基于文本挖掘机器学习技术的“智慧政务”应用

摘要

本文旨在利用自然语言处理技术对“智慧政务”问题进行文本挖掘。基于机器学习中的多项式朴素贝叶斯算法对群众问政留言合理分类，利用 TF-IDF 方法选取文本关键词项进行留言间的相似度计算，将相关部门对群众留言的答复意见进行归类，由相似度、点赞数和反对数定义出合理的热度评价指标并进行了热度排名。为提升政府的管理水平和施政效率起到推动作用。

针对问题一，将附件二数据首先进行欠抽样、去重、去除脱敏字符、去除停用词等自然语言处理技术对数据预处理，用 TF-IDF 方法提取词频，生成词向量。将处理好的数据 80%作为训练集，20%作为测试集，使用多项式朴素贝叶斯、支持向量机(SVM)两种分类模型，用 python 语言编程进行数值实验，进行测试集的一级标签分类，依据查全率，准确率、F1-Score 等分类评价结果选取评价模型建立关于留言内容的一级标签分类模型。

针对问题二，对留言数据进行预处理, 利用 TF-IDF 模型寻找数据集中具有较高 TF-IDF 值的重要词项，自定义词典分析数据之间的语义相似度,由留言之间相似度转化为衡量留言文本关键词相似度问题。结合词项相似度加权树以及文本语义相似度定义将某一时段内反映特定的问题的留言进行归类，通过加权定义出合理的热度评价指标并进行热度排序。

针对问题三，选取答复意见的六个角度进行衡量答复意见质量，即回复的完整性、一致性、可解释型、相关度、可解释型、时效性。通过 K-means 分类、文本相似度计算、极性得分、ARI 等多种方法得出不同主题下不同角度的评价值，利用了多元线性回归模型将不同指标整合，确定各个指标的权值并得到总的评价值，以此作为答复意见质量的评价方案。

关键词：文本挖掘、机器学习、多项式朴素贝叶斯、IF-IDF、文本相似度、多元线性回归

Abstract

This article aims to use natural language processing technology to carry out text mining on the "smart government" problem. Based on the polynomial naive Bayes algorithm in machine learning to classify the masses 'political messages reasonably, use the TF-IDF method to select text keywords to calculate the similarity between the messages, and classify the responses of the relevant departments to the masses' messages. Reasonable heat evaluation indexes are defined by similarity, likes and dislikes, and heat rankings are performed. It plays a role in promoting the government's management level and governance efficiency.

To solve the first problem, The data in Annex 2 is first preprocessed by natural language processing techniques such as undersampling, deduplication, desensitization characters, and stop words removal, and the word frequency is extracted by the TF-IDF method to generate word vectors. Use 80% of the processed data as the training set and 20% as the test set. Use polynomial naive Bayes and support vector machine (SVM) two classification models. Use Python language programming to perform numerical experiments and perform the first-level labeling of the test set. Classification, according to the results of classification evaluation such as recall rate, accuracy rate, F1-Score, select an evaluation model to establish a first-level label classification model about the content of the message.

To solve the second problem, Preprocess the message data, use the TF-IDF model to find important terms with higher TF-IDF values in the data set, and use a custom dictionary to analyze the semantic similarity between the data. Keyword similarity problem. Combining the term similarity weighting tree and the text semantic similarity definition, the messages that reflect a specific problem within a certain period are classified, and a reasonable heat evaluation index is defined and ranked by weighting.

To solve the third problem, Select the six angles of reply opinions to measure the quality of reply opinions, namely the completeness, consistency, interpretability, relevance, interpretability and timeliness of the reply. Through various methods such as K-means classification, text similarity calculation, polarity score, ARI, etc., the evaluation values of different angles under different topics are obtained. The multiple linear regression model is used to integrate different indicators, determine the weight of each indicator and obtain The total evaluation value is used as the evaluation plan for the quality of the response.

Keywords: Text mining, machine learning, polynomial naive Bayes, IF-IDF, text similarity, multiple linear regression

目录

1 问题分析

习近平总书记在多个场合反复强调，互联网是我们面临的“最大变量”。在互联网时代，网络问政正在逐步变为各级政府治国理政的一种常态，也成为政府获取民意，问计于民、问需于民、问政于民最有效的途径和手段。随着各类民意相关的文本数据量不管攀登，人工进行留言分类和归类已经不切实际，因此利用大数据建立高准确率的模型以及合理的评价标准对留言进行分类与归类是本文所要解决的内容。

对于问题一，针对利用自然语言数据预处理后的附件二的群众留言用于基于问题一建立的多项式朴素贝叶斯模型和 SVM 模型进行机器学习并选取部分群众留言进行分类测试，一种为直接多分类另一种模型将多分类问题转化为多个二分类并利用 F-Score 对分类结果进行评价，选取评价指数最高的模型建立关于留言内容的一级标签分类模型。

对于问题二，自然语言处理技术对留言数据进行预处理后建立自定义词典提高分词准确性，然后利用 TF-IDF 模型寻找文本中具有较高 TF-IDF 值的重要词项。以分词结果为依据自定义语料库用于分析数据之间的语义相似度，将问题由留言之间相似度转化为衡量留言文本关键词相似度问题。

对于问题三，本文选取信息量、相关度、完整性、一致性、有效性和可解释性作为评价答复意见质量的指标，利用不同种方法计算得到每个指标的评价值。建立多元线性回归模型计算得到各指标之间权重，得到一个总评价方案。

2 模型假设

为了便于问题的研究，对题目中某些条件进行简化及合理假设：

假设 1 训练集与测试集中所有留言数据样本不考虑其出现顺序。

假设 2 留言数据的长度与标签类别不相关。

假设 3 关键词向量中词项顺序对文本相似度计算无影响。

假设 4 每条文本数据间相互独立。

3 符号说明

符号	含义
$D=\{d_1,d_2,d_3,\cdots d_n\}$	留言数据集
$C=\{c_1,c_2,c_3,\cdots c_7\}$	一级分类标签集
$T=\{t_1,t_2,t_3,\cdots t_n\}$	一级分类下类别特征集合
$v=(w_1,w_2,\cdots w_n)$	各条留言中特征词向量

4 数据处理

对于不同问题需求以及数据样本采取不同的数据预处理，下文将本文所解决的全部问题的数据处理方法一一列出，针对问题一、二、三进行不同的处理方式。

4.1 自然语言处理技术下数据预处理

1)附件数据去重

根据附件 2、3、4 提供的留言数据以及相关部门答复意见通过 Excle 筛选功能发现有部分文本存在重复现象，如图所示：

留言编号	留言用户	留言主题	留言时间	留言详情	一级标签
52113	U0002991	县客运市场太乱，无人整	18/5/29 23:17:	长期不按规定站外	交通运输
56969	U000583	西渡至碧崖客运班车乱	18/4/5 19:12:0	费就到了12.00元，	交通运输
56970	U000583	西渡至碧崖客运班车乱	18/4/5 19:10:0	费就到了12.00元，	交通运输
61180	U0002991	县客运市场太乱，无人整	18/5/29 23:15:	长期不按规定站外	交通运输
154762	U0004057	县洣水河流域桥梁老化严	19/7/11 10:31:	希望政府能彻底整	交通运输
6103790	U0004057	县洣水河流域桥梁老化严	19/7/11 10:32:	希望政府能彻底整	交通运输
56988	A00085407	渡镇咸水学校师资力量严	18/3/8 18:01:4	介。但今天想想与	教育文体
254446	U0003683	渡镇咸水学校师资力量严	18/3/8 18:01:4	。但今天想想与看	教育文体
32994	U000651	颐景园山体公园游乐设	19/8/28 10:15:	景观池，导致临颐景	商贸旅游
40639	U000651	颐景园山体公园游乐设	19/8/28 10:15:	景观池，导致临颐景	商贸旅游
154000	U0002094	山路建设家园的电梯有扌	19/3/7 14:13:	有一次从6层掉到2层	商贸旅游
154009	U0002094	山路建设家园的电梯有扌	19/2/21 17:42:	有一次从6层掉到2层	商贸旅游

图 1 附件 2 重复留言举例

留言主题	留言时间	留言详情
市能否设立南塘城轨公交站？	2019/10/31 21:19:59	南塘小学，A市一中城
市能否设立南塘城轨公交站？	2019/10/31 21:17:22	南塘小学，A市一中城
市能否设立南塘城轨公交站？	2019-10-31 21:19:59	南塘小学，A市一中城
利用绿色网络按摩平台进行网	2019/12/6 10:52:07	很多商家根本都没有
利用绿色网络按摩平台进行网	2019/12/6 10:45:26	很多商家根本都没有
对在A7县诺亚山林小区门口设	2019/8/16 8:36:38	充斥着无数的病菌。
对在A7县诺亚山林小区门口设	2019/8/16 8:37:43	良的家里，实际却是
医药职业中等专业学校频繁	2019/12/15 18:29:21	单招，第二次又说让我
医药职业中等专业学校频繁	2019/12/15 18:25:03	单招，第二次又说让我
恢复八一路等市区主道的自行	2019/9/27 14:12:08	。此外，东风路的省
恢复八一路等市区主道的自行	2019/9/27 14:20:04	。此外，东风路的省
驾驶证期满换证，一个星期了	2019/4/26 15:28:42	都快一个星期了都没
驾驶证期满换证，一个星期了	2019-04-26 15:28:42	都快一个星期了都没
对在A7县诺亚山林小区门口设	2019/7/8 10:39:54	到干扰。本小区的业主

图 2 附件 3 重复留言举例

利用 Python 中 Pandas 库中的 drop_duplicates 函数进行文本数据去重，针对不同问题进行处理时附件 3、4 需要考虑时间与文本两者皆重复。去重结果如图所示：

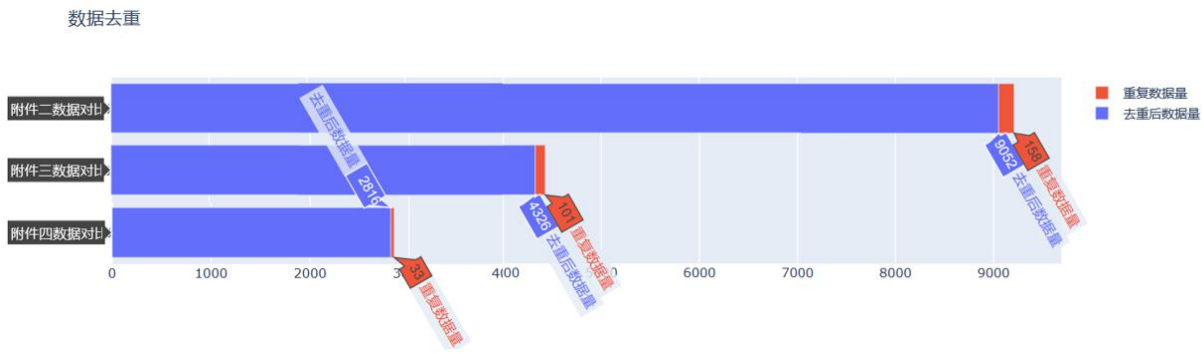

图 3 数据去重结果

2)附件 2 留言数据按一级标签分类

将去重后的数据进行比例输出，为下一步采取欠抽样或过抽样提供数据支持。

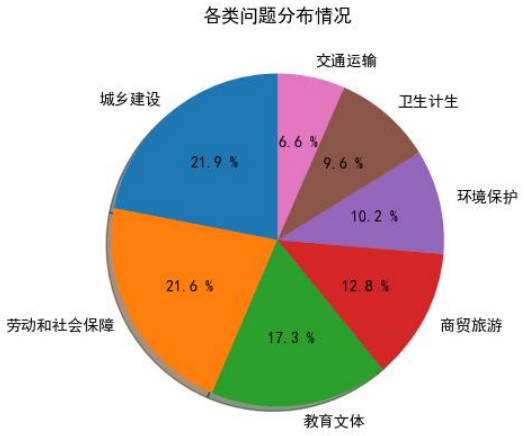

图 4 一级标签下的分类比

3)附件 2 欠抽样

在一级标签下的分类比例可知七大类一级标签存在数据占比不平衡且两两比例相差较小，为解决这一问题本文选取欠抽样平衡分类比。欠抽样是通过减少多数类样本来提高少数类的分类性能，以此解决不同标签下数据不平衡带来的影响。

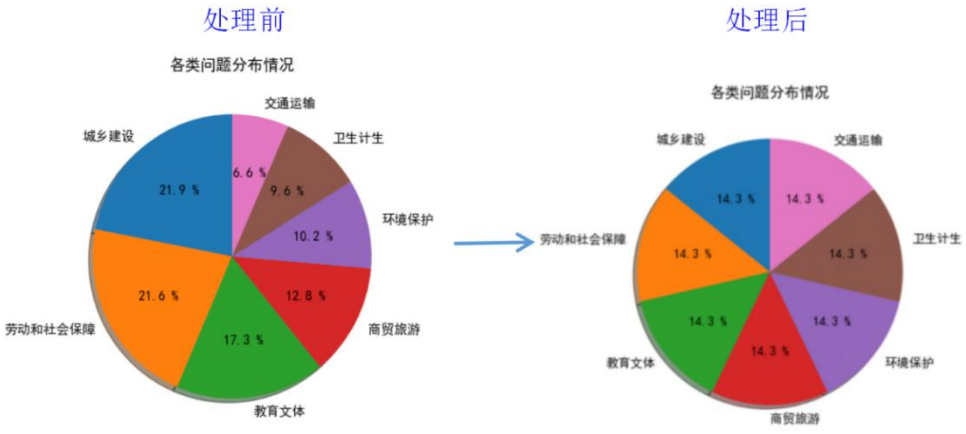

图 5 欠抽样前后占比

4)附件数据清洗

将需要进一步使用的文本数据利用 Python 软件辅助去除空格，去除脱敏字符包括银行卡号、身份信息、日期、电话等对某些敏感信息通过脱敏规则进行数据的变形，实现敏感隐私数据的可靠保护。

5)附件文本中文分词

以词作为基本单元，利用 Python 结巴分词（jieba）对中文文本进行词语的切分，识别出各语句的重点内容。

6)附件文本停用词过滤

将留言中常用的功能性词语是限定词，如“的”、“一个”、“这”、“那”等。仅仅是用于协助一些文本的名词描述和概念表达且没有太多的实际含义。

表 1 数据处理举例

留言数据：
“谢芳家人工资需要涨点，因为她是英雄，更是群众心目中可尊可爱女杰。”
分词之后：
“谢芳”“家人”“工资”“需要”“涨”“点”“因为”“她”“是”“英雄”“更”“是”“群众”“心目”“中”“可尊”“可爱”“女杰”
去除脱敏词后：
“家人”“工资”“需要”“涨”“点”“因为”“她”“是”“英雄”“更”“是”“群众”“心目”“中”“可尊”“可爱”“女杰”
停用词过滤后：
“家人”“工资”“涨”“点”“她”“英雄”“群众”“心目”“可尊”“可爱”“女杰”

7)附件 2 数据抽取

在问题一中随机抽取欠抽样处理后的留言数据的 80%作为机器学习样本，其余作为测试集样本，用于基于问题一建立的多项式朴素贝叶斯模型和 SVM 模型进行训练。

4.2 文本数据集向量表示

针对问题一将处理好的留言数据 80%作为训练集，20%作为测试集，分别进行词频转换并进一步 TF-IDF 权重向量。针对问题二将处理好的留言文本利用 TF=IDF 权重选出每条留言中关键词项组成向量。

4.1.1 词频转换

采用 sklearn 库中的 CountVectorizer 类进行特征词的量化计算，先根据训练集中所有数据样本，不考虑其出现顺序，只将训练文本中每个出现过的词汇单独视为一列特征，构成一个词汇表[1]，然后在 Python 软件中使用 fit__transform 函数将留言特征词的信息转换为词频矩阵。矩阵元素 a[i][j] 表示 j 词在第 i 个留言下的词频。即各个词语出现的次数，并统计留言数据中的关键字。

4.1.2 TF-IDF 权重向量

TF-IDF 权重向量原理为 $IF \times IDF$ ，解释如下表：

表 2 TF-IDF 权重向量解释

	TF（词频）	IDF（逆向文件频率）
含义	一个给定词语 t 在一条给定留言 d 中出现的频率	一个词语 t 对于整个留言数据集的重要性的度量
效果	词语 t 对留言 d 来说越重要，TF 越低，则词语 t 对留言 d 来说越不重要	包含词语 t 的留言越少，IDF 越大，则词语 t 在整个留言数据集上具有很好的类别区分能力
表达式	$TF_{i,j}=\frac{n_{i,j}}{\sum_{k=0}^{n} n_{k,j}}(0 \le k \le n)$	$IDF_i=\log\frac{\lvert D\rvert}{\lvert\{j:t_i \in d_j\}\rvert}$
表达式字母解释	$n_{i,j}$ 是词语 t_i 在留言 d_j 中的出现次数，分母则是在留言 d_j 中所有词语的出现次数之和	$\lvert D\rvert$ 是留言数据集中所有留言总数，分母是包含词语 t_i 的所有留言数

TF 刻画了词语 t 对某个留言数据的重要性，IDF 刻画了词语 t 对整个留言数据集的重要性。总的来说 $IF \times IDF$ 的值越大，代表这个权重就越大，这个词对于留言数据集就越重要。

5 问题一的分析与求解

问题一需要解决的问题是将群众的留言的内容按照划分体系的三类标签中一

级标签进行分类并建立分类模型。本文基于问题一建立了多项式朴素贝叶斯和SVM两种模型，根据F-Score对分类的评价标准进一步确定一级标签分类的模型。

5.1 应用多项式朴素贝叶斯模型

5.1.1 模型建立

基于问题一建立的多项式朴素贝叶斯模型[2]中，留言数据集被看作是一系列单词组成的序列，并且假设留言的长度与类别不相关，且每条留言中的词与其在Excle中的顺序位置也无关。

留言d_k于类别c_i时特征项t_i出现一次的概率为$P(t_j \mid c_i)$，则出现m_k次的概率为$P(t_j \mid c_i)^{m_k}$假设m共有个特征项，则有:

$$P(d_k \mid c_i) = m! \prod_{j=1}^{m} \frac{P(t_j \mid c_i)^{m_k}}{m_k!}$$

在多项式朴素贝叶斯中$P(t_j \mid c_i)$采用上文计算的词频进行计算:

$$P(t_j \mid c_i) = \frac{\sum_{k=1}^{|D|} N(t_j, d_k)}{\sum_{j=1}^{|V|} \sum_{k=1}^{|D|} N(t_j, d_k)}$$

式中，$\sum_{k=1}^{|D|} N(t_j, d_k)$表示特征项$t_j$在一级标签类别$c_i$的各留言中出现的频数之和，$\sum_{j=1}^{|V|} \sum_{k=1}^{|D|} N(t_j, d_k)$表示所有特征项在一级标签类别$c_i$中出现的频数之和。为避免出现零概率，通常加入平滑因子：

$$P(t_j \mid c_i) = \frac{\sum_{k=1}^{|D|} N(t_j, d_k) + 1}{\sum_{j=1}^{|V|} \sum_{k=1}^{|D|} N(t_j, d_k) + |V|}$$

式中，$|V|$是训练集的词汇表中的单词数量。

5.1.2 模型结果与分析

利用 Python 软件实现该模型的测试效果如下：

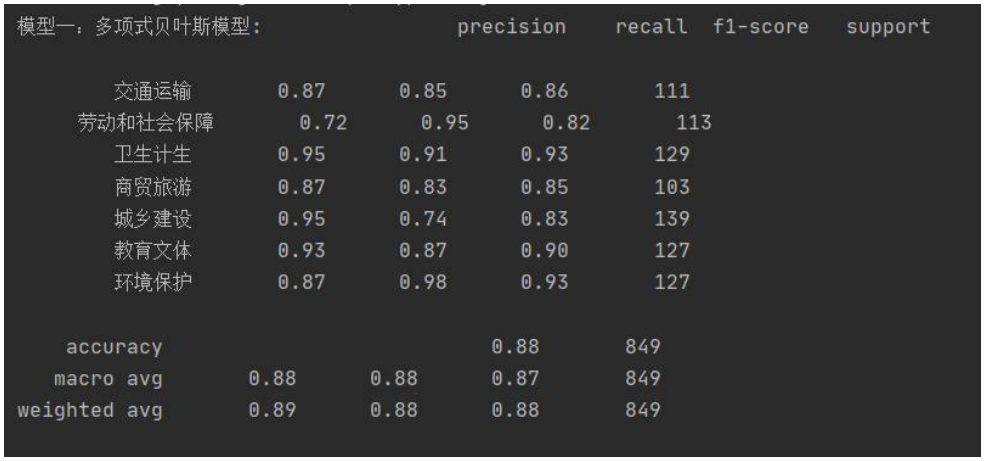

图 6 多项式朴素贝叶斯模型结果

程序中依次将七类一级分类标签筛选出来，获取并输出每一类的查准率（precision）、查全率（recall）、F1-score 分数以及从测试数据中分出的每类的数量。最后分别求取精确平均值以及宏平均、微平均，三个数据对比，以得到更为准确的判断数据。

可以看到图中七类中 F1-score 最低为 0.83，最高为 0.93，平均为 0.88，性能比较稳定，并且分类的精确度也不低。

5.2 应用 SVM 支持向量机模型

问题一属于多分类问题但面对多分类问题是存在一些分类难度，因此本节选用 SVM 二分类模型，将多分类问题转化成多个二分类问题。例如一类是城乡建设另一类为其余六类一级标签，以此进行七次模型运用。

5.2.1 模型建立

基于问题一的 SVM 模型原理[3]是对于一组给定的样本集 D，样本数据是隶属于两个不同类别的样本数据，通过训练构建一个间隔超平面，同时在它的两边建立两个和它平行且有一定距离的超平面，尽可能最大化距离，这样最终能得到总误差最小的分类结果。除此之外，在低维空间中的样本可利用核函数映射到高维空间，并在其中使原本线性不可分的样本找到最优线性分类超平面。二维空间最优分类超平面如图所示：

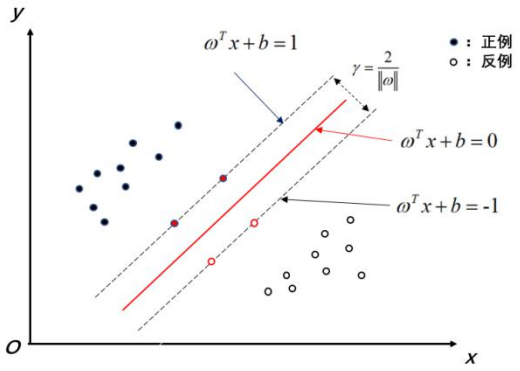

图 7 二维空间超平面

划分超平面的式子为：

$$\omega^T x+b=0$$

其中超平面方向由法向量$\omega=(\omega_1,\omega_2,\cdots\omega_d)$决定，超平面与原点间距离有位移项 b 决定。两个正反类到超平面距离（间隔）之和为$\gamma=\frac{2}{\|\omega\|}$，使其值最大化就能最大间隔划分分类超平面分类误差就越小。核函数$K(d,d_t)$将用特征项权表示的待分类向量d和支持向量d_t映射到高维线性空间，点积(d,d_t)的计算变成核函数计算，再转化对偶问题计算，从而构成$\sum\alpha_i y_t K(d,d_t)$。最终分类器决策函数如下：

$$f(d)=\mathrm{sgn}\sum_{i=1}^{p}\alpha_i y_t K(d,d_t)$$

最终由决策函数值判定文本地的所属类别。

5.2.2 模型的结果与分析

利用 Python 软件实现该模型的测试效果如下：

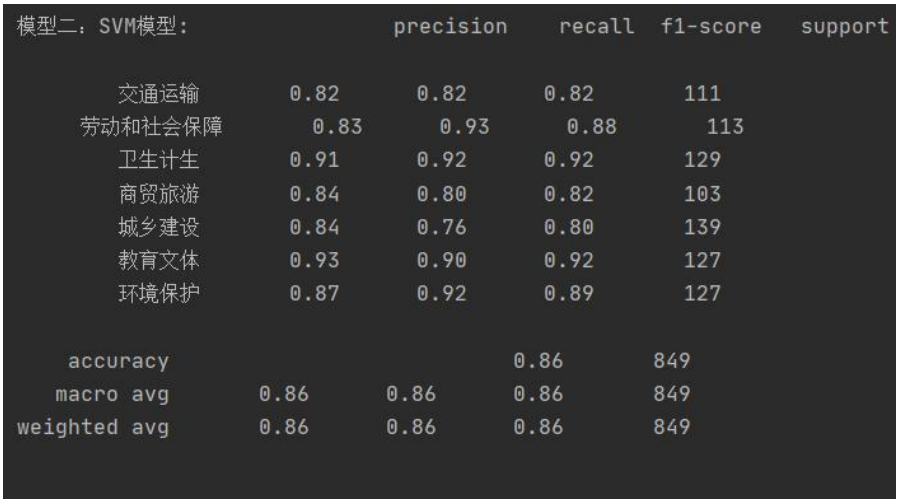

图 8 支持向量机模型结果

程序中依次将七类筛选出来，获取并输出每一类的查准率（precision）、查全率（recall）、F1-score 分数以及从测试数据中分出的每类的数量。最后分别求取精确平均值以及宏平均、微平均，三个数据对比，以得到更为准确的判断数据。

可以看到图中七类中 F1-score 最低为 0.82，最高为 0.92，平均为 0.86，性能比较稳定，并且分类的精确度也不低。

5.3 F-Score 评价两种分类模型

依据题意选取 F-Score 对分类方法进行评价：

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

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。由多次测试得到：

表 3 模型 F-Score 评价值对比

多项式朴素贝叶斯模型下分类	F-Score 评价值	SVM 模型下分类	F-Score 评价值
交通运输	0.86	交通运输	0.82
劳动和社会保障	0.82	劳动和社会保障	0.88
卫生计生	0.93	卫生计生	0.92
商务旅游	0.85	商务旅游	0.82
城乡建设	0.83	城乡建设	0.80
教育文体	0.90	教育文体	0.92
环境保护	0.93	环境保护	0.89
均值	0.88	均值	0.86

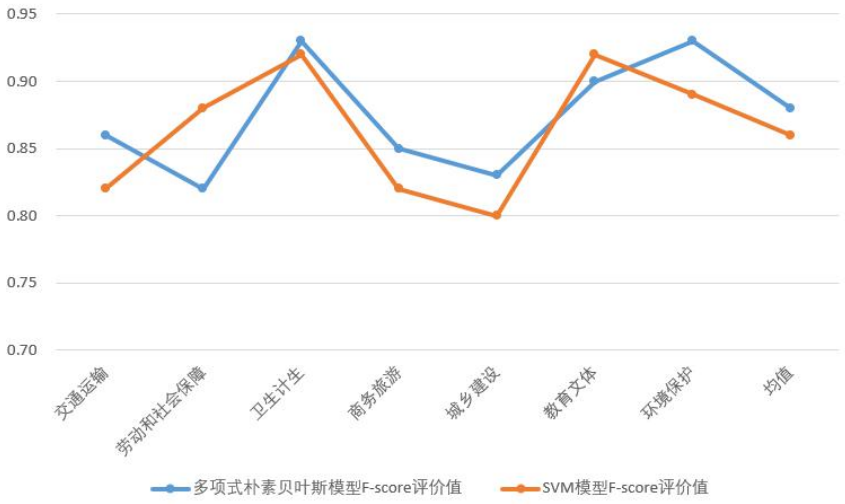

图 9 两种模型比较

在多次的测试后依据 F-Score 评价标准，SVM 模型的分类准确度、F1-score 分数均低于多项式朴素贝叶斯模型，因此最终我们决定采用多项式朴素贝叶斯模型作为问题一一级标签分类模型。

6 问题二的分析与求解

问题二是将留言中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，依据合理的热度评价指标将热度前五名的热点问题总结。将问题由留言之间相似度转化为衡量留言文本关键词相似度问题，并计算得到$n \times n$的相似度矩阵（n为留言条数）。由各条留言相似度进行留言主题的归类。

6.1 关键词项选择

留言数据自然语言预处理完成后，需要对整个留言集中每一条留言的词项进行 TF-IDF 值的计算。为了减小文本特征向量的维度，把每一条留言中词项的 TF-IDF 值进行排序，从中选取 TF-IDF 值大于 p（p 为百分比）的名词和动词词项作为关键词项，选用 TF-IDF 值大于 p 关键词项向量作为本条留言的特征表示，为下文进行文本的相似度计算做准备。

6.2 构建用户词典

- 读入留言数据，为方便下面的操作我们把数据导入 txt 文件中。
- 为了使用词频分析更准确地提取数据中出现次数较多的关键字，对数据进行第

一遍去重。

• 使用 Python jieba 库中的 extract_tags 和 textrank 分别对数据进行基于 TF-IDF 算法和基于 TextRank 算法的关键字提取，显示出出现频率最高的前二十个词语。

图 10 extract_tags 方法结果

图 11 textrank 方法结果

可以看到多数为地名如社区、小区，如果直接进行分词可能会影响到对句意判断的准确性。

• 使用正则表达式，以高频词汇为依据，匹配各类地名，如某某小区、某某街道，第二次去重后保存为 txt 文件作为自定义字典。

6.3 文本相似度的计算

在得到每条留言的特征向量之后，计算每条留言与其余各条留言的相似度计算。因为每天留言的关键词项代表了每条留言中最重要的信息，因此留言之间相似度转化为衡量留言文本关键词相似度问题来计算。因此，留言之间的相似度就转换为留言中关键词项向量间的相似度。

两条留言文本相似度比较举例[4]：

设 v_i, v_j 是两篇不同文本的关键词向量，其中 $v_i = (w_1, w_2, \cdots w_n)$，

$v_j = (w_1, w_2, \cdots w_n)$。定义留言文本相似度为：

$$TextSim(v_i, v_j) = wf \times VectSim(v_i, v_j)$$

式中 wf 表示关键词向量 v_i 和 v_j 之间相似度的加权因子，$VectSim(v_i, v_j)$ 表示关键词向量 v_i 和 v_j 之间的相似度。如果两条留言中彼此相似度较高的词项越多，而词项所占的 TF-IDF 值在各留言文本中比例越高，说明这些词项更能反映它们在

留言文本中的重要性，因此我们根据关键词向量中满足相似度阈值条件的关键词项的TF-IDF值在整篇文本TF-IDF值总和中所占的比例进行加权，具体的加权因子公式为：

$$wf = 1 + ave(i,j) \times (\sqrt{VectSim(v_i, v_j)} - VectSim(v_i, v_j)$$

$$ave(i,j) = \frac{1}{2}\left(\frac{\sum_{k \in \Lambda_i} TFIDF(w_{ik})}{\sum_{k=1} TFIDF(w_{ik})} + \frac{\sum_{l \in \Lambda_j} TFIDF(w_{jl})}{\sum_{j=1} TFIDF(w_{jl})}\right)$$

式中$TFIDF(w_{ik})$表示关键词项w_{ik}的TF-IDF值。式中集合Λ_i和Λ_j定义如下：

$$\Lambda_i = \{k : 1 \le k \le m, \max\{Sim(w_{ik}, w_{jl})\} \ge \mu\}$$ （同理Λ_j）

式中$Sim(x,y)$是数据点x和y之间的相似度，应满足以下条件:

当且仅当x=y时，$Sim(x,y) = 1 (0 \le Sim(x,y) \le 1)$。

对于所有的x和y，$Sim(x,y) = Sim(y,x)$。

6.4 问题求解与分析

由于每一条留言主题需要与由全部留言主题构成的语料库进行比较，以求得特征词，为了降低时间复杂度，先读取全部的留言主题，把这些留言主题处理成语料库，得到TF-IDF值。然后遍历附件三中的留言主题，由于附件三中的留言编号不是连续递增，为了方便，在实际处理数据时，新增了一列从0开始，公差为1的递增数列来对留言进行编号。结合留言主题下留言条数与点赞反对数合理建立留言问题的热度评价标准。

6.4.1 留言主题文本相似度排名

在遍历留言主题时，先创建了Hotspot和HotspotNum两个文件夹，每个文件夹中都有从00000.txt,00001.txt,……，04325.txt共计4326个文件，HotspotNum文件夹中，每个文件的内容是与文件名对应主题相似度高的主题对应的编号，Hotspot文件夹中，每个文件的内容与文件名对应主题相似度高的主题对应的编号和主题。程序运行完成后，因为所有序号都是按照五位进行输出，因此文件大小越大，与其相似的主题也就越多，这个问题是热点问题的概率也就越大，我们对HotspotNum文件夹中的文件按照大小排序，再根据这个顺序，结合Hotspot文件夹中的数据对其进行去重，得到了十个留言次数最多的问题，

利用Python实现模型将热点问题主题的范围缩小的前十个主题，依据文本相似度的计算得到前十名的热点问题如下：

问题ID	条数	问题描述
1	43	A市伊景园滨河苑定向售房，捆绑车位销售
2	38	A市A2区丽发新城小区旁搅拌站扰民
3	26	A市人才购房补贴问题
4	23	A市A5区魅力之城小区小区临街餐饮店油烟噪音扰民
5	15	A7县多个小区存在麻将馆扰民
6	10	A市经开区泉星公园项目规划需优化
7	10	咨询A市住房公积金贷款问题
8	9	A1区朝阳社区摊位扰民
9	9	A市经济学院强制学生实习
10	7	A市北辰定江洋小区住改商问题

图 12 相似度降序排名

6.4.2 热度评价指标

基于排名前十名的的热点问题中，计算出每个热点问题主题下留言的点赞数与反对数之和生成降序排名，选取前五为排名前 5 的热点问题。具体结果详见“热点问题表”、“热点问题留言表”。

问题ID	热度指数	问题描述
1	90	A市A2区丽发新城小区旁搅拌站扰民
2	67	A市伊景园滨河苑定向售房，捆绑车位销售
3	60	A市A5区魅力之城小区小区临街餐饮店油烟噪音扰民
4	58	A市人才购房补贴问题
5	34	A7县多个小区存在麻将馆扰民
6	31	A市经开区泉星公园项目规划需优化
7	23	A市经济学院强制学生实习
8	14	A市北辰定江洋小区住改商问题盼
9	11	咨询A市住房公积金贷款问题
10	10	A1区朝阳社区摊位扰民

图 13 热点问题关注数排名

7 问题三的分析与求解

7.1 留言答复质量评价角度

评价留言答复的质量有不同角度，如下表所示：

表 4 衡量答复内容指标

指标	解释
信息量	从内容上保证答复的质量，以答复长度衡量包括字数和词语数量
相关度	答复内容与答复留言主题的相关性
完整性	答复内容是否包含数值型和文本型

	数据（完整的答复标记为 1，不完整标记为 0）
一致性	判断答复内容是否与留言问题内容一致
有效性	判断答复内容是否可取
可解释性	答复内容可以被理解采纳的程度

- 信息量：信息量包括句量、词量、字量等，一般认为答复的信息量越大答复的质量就越高，本文采用答复的句子长度衡量信息量。
- 相关度：将答复意见中每一个答复意见对应词典库中的一个文档，文档中词项内容用向量表示，引用问题二文本相似度计算方法计算“答复意见—留言主题”相似度计算。
- 完整性：将不同主题的答复意见分词选取出针对不同答复主题的特征词向量，采用基于长度的 K-means 算法[5]，得到属于该类主题的词项数与不属于该主题词项数，利用准确率、召回率、F1 值衡量答复内容完整性，具体流程如下图：

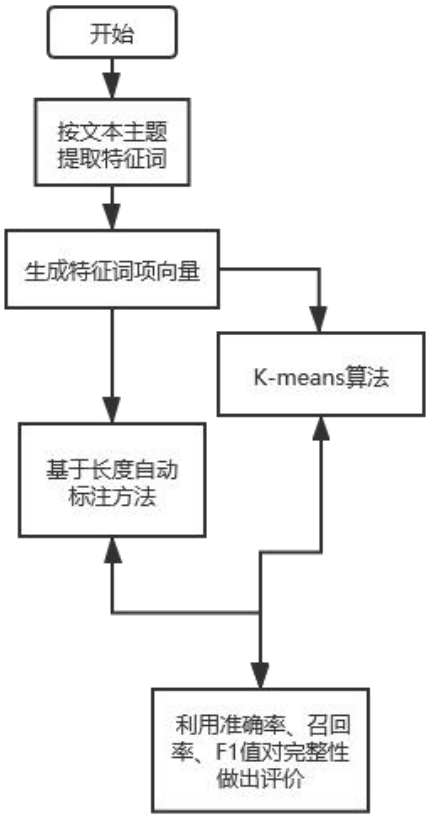

图 14 完整性评价流程图

- 一致性：答复内容评分记为 X_1，留言内容评分记为 X_2，如果两种评价一致

$(X_1 = X_2)$ 时一致性得分 $P=1$，不一致时 $P=0.5$。具体计算内容如下：

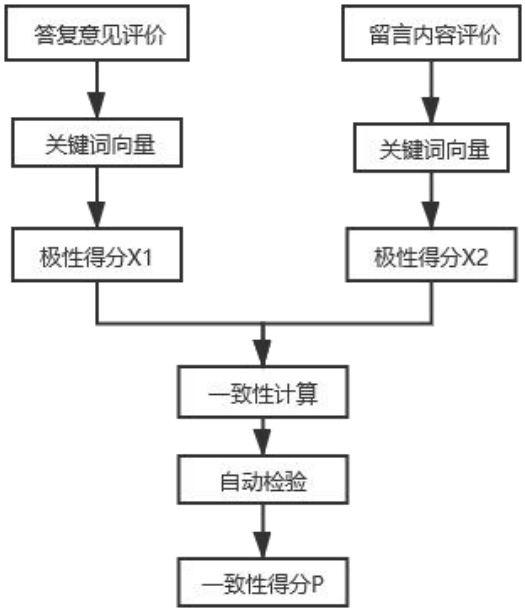

图 15 一致性计算流程图

• 可解释性：可解释性又可理解为可读性，本文选用自动化可读性指数 ARI 衡量，ARI 依赖于文本的字符数，计算 ARI 时采用公式[5]：

$$ARI = 4.71*(\frac{总字符数}{总字数}) + 0.5*(\frac{总字数}{总句数}) - 21.43$$

7.2 确定答复评价角度权重

由 7.1 可知评价留言答复质量的角度有很多，对于上文所提到的衡量指标之间存在相应的权重问题，以此便于同时从多个角度对留言的答复进行衡量答复意见的质量。

7.2.1 建立多元回归模型

多元线性分析是指因变量与多个自变量之间的线性关系。假设 Y 是因变量（答复意见质量），$X_1, X_2, X_3, \cdots X_n$ 为自变量（评价答复意见角度），则多元线性回归模型可表示为：

$$Y = a_0 + a_1 X_1 + a_2 X_2 \cdots + a_n X_n + e$$

式中，a_0为常数项，$a_1 a_2 \cdots a_n$为回归系数。利用此式对多角度下的答复意见质量进行评价，本文将不进行结果分析。

8 参考文献

[1]吕彦[1], 程树林[1]. 商品自动分类的贝叶斯方法及 Python 实现[J]. 安庆师范大学学报:自然科学版, 2019(2):66-69.
[2]李腾飞. 基于多项式朴素贝叶斯算法的垃圾邮件过滤器的设计与实现[J]. 科技资讯, 2018, 16(33):7-8+10.
[3]郭超磊. 基于 SA-SVM 的中文文本分类研究[D]. 2019.
[4]黄承慧, 印鉴, 侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法[J]. 计算机学报, 2011(05):98-106.
[5]郭银灵. 基于文本分析的在线评论质量评价模型研究[D]. 2017.