基于机器学习算法的“智慧政务”文本挖掘

摘要

本文对“智慧政务”中文本挖掘的实现方式进行了讨论。近年来，由于政府了解民意的渠道逐步通畅，与民意相关的文本数据量也不断攀升，这给依靠人工来进行留言分类及处理的相关部门带来了极大挑战。而同时，使用人工进行这些工作具有效率低，错误率较高，成本大等特点，因此构建一个智慧文本处理系统就具有相当重要的意义。本文使用了机器学习中的多种算法，构建了一套行之有效的文本处理系统，具有一定的应用价值。

针对问题一：本文使用了三种机器学习分类算法：朴素贝叶斯分类算法、k 近邻算法、支持向量机算法，分别对群众留言进行了分类，均取得了一定的效果。而同时，考虑到原本文本的分类已经有了一套分类的标签，因此也构造了以该标签进行分类的分类方式。由于这种方法具有一定的局限性，所以选择将其与以上三种机器学习算法进行线性表决，表决之后的结果作为最终结果进行输出。

针对问题二：问题二实际上是一个关于文本聚类和文本识别的问题。因此首先对文本数据进行了处理，得出了留言的相关性矩阵，再使用广度优先遍历的方式将得到的相关性矩阵进行聚类，从而得出热度最高的群众反映的问题。针对时间、地点，主要事实的辨别，本文采用了文本比对的方式，清洗掉无关紧要的词汇之后进行比对找出可能的时间、地点、主要事实，将结果以较简洁的方式表达出来，最后将结果写入到 excel 文档中。

针对问题三，本文从留言回复的相关性、完整性、可解释性方面对问题进行了评价。为了将留言的特点可视化，本文定义了三种因子，即相关性因子、完整性因子和可解释性因子，对文本进行处理后分别计算，将结果综合考虑后输出，在实际应用中，可以通过调整三种因子的权重来获得较为满意的评价指标。

关键词：文本挖掘、朴素贝叶斯分类算法、k 近邻算法、支持向量机算法、广度优先聚类

ABSTRACT

This paper discusses the way to realize the Chinese text mining of "intelligent government affairs". In recent years, as the channels for the government to understand public opinions are gradually unobsused, the amount of text data related to public opinions is also increasing, which brings great challenges to the relevant departments that rely on human beings to classify and process comments. At the same time, it has the characteristics of low efficiency, high error rate and high cost, so it is of great significance to build a smart text processing system. This paper USES a variety of algorithms in machine learning to build an effective text processing system.

Aiming at problem 1: this paper USES three kinds of machine learning classification algorithms: naive bayes classification algorithm, k-nearest neighbor algorithm, and support vector machine algorithm. Meanwhile, considering that the classification of the original text already has a set of classification labels, the classification method of classification by this label is also constructed. Since this method has certain limitations, it is selected to conduct linear voting with the above three machine learning algorithms, and the result after the voting is output as the result, and a certain effect is obtained.

Aiming at problem2 : problem two is actually a text clustering and text recognition problem. Therefore, firstly, the text data is processed, and the correlation matrix of messages is obtained. Then, the correlation matrix is clustered by breadth-first traversal, so as to get the most popular problems. For the discrimination of time, place and main facts, this paper adopts the method of text comparison. After cleaning the irrelevant words, the paper finds out the possible time, place and main facts by comparison, and then expresses the results in a more concise way. Finally, the results are written into the excel document.

Aiming at problem 3, this paper evaluates the relevance, completeness and interpretability of message reply. In order to visualize the characteristics of comments, this paper defines three factors, namely the correlation factor, the integrity factor and the interpretability factor. After processing the text, they are calculated respectively, and the results are output after comprehensive consideration. In practical application, the weight of the three factors can be adjusted to obtain a satisfactory evaluation index.

Keywords: text mining, naive bayesian classification algorithm, k-nearest neighbor algorithm, support vector machine algorithm, breadth first clustering

目录

第一章 问题重述

1.1 问题背景

近年来，随着微信、微博、市长信箱、阳光热线等网络问政平台逐步成为政府了解民意、汇聚民智、凝聚民气的重要渠道，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。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。

1.2 需要解决的问题

1.2.1 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。本问题即旨在建立关于留言内容的一级标签分类模型。

模型分类效果评估如下（使用 F-Score 方式对分类方法进行评价）：

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

其中 P_i 为第 i 类的查准率，R_i 为第 i 类的查全率。

1.2.2 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。本问题旨在建立一个留言分类模型，定义合理的热度评价指标，并给出评价结果，按相应的的格式给出排名前 5 的热点问题，并保存为文件“热点问题表.xls”。按相应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”。

1.2.3 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。

第二章 机器学习分类算法介绍及评估

2.1 朴素贝叶斯算法[1]

2.1.1 朴素贝叶斯算法概述

朴素贝叶斯分类（NBC）是以贝叶斯定理为基础并且假设特征条件之间相互独立的方法，先通过已给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入 X 求出使得后验概率最大的输出 Y。

2.1.2 朴素贝叶斯算法性能分析

朴素贝叶斯算法逻辑性较为简单，健壮性较好，但是当数据集的属性存在相互关联的时候，分类效果可能会降低。

采用鸢尾花(iris)数据集（机器学习某常用数据集）中的少量数据，使用朴素贝叶斯算法进行预测，调整参数，得到的结果如下：

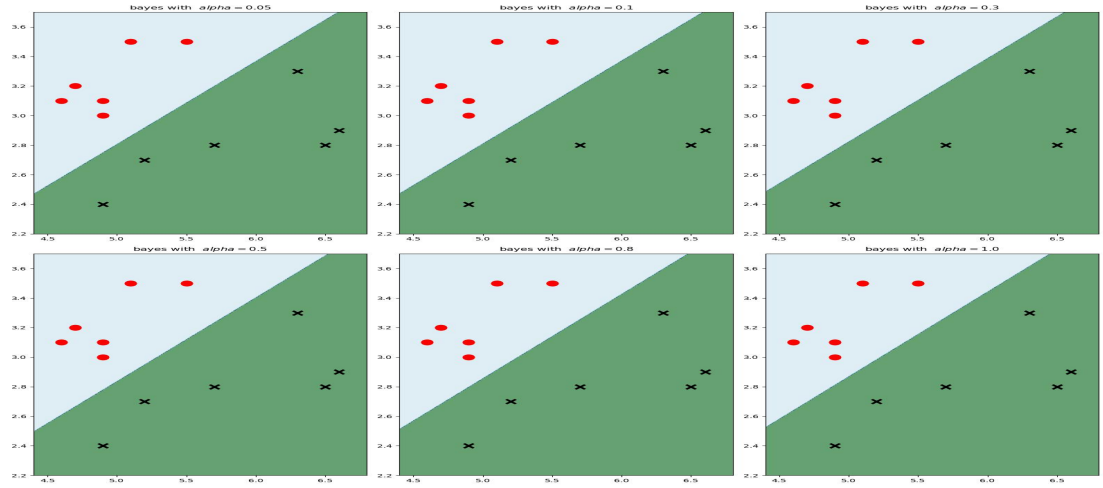

图 1 朴素贝叶斯算法调整参数 alpha 运行得到的结果

由图 1 可以看出，朴素贝叶斯算法的性能较为优良，正确地完成了分类任务。

2.2 k 近邻算法[2]

2.2.1 k 近邻算法概述

所谓 K 近邻算法，即是给定一个训练数据集，对新的输入实例，在训练数据集中找到与该实例最邻近的 K 个实例（也就是上面所说的 K 个邻居）， 这 K 个实例的多数属于某个类，就把该输入实例分类到这个类中。

2.2.2 k 近邻算法性能分析

KNN 算法本身简单有效，它是一种 lazy-learning 算法，分类器不需要使用训练集进行训练，训练时间复杂度为 0。KNN 分类的计算复杂度和训练集中的文档数目成正比，也就是说，如果训练集中文档总数为 n，那么 KNN 的分类时间

复杂度为 O(n)。

采用鸢尾花(iris)数据集中的少量数据，使用 k 近邻算法进行预测，调整参数，得到的结果如图 2 所示：

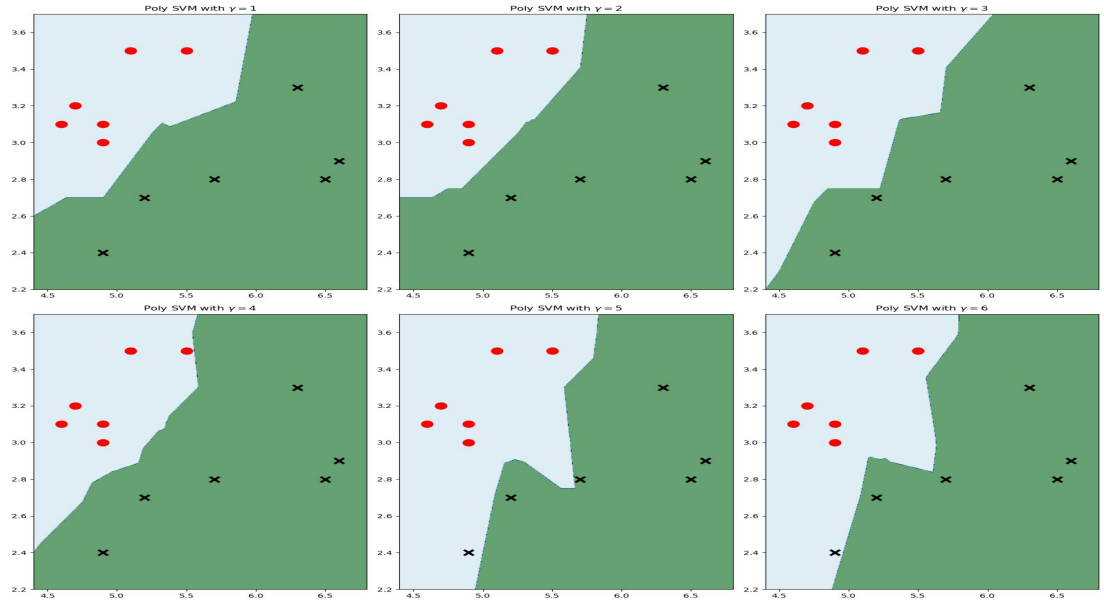

图 2 使用 k 近邻算法调整参数 n_neighbors 得到的结果

从该预测结果可以看出，k 近邻算法的分界面并不平整，这是由 k 临近算法的特性决定的。

2.3 支持向量机算法[3]

2.3.1 支持向量机算法概述

支持向量机（Support Vector Machine, SVM）是一类按监督学习（supervised learning）方式对数据进行二元分类的广义线性分类器（generalized linear classifier），其决策边界是对学习样本求解的最大边距超平面（maximum-margin hyperplane）

2.3.2 支持向量机算法性能分析

对于非线性问题，支持向量机算法通过非线性变换转换到高维的特征空间，在高维特征空间中构造线性判别函数来实现训练样本分类，巧妙地避免了“维数灾难”问题，其算法复杂度与特征空间的维数无关。通过变换不同的核函数，同样可以得到不同的分类效果。

采用鸢尾花(iris)数据集中的少量数据，使用支持向量机算法进行预测，调整参数，得到的结果如图 3、4、5 所示：

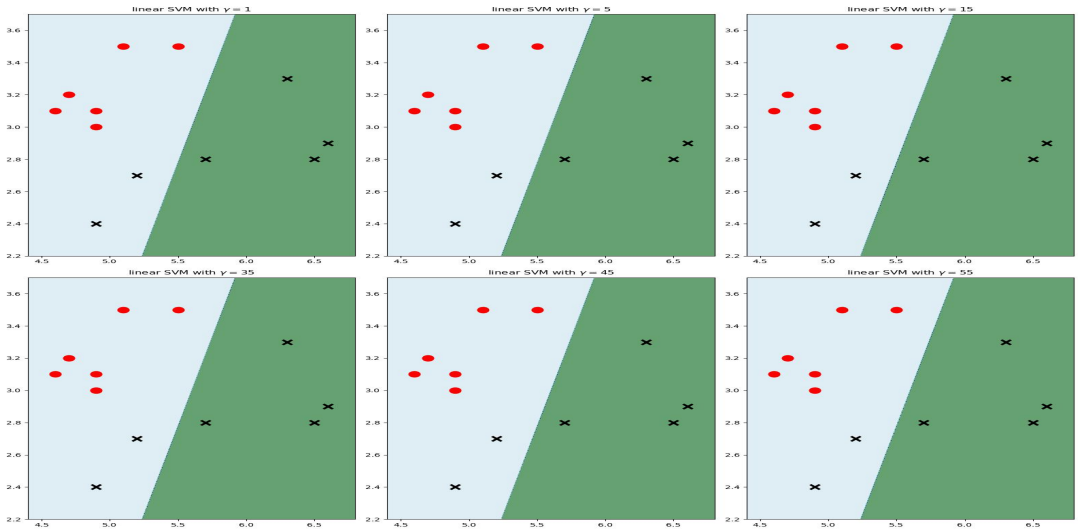

图 3 使用线性核函数，调整参数 grama 进行预测的结果

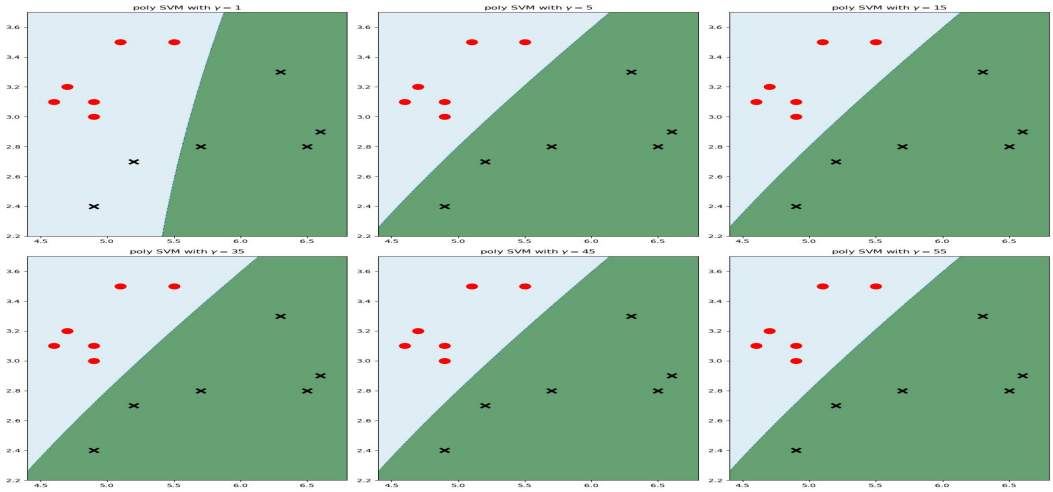

图 4 使用多项式核函数，调整参数 grama 进行预测

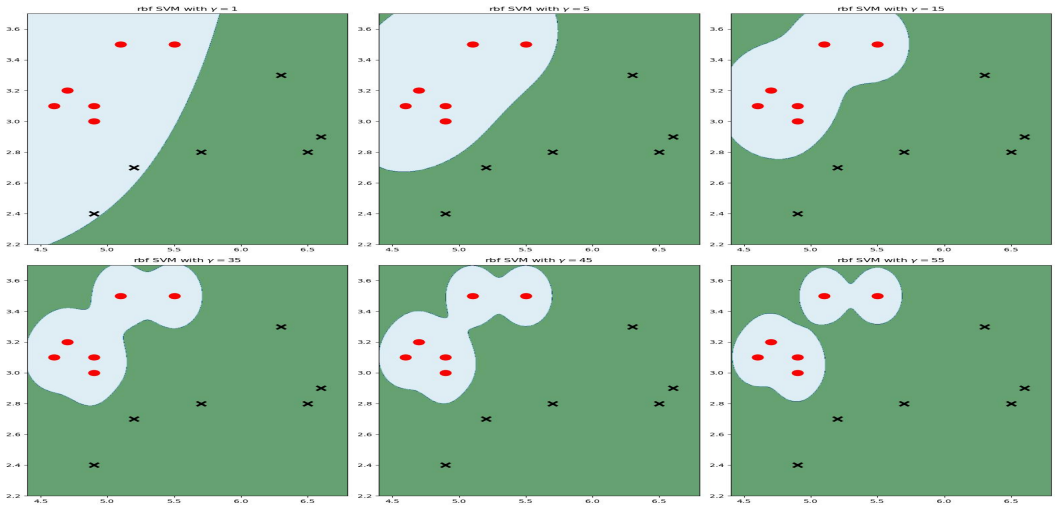

图 5 使用高斯核函数，调整参数 grama 进行预测的结果

由上述训练结果的可视化以及对函数本身的特征分析可以看出，线性核函数的分界面较为平整，当样本线性可分的程度比较大的时候优先考虑线性核函数；多项式核函数可以将低维的输入空间映射到高维，但是可能会面临计算量较大的问题；高斯核函数局部性较强，一般使用性能也较好，但是若是参数选取不当，可能会造成过拟合的问题。

第三章 问题一的分析与求解

3.1 问题一求解总体流程

3.1.1 问题一求解总体流程概述

分析问题一不难看出，这是一个文本分类问题，数据集为一个 excel 文档，其中包含了群众的留言记录。所以考虑通过先对数据进行预处理，使其符合机器学习分类算法所要求的格式（本问题采用 python 语言 sklearn 库中的各机器学习算法），然后进行机器学习分类，对分类结果进行表决，从而得到预测值。

3.1.2 问题一算法流程图

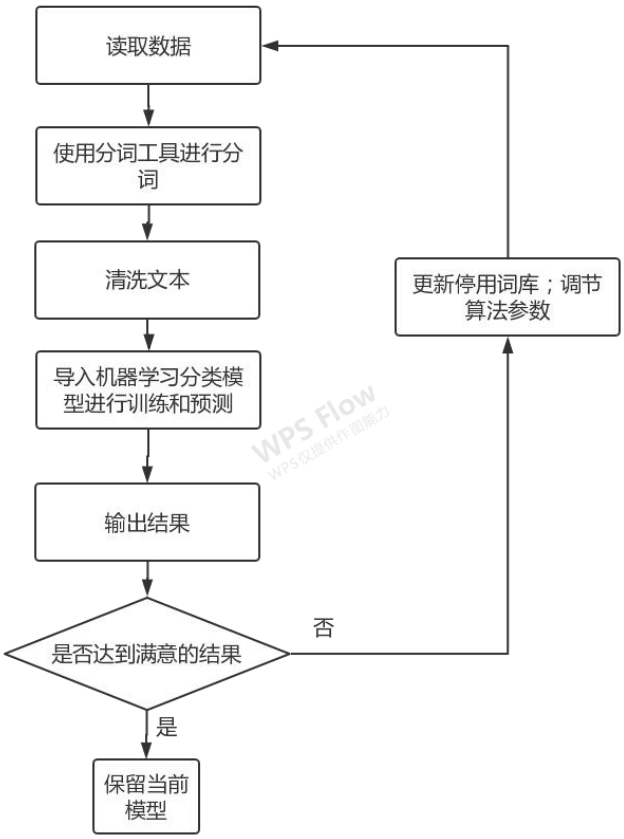

图 6 问题一算法流程图

3.2 数据预处理

通过对数据（格式为 excel）的读取，得到了一组文本及所对应的标记，然而这样的文本并不能够直接用于分类，因为机器学习分类算法不能够直接对一整段文本进行处理，要先经过分词、洗词等步骤，流程如下：

3.2.1 分词

分词，即将一整段文本分割成一个个的词或者词组，这样的文本能够被机器学习算法所识别从而进行预测。中文分词采用结巴分词，其分词效果在本案例中表现较为优良，效果如下：

分词前：A 市西湖建筑集团占道施工有安全隐患

分词后：A 市 西湖 建筑 集团 占 道 施工 有 安全隐患

3.2.2 清洗无用的词语和符号

经过对分词之后数据的观察，发现有很多数据并不能够带来有效信息，反而会干扰学习，所以进行分词后的短语进行进一步的清洗，去掉无用的词语。无用词语有如下几个方面：

（1）一般的停用词库中所包含的词语。如“不仅”、“不但”以及各类标点符号等，一般的清洗文本用停用词库即可，但是要注意，有些特殊符号在本问题中可能有其作用，如书名号等，所以考虑对一般的停用词库进行一定的删减。

（2）特殊的符号如换行符，制表符等。这些符号的存在可能会干扰分词及机器学习，所以在洗词的过程中要将其清洗掉。

（3）在本问题中经常出现但是对于文本分类没有太大帮助的词语，如“用户”、“你好”等词语，这些词语并不包含在一般停用词库中，要进行手动添加。

清洗文本示例如下：

清洗文本前：A 市 西湖 建筑 集团 占 道 施工 有 安全隐患

清洗文本后：西湖 建筑 集团 占 道 施工 安全隐患

可以看出“A”、“市”、“有”已经被剔除，其中“A”、“市”是手动添加在停用词表中的，因为这对文本的分类并没有太大关系，而“有”在一般停用词库中。

3.3 基于机器学习算法进行文本识别

本题目使用了三种机器学习分类的算法，这三种算法均来自 sklearn 库，对这三种算法的应用方式讨论见下:

3.3.1 贝叶斯分类算法

使用贝叶斯分类算法进行文本分类，使用不同的模型并调整参数以及 alpha（alpha:先验平滑因子，默认等于 1，当等于 1 时表示拉普拉斯平滑），得到的

结果如表 1、2 所示：

表 1 使用多项式模型（MultinominalNB）调参测试结果

标号	实验次数	Alpha	平均总精确率	平均 F1
1	3	0.1	0.856	0.832
2	3	0.5	0.735	0.688
3	3	1.0	0.638	0.589

表 2 使用伯努利模型（BernoulliNB）调参测试结果

标号	实验次数	Alpha	平均总精确率	平均 F1
1	3	0.1	0.853	0.812
2	3	0.5	0.611	0.632
3	3	1.0	0.453	0.395

从表 1、2 可以看出，使用多项式模型在本案例中能够取得更好的效果，且在参数 alpha 取 0.1 的时候能够取到较好的效果。

3.3.2 k 近邻算法

使用 k 近邻算法进行文本分类，调整参数 n_neighbors，结果如表 3 所示：

表 3 使用 k 临近算法进行调参得到的结果

标号	实验次数	n_neighbors	平均总精确率	平均 F1
1	3	2	0.786	0.793
2	3	8	0.734	0.759
3	3	20	0.763	0.775

从表 3 中可以看出，k 临近算法进行调参对该问题的结果影响并不大。

3.3.3 支持向量机分类算法

支持向量机分类算法具有不同的核函数，这些核函数的特性如上文讨论所示，分别采用不同的核函数，其余参数取默认值，进行计算，结果如表 4 所示：

表 4 使用不同的核函数的运行结果

标号	实验次数	核函数	平均总精确率	平均 F1
1	3	线性核函数	0.883	0.875
2	3	高斯核函数	0.768	0.759
3	3	多项式核函数	0.635	0.621

可以看出，线性核函数在本问题中的表现较好，所以在本问题中选择线性核函数进行预测。

3.3.4 三种算法的联合比较

经过多次调整参数，三种算法各自的最优表现如表 5 所示：

表 5 三种学习算法的最优表现对比

标号	算法	平均总精确率	平均 F1
1	贝叶斯分类算法	0.865	0.838
2	k 近邻算法	0.856	0.876
3	支持向量机算法	0.894	0.893

可以看出，支持向量机算法在本问题中的表现最好。

3.4 算法的进一步改进

3.4.1 多个分类器进行线性表决

考虑到本问题中使用了多个算法对结果进行预测，如果仅仅取一种分类方式作为结果则会浪费其余的方式，所以将三种算法分类器得到的结果进行线性表决，即若两个或两个以上的分类器预测结果相同，则取该预测结果，否则，取支持向量机算法的预测结果（其表现最好）以这一结果作为最后的结果。

3.4.2 考虑利用原本存在的分类标签

在给出的数据中同样给出了一、二、三级分类标签，利用这一标签对文本进行比对同样可以得出一个评判结果，但是若是留言中不含有分类标签中的词语，就会造成无法判别的情况。基于此，可以将这种分类方式和以上三种学习器组合使用，从而提高分类的精度。

3.5 结果评估

对本问题的解决，本文使用了三种算法，使用 F-score 评判方式，运行结果的 F1 稳定在百分之八十五以上，收到了较好的效果。

第四章 问题二的分析与求解

4.1 问题二求解总体流程

4.1.1 问题二求解总体流程概述

问题二其本质上是一个文本聚类和文本识别的问题，但是文本聚类的标签并不确定。可以通过计算文本之间的匹配度得出文本之间的匹配度矩阵，进而采用广度优先遍历可能为一类的留言，从而完成聚类。时间、地点以及问题描述可以

通过对高频词的获取近似取得。

4.1.2 问题二算法流程图

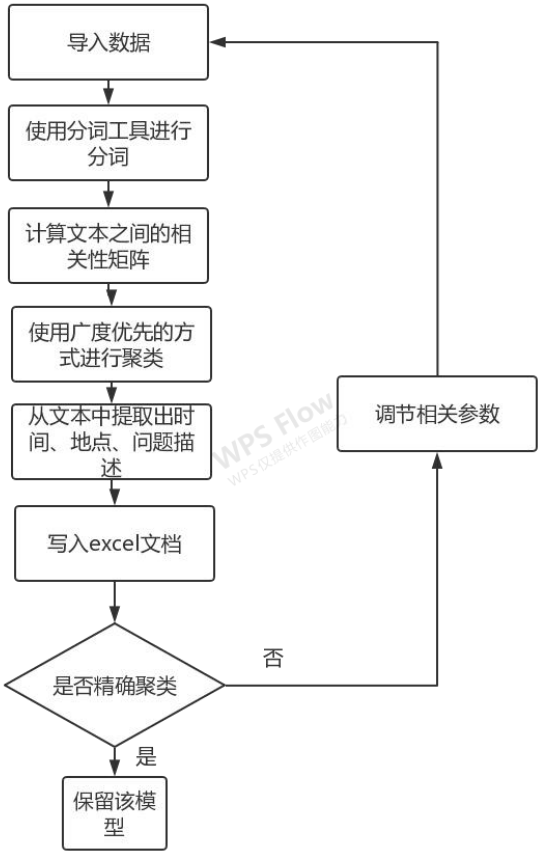

图 7 问题二算法流程图

4.2 数据预处理

摘取相似的留言标题如表 6 所示：

表 6 一类相似的留言样例

标号	留言标题
1	A 市经济学院体育学院变相强制实习
2	A 市经济学院强制学生实习
3	A 市经济学院强制学生外出实习

从中可以看出，有在这些标题的文本中有几个词语是完全相同的，如“A 市”、“经济学院”、“强制”、“实习”，说明同一热点问题的文本匹配度很高，所以可以采用进行先分词而后进行文本相似度的计算：

4.2.1 分词

与问题一相似，同样使用结巴分词。

4.2.2 清洗无关紧要的文本

这一步骤与问题一也基本相同。但是有一点要注意，由于问题一与问题二的目标并不相同，问题一和问题二清洗掉的文本不应相同，如在问题一中，例如地点、时间等性质的词语与文本的类别之间的关系并不强，可以清洗掉，而在问题二中，时间、地点等词语是聚类的关键，不能够被白白清洗掉。

4.3 进行文本相关度的计算

从表 6 可以看出，相似的标题中有相当数量的词重复，所以只需要使用标题就能够很好地完成聚类。

文本的相关度有如下几种计算方式：

（1）对分词后得到的的列表 A、B，A 相对于 B 的相关度定义为：

$$relevance = \frac{\sum_{1}^{len(A)} I_i}{len(A)}$$

其中 $len(A)$ 是列表 A 的长度，而 I_i 定义为：

$$I_i = \begin{cases} 1 & \text{A[i]在B中} \\ 0 & \text{A[i]不在B中} \end{cases}$$

由此得到的相关性矩阵为：

$$R_{n\times n} = \begin{matrix} 1 & re_{1,2} & \dots & re_{1,n} \\ re_{2,1} & 1 & \dots & re_{2,n} \\ \dots & \dots & \dots & \dots \\ re_{n,1} & re_{n,2} & \dots & 1 \end{matrix}$$

（2）考虑这样一种情况：

表 7 另一种留言文本情况

标号	留言标题
1	请加快创建 A 市食品安全城市建设步伐
2	请加快 A 市大王山世界旅游度假地建设力度
3	请 A 市加快地铁的建设

可以看出，这些明显并不属于同一个热点问题。有些词语也相同，如“请”、“A 市”、“加快”、“建设”，在聚类中较为容易把这些聚到一类，经过考察发现，其中普遍原因是时间地点的相同，而发生的事件不同，而时间地点等通常都处在文本的前半部分，所以对分词后得到的的列表 A、B，A 相对于 B 的相关度定义为：

$$relevance = \frac{\sum_{1}^{len(A)/2} I_i + 2\sum_{\text{len(A)/2+1}}^{len(A)} I_i}{len(A)}$$

$$I_i = \begin{cases} 1 & \text{A[i]在B中} \\ 0 & \text{A[i]不在B中} \end{cases}$$

得到的相关性矩阵为：

$$R_{n\times n} = \begin{matrix} re_{1,1} & re_{1,2} & \dots & re_{1,n} \\ re_{2,1} & re_{2,2} & \dots & re_{2,n} \\ \dots & \dots & \dots & \dots \\ re_{n,1} & re_{n,2} & \dots & re_{n,n} \end{matrix}$$

注意这里的对角线数值不再为 1，而是在 1.5 左右。

（3）先统计所有文本的词集合及其频率，而后对每个文本，都构建一个向量，向量的值是词语在本文档中出现的次数。

以表 6 中的数据为示例，对表 6 中的数据进行分词，假设结果为表 8：

表 8 一种群众留言分词示例

标号	留言标题
1	A 市 经济学院 体育学院 变相 强制 实习
2	A 市 经济学院 强制 学生 实习
3	A 市 经济学院 强制 学生 外出 实习

对表中的分词结果组成集合，结果如下：

A 市 经济学院 强制 学生 外出 实习 体育学院 变相

则构建的向量为表 9：

表 9 使用分词结果构造的向量

标号	A 市	经济学院	强制	学生	外出	实习	体育学院	变相
1	1	1	1	0	0	1	1	0
2	1	1	1	1	0	1	0	0
3	1	1	1	1	1	1	0	0

计算这些变量之间的距离，构成相关性矩阵（距离越小，相关度越高）：

得到的相关矩阵为：

$$R_{n\times n} = \begin{matrix} re_{1,1} & re_{1,2} & \dots & re_{1,n} \\ re_{2,1} & re_{2,2} & \dots & re_{2,n} \\ \dots & \dots & \dots & \dots \\ re_{n,1} & re_{n,2} & \dots & re_{n,n} \end{matrix}$$

其中，第三种方法进而可以使用 k-means 算法进行聚类，为了代码的简洁性

和运算的速度，本文采用第二种方式进行相关矩阵的计算。

4.4 使用广度优先的方式进行聚类

4.4.1 对相关性矩阵进行处理

相关性矩阵中每一个元素 re_{ij} 表示了两个文本之间的相关系数。相关系数较小的两个文本之间不会被聚类，所以选取某个阈值，大于该阈值的两个文本之间被认为相关性较强，联系保留，否则，联系舍弃。经过这样一个处理，实质上得到类似这样的一幅图：

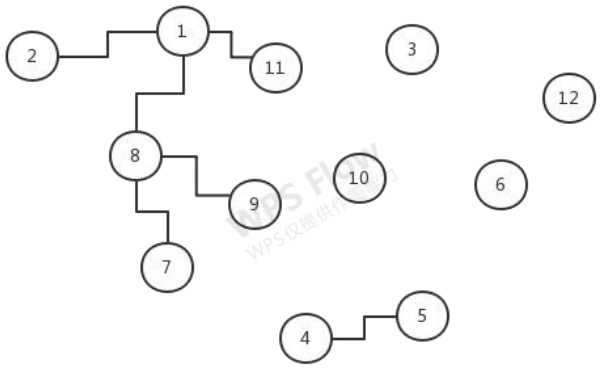

图 8 相关性矩阵处理后得到的结果

这张图是部分连通的。

4.4.2 利用广度优先进行聚类

采用广度优先遍历的方式遍历这张图，将有联系的文本聚为一类后进行输出。

4.5 聚类结果进一步处理及显示

4.5.1 按热度对聚类结果排序

首先定义热度因子：

$$heat = records \times (praise + against)$$

其中 records 为该类记录条数，praise 为点赞数，against 为反对数，之所以使用‘+’，是因为无论是点赞还是反对都是民意的反映

对已经聚好类的各个热点问题，计算热度因子后进行排序，选出其中的前五个，即为热点问题。

4.5.2 时间范围的显示

留言时间在数据集中绝大部分为纯文本形式表示，因此首先使用 python 中的 time 模块将时间转化为时间戳格式，而后进行比较输出。

4.5.3 地点、人物及问题描述的显示

1. 地点及人物的提取

地点、人物的提取要求对文本进行有效地识别，剔除非相关性词语如“请”“你好”等，并保留相关词语。其显示方式有如下两种：

（1）考虑到时间、地点通常位于留言标题的前半部分，因此可以对文本进行划分，选取前半部分的词语，统计词频，而后输出，则大概率能够捕捉到时间、地点。

（2）对划分后的词进行词性标注后选出其中具有名词性的词语，选出前几个输出、大概率能够得到时间地点。

2.问题描述的提取

由于群众的留言多种多样，问题的的提取较为复杂，为了简单化，选取留言类中第一个留言的标题作为问题描述，事实证明其效果尚可。

4.6 结果评价

使用全部数据进行测试结果如表 10、11 所示：

表 10 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
1	1	9292	2019/1/3 10:34:29 至 2020/1/7 22:50:33	A 市地铁 1 号线	A 市地铁 1 号线北延二期有规划建至丁字湾街道，能否先行调研二期工程？
2	2	3108	2019/2/21 18:45:14 至 2019/3/1 22:12:30	A 市 58 车贷案	请书记关注 A 市 A4 区 58 车贷案
3	3	2688	2019/8/23 14:21:38 至 2019/9/6 18:36:16	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？
4	4	2666	2019/1/2 17:25:52 至 2019/11/1 15:50:17	请 A 市加快高铁西站	请 A 市加快高铁西站的建设力度
5	5	2624	2019/1/2 17:25:52 至 2019/11/1 15:50:17	请加快 A 市	请加快 A 市一圈二场三道步伐力度

表 11 热点问题留言明细表

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
1	20	A000	A 市地铁 1 号线	2019/9/23		0	4

	06064	106846	北延二期有规划建至丁字湾街道，能否先行调研二期工程？	19:28:24	地铁 1 号线北延二期有规划建至丁字湾街道，暂 1 期还没开建，能否先行调研二期工程.....		
1	2372121	A0001067197	A 市地铁 1 号线北延线何时开工？	2019/9/8 11:35:46	地铁 1 号线北延从 2017 年规划，到说 2018 开工建设，又遇一个楼盘的投诉调整为预开工，2018 年底发布预备 2019 年 9 月开工。到现为止一点动静都没....	0	3
...	...	...	...	...	...	...	..
2	1943443	A0001061161	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。	2	1943
...	...	...	...	...	...	...	...

可以看出，结果还是较为令人满意的。

第五章 问题三的分析与求解

5.1 答复意见的评价指标定义

答复意见主要从以下几个方面进行评价：

相关性：度量答复和所问的问题的相关程度的指标。文本之间的相关系数与相关性的定义较为相似，设留言分词后为列表 A，对回复进行分词后为列表 B，在本文中定义回复的相关性系数（relevance）为：

$$relevance = \frac{\sum_{1}^{len(A)} I_i}{len(A)}$$

$$I_i = \begin{cases} 1 & \text{A[i]在B中} \\ 0 & \text{A[i]不在B中} \end{cases}$$

由上面的公式可知，相关性系数的取值范围为 0-1。

完整性：即不能够忽略重要的相关因素。由于该性质度量比较困难，所以使用回复与留言的长度之比来定义完整性，可以预见，留言的回复的长度与其完整性呈正相关。设留言分词后为列表 A，对回复进行分词后为列表 B，进行如下计算：

$$rate = \frac{len(B)}{len(A)}$$

考虑到 rate 大部分集中在 1 左右，对其进一步处理，得到完整性系数（complete）为

$$complete = \frac{2^{rate-1}}{10}$$

得到的结果大部分处于 0-1 之间。

可解释性：答复具有一定的原则且能够被群众所理解。针对这一性质，对于可能具有高解释性的回复提出如下原则：

（1）对于回复群众咨询的问题的回复，最好能够给出回复依靠的根据，如《A 市轨道交通线网规划修编规划》，《A6 区“一圈两场三道”专项规划》等，或者表达该类意思，如“根据”、“因为”等，或者应当具有指向性的词语，如“详情”、“咨询”等

（2）对于解决群众反映的问题的回复，应当汇报工作的情况，回复中应当具有以下词语“已经”、“正在”、“我局”、“高度重视”等

定义可解释性系数（clear）为：

$$clear = \frac{S}{5}$$

其中 s 为含有类似如上所述的词的个数。

综合评价：考虑到上述三个因子大部分都处于 0-1 之间，直接相乘的话范围波动过大且不合理，

所以进行如下计算：

$$overall = \frac{10}{7}\left[(1+relevance)(1+complete)(1+clear)-1\right]$$

定义综合评价因子（overall_rate）为：

$$overall\_rate = \begin{cases} overall & \text{overall} < 10 \\ 10 & \text{overall} \geq 10 \end{cases}$$

这样可以保证评价因子在 0-10 之间。

5.2 数据的处理及指标计算

5.2.1 数据的处理

首先对留言（或其标题）和回复进行分词，只是不需要进行洗词。为了计算可解释性系数，还需要对文本进行处理，选出其中的高频回复词汇，从中挑出比较具有代表性的词汇组成列表，用于可解释性的检测，利用所给的文本找出的具有代表性的词汇如表 12 所示：

表 12 具有代表性的可解释性词汇

回复类型	具有代表性的词汇
回复群众咨询的问题的回复	情况 回复 收悉 相关 《 》 支持 留言 监督 办理 理解 部门 规划 街道 业主 政策 关心 咨询 社区 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 登记 条件 符合 进一步 通知
解决群众反映的问题的回复	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 组织 关心 人员 政府 施工 发展 关注 改造 高度重视 投诉 意见 城市 经营 实施 安排 解决 核实 收费 工程 整改 人民政府 符合

5.2.2 指标的计算

按照上文定义的方式进行计算即可。其中综合评价因子中的 s 即为文本中含有如表 12 所示的词的个数。

5.3 结果的评估

选取全部数据（附件四）进行评估，一部分结果如表 13 所示：

表 13 对回复的评估结果

序号	评价值	序号	评价值	序号	评价值
1	0.964	8	2.244	15	2.205
2	1.876	9	1.121	16	2.721
3	2.929	10	1.432	17	0.716
4	2.143	11	1.092	18	0.944
5	2.345	12	10	19	1.783
6	1.792	13	1.646	20	0.736
7	2.891	14	0.079		

对全样本进行计算的均值为 1.23。

第六章 总结与展望

本文聚焦于“智慧政务”中的文本挖掘，目的是为了建立一套行之有效的文本识别、分类系统，最终得到如下成果与结论：

关于问题一，本文建立了多种机器学习算法进行综合的预测模型，分类正确率、F-score 多次运行平均能够维持在 85%以上，基本上可以用于留言分类的应用。在实际应用中，可以尝试对停用词表进行进一步的完善，以增强分类的精确性。此外，该解决方案同样具有较好的延展性，不仅可以用于政务系统文本分类，还可以用于其他二分类或者多分类问题，如假新闻的判别等。

关于问题二，本文构建了对群众反映的热点问题进行聚类与判别的方式，较为圆满地完成了聚类及显示的任务。在实际应用中，可以通过对文本的相关系数矩阵进行处理控制聚类的程度。当然，本文对于事件概况等文本提取还有一定的欠缺，在实际应用中，可以通过机器学习预测算法通过大样本学习，从文本中提取出需要的内容。

关于问题三，本文构建了一整套的评价体系，用于评价留言回复的质量，并成功地对题目所给数据的回复进行了评价。在实际应用中，可以通过增加评价指标和改变各个指标的权重来达到想要的评价效果。

参考文献

[1]百度百科.“朴素贝叶斯”词条[EB/OL].https://baike.baidu.com/item/朴素贝叶斯/4925905#reference-[4]-992724-wrap
[2]百度百科.“k 近邻算法”词条[EB/OL].https://baike.baidu.com/item/k 近邻算法/9512781?fr=aladdin
[3]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