...
(0,12129)	2	(0, 12130)	1
...	...	...	...
(418,10691)	1	(418,10963)	1
(418,11403)	1	(418,11582)	1

为了降低一些频繁出现的没有意义的词而掩盖了那些很少出现却十分有用的词的情况，我们接下来选择使用 TfidfVectorizer 挖掘更有意义的特征。先选择把原始的特征词转换成 tf-idf 的特征矩阵，根据训练集新生成的词典和逆文档词频从而产生出高权重的 tf-idf

td-idf 特征矩阵					
1	(0,12794)	0.20299592450145734	2	(0,12657)	0.09608977174818618
3	(0, 12302)	0.20299592450145734	4	(0,12194)	0.1726555603794361
...	...	...	...		
17	(0, 9638)	0.20299592450145734	18	(0, 9332)	0.20299592450145734
...	...	...	...		
39	(0, 4700)	0.20299592450145734	40	(0, 4649)	0.14816011806850232
...	...	...	...	...	...

四、问题二的分析与求解

4.1 问题二算法流程图



4.2 数据处理

根据对附件 3 的数据观察，留言主题相较于留言详情更易于分割出有效的地点和人群，所以我们选择留言主题作为文本分类的数据。同样这次数据需要预处理用以清理掉意义不大的“脏数据”，使留言分类更加准确。

我们与问题一采用同样的 jieba 分词和 N-Gram 清洗数据方法，但是依据数据对停用词库进行了改变。根据下表数据对比可知，文本中存在部分的字母地点，这字段是我们对特定地点以及特定人群进行留言分类的重要依据，因此我们把字母和部分地点词从停词库中清除。

A5 区劳动东路魅力之城小区油烟扰民
A 市经济学院寒假过年期间组织学生去工厂工作

L 市物业服务收费标准应考虑居民的经济承受能力
A 市江山帝景新房有严重安全隐患

4.3 构造文本特征

在不考虑分词与文本中上下文的关系情况下，我们关心的是每个出现过的词的权重。我们将分好的训练集和测试集单独进行如下步骤。将去除停用词后那些不重复的特征词单独作为一列特征词汇集合成为词表使用 CountVectorizer 类把所有训练文本中的特征词转换为词频矩阵

4.4K-means 聚类

k 均值聚类算法（k-means clustering algorithm）是一种迭代求解的聚类分析算法，其步骤是随机选取 K 个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。聚类中心以及分配给它们的对象就代表一个聚类。每分配一个样本，聚类的聚类中心会根据聚类中现有的对象被重新计算。这个过程将不断重复直到满足某个终止条件。终止条件可以是没有（或最小数目）对象被重新分配给不同的聚类，没有（或最小数目）聚类中心再发生变化，误差平方和局部最小。

4.5 热点指数

热度评价依据为点赞数加上反对数，在数据中我们发现，点赞所在列和反对所在列的顺序并不固定，我们考虑了两个方案，
（1）对热度系数结果取绝对值
（2）在计算热度系数之前，我们先判断哪个是点赞列哪个是反对列 ，在进行相减。
最后考虑到有小概率出现反对数大于点赞数，使得绝对值结果较大却并不是我们需要的热点问题情况造成误差，我们最后采取了第二种方案

4.6 时间范围

对热点事件的时间范围我们采取逐步压缩范围的方法，先将最小值定为最大，最大值设为最小，然后逐一循环每类中的所有时间数据与最大值最小值不断替换，实现范围的最终确定。
为了解决在读取时间数据时可能会出现时间格式的不同的情况，我们选择分别判断时间类型，分别采用两种不同的方式将时间转换为时间戳，方便我们进行大小比较。

第一类:可以正确的时间格式字符串读入，采用 timeme = time.strptime()转换为时间戳
2019/2/14 20:00:00

1550145600

第二类:以浮点数读入，先使用 xlrd.xldate.xldate_as_datetime(,0)函数转换为我们需要的时间格式，在转换为时间戳

43419.6716666667

2018-11-15 16:07:12

1542269232

处理完时间数据存储到时间列表中

2019/1/8 10:26:25
2019/7/7 7:28:06
2019-02-26 15:22:05
2019-04-26 15:28:42
2019-05-14 11:22:13
2019-05-30 17:32:02

4.7 获取地点/人物

根据对数据的对比观察，发现我们所要获取的地点信息绝大多数为英文字母开头 且在句子的前半部分 故采取以下办法。先对留言主题分词后的词语进行逐个判断 ，判断是否含有字母，当找到地点字母后，向后排查地点，与留言详情进行匹配，在大概率获取得到我们所需要的信息后跳出此次循环防止向后寻找到更多多余信息

热度排名	问题 ID	热度指数	时间范围	地点/人群	问题描述
1	1	2388	1970/04/26 17:46:40 至 2019/12/17 18:40:00	A2 天下公司	A2 区富惠天下公司涉嫌非法集资一事为何迟迟不立案
2	2	2159	1970/04/26 17:46:40 至 2020/01/06 07:48:01	A5 区洞井镇	A5 区洞井镇目前仅有一所民办的同升湖高中学校
3	3	1795	1970/04/26 17:46:40 至 2019/12/25 02:12:23	A7 优步花园	请督促 A7 县星沙街道办、县住建局积极处理恒大翡翠华庭相关问题
4	4	709	1970/04/26 17:46:40 至 2019/12/02 16:36:16	A4 市绿地	A4 区绿地海外滩小区距渝长厦高铁太近了
5	5	613	1970/04/26 17:46:40 至 2020/01/15 07:56:21	A2 黄兴路步行街	A2 区黄兴路步行街大古道巷住户卫生间粪便外排

问题ID	留言编号	留言用户	留言主题	留言时间	留言详情	点赞数	反对数
1	191311836	A0003583	A2区富惠天下公司涉嫌非法集资一事为何迟迟不立案	2019/3/12 20:58:45	西地省富惠天下商务信息服务有限公司于2014年在A2区注册成立，并通过富惠天下互联网平台发布项目，向社会吸引投资人投资理财。2018年11月，该公司经A2区金融办同意，发出良性退出方案。但至今未按方案兑付投资人本金和利息，且该公司发布项目也涉嫌虚造。今年2月，有几十位投资人先后到A市A2区公安分局城南路派出所报案，且提交了报案材料，但至今未得到立案的消息。该公司负责人人至今未向投资人履行协议和良性退出方案的承诺。当防止该公司负责人转移资产，逃避责任，恳请责成A2区公安分局迅速立案查处，维护广大投资人权益。	0	0
2	194343	A0001061	承办A市58车贷案警官应跟进关注留言	2019/3/1 22:12:30	胡书记：您好！58车贷案发，引发受害人举报投诉，也引起市领导的重视，公布了受害人的留言，使受害人深受感动，也看到了希望。但是，A市A4区经侦并没有跟进市领导的留言，案件调查进展报告、司法审计报告还是没有公布。鉴于此情，垦请胡书记关注此案，督促A4区经侦，领会市领导意图，尽快跟进领导留言。	733	0
3	206770	A0009198	A2区暮云派出所报案1个月，反复推脱不予立案	2019/10/11 11:32:44	您好！很抱歉冒昧的打扰您了，实属无奈之举。今年9月12日在A2区格兰小镇135栋**********拨打110报警电话报警，暮云派出所的民警前来接警。我们因被格兰小镇135栋业主肖钢诈骗而报警，警方未第一时间出警，期间三次打来电话称能不能不来现场，在我们再三要求及多次保健下才姗姗来迟。警方到达现场后并未收集我们说提及的证物（电脑等）导致我们到警局后嫌疑人的妻子潜回家伙同他人转移.我们当场提交了我们所准备的材料（打款截图、聊天记录等）之后被要求回家等消息，大概2个星期左右没得到任何答复，没有任何进展.多次前往派出所问询，民警都不予理睬.直到9月25日我们到市局的信访部门寻求帮助后，在信访部门领导的帮助下26日暮云派出所才安排了一个叫蒋黔楚的所长接待我们，并说明了立案条件，让我们准备资料，并承诺立案条件达到（诈骗金额120万以上，被骗人数30人）就能马上立案侦查。第二天被骗人员到所里报警、录口供，并且补交了材料后离开.直到今天，10月11日依然没有任何消息，打电话去询问一直被各种推诿.立不立案依然不给答复。暮云派出所的所作所为令民众寒心！请求帮忙让犯罪行为得到惩罚，让民众的冤有处申！	1	0

五、问题三的分析与求解

5.1 答复意见的评价指标

主要从一下 3 个方面来进行评价

5.1.1 相关性：

为了对留言回复的相关性进行评价，需要从附件四的问题详情和留言回复中分别提取出特征词，使用词频或者 TFidf 构造出词袋模型，构建样本间的相关性矩阵并对数字映射后的特征做余弦相似度匹配。匹配公式如下:

$$\cos\theta = \frac{a \cdot b}{||a|| \cdot ||b||}$$

其中a，b分别代表两个文本特征词

5.1.2 完整性：

在不通读句子的前提下，难以判断留言回复是否完整，故不妨选用留言长度与回复长度之比来定义完整性。比值越大说明完整性越差

$$K = \frac{len\ （s）}{len\ （m）}$$

其中s,m代表文本长度

5.1.3 可解释性：

可解释性表示在观察的基础上进行思考,合理地说明事物变化的原因，事物之间的联系，或者是事物发展的规律。就本题而言，可解释性指的是留言回复是否科学，是否权威，是否起到作用，表示这些特征的词语比如“根据”“由于“正在“重视””解决”等等，这些特征词构成特征词集，并与留言回复分词后的集合对比，回复分词集合含有的特征词数越多，该留言的可解释性越强

$$t = \frac{s}{10}$$

其中s代表留言回复词集与特征词集的符合个数

自定义特征词表

情况	办理	理解	部门	规划	街道	根据	组织	关心	人员	业主	关心	咨询	社区	登记
学生	服务	申请	政府	标准	教育	事件	建议	居民	提供	发展	关注	投诉	意见	方案
条件	符合	通知	回复	收悉	相关	《》	支持	留言	监督	政策	由于	处理	工作	情况
建设	项目	我局	调查	现将	管理	办理	监督	理解	规划	街道	政府	施工	发展	关注
经营	实施	安排	解决	核实	收费	工程	整改	符合	人民	政府	重视	进一步		

5.2 标准化处理与计算

将上述得到的数据进行标准化处理，采用规范化方法，公式如下:

$$y_i = \frac{x_i - \min\limits_{1 \le j \le n} \{x_j\}}{\max\limits_{1 \le j \le n} \{x_j\} - \min\limits_{1 \le j \le n} \{x_j\}}$$

其中x_i表示三类特征得出的数据值

标准化后得到的结果在【0,1】之间，为了保证结果精确，超过 1 的数值仍按 1 计算，之后再进行线性加权求和，得到对留言回复的综合评价。

序号		相关性	完整性	可解释性
1	规范化前	0.3686635944700461	0.8378378378378378	48
	规范化后	0.2589041095890411	0.8378378378378378	1.8045112781954886
	评估结果	1		
2	规范化前	0.2076923076923077	0.7344632768361582	22
	规范化后	0.2589041095890411	0.7344632768361582	0.8270676691729323
	评估结果	0.6068116851993772		
3	规范化前	0.288	0.6578947368421053	17
	规范化后	0.3590136986301363	0.6578947368421053	0.6390977443609023
	评估结果	0.5520020599443815		
4	规范化前	0.3076923076923077	0.36312849162011174	18
	规范化后	0.3835616438356164	0.36312849162011174	0.6766917293233082
	评估结果	0.4744606215930121		
4	规范化前	0.4339622641509434	0.5760869565217391	7
	规范化后	0.540966658051176	0.5760869565217391	0.2631578947368421
	评估结果	0.46007050310325237		

参考文献

[1]百度百科.“支持向量机”词条[EB/OL]https://baike.baidu.com/item/支持向量机/9683835?fr=Aladdin

[2]《Python 数据分析与挖局实战》，作者张良均、王路、谭立云等

[3]百度百科.“K-means 聚类算法”词条[EB/OL]http://baike.sogou.com/v8674815.htm?fromTitle=k-means