第八届“泰迪杯”
大学生数据挖掘挑战赛

题目名称：“智慧政务”中的文本挖掘应用

“智慧政务”中的文本挖掘应用

摘　要：近年来，科技的迅速发展，使得微信、微博、市长信箱、阳光热线等网络问政平台广泛应用，这成为政府了解民意、汇聚民智、凝聚民气的重要渠道，但随着与各类社情民意相关的文本数据量的不断攀升，给以往主要依靠人工来进行留言划分和热点整理的部门带来了极大挑战。于此同时，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。本文立足于此，基于数据挖掘技术对网络问政平台上的数据进行内在信息的挖掘与分析。

在本次数据挖掘过程中，我们首先通过 python 工具对获取到的数据进行数据预处理、分词、停用词过滤等操作，实现对评论数据的清洗和优化，有助于后面模型的建立。

接着，本小组采用多种方法来进行数据挖掘模型的构建，对文本进行多方面多维度的分析。其中，我们通过高斯贝叶斯模型和支持向量机模型实现对文本的分类，并探索在不同情况下，各模型分类效果的不同；通过基于 K-means 算法的文本聚类模型实现留言的聚类，通过隐马尔可夫模型（HMM）结合维特比算法实现对中文实体的识别和提取；通过结合实际情况和数据构建的热度评价模型来实现问题热度的评价和量化，最后通过 TextRank 的关键词摘要算法来构建回复内容的评价体系。

关键词：留言内容；文本分析；多项式贝叶斯模型；SVM 模型；HMM 的维特比算法；TextRank

Application of text mining in "intelligent government affairs"

Abstract： In recent years, the rapid development of science and technology, make WeChat, weibo, mayor mailbox, sunshine hotline ask ZhengPing machine is widely used, which become the government understand the importance of public opinion, gathering intelligence, condensed bull channel, but as the amount of text data related to all kinds of public opinion of the rising, leave a message to past mainly rely on artificial to divide and hot finishing department brought a great challenge.At the same time, with the development of big data, cloud computing, artificial intelligence and other technologies, the establishment of intelligent government affairs system based on natural language processing technology has become a new trend of social governance innovation and development, which plays a great role in promoting the management level and efficiency of the government.This paper based on the data mining technology based on the network platform for the internal information mining and analysis of the data.

In this data mining process, we first carried out data preprocessing, word segmentation, stop word filtering and other operations on the acquired data through python tools to achieve the cleaning and optimization of comment data, which is conducive to the establishment of the following model.

Then, the team adopts a variety of methods to construct the data mining model and conduct multi-dimensional analysis of the text among them, we use the Gaussian Naive Bayes model and the Support Vector Machine model to realize the classification of text, and explore the different classification effects of each model in different situations.The text clustering model based on k-means algorithm is used to realize the clustering of messages. The Hidden Markov Model (HMM) and viterbi algorithm are used to realize the recognition and extraction of Chinese entities.The heat evaluation model built by combining the actual situation and data is used to evaluate and quantify the heat of the problem. Finally, TextRank's keyword summarization algorithm is used to construct the evaluation system of the response content.

Key words： Message content； text analysis； Multinomial Naive Bayes； SVM； HMM viterbi algorithm； TextRank

目　录

一、 背景介绍

近年来，科技的飞速发展为人们生活带来了极大的便利，居民在线化程度不断提高，从居民反映问题角度来说，人们不再依靠手写信或亲自走访有关部门进行提议和投诉，取而代之的是，微信、微博、市长信箱、阳光热线等网络问政平台的逐步应用，这成为政府了解民意、汇聚民智、凝聚民气的重要渠道，但随着各类社情民意相关的文本数据量不断攀升，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。于此同时，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，对提升政府的管理水平和施政效率具有极大的推动作用。所以，将文本挖掘技术应用于“智慧政务”中已经迫在眉睫。

二、 挖掘目标

本次挖掘针对收集自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，利用自然语言处理和文本挖掘的方法，在对文本数据进行数据清洗、中文分词、停用词过滤、同义词转换后，通过建立多项式贝叶斯模型、SVM 模型 、隐马尔代夫模型（HMM）、文本聚类模型等，实现对文本留言意见的分类和热点分析，并对相关部门的答复意见做出评价，以期望尽己所能为智慧政务的发展做出一点贡献，详细目标如下：

在问题一群众留言分类中，通过使用 tf-idf 权重向量，采用文本分类领域较为成熟的三种模型（多项式贝叶斯、高斯贝叶斯、SVM 模型），探索在不同情况下各模型的分类效果，进而比较分析确定留言分类的模型。

在问题二热点问题挖掘中，基于 k-means 算法进行文本聚类，并综合考虑数据各参数情况，结合现实，构建热点问题的评价指标，最终基于隐马尔可夫模型(HMM)的维特比算法尝试对问题中的地名进行识别，提供一套完整的热点问题评价和提取体系。

在问题三答复意见的评价中，基于回复内容的情况，联系现实，从时效性、完整性、相关性方面入手，构建一套回复评价体系并实现。

三、 问题解决

1 总体流程

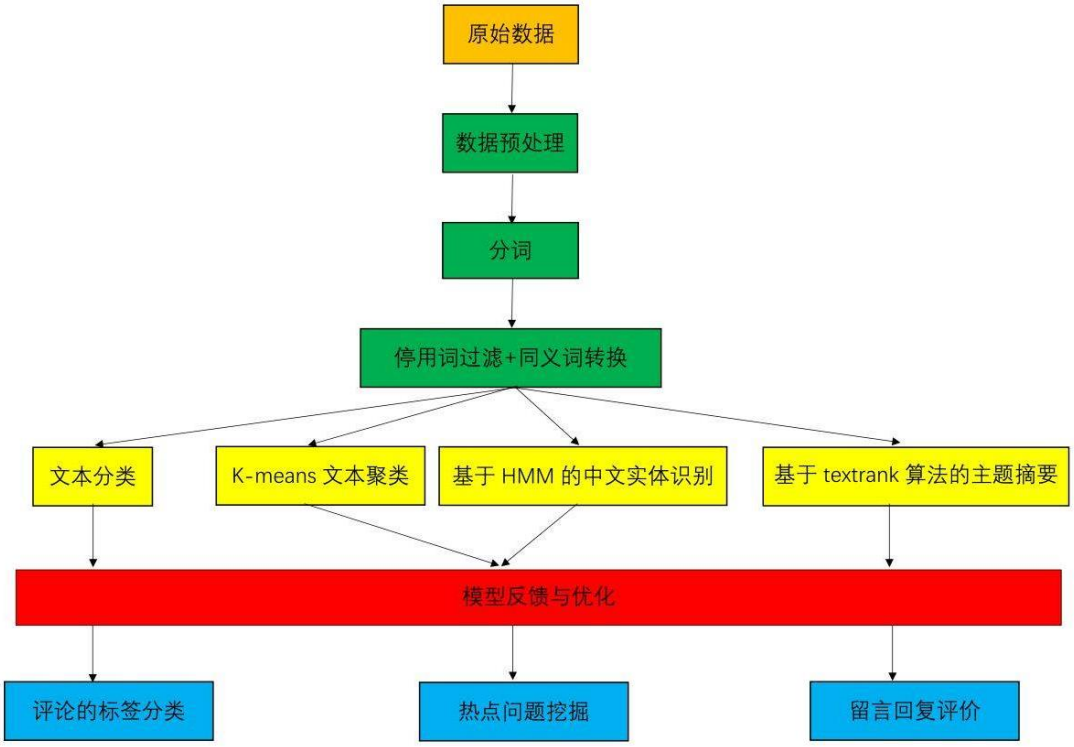


图 1　总体流程图

本论文的分析流程主要分为以下几个部分：

第一步：获取原始数据，部分数据为自行爬取或者训练；

第二部：对数据进行基本的处理，包括数据预处理，文本分词，停用词过滤和同义词转换等操作；

第三步：通过多种算法和多种模型对处理后数据进行多维度分析；

第四步：根据模型给出的结果进行反馈和参数调优，以求得到更好的效果；

第五步：对结果进行分析和价值信息的挖掘，对存在的问题提出意见和建议。

2 具体步骤

2.1 数据说明

本文主要的数据为赛事方给，内容网络问政平台的群众留言，其他数据例如HMM 中文分词语料为网上爬取，数据统计自人民日报语料库，相对于其他数据该数据内容完整，语料覆盖面广，另外词语的词性标注已经完成，可以更好地进行接下来其他参数的训练。

2.2 数据预处理

得到初始数据，首先要进行数据的预处理。群众留言中和回复内容中有许多价值较低的内容，如果不进行数据的预处理，不光会使得到的数据不真实并且对我们后期的模型优化和建议也有着误导的影响。

下面我们将从数据唯一性、统一性方面对数据进行处理，以便于后续有效深入的数据分析与挖掘。

2.2.1 数据唯一性—文本去重

留言文本去重，即删除留言数据中同一用户针对同一个问题做出的相同留言，此问题进行文本去重的主要原因如下：

① 若一个用户对同一个问题做出多条相同留言，难免会带有感情色彩，这些评论一般都是没有意义的。

② 进行文本去重会提高分类模型的准确度和真实性，比如示例数据-附件1 中存在多条“请给 A5 区木莲路冯家冲安置小区开通天然气和消防水源”，这样的留言是会影响分类的模型，从而影响准确度，所以予以删除。

通过分析数据发现，留言平台群众留言的数据比较多，所有针对同一个问题进行留言反映的情况很多例如示例数据-附件 3 中出现“A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气”和“A5 区劳动东路魅力之城小区一楼的夜宵摊严重污染附近的空气，急需处理!”这两条留言都是反映魅力之城小区夜宵摊污染空气问题的，但是两者的文本相似度较高，如果误将两条留言当作重

复处理显然是不行的，所以在比较了多种文本去重方法之后，我们小组打算采用完全重复去重方式，即将留言内容完全重复的数据进行剔除，其他情况都保留。

通过之后的数据分析，我们发现效果还是不错的。

2.2.2 数据统一性—统一时间戳格式

保持同一类型数据的统一性对于数据的统一操作起着至关重要的作用，以下是全部数据-附件三中的部分数据。

表 1　留言格式对比表

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
188006	A000102948	A3区一米阳光婚纱艺术摄影是否合法纳税了？	2019/2/28 11:25:05	，因为地处居民楼内	0	0
188007	A00074795	咨询A6区道路命名规划初步成果公示和城乡门牌问题	2019/2/14 20:00:00	0年都未曾更换过，	0	1
188031	A00040066	反映A7县春华镇金鼎村水泥路、自来水到户的问题	2019/7/19 18:19:54	，且天还没黑就开	0	1
188039	A00081379	A2区黄兴路步行街大古道巷住户卫生间粪便外排	2019/8/19 11:48:23	行清扫。没有解决	0	1
188059	A00028571	A3区中海国际社区三期与四期中间空地夜间施工噪音扰	2019/11/22 16:54:42	投诉业主，态度强	0	0
319659	A023956	A市江山帝景新房有严重安全隐患	2019-05-30 17:34:02	天气后过道全部是水	0	0
321736	A9992521	A市能不能提高医疗门诊报销范畴	2019-06-12 08:23:01	，小孩体弱多病各种	1	0
323034	A012414	L市物业服务收费标准应考虑居民的经济承受能力	2019-06-19 17:46:24	清运费按不超过30元	0	0
323149	A1241141	请给K3县乡村医生发卫生室执业许可证	2019-06-20 20:38:47	是证件下来啊。有些	0	0
336608	A0005623	希望西地省把抗癌药品纳入医保范围	2019-09-08 21:01:59	明有药可治，就是买	0	0

我们可以截取的两部分的“留言时间”格式是不同的，通过 python 进行格式的判定可以得到一部分是字符串格式，另一部分是 timestamp(时间戳)格式，因此格式不同对于接下来时间的运算和比较就会出现问题，所以进行时间数据格式的统一是必要的，这里我们利用的是 python 的格式判断函数 isinstance()然后将 timestamp 格式的数据都转换成字符串，便于后续的处理。

2.3 文本分词

中文分词是中文信息处理的重要的基础环节，可以从以下几点来认识，首先，“词”是组成句子的基本单位，要对句子进行分析，首先要对“词”进行分析，只有在这个基础上才能谈得上进一步做其他的处理；其次，计算机有关汉语言的很大一部分是以机器词典（给出词的各项信息，包括句法信息，语义信息，甚至语用信息等）的形式存储的，项目接下来用的文本分类和文本聚类模型以及文本主题摘要都是在词的基础上进行的，因此进行文本分词是接下来一切工作的基础。

我们小组采用的是结巴分词工具，它是基于前缀词典实现高效的词图扫描，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)，继而采用动态规划查找最大概率路径，找出基于词频的最大切分组合，在对于未登录词方面，采

用了基于汉字成词能力的 HMM 模型，使用 Viterbi 算法。是目前中文分词领域较为成熟和稳定的工具。

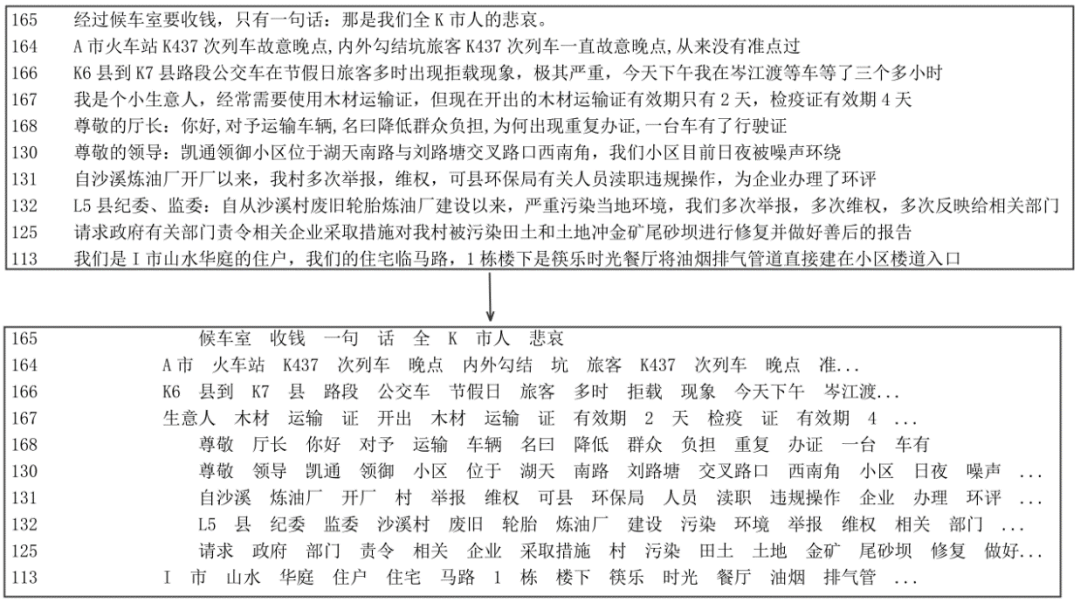


图 2　结巴分词

2.4 停用词过滤

在进行中文分词之后，我们将句子转换为一个个词语的集合，但是我们不难发现，在这个集合中存在少许“的”“了”“呢”“吗”“就”“才”之类对留言的意思影响不大的词语，这是因为在一段文本中有各种各样的词，有些能够作为关键字，有些显然不能作为关键字，这类明显不能作为关键字的词就是停用词。停用词主要包括了语气助词、 副词、介词、连词等，通常自身并无明确意义，只有将其放入一个完整的句子中才有一定作用的词语，这样的词不仅无法准确的表达文章的意思，难以提高关键字的准确性，同时还会降低处理的效率。因此在进行具体问题的解决之前，我们需要将这些停用词进行过滤。另外，停用词表要根据项目所需适时调整，比如在此次项目中，留言列表出现大量 A 区等词语，将其过滤有利于提高文本分类聚类的准确性。

本文主要基于根据自身需要所建立的停用词表，采用字符匹配的方式扫描分词词典进行删除。结果示例如下：

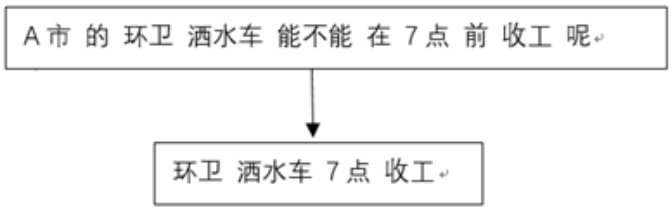


图 3　去除停用词

2.5 同义词转换

同义词转化是自然语言处理中比较重要的环节，同义词转换的好坏对于评价参数例如召回率等很大的影响，本项目采用同义词转换的原因主要有以下两方面：

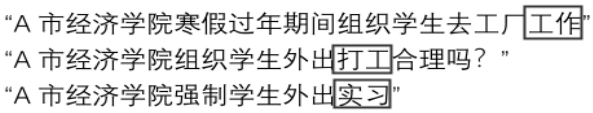


图 4　同义词示例 1

①在进行文本聚类时，同义词转换能大大提高文本聚类的性能。附件三-示例数据中有下面三条数据：

根据信息我们可以看出，三条数据都是反映 A 市经济学院学生被强制实习的情况，但是说法的不同造成三条分此后的结果分别是反应“工作”，“打工”和“实习”，所以这时候就需要同义词替换将称谓上的差异化进行一个弥补，以得到更好的聚类效果。

②同义词替换也能提高文本相似度算法的真实性和可靠性。对于问题三中，需要对回复意见进行评价，文本之间的相似度一定程度上展示了回复内容的相关性和完整性。拿下面这个例子来说：

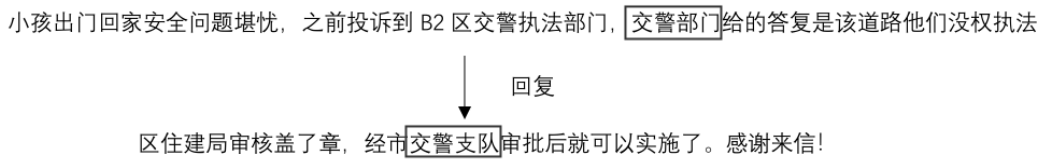


图 5　同义词示例 2

交警部门和交警支队在这里指的是同一个部门，但是基于 textrank 算法提取关键词后两者没有相同的关键词，造成留言和回复的相关性较低，这对回复的评价起了误导作用，因此利用同义词替换将“交警部门”和“交警支队”等同起

来，则可以很好的解决这一问题。

对于同义词转换的方法，很多人会想到使用 word2vec 来挖掘同义词，但是根据 word2vec 的原理可知，其挖掘的 topK 相近词其实是基于共现关系的相关词，并不是只有近义词，因此会出现很大的噪声。拿“富裕”来说，在一篇文章中利用 word2vec 寻找富裕的关键词。

```
1  ‘富裕’近义词:
2  富裕:1.0
3  富足:0.734253
4  穷困:0.714261
5  中产阶级:0.712797
6  贫穷:0.711762
7  富有:0.681186
8  富人:0.656687
9  有钱:0.633914
10 贫苦:0.632235
11 殷实:0.604537
```

图 6　同义词示例 3

如上所示，虽然富裕的相近词有富有，富足，但是也返回了穷困，贫穷等反义词，因此使用完全无监督的 word2vec 挖掘近义词效果并不好。

因为不同的同义词字典有不同文本擅长领域，考虑到我们群众留言分类的特殊性，我们采取哈工大的同义词词林以及实际留言中涉及到的同义词结合的方式完成同义词字典的完善。

```
魅力之城    魅城
经济学院    经济院
交警支队    交警部门    交警执法部门
住建局      建设局
干净        卫生        洁净
收费标准    收费方案
待遇        报酬        酬劳        薪金
公寓        房屋        楼房        住房
督察        督促        监督
农村        村委
法律        法规        实施条例    规则
拆迁        征拆
道路        公路
建设        建筑        修建
污染        脏乱
证照        证件        资格
```

图 7　同义词转换

2.6 群众留言分类

群众留言分类是根据用户的留言内容确定留言的一级标签，是典型的文本分类问题，本文采用文本分类领域较为成熟的三种模型（多项式贝叶斯、高斯贝叶斯、SVM 模型），采用 tf-idf 权重向量，寄希望于探索在不同情况下各模型的分类效果，继而通过比较分析确定留言分类的模型。

2.6.1 统计各标签分布情况

在进行正式分类之前，我们首先需要了解各标签的分类情况，通过 python 的 plot 函数对各标签的分布进行了可视化呈现。

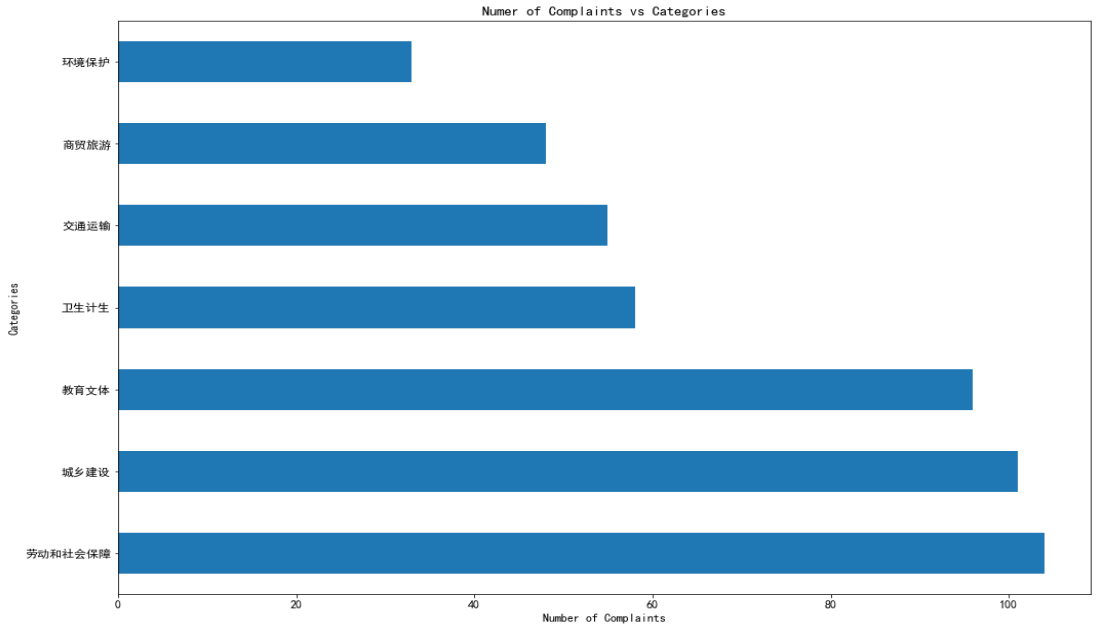


图 8　各标签分布情况

通过图表我们可以看出，各类标签数量的分布差别还是很大的，其中以环境保护类留言最少，少于劳动和社会保障的 1/2。这对于接下来样本的选取提供了依据。

2.6.2 训练生成词向量

词向量技术可以把自然语言中的词语转化为稠密的向量，相似的词会用相似的向量表示，通过转化，能更方便挖掘文字中词语和句子之间的特征。

TF-IDF 是一种常常被用于文本挖掘和资讯检索的加权技术。此权重是一个

统计量度，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。

词频 (TF) 指的是某一个给定的词语在该文件中出现的次数。这个数字通常会被归一化，以防止它偏向长的文件，其计算公式为：

$$TF_w=\frac{\text{在某一类中词条 w 出现的次数}}{\text{该类中所有的词条数目}}$$

IDF 的主要思想是：如果包含词条 w 的文档越少，IDF 越大，则说明词条具有很好的类别区分能力。某一特定词语的 IDF，可以由总文件数目除以包含该词语之文件的数目，再将得到的商取对数得到：

$$IDF=\log\left(\frac{\text{词库中的文档总数}}{\text{包含词条 w 的文档数}+1}\right)$$，分母加 1 的原因是避免分母为 0

最终的权重为两者之积：

$$TF\text{-}IDF=TF*IDF$$

2.6.3 文本分类

2.6.3.1 多项式贝叶斯

多项式贝叶斯是朴素贝叶斯分类器的一种针对文本分类的常见模型，相对于伯努利模型来说，他更侧重于对词语进行分类。其原理主要是基于朴素贝叶斯，通过句子中各单词在各类中出现的概率来计算句子属于各类的概率。

多项式贝叶斯的计算公式如下：

$$P(x_i \mid y_k)=\frac{N_{y_k i}+\alpha}{N_y+\alpha n}$$

其中 Nyki=Σx 表示在训练集 T 中类 yk 具有特征 i 的样本的数量，Ny=$\sum_{i=1}^{|T|} Nyi$,表示训练集 T 中类 yk 的特征总数。平滑系数 a>0 防止零概率的出现，当 a=1 称为拉普拉斯平滑，而 a<1 称为 Lidstone 平滑。

2.6.3.2 高斯贝叶斯

高斯贝叶斯是在朴素贝叶斯的基础上改进与优化后的模型，相对于朴素贝叶斯来说，他能处理连续数据，并加入了一种经典假设“与每个类相关的连续变量的分布是基于高斯分布的”，因此高斯贝叶斯的计算公式如下

$$P(x_i = v \mid y_k) = \frac{1}{\sqrt{2\pi\sigma_{y_k}^2}} \exp\left(-\frac{(v-\mu_{y_k})}{2\sigma_{y_k}^2}\right)$$

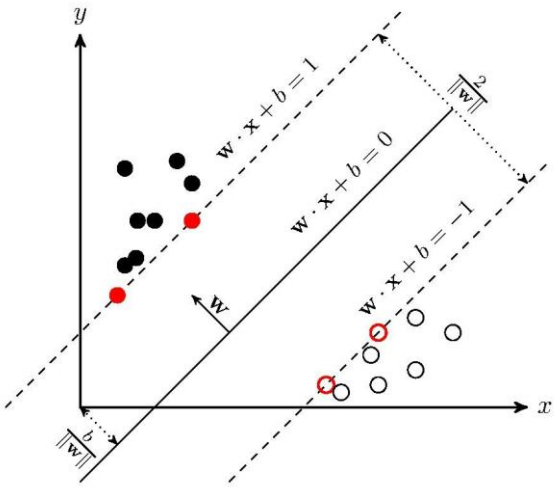
其中， μ_{yk}， σ^2_{yk} 表示全部属于类 y_k 的样本中变量 x_i 的均值和方差

2.6.3.3 支持向量机（SVM）

支持向量机（support vector machines，SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM 还包括核技巧，这使它成为实质上的非线性分类器。SVM 的的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最小化问题。SVM 的的学习算法就是求解凸二次规划的最优化算法。

① 线性 SVM 原理

SVM 学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。如下图所示，$w \cdot x + b = 0$ 即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个（即感知机），但是几何间隔最大的分离超平面却是唯一的。



阐述算法原理之前，先给出一些定义。假设给定一个特征空间上的训练数据集：

$$T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$$

其中，$x_i \in R^n$, $y_i \in \{+1, -1\}$, $i=1,2\cdots N$, x_i 为第 i 个特征向量，y_i 为类标记，

当它等于+1 时为正例；为-1 时为负例。再假设训练数据集是线性可分的。

因此对于给定的数据集和超平面 $w \cdot x + b = 0$，定义超平面关于样本点的几何间隔为：

$$\gamma_i = y_i \left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \cdot \boldsymbol{x_i} + \frac{b}{\|\boldsymbol{w}\|} \right)$$

超平面关于所有样本点的几何间隔的最小值为：

$$\gamma = \min_{i=1,2...,N} \gamma_i$$

实际上这个距离就是支持向量到超平面的距离。

根据以上定义，SVM 模型的求解最大分割超平面问题可以表示为以下约束最优化问题。

$$\max_{\boldsymbol{w},b} \ \gamma$$

$$s.t. \quad y_i \left(\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \cdot \boldsymbol{x_i} + \frac{b}{\|\boldsymbol{w}\|} \right) \geq \gamma \, , i = 1, 2, \ldots, N$$

通过化简，SVM 模型的求解最大分割超平面问题又可以表示为以下约束最优化问题:

$$\min_{\boldsymbol{w},b} \ \frac{1}{2} \|\boldsymbol{w}\|^2$$

$$s.t. \ \ y_i \left(\boldsymbol{w} \cdot \boldsymbol{x_i} + b \right) \geq 1, \ i = 1, 2, \ldots, N$$

这是一个含有不等式约束的凸二次规划问题，下面我们对其使用拉格朗日乘子法求其对偶问题。

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \left(\boldsymbol{x_i} \cdot \boldsymbol{x_j} \right) - \sum_{i=1}^{N} \alpha_i$$

$$s.t. \quad \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\alpha_i \geq 0, \ i = 1, 2, \ldots, N$$

对于如上形式，我们可以采用序列最小优化（SMO）等优化算法得到 α^*，继

而根据 α^*求解出 w 和 b,进达到我们最初的目的：找到超平面，即”决策平面”。

前面的推导都是假设满足 KKT 条件下成立的，KKT 条件如下：

$$\begin{cases} \alpha_i \geq 0 \\ y_i\left(\boldsymbol{w_i} \cdot \boldsymbol{x_i} + b\right) - 1 \geq 0 \\ \alpha_i\left(y_i\left(\boldsymbol{w_i} \cdot \boldsymbol{x_i} + b\right) - 1\right) = 0 \end{cases}$$

另外，根据前面的推导，还有下面两个式子成立

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x_i}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

由此可知在 α^*中，至少存在一个 $\alpha_j^*>0$,对此 j 有

$$y_j\left(\boldsymbol{w}^* \cdot \boldsymbol{x_j} + b^*\right) - 1 = 0$$

因此我们可以得到

$$\boldsymbol{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \boldsymbol{x}_i$$

$$b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i\left(\boldsymbol{x_i} \cdot \boldsymbol{x_j}\right)$$

对于任意训练样本$\left(\boldsymbol{x_i}, y_i\right)$，总有$\alpha_i = 0$ 或者$y_j\left(\boldsymbol{w} \cdot \boldsymbol{x}_j + b\right) = 1$ 。

若 $\alpha_i = 0$ ，则该样本不会在最后求解模型参数的式子中出现。若 $\alpha_i > 0$ 则必有 $y_j\left(\boldsymbol{w} \cdot \boldsymbol{x_j} + b\right) = 1$，所对应的样本点位于最大间隔边界上，是一个支持向量。这显示出支持向量机的一个重要性质：训练完成后，大部分的训练样本都不需要保留，最终模型仅与支持向量有关。

到这里都是基于训练集数据线性可分的假设下进行的,但是实际情况下几乎不存在完全线性可分的数据，为了解决这个问题，引入了“软间隔”的概念，即允许某些点不满足约束：

$$y_j\left(\boldsymbol{w} \cdot \boldsymbol{x_j} + b\right) \geq 1$$

采用 hinge 损失，将原优化问题改写为：

$$\min_{\boldsymbol{w},b,\xi_i} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$s.t. \ \ y_i\left(\boldsymbol{w}\cdot\boldsymbol{x_i}+b\right) \geq 1-\xi_i$$

$$\xi_i \geq 0\ ,\ i=1,2,\ldots,N$$

其中为 ξ_i“松弛变量”，$\xi_i = \max\left(0, 1-y_i\left(\boldsymbol{w}\cdot\boldsymbol{x_i}+b\right)\right)$，即一个 hinge 损失函数。每一个样本都有一个对应的松弛变量，表征该样本不满足约束的程度。C>0 称为惩罚参数， C 值越大，对分类的惩罚越大。跟线性可分求解的思路一致，同样这里先用拉格朗日乘子法得到拉格朗日函数，再求其对偶问题。

综合以上讨论，我们可以得到线性支持向量机学习算法如下：

输入：训练数据集 $T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_N,y_N)\}$，其中 $\boldsymbol{x}_i\in\mathbb{R}^n$， $y_i\in\{+1,-1\},i=1,2,\ldots N$

输出：分离超平面和分类决策函数

1）选择惩罚参数 C > 0 ，构造并求解凸二次规划问题

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\left(\boldsymbol{x_i}\cdot\boldsymbol{x_j}\right)-\sum_{i=1}^{N}\alpha_i$$

$$s.t. \quad \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$0\leq\alpha_i\leq C,\ i=1,2,\ldots,N$$

得到最优解 $\boldsymbol{\alpha}^*=\left(\alpha_1^*,\alpha_2^*,\ldots,\alpha_N^*\right)^T$

2）计算

$$\boldsymbol{w}^*=\sum_{i=1}^{N}\alpha_i^* y_i\boldsymbol{x}_i$$

选择 α^*的一个分量 α_j^*满足条件 $0<\alpha_j^*<C$，计算

$$b^*=y_j-\sum_{i=1}^{N}\alpha_i^* y_i\left(\boldsymbol{x_i}\cdot\boldsymbol{x_j}\right)$$

3）求分离超平面

$$\boldsymbol{w}^*\cdot\boldsymbol{x}+b^*=0$$

分类决策函数：

$$f(\boldsymbol{x}) = sign\left(\boldsymbol{w}^* \cdot \boldsymbol{x} + b^*\right)$$

② 非线性 SVM 算法原理

对于输入空间中的非线性分类问题，可以通过非线性变换将它转化为某个维特征空间中的线性分类问题，在高维特征空间中学习线性支持向量机。由于在线性支持向量机学习的对偶问题里，目标函数和分类决策函数都只涉及实例和实例之间的内积，所以不需要显式地指定非线性变换，而是用核函数替换当中的内积。核函数表示，通过一个非线性转换后的两个实例间的内积。具体地 $K(x,z)$ 是一个函数，或正定核，意味着存在一个从输入空间到特征空间的映射 $\phi(x)$ ，对任意输入空间中的 x, z ，有 $K(x,z) = \phi(x) \cdot \phi(z)$

在线性支持向量机学习的对偶问题中，用核函数 $K(x,z)$ 替代内积，求解得到的就是非线性支持向量机

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*\right)$$

综合以上讨论，我们可以得到非线性支持向量机学习算法如下：

输入：训练数据集 $T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$ 其中，$\boldsymbol{x}_i \in \mathbb{R}^n$ ， $y_i \in \{+1, -1\}, i = 1, 2, \ldots N$

输出：分离超平面和分类决策函数

1）选取适当的核函数 $K(x,z)$ 和惩罚参数 $C > 0$ ，构造并求解凸二次规划问题

$$\min_{\boldsymbol{\alpha}} \ \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x_i}, \boldsymbol{x_j}) - \sum_{i=1}^{N}\alpha_i$$

$$s.t. \quad \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$0 \le \alpha_i \le C,\ i = 1, 2, \ldots, N$$

得到最优解 $\boldsymbol{\alpha}^* = \left(\alpha_1^*, \alpha_2^*, \ldots, \alpha_N^*\right)^T$

（2）计算

选择 α^*的一个分量 α_j^*满足条件 $0<\alpha_j^*<C$，计算

$$b^* = y_j - \sum_{i=1}^{N} \alpha_i^* y_i K(\boldsymbol{x_i}, \boldsymbol{x_j})$$

（3）分类决策函数：

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_i K(x, x_i) + b^*\right)$$

在这里需要用到一个常用的核函数——高斯核函数

$$K(x, z) = \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right)$$

对应的 SVM 是高斯径向基函数分类器，在此情况下，分类决策函数为

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i^* y_i \exp\left(-\frac{\|x - z\|^2}{2\sigma^2}\right) + b^*\right)$$

2.6.3.4 三种模型预测结果分析

将全部数据按照训练集/测试集=4/1 的比例进行数据拆分，对训练集数据进行词频转换和 tf-idf 权重向量函数转化后进行以上三种模型的预测，预测结果如下：

模型一：多项式贝叶斯模型：

	precision	recall	f1-score	support
0	0.94	0.15	0.25	116
1	0.71	0.98	0.82	427
2	1.00	0.41	0.58	153
3	0.94	0.34	0.50	253
4	0.56	0.93	0.70	384
5	0.87	0.86	0.86	299
6	0.97	0.65	0.78	179
accuracy			0.73	1811
macro avg	0.86	0.62	0.64	1811
weighted avg	0.80	0.73	0.70	1811

模型二：高斯贝叶斯模型：

	precision	recall	f1-score	support
0	0.70	0.36	0.48	116
1	0.74	0.76	0.75	427
2	0.73	0.59	0.65	153
3	0.66	0.53	0.58	253
4	0.61	0.71	0.66	384
5	0.64	0.80	0.71	299
6	0.79	0.73	0.76	179
accuracy			0.68	1811
macro avg	0.69	0.64	0.65	1811
weighted avg	0.69	0.68	0.67	1811

模型三：SVM 模型：

	precision	recall	f1-score	support
0	0.84	0.71	0.77	116
1	0.89	0.95	0.92	427
2	0.84	0.88	0.86	153
3	0.86	0.74	0.79	253
4	0.82	0.88	0.85	384
5	0.93	0.90	0.91	299
6	0.88	0.91	0.89	179
accuracy			0.87	1811
macro avg	0.87	0.85	0.86	1811
weighted avg	0.87	0.87	0.87	1811

图 9　预测结果 1

通过结果我们可以发现，三种模型的 f1-score 的加权平均值分别为 0.70、0.67、0.87。显然 SVM 的预测效果要好于其他两种模型。

考虑到多项式贝叶斯和高斯贝叶斯是基于词语的频率进行概率预测的，所以由于数据各标签数量的分布不均，势必会对模型的预测造成影响。因此我们对数据进行欠抽样操作，每类抽取等数量样本进行训练。

训练结果如下图所示：

模型一：多项式贝叶斯模型：

	precision	recall	f1-score	support
0	0.91	0.84	0.87	110
1	0.72	0.90	0.80	84
2	0.92	0.92	0.92	99
3	0.92	0.77	0.84	101
4	0.85	0.68	0.75	115
5	0.88	0.91	0.90	99
6	0.79	1.00	0.88	92
accuracy			0.85	700
macro avg	0.86	0.86	0.85	700
weighted avg	0.86 ↑	0.85 ↑	0.85 ↑	700

模型二：高斯贝叶斯模型：

	precision	recall	f1-score	support
0	0.76	0.59	0.66	110
1	0.64	0.69	0.67	84
2	0.80	0.76	0.78	99
3	0.64	0.60	0.62	101
4	0.61	0.57	0.59	115
5	0.75	0.84	0.79	99
6	0.73	0.93	0.82	92
accuracy			0.70	700
macro avg	0.70	0.71	0.70	700
weighted avg	0.70 ↑	0.70 ↑	0.70 ↑	700

模型三：SVM 模型：

	precision	recall	f1-score	support
0	0.84	0.84	0.84	110
1	0.83	0.88	0.86	84
2	0.90	0.87	0.88	99
3	0.76	0.77	0.76	101
4	0.79	0.74	0.76	115
5	0.88	0.87	0.87	99
6	0.88	0.91	0.89	92
accuracy			0.84	700
macro avg	0.84	0.84	0.84	700
weighted avg	0.84 ↓	0.84 ↓	0.84 ↓	700

图 10　预测结果 2

从上图我们可以看到，通过欠抽样操作，多项式贝叶斯模型的 f1-score 提

高到了 0.85，高斯贝叶斯模型的 f1-score 提高到了 0.70，但涨幅不高，相对来说效果欠佳。而 SVM 模型的 f1-score 由于训练集和数据集的规模下降从 0.86 降低到了 0.84。

通过数据的结果可以看出，由于上述三种模型在文本分类中的原理不同，多项式贝叶斯由于依靠词语在训练集中的频率进行分类的判定，造成对于数据不平衡的敏感性方面，多项式贝叶斯模型高于 SVM 模型，但是在保证数据平衡或者通过欠抽样方式的前提下，多项式贝叶斯的分类效果还是比较好的。而对于 SVM 分类模型，稳定是它的主要特征，通过保证训练集的数量，同样可以达到比较好的效果。

2.6.4 模型改进和参数的优化

通过分类效果的比较，我们首先对 SVM 和多项式贝叶斯模型进行调参和优化，通过 klearn.model_selection 模块中的 GridSearchCV 函数对模型的各参数进行寻优，来实现模型的最佳效果，通过调优和比较分析，我们最终确定了使用的分类模型为 svm.linearSVC()实现的线性分类支持向量机模型，此模型基于 iblinear 库实现，有多种惩罚参数和损失函数可供选择，数据量大时也可以很好地进行归一化，既支持稠密输入矩阵也支持稀疏输入矩阵，比较适合数据量大的多分类问题。

2.7 热点问题挖掘

热点问题挖掘是目前自然语言处理领域比较热门的问题，对于网络问政平台来说，通过及时发现热点问题，有助于相关部门进行有针对性地处理，提升服务效率。另一方面将群众反映最为强烈的问题挖掘出来并进行优先处理也能体察民情，拉近政府和群众的关系，维护社会的稳定。

对于热点问题的挖掘主要是基于 k-means 算法进行文本聚类，通过综合考虑数据各参数情况，结合现实，构建了热点问题的评价指标，最终基于隐马尔可夫模型(HMM)的维特比算法尝试对问题中的地名进行识别，提供了一套完整的热点问题评价和提取的体系。

2.7.1 tf-idf 权重向量转化

这里的 tf-idf 权重向量转化和文本分类中的类似，在此不作赘述。

2.7.2 确定最优聚类个数

k-means 算法的缺点在于无法确定聚类的个数，而聚类个数的确定一定程度上决定着聚类的效果，因此我们采用手肘法和轮廓系数（silhouette_score）法来确定聚类的个数。

①手肘法

手肘法的核心指标是 SSE(sum of the squared errors，误差平方和)：

$$SSE = \sum_{i=1}^{k} \sum_{c \in C_i} |p - m_i|^2$$

其中，Ci 是第 i 个簇，p 是 Ci 中的样本点，mi 是 Ci 的质心（Ci 中所有样本的均值），SSE 是所有样本的聚类误差，代表了聚类效果的好坏。随着聚类数 K 的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和（SSE）自然会逐渐变小。但 K 增大到一定程度时，K 增大对 SSE 减小的作用越来越小，因此 K—SSE 曲线呈现手肘状，拐点附近的 K 值通常为适当的分群数量。

如下是我们根据附件 3_示例数据进行的 SSE 曲线：

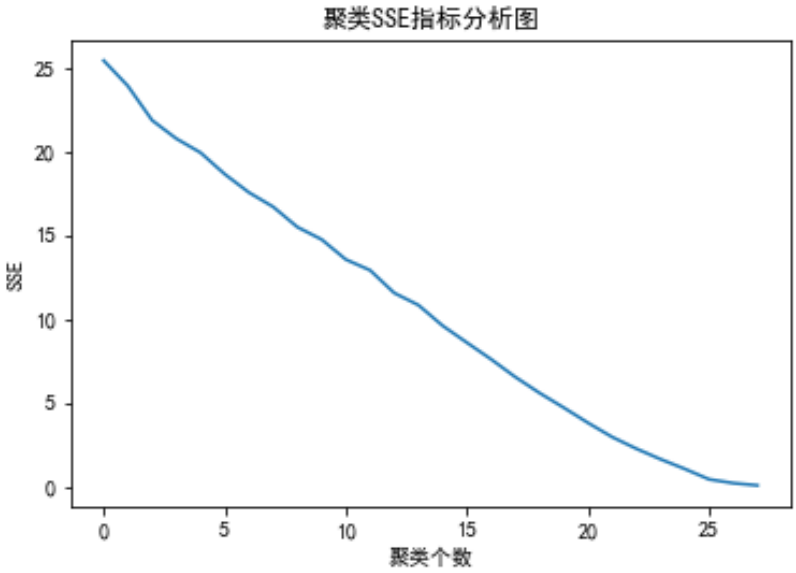


图 11　聚类 SSE 指标分析图

图表显示的 SSE 曲线接近一个弧线，没有比较明显的手肘点，所以采取下面的轮廓系数方法进行个数的确定。

②轮廓系数法

轮廓系数是评判聚类好坏的标准，结合类内聚合度以及类间分离度两种指标来计算得到。平均轮廓系数的取值范围为[-1,1]，且簇内样本的距离越近，簇间样本距离越远，平均轮廓系数越大，聚类效果越好。轮廓系数法就是基于平均轮廓系数对聚类效果进行评估的。

同样将附件三_示例数据进行聚类，求得轮廓系数曲线如下图所示：

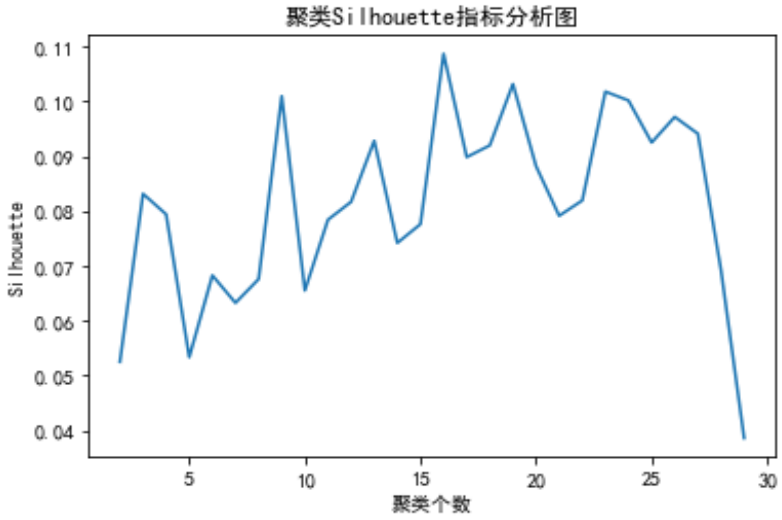


图 12　聚类 Silhouette 指标分析图

图像显示出比较大的振幅，当 k=16 时得到的轮廓系数值最大，因此接下来的聚类我们可以将聚类数确定为 16。

2.7.3 K-means 算法聚类

由于具有出色的速度和良好的可扩展性，Kmeans 聚类算法算得上是当前使用最普遍的聚类方法。这个算法通常是以局部最优结束，其原理也比较简单，很容易实现，收敛速度快，当结果簇是密集的，而簇与簇之间区别明显时，它的效果较好，主要需要调参的参数仅仅是簇数 k。

以下是示例数据的部分聚类结果：

表 3　示例数据聚类结果

问题ID	留言编号	留言用户	留言主题
0	336608	A0005623	希望西地省把抗癌药品纳入医保范围
1	360102	A1234140	A5区劳动东路魅力之城小区底层餐馆油烟扰民
1	360105	A120356	A5区魅力之城小区一楼被搞成商业门面，噪音扰民严重
1	360100	A324156	魅力之城小区临街门面油烟直排扰民
1	360101	A324156	A5区劳动东路魅力之城小区油烟扰民
2	360114	A0182491	A市经济学院体育学院变相强制实习
2	360111	A1204455	A市经济学院组织学生外出打工合理吗？
2	360113	A3352352	A市经济学院强制学生外出实习
2	360112	A220235	A市经济学院强制学生实习
2	360110	A110021	A市经济学院寒假过年期间组织学生去工厂工作

通过结果我们可以看到聚类的效果还是比较好的。

2.7.4 模型反馈与优化

因为自然语言处理的复杂性与中文称呼的多样性，造成对于同一问题的说法不尽相同，模型的反馈主要集中在同义词的转换和词典的补充。上述已说明具体方法，在此不做赘述。

2.7.5 热度模型的构建

热度模型的构建是挖掘热点问题的关键，通过研究和结合留言平台的实际情况，我们对于热度从以下参与性、普及性、时效性以及聚类模型的缺陷四个方面展开：

①参与性

参与性是指相关问题群众参与留言的程度，在这个问题上反映在同一类问题的相关留言数。考虑到网络问政平台的特殊性和严肃性，大部分人对于留言的发送更加谨慎和负责，且相对于对别人的留言点赞来说，一条留言的产生需要更多的时间和精力。基于此，我们认为参与性是衡量问题热度的一个重要的方面。

②认同性

认同性主要指群众对一个问题认同的程度，反映到实际问题上即为问题的点赞数。点赞数虽然相对于群众留言来说对于某一问题的反映程度相对较低，但考虑到一般人群只会对自己表示认同留言进行点赞，所以这一指标也是问题热度衡量模型中不可或缺的因素。

③时效性

时效性是指反映同一问题的时间维度上的密集性，如果一段时间内许多群众针对一类问题进行了很集中的反映和留言，那么说明问题出现的比较突然，同时与群众的矛盾也较为尖锐。下图为附件 3_全部数据中的部分数据：

表 4　数据的时效性体现

205	236303	A0002242	A市万科魅力之城有的房屋楼板严重开裂	2019/10/29 16:57:02
205	198084	A0002242	A市万科魅力之城小区近百户楼板开裂墙面开裂	2019/10/23 15:01:30
205	205168	A0002242	A市万科魅力之城近百户房屋楼板、墙面开裂!	2019/10/24 10:22:16
205	233338	A0002242	A市万科魅力之城楼板和墙面开裂，请政府管一管吧	2019/11/13 10:58:25

数据显示，针对 A 市万科魅力之城的房屋楼板开裂问题在不到一个月的时间内反映了多次，这充分体现了问题的迫切性和突然性。所以充分考虑问题的时效性也能在一定程度上完善热度问题评价模型。

④k-means 算法是聚类领域里面较为成熟稳定的算法，但是由于其基于距离作为相似度评价的指标，另外由于 k 的估算很难使聚类结果达到很完美的情况，造成基于 k-means 算法的文本聚类存在着某些聚类群中文本相关性不大，甚至根本无关的情况。而这些在一起无效的聚类势必会影响热点问题的挖掘。如下图：

表 5　聚类结果问题表

221	202928	A0006335	请拆除A2区金线街40号的违章建筑
221	242745	A0006372	A市一学生中考提前科的成绩怎么变成C了？
221	282665	A8801132	A3区A3区大道晚上10点后有很多货车通行，附近居民无法入睡
221	262531	A0005906	西地省普通话水平测试网上报名好难
221	230618	A0004365	A3区东方红路嘉顺苑优莱克超市后面有老虎机
221	221643	A0002859	请督促A市中院按诉讼程序开庭、尽快判决
221	247257	A0008033	A7县东六路（远大路至映霞路）路面破损不堪，严重影响行车安
221	284408	A0009183	西地省高速为何同工不同酬？

基于此，我们小组考虑有必要先将聚类结果进行可靠性检验，并赋予权重体现到最终的权重模型中，具体是通过利文斯顿距离计算同一类文本的相似度均值，并通过权重参与最后热度计算。

综上，考虑各方面因素后，我们小组建立了下面的热度评价模型：

$$Score(v) = \lambda_1(\sum_{i\in v} x_i - \sum_{i\in v} y_i) + \frac{\lambda_2}{T_{max}-T_{min}+1} + \lambda_3 \frac{\sum_{i\in v}\sum_{j\in v} D(i,j)}{n(n+1)}$$

其中，v 代表聚类之后的某一簇群；γ1、γ2、γ3 分别代表权重系数；Xi 代表对象 i 点赞数，Yi 代表反对数；Tmax,Tmin 分别代表该群时间戳的最大值和最小值；D(i,j)代表 i,j 文本的利文斯顿距离。

其中时效性方面因为是 1/（Tmax-Tmin）为了防止分母出现零的情况，采取类似拉普拉斯平滑处理的方式在分母加一，因为单位是秒，所以不会对结果产生较大的影响。

2.7.6 提取主流问题描述

①基于 LDA 的主题模型

LDA 模型是 Blei 于 2003 年提出的三层贝叶斯主题模型，通过无监督的学习方法发现文本中隐含的主题信息，目的是要以无指导学习的方法从文本中发现隐含的语义维度-即“Topic”或者“Concept”。隐性语义分析的实质是要利用文本中词项(term)的共现特征来发现文本的 Topic 结构，这种模型在文本聚类、主题挖掘、相似度计算等方面都有广“泛的应用，相对于其他主题模型，其引入了狄利克雷先验知识，因此，模型的泛化能力较强，不易出现过拟合现象。其次，它是一种无监督的模式，只需要提供训练文档,它就可以自动训练出各种概率，无需任何人工标注过程，节省大量人力及时间。

虽然 LDA 主题模型有很多优势和特点，但是不适合将模型用在网络问政平台留言问题上，首先聚类后的主题群体比较少，多数簇群只有几个对象，这就造成基于 LDA 的主题模型无法进行训练，并且由于样本较少会造成误差较大；另外，我们本阶段目的是对于一个簇群提取出一个主题短句，而 LDA 主题模型是在分词的基础上基于词语进行主题的提取，造成最终无法形成一段完整的主题描述。

②基于主流话题的主题模型

考虑到自然语言处理首先遵从的适应性原则，我们首先剖析了网上问政平台上留言主题描述的特征：首先是主题描述简短，多数只有一句话。其次，群内主题描述相似度极高，很多只有几个词语上的微小差别。结合这些特征我们小组采用了一种基于主流话题的主流模型进行主题的挖掘。

基于主流话题的主题模型，顾名思义，即通过选择群内所有留言中最具有代

表性的话题作为群的主题描述，该方法适合聚类后群内主题高度相似的情况。通过主流话题作为群内主题一方面可以较准确的对主题进行表示，另一方面，思路清晰，程序简单高效。

提取主流话题的主题模型的核心在于如何定义主流话题，基于此我们构建了一种主流话题的确定模型：

$$Representation(v) = min_{i \in v}(\sum_{j \in v} D(i,j) - (x_i - y_i))$$

通过主流话题提取模型识别出来的主题大部分能成功概括簇群主题，但是遇到上文提到的聚类不成功的簇群时效果会比较差，这还有待后面的改进和优化。

2.7.7 命名实体识别

近年来，隐马尔可夫模型（HMM）在中文自然语言处理中有非常重要的用途，广泛地应用于中文分词、命名实体识别、词性标注、信息抽取等方面。HMM 有成熟的学习算法，训练识别速度快，适用于一些对实时性有要求以及像信息检索这样需要处理大量文本的应用，而在此次项目之中，我们需要从大量文本中识别出人群地点等实体，所以本文采用 HMM 作为命名实体识别训练模型，旨在提高识别速度和准确性。

采用隐马尔可夫模型进行命名实体识别主要是求观察序列的背后最可能的标注序列，即根据输入的一系列单词，去生成其背后的标注，从而得到实体。主要原理如下：

HMM 中，有 5 个基本元素：{N,M,A,B, π}，这 5 个基本元素的意义为：

N:状态的有限集合，指每一个词语背后的标注。

M:观察值的有限集合，指每一个词语本身。

A:状态转移概率矩阵，指某一个标注转移到下一个标注的概率。

B:观测概率矩阵，指在某个标注下，生成某个词的概率。

Π:初始概率矩阵，指每一个标注的初始化概率。

而以上的这些元素，都是可以从训练语料集中统计出来的。最后，我们根据这些统计值，应用维特比（viterbi）算法，计算词语序列背后的标注序列。

本文采用根据自身数据特点稍作改进的人民日报语料库作为预料文件。在进行地点识别的任务之中，首先将状态定义为以下集合：

表 6　状态定义集合

角色	意义	例子
A	上文	**解决** A4 区凯乐国际城周边道路乱停车问题
B	下文	施工噪音**扰民**
X	连接词	A 区**和** B 区
C	特征词的一般性前缀	**经济**学院
F	特征词的人名前缀	**味可可**小吃
G	特征词的地名性前缀	交通银行**北京分行**
K	特征词的机构名、品牌名前缀	**自梦强数码**科技有限公司
I	特征词的特殊性前缀	**中央**电视台
J	特征词的简称性前缀	**巴**政府
D	地点特征词	设立南塘城轨**公交站**
Z	非机构成分	
L	方位词	**东　南**
M	数量词	58
P	数量+单位（名词）	**二期**
W	特殊符号，如括号，中括号	**（）　【】**
S	开始标志	**始##始**

用 HMM 来实现的命名实体识别算法，关键在于标签的自定义，需要人工定义尽可能多的标签，在本文中，我们将尽可能多的地点进行了人工定义，然后在训练语料集里面自动标注这些标签，标注完语料集，再进行生成 HMM 中的转移概率、初始概率、发射概率。

2.7.8 结果分析

通过上述构建的模型体系，得到热度排名前 5 的问题，分别为“请书记关注 A 市 A4 区 58 车贷案”、“A 市 A5 区汇金路五矿万境 K9 县存在一系列问题”、“反映 A 市金毛湾配套入学的问题”、“A4 区绿地海外滩小区距长赣高铁最近只有 30 米不到”、“A 市富绿物业丽发新城强行断业主家水”，具体信息见表 7：

表 7　热点问题排名表（前五）

热度排名	问题ID	热度指数	时间范围	地点/人群	问题描述
1	157	2346	2019/5/21 17:52:35至2019/5/21 17:52:35	A市 A4区	请书记关注A市A4区58车贷案
2	332	2107	2019/4/19 20:04:05至2019/4/19 20:04:05	A市 A5区 汇金路	A市A5区汇金路五矿万境K9县存在一系列问题
3	2053	1759	2019/11/18 12:23:22至2019/12/15 12:32:1	A市 金毛湾	反映A市金毛湾配套入学的问题
4	2680	671	2019/3/11 13:08:05至2019/3/11 13:08:05	A4区 绿地海外滩小区	A4区绿地海外滩小区距长赣高铁最近只有30米不到，合理吗？
5	2881	244	2019/10/7 1:25:25至2019/10/7 1:25:25	A市 富绿物业丽发新城	A市富绿物业丽发新城强行断业主家水

以问题“A 市 A5 区汇金路五矿万境 K9 县存在一系列问题”为例，分析热点问题留言明细表，从表 8 可以看出，共有六个用户针对此问题做出留言，其中

编号为 208636 的留言共收到 2097 点赞数，表明该问题在群众中的认同度很高，由此可见，A 市 A5 区汇金路五矿万境 K9 县存在的问题亟待解决。

表 8　热点问题明细表

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
332	275491	A00061339	A市五矿万境K9县负一楼面积缩水	2019/9/10	关于五矿万境·I	0	0
332	262599	A000100428	A市五矿万境K9县房屋出现质量问题	2019/9/19	我是西地省A市王	0	0
332	208069	A00094436	A5区五矿万境K9县的开发商与施工方建房	2019/5/5	1本人是A5区洞井	2	0
332	208636	A00077171	A市A5区汇金路五矿万境K9县存在一系列问	2019/8/19	我是A市A5区汇金	2097	0
332	215507	A000103230	A市五矿万境K9县存在严重的消防安全隐患	2019/9/12	预交房23栋没有	1	0
332	252650	A00010531	A市五矿万境K9县交房后仍存在诸多问题	2019/9/11	尊敬的相关部门	0	0

2.8 答复意见的评价

对于网上问政平台的留言，认真及时的回复是相关部门的责任也是义务，通过建立合理的回复评价体系对回复的质量进行评价，对于政府部门及时发现问题、不断提高办公效率有着积极的影响。

我们小组基于回复内容的情况，联系现实，从时效性、完整性、相关性方面入手，构建了一套回复评价体系，并对回复的质量进行了量化评价和针对性建议。

2.8.1 基于 Textrank 的中文摘要提取

TextRank 由 Mihalcea 与 Tarau 于 EMNLP’04 [1]提出来，其思想非常简单：通过词之间的相邻关系构建网络，然后用 PageRank 迭代计算每个节点的 rank 值，排序 rank 值即可得到关键词。

Textrank 算法的基础是 PageRank，PageRank 是用来解决网页排名的算法，网页之间的链接关系即为图的边，迭代计算公式如下：

$$PR(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} PR(V_j)$$

其中，PR(Vi)表示结点 Vi 的 rank 值，In(Vi)表示结点 Vi 的前驱结点集合，Out(Vj)表示结点 Vj 的后继结点集合，d 为 damping factor 用于做平滑。

而 Textrank 就是基于 PageRank 的思想，某一个词与其前面的 N 个词、以及

后面的 N 个词建立图相邻关系，并且考虑到不同词对可能有不同的共现（co-occurrence）TextRank 将共现作为无向图边的权值。所以，TextRank 的迭代计算公式如下：

$$WS(V_i) = (1-d) + d * \sum_{j \in In(V_i)} \frac{w_{ji}}{\sum_{V_k \in Out(V_j)} w_{jk}} WS(V_j)$$

可以看出，该公式仅仅比 PageRank 多了一个权重项 Wji，用来表示两个节点之间的边连接有不同的重要程度。

我们基于 TextRank 对附件四中留言详情和回复内容进行关键字的提取，作为接下来模型计算的数据准备。提取效果如下图所示：

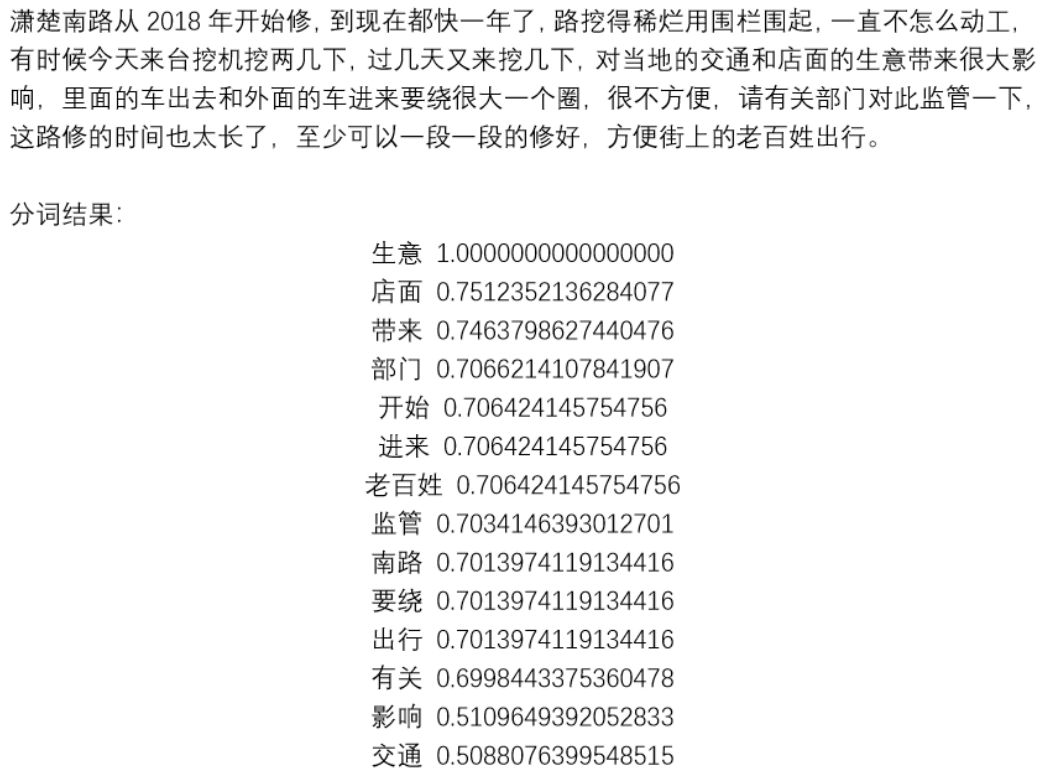


图 13　提取结果展示

2.8.2 模型的构建

对于留言回复内容评价指标的构建我们主要从时效性、完整性、相关性三个方面进行考虑：

①完整性和相关性

完整性和相关性都是针对回答内容进行衡量，完整性主要指回答的内容是否涵盖了留言所反映的问题，他强调回答内容的全面性。例如下文回复的内容：

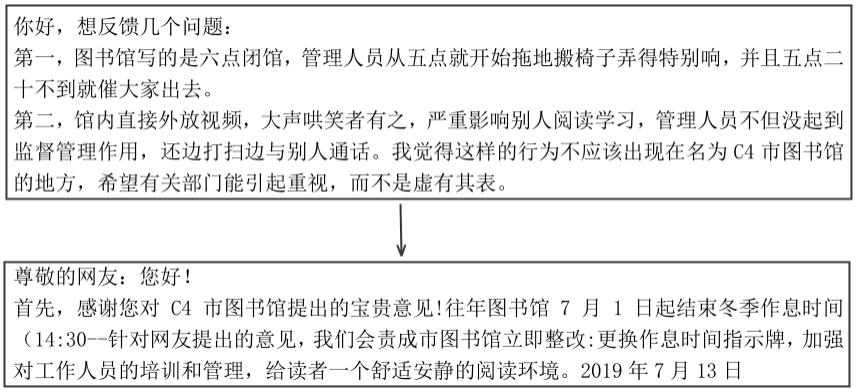


图 14　留言回复示例 1

相关性是指回答的内容是否和留言的内容有关，他强调内容要真实有效并且能够帮助到群众，例如下文的例子是很显然没有达到相关性要求的。

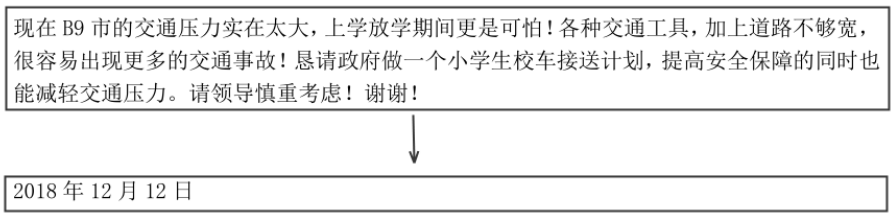


图 15　留言回复示例 2

因为留言详情和回复的侧重点不同，即使针对同一问题，在文本的相似度方面也会出现很大的差异，基于这种情况，我们采取基于 TextRank 的文本主题摘要的方法，对两段文本的相同或者相似关键字通过权重的方式进行累加，并通过权值系数作为文本相关性和完整性的得分。

②时效性

群众留言所反映的内容涉及多方面多领域，存在许多的问题是迫切需要解决的，例如上文中提到的魅力之城小区楼板开裂问题需要在短时间内对留言进行

回复并采取解决措施，所以时效性是衡量回复质量的一个重要的方面。我们基于这方面的考虑，通过时间戳计算出留言时间和回复时间的差值，并根据实际情况将天数进行分段，根据所处的区间进行差异的赋权作为相关性得分。

通过完整性相关性和时效性的关系，两部分在反映回复内容质量方面是相辅相成的，所以基于此我们将两者的得分以相乘的形式进行计算得到我们最终对于回复内容的量化得分，模型如下图所示：

$$Score(x) = \lambda f(T_r - T_s) \times \sum_{i \in w_r} \sum_{j \in w_s} g(i,j)$$

其中γ为权值系数，Tr、Ts 分别表示回复时间和留言时间，Wr 和 Ws 分别代表基于 TextRank 提取的回复内容和留言详情的关键词词组，F 函数如下：

$$f(x) \begin{cases} 2 & x \leq 5 \\ 1 & 5 < x \leq 15 \\ 0.5 & 15 < x \leq 30 \\ 0.2 & x > 30 \end{cases}$$

G 函数如下，其中 x,y 分别代表关键词，weight 为相应权重。

$$g(x,y) \begin{cases} weight(x) \times weight(y) & x = y \\ 0 & x \neq y \end{cases}$$

2.8.3 回复评价和建议

通过回复评价模型的构建，我们可以对回复进行多方面分值的量化。继而我们根据各方面的表现结合实际情况对每条回复进行了针对性的评价和建议，部分如下：

表 9　回复评价和建议

37459	A0	请问	2019/1/13 1:56:01	请问，带小孩去打疫苗要带什么	2019年1月14日	2019/1/14 16	0 评价内容效果较差，回复很及时，总体效果比较差。
37467	A0	反映	2019/1/4 16:10:28	近段时间夜晚和国假日醴官公路	网友：您好！收到您反映	2019/1/9 16:	310 评价内容效果较好，回复很及时，总体效果较好。
37474	A0	B9市	2018/12/25 14:41:15	我父母B9市农村户口，在家种地	尊敬的网友：网友反映的	2018/12/30 1	208 评价内容效果较好，回复很及时，总体效果较好。
37482	A0	建议	2018/12/7 18:48:01	现在B9市的交通压力实在太大，	2018年12月12日	2018/12/13 1	0 评价内容效果较差，回复比较及时，总体效果比较差。
37483	A0	关于	2018/12/6 8:34:31	B9市的医保今年是在网上办理，	根据《B9市2019年度城乡	2018/12/12 1	0 评价内容效果较差，回复比较及时，总体效果比较差。
37485	A0	B9市	2018/12/2 16:10:52	2018年12月2日晚10点30分，B9市	尊敬的网友：经查，欧洲	2019/1/7 15:	32.8 评价内容效果较好，回复时间过长，需要改正，总体效果较好。
39224	A0	反映	2018/1/21 0:18:04	领导您好！本人发现B6县多	网友：你好！你的信件已	2018/6/29 23	31.4 评价内容效果较好，回复时间过长，需要改正，总体效果较好。

通过从时效性、相关性、总体效果给出的文字建议，希望对相关部门和人员对于回复工作的改进起到一定的参考作用。

2.8.4 结果分析

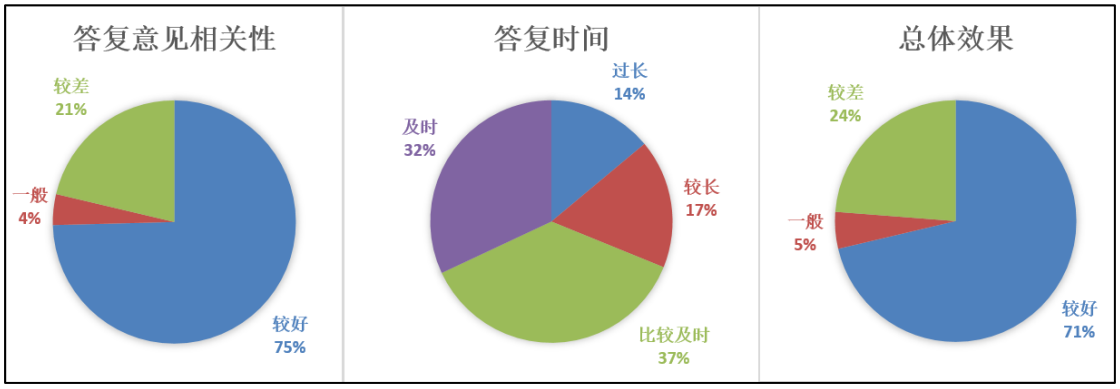


图 16　留言评价结果分析

从图中可以看出，无论是从答复意见的相关性还是答复时间来看，相关部门对大部分留言的答复情况较好，能根据留言详情在 15 天内做出相关回复，但是仍然有 30%左右的留言存在相关性较低，回复时间较长的现象。近年来，随着智慧政务地不断发展，微信、微博、市长信箱、阳光热线等网络问政平台的逐步应用，留言的在线化已是常态，但是如果相关部门回复不及时甚至根本没有回复，将会导致这些栏目成为摆设，另外，若百姓的留言不能得到及时且有效地回复，他们将反复留言，这不仅会产生反作用，还会影响政府的公信力，也就失去了智慧政务本身存在的意义和价值，所以为了避免留言板成为形式主义，相关部门应该加强重视，安排专人负责，及时整理、回复、反馈、解决问题，正如习近平总书记曾经告诫领导干部那样：“领导干部要经常上网，回应网民关切，对网民要多些包容和耐心，及时吸纳建设性意见，认真研究和吸取善意批评。” 希望相关部门也要多上网看看，积极弘扬社会正能量，重视网络上的民意反映，处理好关乎群众切身利益的各类事项，共建和谐稳定的智慧政务平台，促进社会更好地发展！

四、 结论

本文通过对处理过的网上政务留言信息通过建立多种数据挖掘模型，得到了具有一定价值的结果，实现了对留言内容的一级标签分类，并通过文本聚类、命名实体识别，热度评价等模型实现了热点问题的挖掘，最终基于 TextRank 文本摘要算法结合回复评价模型对回复内容进行了评价和针对性建议。总体来说，我们解决了大部分的问题，并通过结果反馈显示，效果较为满意。

但是我们基于文本挖掘和分析过程中还是出现了一些不尽如人意的地方，例如基于 k-means 算法的文本聚类结果存在一些簇群内部相关性很低的问题，HMM 的中文命名实体识别准确率较低，回复评价问题中一些回复较好的留言被误判。这些问题一部分是因为中文语言处理自身存在的缺陷，也有一部分是因为我们构建模型的方法存在一些不足。因此这些问题也是我们以后将要继续研究和探讨的地方。

五、 参考文献

[1]石凤贵.基于 TF-IDF 中文文本分类实现[J].现代计算机,2020(06):51-54+75.

[2]张明. 表示学习词向量提取及其在情感分析中的应用研究[D].江南大学,2019.

[3]俞鸿魁. 基于角色标注的中文机构名识别[C]. Northeastern University、Tsinghua University、Chinese Information Processing Society of China、Chinese Languages Computer Society, USA.Advances in Computation of Oriental Languages--Proceedings of the 20th International Conference on Computer Processing of Oriental Languages.Northeastern University、Tsinghua University、Chinese Information Processing Society of China、Chinese Languages Computer Society, USA:中国中文信息学会,2003:91-99.

[4]张伦干. 多项式朴素贝叶斯文本分类算法改进研究[D].中国地质大学,2018.

[5]李海磊,杨文忠,李东昊,温杰彬,钱芸芸.基于特征融合的 K-means 微博话题发现模型[J].电子技术应用,2020,46(04):24-28+33.

[6]韩琮师,李旭健.改进的 K-means 算法研究[J].软件,2020,41(03):21-23.

[7]祝继锋. 基于 SVM 和 HMM 算法的中文机构名称识别[D].吉林大学,2017.

[8]邵岚,姚艳松,李宣.基于互联网大数据的自然语义分析研究[J].网络安全技术与应用,2019(01):31-32.

[9]吴广财. HMM 增量学习算法在中文命名实体识别中的应用研究[D].华南理工大学,2011.

[10]李玮瑶,赵凯.基于特征提取的网络热点事件挖掘算法[J].计算机与现代化,2015(05):17-20.

[11]刘惠,赵海清.基于 TF-IDF 和 LDA 主题模型的电影短评文本情感分析——以《少年的你》为例[J].现代电影技术,2020(03):42-46.

[12]陈嘉钰,李艳.基于 LDA 主题模型的社交媒体倦怠研究——以微信为例[J].情报科学,2019,37(12):78-86.

[13]李航.统计学习方法[M].北京：清华大学出版社，2012.

[14]周志华.机器学习[M].北京：清华大学出版社，2016.

[15]Gilbert R C , Trafalis T B , Adrianto I . Support Vector Machines for Classification[M]// Encyclopedia of Operations Research and Management Science. John Wiley & Sons, Inc. 2011.