

“智慧政务”中的文本挖掘应用

摘要

智慧政府的初始概念由智慧城市发展而来，而智慧政府就是有效利用互联网信息与通讯技术手段，进而对社会上各种相关的群体包括对政府进行治理方面的各种数据内容进行整合汇总，对公共活动所需求的重要信息进行梳理并分析判断、科学决策、对此提出合理化的建议。随着新一轮的信息技术与社会经济的大融合，政府网站作为政府对外宣传展示、为民服务以及同人民群众沟通的重要窗口，是电子政务建设的关键环节，亟需运用新技术以实现服务新。通过各类信息的汇集，在政务服务网终端服务器上已经储存了大量的数据。如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的“加工能力”，通过“加工”实现数据的“增值”。

本文研究了在文本分类问题中通过已经给定的训练样本数据集(一般情况下是指已经被人工判断并已经标注好类别信息的文本)和文本分类的体系，通过提取文本 TF-IDF 特征值建立朴素贝叶斯模型和 LinearSVC 模型来判断还未被分类的特定类别文本内容，在完成相应判断之后再根据判断的结果分到已经给定的相对应的预定文本类别中。同时还对留言详情的文本进行文本特征提取后进行 K-means 聚类来挖掘热点问题再通过构建合理的热度指数排序后建立热点问题表。最后我们还通过对政务留言回复文本的相似度计算了相关性、规范性、时效性等一系列指标最终生成答复质量评价指标。

关键词：文本多分类；文本聚类；文本相似度；朴素贝叶斯；TF-IDF；LinearSVC

Abstract

The initial concept of smart government is developed from smart cities, and smart government is to effectively use Internet information and communication technology to integrate and aggregate various relevant groups in society, including various data content of government governance. , Sort out and analyze important information required for public activities, analyze and judge, make scientific decisions, and make rational suggestions for this. With the new round of integration of information technology and social economy, the government website, as an important window for government publicity and display, serving the people and communicating with the people, is a key link in the construction of e-government, and new technologies are urgently needed to achieve New service. Through the collection of various types of information, a large amount of data has been stored on the terminal server of the government service network. If big data is compared to an industry, then the key to achieving profitability in this industry lies in improving the "processing capacity" of the data and realizing the "value added" of the data through "processing".

This paper studies the text classification problem through the training data set that has been given and text classification system, by extracting the text TF-IDF feature value Establish a simple Bayesian model and a LinearSVC model to judge the text content of the uncategorized categories that have not yet been classified, and then divide them into the corresponding predetermined text categories that have been given after the corresponding judgment is completed. At the same time, the text features of the message details are extracted, then K-means clustering is used to mine hot spots, and then a hot spot list is established by constructing a reasonable ranking of the heat index. Finally, we also calculated a series of indicators such as relevance, standardization, and timeliness through the similarity of the reply texts of government affairs messages, and finally generated response quality evaluation indicators.

Keywords: text multi-classification; text clustering; text similarity; naive Bayes; TF-IDF; LinearSVC

目录

1. 问题重述.....	1
2. 模型假设.....	1
3. 符号说明.....	1
4. 数据预处理.....	2
4.1 数据清洗.....	2
4.2 中文文本分词.....	3
4.2.1 创建自定义词典.....	3
5. 群众留言分类.....	4
5.1 指标的确立.....	4
5.2 模型选择.....	4
5.3 基本原理介绍.....	6
5.3.1 文本分类.....	6
5.3.2 TF-IDF 特征值.....	7
5.3.3 卡方(chi2)检验.....	8
5.4 模型引入.....	9
5.4.1 朴素贝叶斯分类.....	9
5.4.2 线性支持向量机.....	10
5.5 判断结果分析.....	11
5.6 模型评价.....	13
6. 热点问题挖掘.....	15
6.1 指标的确立.....	16
6.2 文本聚类.....	16
6.2.1 文本特征处理.....	16
6.2.2 K-means 聚类.....	17
6.3 定义热度指数.....	17
6.4 热点问题表.....	18
7. 答复意见评价.....	19
8. 模型评价.....	21
9. 参考文献.....	22

1. 问题重述

问题一：为了解决群众问题，设置网络问政平台让群众进行反馈，在处理网络问政平台的群众留言时，工作人员需要按照相对应的体系对留言进行分类，方便将群众留言分配至相应的职能部门处理。现在很多电子政务系统还是依靠人工根据经验进行分类处理，这样存在工作量大、效率低、出错率高等问题。根据已经给定的数据来建立一级标签分类模型，通过对已分类数据的训练对未分类数据进行自动分类。

问题二：某一段时间内群众集中反映的某一问题称为热点问题，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”。若能及时发现热点问题，能够使得相关部门进行有正对性的处理，提升服务效率。将“用户留言详情表”的数据根据某一时段内反映的问题内容进行归类，定义出合理的热度评价指标，并给出评价结果，整理出具有热度排名、问题 ID、热度指数、时间范围、地点/人群、问题描述的“热点问题表”。根据“热点问题表”整理出“热点问题留言明细表”。

问题三：根据相关部门对留言的答复意见，可以从答复的相关性、完整性、可解释性等角度，对答复意见的质量制定出一套评价方案。

2. 模型假设

1. 假设分词没有误差
2. 假设聚类簇的数目合理
3. 假设 K-means 聚类准确度与数据量多少无关
4. 假设 TF-IDF 默认阈值合理

3. 符号说明

符号	说明
TF	词频

IDF	逆文本频率指数
t	词条内容
m	文本数
$N(x)$	含有词 x 的文本数
$P(D)$	训练样本数据 D 的先验概率
$P(h)$	一个假设 h 在位悬链钱的初始概率
P_i	第 i 类的查准率
R_i	第 i 类的查全率
H	热度指数
Q	每一个簇中评论的个数
C	每簇类留言中的点赞数之和
Z	每簇类留言中的反对数之和
N	答复意见质量
χ	复意见与留言详情的相关性
K	答复意见的规范性
O	答复的时效性

4. 数据预处理

4.1 数据清洗

由于数据来源于群众留言，而群众留言具有语言不规范、错别字与表达不到位的特点，如：将“我们市”、“A 市”、“该市”所描述的含义一致，为提高文本数据的可信度与后续模型的准确度，首先去除用户评论数据中的空值、重复值与符号等内容。由于数据量达到上千条，本研究采用 Python 编程语言通过编写程序对 excel 文件进行自动化处理。

4.2 中文文本分词

分词是指将一句完整的话使用某种方法进行拆分形成词集，该词集将作为后续文本挖掘的基本单位。本实验中，将使用 jieba 库对使用的中文文本数据运用计算机自动切分，使其按照用空格的形式将词与词分开，这样更方便计算机识别文本进行计算。

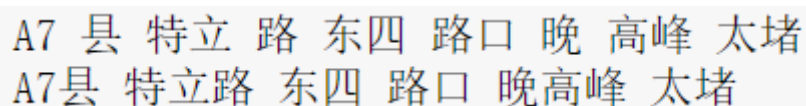
4.2.1 创建自定义词典

在使用了 jieba 分词后虽然达到了分词的目的，但效果并不理想，这是由于在特定场景中的特定专业词汇与一些路标名词在词库中是没有的，如“晚高峰”，若直接分词，则会形成“晚”和“高峰”两个词，并且我们会发现，分词结果中缺少了一些英文字表地标的名称，如“A 市”在分词结束后则会变为“市”，这显然达不到我们预期的效果，所以需要对词典进行进一步优化，这些则需要手动添加。

自定义词典可以用来添加 jieba 词库中不存在的词，自定义词典的方法为直接加入文本。增加新词的示例如下：

```
Data = 'A7 县特立路与东四路口晚高峰太堵'
jieba.add_word('晚高峰')
jieba.add_word('特立路')
jieba.add_word('A7 县')
```

通过自定义词典添加了新词后，得到的两种分词结果如下图：



A7 县 特立 路 东四 路口 晚 高峰 太堵
A7县 特立路 东四 路口 晚高峰 太堵

图 4-2-1 分词结果图

由结果可以看出，在增加了新词后，分词效果较之前更加准确了。

5. 群众留言分类

对群众留言分类问题的思路为：首先对留言进行数据预处理，即分词与转换为 TF-IDF 值，处理结束后对数据进行划分训练集和测试集，选择模型并使用训练集数据建模进行文本分类构成文本分类器，进一步训练优化模型使模型达到最优。将测试集数据导入训练好的模型，对模型进行调整使其达到最优，最后使用 F-Score 值对分类模型进行评价。

5.1 指标的确立

政府对留言的处理情况是群众衡量政府办公质量的一个重要标准，由于政府每日接受留言信息量大，需要对每条留言进行分类以交付管部门处理，诶提高其工作效率，获取了使用政府问政平台上的留言数据，得到下表所示的政府留言数据说明表 5-1：

表 5-1 政府留言数据说明表

变量名	详细说明	取值范围
留言编号	定量变量	0-336810076
留言用户	文本数据	/
留言主题	文本数据	/
留言时间	单位：年/月/日	2010/11/2-2020/1/8
留言详情	文本数据	/
一级标签	定性变量 共七个水平	城乡建设、环境保护、交通运输、教育文体、商贸旅游、计生卫生、劳动和社会保障

5.2 模型选择

由于政府收到的留言量大，为方便相关部门及时解决问题，需要对留言进行分类以提高政府工作效率。文本分类的常用方法包括朴素贝叶斯、卡方(chi2)检

验和线性支持向量机。在对群众留言分类问题进行分析和对模型优劣及特点进行比较选择使用朴素贝叶斯和线性假设的支持向量机分类器 LinearSVC 作为我们分类器，分别进行分类处理，后对两者准确度进行计算。

朴素贝叶斯的主要优点有：

- 1) 朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。
- 2) 对小规模的数据表现很好，能个处理多分类任务，适合增量式训练，尤其是数据量超出内存时，我们可以一批批的去增量训练。
- 3) 对缺失数据不太敏感，算法也比较简单，常用于文本分类。

朴素贝叶斯的主要缺点有：

- 1) 理论上，朴素贝叶斯模型与其他分类方法相比具有最小的误差率。但在属性个数比较多或者属性之间相关性较大时，分类效果不好。
- 2) 需要知道先验概率，且先验概率很多时候取决于假设，在某些时候会由于假设的先验模型的原因导致预测效果不佳。
- 3) 由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。
- 4) 对输入数据的表达形式很敏感。

支持向量机优点：

- 1) 使用核函数可以向高维空间进行映射
- 2) 使用核函数可以解决非线性的分类是一种有坚实理论基础的新颖的适用小样本学习方法。它基本上不涉及概率测度及大数定律等，也简化了通常的分类和回归等问题。
- 3) 分类思想很简单，就是将样本与决策面的间隔最大化
- 4) 分类效果较好，有优秀的泛化能力。

支持向量机缺点：

- 1) SVM 算法对大规模训练样本难以实施，当数据量很大时该矩阵的存储和计算将耗费大量的机器内存和运算时间。
- 2) 用 SVM 解决多分类问题存在困难，经典的支持向量机算法只给出了二类分类的算法，而在实际应用中，一般要解决多类的分类问题。
- 3) 对缺失数据敏感，对参数和核函数的选择敏感。如何根据实际的数据模型

选择合适的核函数从而构造 SVM 算法，目前比较成熟的核函数及其参数的选择都是人为的，根据经验来选取的，带有一定的随意性。

5.3 基本原理介绍

5.3.1 文本分类

文本分类的一般定义是指通过已经给定的训练样本数据集(一般情况下是指已经被人工判断并已经标注好类别信息的文本)和文本分类的体系，判断还未被分类的待定类别文本内容，在完成相应判断之后再根据判断的结果分到已经给定的相对应的预定文本类别中。既文本分类是基于种类已设定的体系，按照文本的内容将文本划分到已设定的文本类别中。一般情况下，分类方法首先运用知识工程或机器学习的理论构建分类模型，然后通过模型的分类标准将带分类的文本划分到已设定的文本类别中。

通常情况下一个完整的中文文本分类系统由 6 个核心功能模块构成：

(1) 文本预处理：文本预处理模块中含有剔除停用词及文本信息分词，其中居首要地位的是文本信息分词。

(2) 文本表示：文本代表着文本分类的基础。当前，文本表示模块中经常使用的方法是向量空间模型。

(3) 文本特征选择：维数约简可在文本特征选择模块中实现，此模块能够从文本中提取显著表现出信息种类的特征项。下文将在第 4 章中详细概述特性选择的内容。

(4) 特征权重计算：在文本信息中，特征权重能够突显出特征项所拥有的重要性，又或是能判别出区分能力的级别。

(5) 分类器学习训练：对训练样本集的处理主要是熟练利用统计学里面的采取学习算法，对不同分类器的参数值做出预估，综合上述分析结果，建立一个自动分类器，这个分类器可以集训练、学习训练两部分为一体。

(6) 测试与评价：利用已经建立的分类器，用分类测试的方对检测测试集文档。 确认合适的评价指标，在分类器完结束训练和测试之后评价其使用性能。

当性能与要求相差太远时，按照之前的流程再建立一个这样的分类器。

5.3.2 TF-IDF 特征值

$TF-IDF$ (term frequency - inverse document frequency) 是一种用于信息检索与数据挖掘的常用加权技术。 TF 是词频(Term Frequency)， IDF 是逆文本频率指数(Inverse Document Frequency)。 $TF-IDF$ 还是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。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。 $TF-IDF$ 加权的各种形式常被搜索引擎应用，作为文件与用户查询之间相关程度的度量或评级。除了 $TF-IDF$ 以外，因特网上的搜索引擎还会使用基于链接分析的评级方法，以确定文件在搜寻结果中出现的顺序。

$TF-IDF$ 的主要思想是：如果某个词或短语在一篇文章中出现的频率 TF 高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。 $TF-IDF$ 实际上是： $TF \times IDF$ ， TF 词频(Term Frequency)， IDF 逆向文件频率(Inverse Document Frequency)。 TF 表示词条在文档 d 中出现的频率。 IDF 的主要思想是：如果包含词条 t 的文档越少，也就是 n 越小， IDF 越大，则说明词条 t 具有很好的类别区分能力。如果某一类文档 C 中包含词条 t 的文档数为 m ，而其它类包含 t 的文档总数为 k ，显然所有包含 t 的文档数 $n = m + k$ ，当 m 大的时候， n 也大，按照 IDF 公式得到的 IDF 的值会小，就说明该词条 t 类别区分能力不强。

对于在某一特定文件里的词语来说，它的重要性可表示为：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \dots\dots\dots (5-1)$$

以上式子中分子是该词在文件中的出现次数，而分母则是在文件中所有字词的出现次数之和。

$$IDF(x) = \log \frac{N}{N(x)} \dots\dots\dots (5-2)$$

其中 N 表示语料库中文本数量， $N(x)$ 表示含有词 x 的文本数。如果一个词在

语料库中不出现，那么 $N(x)$ 则为 0，而分母不能为 0，出现计算错误。这里使用平滑处理，使语料库没有出现的词也能得到一个合理的 IDF 值，对上述公式改进后得到：

$$IDF(x) = \log \frac{N+1}{N(x)+1} + 1 \quad \dots\dots\dots (5-3)$$

根据上述公式，我们就得到 $TF-IDF$ 值的计算公式：
 $TF-IDF = TF(x) \times IDF(x)$ 其中 $TF(x)$ 表示词 x 在当前文本中的词频。

5.3.3 卡方(chi2)检验

卡方检验是一种以 X^2 分布为基础的用途广泛的假设检验方法。是一种非参数检验方法。它的无效假设 H_0 为：观察频数与期望频数没有显著性差异。

卡方检验的主要步骤有如下 5 步。

(1) 提出原假设：

H_0 ：总体 X 的分布函数为 $F(x)$ 。

如果总体分布为离散型，则假设具体为

H_0 ：总体 X 的分布律为 $P\{X=x_i\} = p_i, i=1,2,\dots$

(2) 将总体 X 的取值范围分成 k 个互不相交的小区间 $A_1, A_2, A_3, \dots, A_k$ 可取

$A_1 = (a_0, a_1], A_2 = (a_1, a_2], \dots, A_k = (a_{k-1}, a_k]$ 。

(3) 把落入第 i 个小区间的 A_i 的样本值的个数记作 f_i ，成为组频数（真实值），所有组频数之和 $f_1 + f_2 + \dots + f_k$ 等于样本容量 n 。

(4) 当 H_0 为真时，根据所假设的总体理论分布，可算出总体 X 的值落入第 i 个小区间 A_i 的概率 p_i ，于是， np_i 就是落入第 i 个小区间 A_i 的样本值的理论频数（理论值）。

(5) 当 H_0 为真时， n 次试验中样本值落入第 i 个小区间 A_i 的频率 $\frac{f_i}{n}$ 与概率 p_i 应很接近，当 H_0 不真时，则 $\frac{f_i}{n}$ 与 p_i 相差很大。

基于这种思想，皮尔逊进如下检验统计量：

$$\chi^2 = \sum_{i=1}^k \frac{(f_i - np_i)^2}{np_i} \dots\dots\dots (5-4)$$

在 0 假设成立的情况下服从自由度为 k-1 的卡方分布。

5.4 模型引入

5.4.1 朴素贝叶斯分类

基于文本分类的机器学习算法主要包括 KNN，LLSF，决策树和 Bayes 等。朴素贝叶斯算法以贝叶斯理论为基础，当各属性之间相互独立时，朴素贝叶斯算法的准确率达到最高。

朴素贝叶斯算法中使用先验概率，这是判断一个假设条件能否成立的背景。在本文数据挖掘中，用 $P(h)$ 表示一个假设 h 在位悬链钱的初始概率，即假设 h 的先验概率。而在没有先验概率时，可为各假设条件赋一个相同的先验概率。同理，若用 $P(D)$ 表示训练样本数据 D 的先验概率。对于 $P(D|h)$ ，它表示当假设 h 成立时观察到数据 D 的条件概率。在机器学习中，通常需要研究的是 $P(h|D)$ ，即给定一个训练样本数据 D 之后，判断在数据 D 的基础上假设 h 成立的条件概率，也把它叫做后验概率，它表示训练样本数据 D 出现时假设 h 成立的置信度。

根据贝叶斯公式

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \dots\dots\dots (5-5)$$

去掉公式中的 $P(D)$ ，公式中后验概率 $P(h|D)$ 的值就取决于 $P(D|h)P(h)$ 的乘积，即本文挖掘中所要用到的朴素贝叶斯算法的核心思想。

一个数据集中，往往包含多个属性，假设数据 D 中包含 n 个属性 a_1, a_2, \dots, a_n ，基于朴素贝叶斯算法的假设条件：个属性值之间相互独立。即在给定一个具体目标时， a_1, a_2, \dots, a_n 同时发生的概率等于每个属性单独发生概率的乘积，公式表

示为：

$$P(D|h) = P(a_1, a_2, \dots, a_n | h) = \prod_{i=1}^n P(a_i | h) \quad \dots\dots\dots (5-6)$$

则朴素贝叶斯的后验概率公式即可表示为：

$$P(h|D) = P(h) = \prod_{i=1}^n P(a_i | h) \quad \dots\dots\dots (5-7)$$

朴素贝叶斯分类器

文本数据不能直接用于数据挖掘，需要经过处理后得到向量矩阵才可以作为文本分类器训练和测评使用的数据，本实验中选择使用朴素贝叶斯的方法来构建文本分类器。

朴素贝叶斯分类器试讲朴素贝叶斯算法应用于分类器上的设计，是一种基于概率学的算法。根据贝叶斯

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \quad \dots\dots\dots (5-8)$$

特征提取后，被处理的文本数据将会变为特征向量 $X(x_1, x_2, \dots, x_n)$ ，接下来将特征向量 $X(x_1, x_2, \dots, x_n)$ 归类到类别 $C(C_1, C_2, \dots, C_n)$ 中。其表示含义为：文本向量 $X(x_1, x_2, \dots, x_n)$ 分别属于的概率值 (P_1, P_2, \dots, P_j) ，其中 P_j 表示 $X(x_1, x_2, \dots, x_n)$ 归类到 C_j 的概率。 $\max(P_1, P_2, \dots, P_j)$ 的结果就是文本 X 所属的类别。

由于朴素贝叶斯的特惠总能条件独立性假设影响了分类器的性能，所以选择对文本赋予不同的权重，对分类贡献能力较差的文本赋予较低权重，对分类贡献能力较大即不常见的文本赋予较高权重，从而达到提高分类的准确率的目的。

5.4.2 线性支持向量机

SVM 是在解决小样本、非线性、高维的分类和回归问题时具有特有优势的机器学习方法，在 SVM 基础上发展的线性支持向量机 (Linear SVM) 已成为处理文本分类等海量高维稀疏数据的一种有效机器学习方法

支持向量机分类器 (Support Vector Classifier) 是根据训练样本的分布，搜索所以可能的线性分类器中最佳的那个，决定分类边界位置的样本并不是所有

训练数据，是其中的两个类别空间的间隔最小的两个不同类别的数据点，即“支持向量”。从而可以在海量甚至高维度的数据中，筛选对预测任务最为有效的少数训练样本。LinearSVC 算法是在底层基于 LIBLINEAR 库实现，具有多种惩罚参数和损失函数可供选择，当训练集实例数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。LinearSVC 采用了线性 SVM 算法。

5.5 判断结果分析

文本分类一般可以分为二分类、多分类、多标签分类三种情况，二分类是指将一组文本分成两个类(0 或 1)，比较常见的应用如垃圾邮件分类、电商网站的用户评价数据的正负面分类等，多分类是指将文本分成若干个类中的某一个类。根据本题的要求需要将留言详情文本进行一级标签进行分类处理，其中共有 7 类一级标签。

首先，对已经清洗和分词好的数据进行 TF-IDF 特征值计算。然后使用 `sklearn.feature_extraction.text.TfidfVectorizer` 方法来抽取文本已计算好的 TF-IDF 的特征值。这里使用了参数 `gram_range=(1, 2)`，表示除了抽取评论中的每个词语外，还需要抽取每个词相邻的词并组成一个“词语对”，如：词 1，词 2，词 3，词 4，(词 1，词 2)，(词 2，词 3)，(词 3，词 4)。这样就大幅度扩展了特征集的数量，有了丰富的特征集才有可能提高我们分类文本的准确度。最终得到的 features 的维度是 (9210, 652705) 这里的 9210 表示样本数据总共有 9210 条留言详情数据，652705 表示其相对对应的特征数量这包括全部评论中的所有词语数和词语对(相邻两个单词的组合)的总数。

其次，使用卡方检验的方法来找出每个分类中的关联度最大的两个词语和词语对。通过卡方检验计算统计样本的实际观测值与理论推断值之间的偏离程度由此推断数据的拟合度和关联度。最终卡方检验得到的结果如下表 5-5-1：

表 5-5-1 卡方检验结果表

关联度 分类标签	Most correlated unigrams		Most correlated bigrams	
交通运输	快递	出租车	的士 司机	出租车 司机
劳动和社会保障	退休	社保	劳动 关系	退休 人员

卫生计生	医生	医院	独生子女 父母	乡村 医生
商贸旅游	传销	电梯	小区 电梯	传销 组织
城乡建设	小区	业主	住房 公积金	公积金 贷款
教育文体	学生	学校	培训 机构	教育局 领导
环境保护	环保局	污染	周边 居民	环保局 领导

一系列的可靠数据结论可以证明找出了每个分类中关联度最强的两个词和两个词语对。这些词和词语对能很好的反映出分类的主题。

最终使用预先选择的朴素贝叶斯分类器 MultinomialNB 和 Linear Support Vector Machine (线性支持向量机)，首先将成功分词的数据转换成词频向量，然后再将词频向量再转换成 TF-IDF 向量，最后训练两个不同的分类器。对准确率进行评估后发现

表 5-5-2 准确率表

LinearSVC	0.875683
MultinomialNB	0.662645

可以看到线性支持向量机的平均准确率达到 87.56%，因此针对平均准确率最高的 LinearSVC 模型查看它的混淆矩阵，进一步观察预测标签和实际标签之间的差异。更好地帮助我们进行模型的优化与调整。

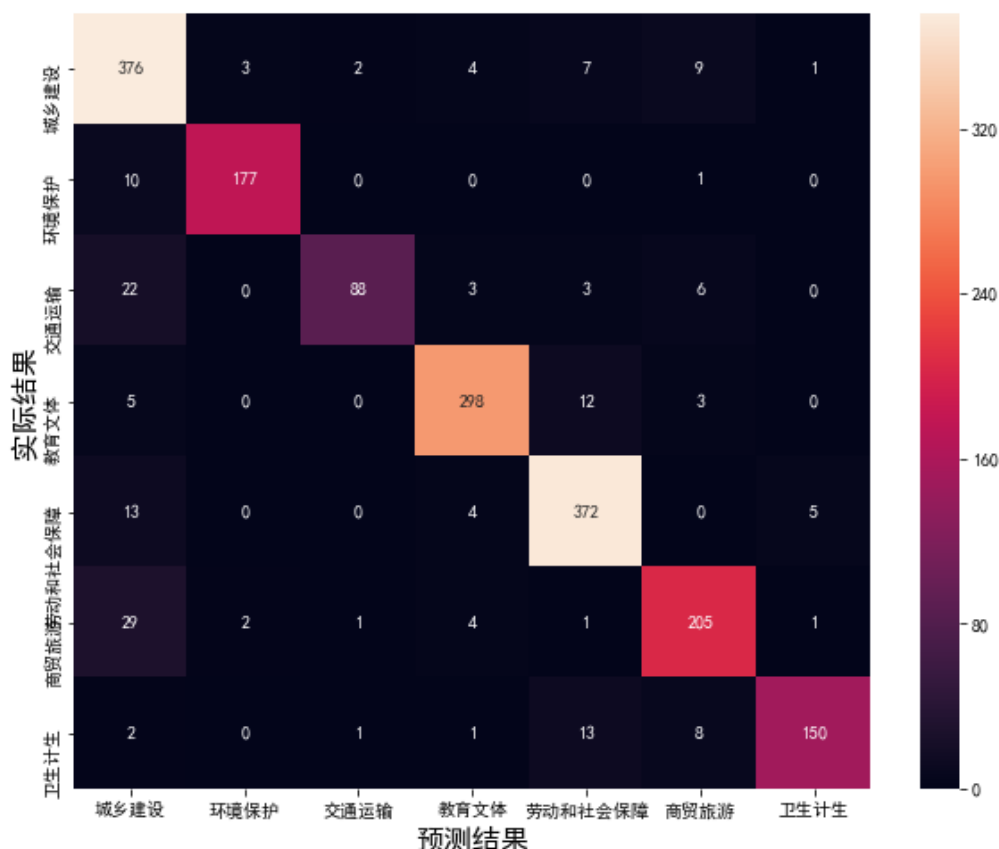


图 5-5-1 预测结果图

如图 5-5-1 所示混淆矩阵的主对角线表示预测正确的数量, 除主对角线外其余都是预测错误的数量。最终 2100 多条测试集数据的结果如图所示城乡建设和劳动社会保障的准确率相对较高, 教育文体和商贸旅游次之, 交通运输最低。因此需要对交通运输相关分词进行调整和优化。

5.6 模型评价

数据挖掘的目的是通过建立模型对网络问政平台上的群众留言进行划分分类标签, 以便于将留言分配给相应的部门进行处理。对留言内容划分一级标签, 这显然是一个分类预测的问题。所以本文选择了朴素贝叶斯和 LinearSVC 模型进行分析预判。基于对上述模型的求解, 对两个模型在该数据集的预测判断上的优劣进行比较, 最终选择出预测结果最佳的 LinearSVC 模型, 并建立混淆矩阵来预测标签和实际标签之间差异。

使用 F-Score 值对模型进行评价

$$F_1 = \frac{1}{n} \sum_{i=1}^n \frac{2P_i R_i}{P_i + R_i} \quad \dots\dots\dots (5-9)$$

其中 P_i 为第 i 类的查准率， R_i 为第 i 类的查全率。 F_1 值为算数平均数除以几何平均数，且 F_1 的数值越大，证明模型的准确率更高。

对模型进行评价后的结果如下：

accuracy 0.9023026315789474					
	precision	recall	f1-score	support	
城乡建设	0.82	0.95	0.88	663	
环境保护	0.96	0.93	0.95	310	
交通运输	0.95	0.70	0.80	202	
教育文体	0.94	0.94	0.94	525	
劳动和社会保障	0.91	0.94	0.93	650	
商贸旅游	0.90	0.83	0.86	401	
卫生计生	0.95	0.85	0.90	289	
accuracy			0.90	3040	
macro avg	0.92	0.88	0.89	3040	
weighted avg	0.91	0.90	0.90	3040	

图 5-6-1 模型评价结果图

可以看出使用 LinearSVC 模型准确率达到 90%。

如果想要模型运行效率高，而对准确率要求不高时，可以选择朴素贝叶斯。
如果对模型准确率要求高，则选择 LinearSVC 模型。

综上比较只是简单粗略的比较，没有对所有因素进行严格控制，并且模型参数较多，没有广泛进行调试，存在着许些不足，有待进一步优化。

6. 热点问题挖掘

将某一时段内群众集中反映的某一问题称作“热点问题”。对于政府部门来说，及时发现热点问题可以大力的节约其信息筛选的人员消耗，对于提高政府部门的工作效率有着极大的意义，并且对于提高政府部门的服务效率也有着显著提高。

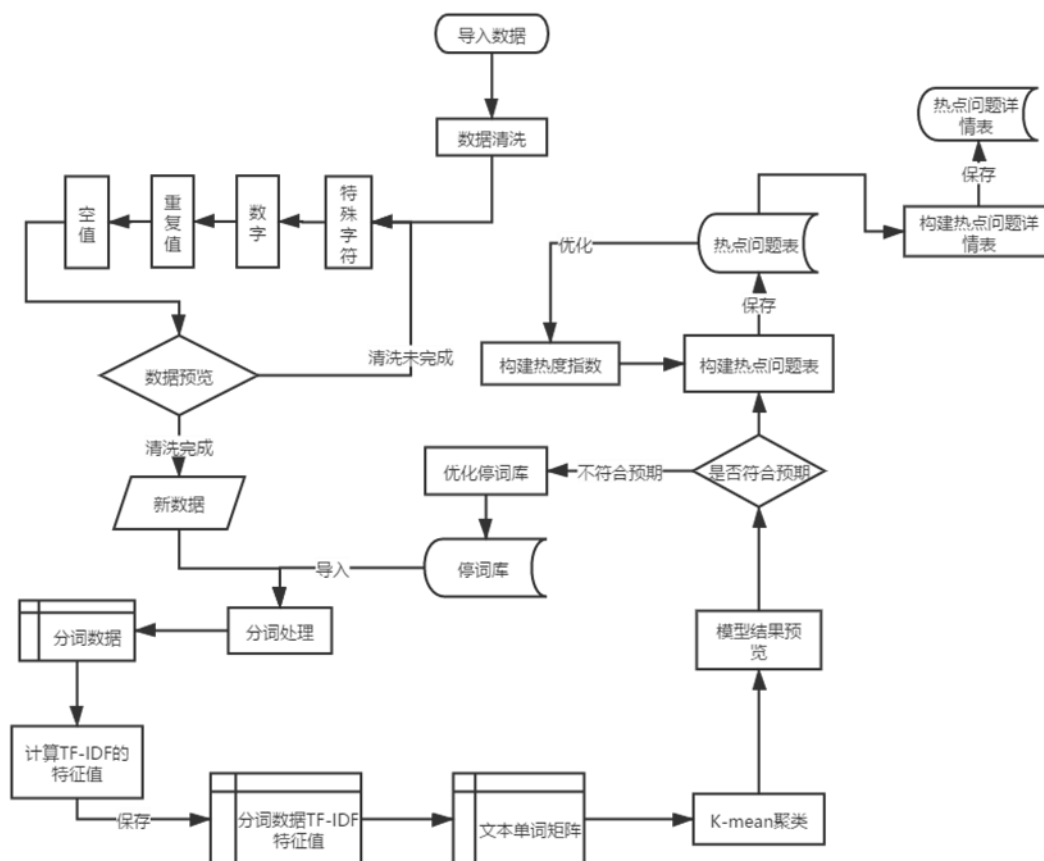


图 6-1 热点问题挖掘流程图

使用前面介绍到的 jieba 中文文本分词方法，对文本进行预处理之后进行词频矩阵的计算，基于词频矩阵计算每个词频向量的 TF-IDF 值，并且生成文本单词矩阵。构建 K-means 聚类模型并且对模型结果进行初步预览和停用库的优化，直到生成的问题簇内单个问题的内容相关度较高为止。同时构建合理的热度指数，并根据热度指数对问题簇进行评分，最后依据热度指数评分排序筛选出排名前五的热点问题表和索引出热点问题详情表。

6.1 指标的确立

政府及时发现热点问题有利于政府及时解决重点问题，获取了使用政府问政平台上的留言数据，得到下表所示的政府留言数据说明表 6-1：

表 6-1 政府留言数据说明表

变量名	详细说明	取值范围
留言编号	连续变量	188006-360114
留言用户	文本数据	/
留言主题	文本数据	/
留言时间	单位：年/月/日	2017/6/8-2020/1/16
留言详情	文本数据	/
反对数	定量变量	0-53
点赞数	定量变量	0-2097

6.2 文本聚类

对于获得的文本数据需要进行整理分析，这里选择使用文本聚类的方法，将文本按其含义聚集在一起，形成簇，一个簇就代表一个话题，所以一个簇内往往包含多个句子。聚类是一种无监督学习算法，可以直接根据数据进行特征处理。

6.2.1 文本特征处理

对于给定的文本数据进行数据清洗和预处理后，利用分词工具 jieba 进行分词，完成文本预处理。文本预处理之后会产生大量的特征词，如果直接使用预处理后的特征词进行挖掘，不仅会造成特征表示上的维度灾难，而且也得不到高质量的聚类结果。因此，需要进一步开展特征提取，从而为后续的挖掘以及最终的聚类带来更好的效果。本次研究使用的用词频-逆文档频率 TF-IDF 来计算文本数据中特征词的权值，按权值大小排序，提取出特征值。

6.2.2 K-means 聚类

K-means 聚类是如今一种非常常用的聚类算法，同时还是经典的划分聚类算法，该算法的优点是时间复杂度低，聚类效果较好。因此，利用 K-Means 算法对经过向量化处理的特征词进行聚类效果较于 LDA 更好，这符合本实验对于热度求解的要求，所以本实验选择 K-means 聚类的方法对文本进行相似度分析。

通常情况下进行 K-means 聚类算法的一般步骤如下：

- 1) 随机选择 n 个簇类中心点；
- 2) 遍历所有数据点，把数据点划分到距离最近的一个簇类中；
- 3) 划分之后就有 n 个簇，计算每个簇类中点的平均值作为新的簇类中心点；
- 4) 重复步骤 2) 和 3)，直到聚类中心不再发生变化，或是迭代的次数达到设定的值。

在本次文本分类的测试中通过一系列可观测的实验结果发现在目前已经测试的簇类个数中当 $n=1000$ 时效果最好，因此后续模型结论都建立在簇类 $n=1000$ 时的情况下。

6.3 定义热度指数

在分析过程中我们不难发现许多的问题反映内容一样但是反映的人和时间有差别。通常情况下我们将这种人群时间集中的问题定义为热点问题，同时在有限的服务资源的情况下优先安排解决热点问题，促使工作效率最大化。及时发现热点问题对于政府的治理有着非常重要的意义。但是选择筛选热点问题前我们必须构建一个衍生数据——热度指数，来帮助我们确定和选择合理的热点问题。这样可以全面地反映当前的民众热点问题。

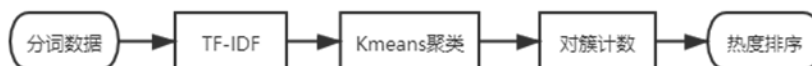


图 6-3 热度指数构建图

在对分词数据进行权值转换，即 TF-IDF 计算后，对文本向量进行 Kmeans 聚类形成一个一个的簇，聚类完成后显示其结果，并对每一个簇进行计数，并定义热度指数，由于部分留言点赞人数大大高于留言的个数，故对留言条数乘以一个

K 值，增大其在热度指数中的权重，最终得到的热度指数公式如下：

$$H = Q \times k + (C - Z) \quad \dots\dots\dots (6-1)$$

其中 H 表示热度指数， Q 表示每一个簇中评论的个数， C 表示每簇类留言中的点赞数之和， Z 表示每簇类留言中的反对数之和。对每一个簇的热度指数进行排序。在文本分析中通过一系列反复试验发现在已测试的 k 值 $k = 200$ 时效果最好，因此后续的结论均建立在该条件下。

6.4 热点问题表

根据定义的热度指数计算生成新列后按照热度指数进行排序最终提取出排名前五的热点问题的热度指数和聚类结果。通过聚类结果人工生成热点问题输出热点问题表如图：

表 6-4-1 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	10013	2019/1/9至2020/1/5	A市A区中海国际社区	噪音污染严重，影响居民休息
2	2	5264	2019/1/1至2019/12/31	A市购房业主	房地产行业口碑差，坑害消费者
3	3	4628	2019/1/2至2019/12/25	A市准业主	购房贷款公积金办理效率低
4	4	4615	2019/2/15至2019/12/26	A7县小区居民	小区麻将馆深夜严重扰民
5	5	4423	2019/2/1至2019/12/6	A市学校学生	学校管理不当，欺骗学生

最终得到前五的热点问题如表所示排名第一的问题是 A 市（区）噪音污染问题，该问题主要集中在 XXX 时间段。同时由热点问题生成的热点问题详情表如下图所示。

表 6-4-2 热点问题详情表

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
1	188059	A00028571	五期与四期中间	9/11/22 16:54	投诉业主，态度强	0	0
1	188399	A00097934	号公馆项目夜	19/7/3 6:23:2	日凌晨点还在施工	0	0
...
1	289832	A00096296	新姚路施工项	9/12/18 2:40	器测量，这个施工	0	0
2	191996	A000108760	区内打架斗殴	9/3/26 10:32	无门，开发商置	0	0
2	197206	A00059278	开裂、电表外	9/11/26 18:20	求提供采购清单、	0	5
...
2	287527	A00019121	场开发商欺骗	9/3/14 17:36	发酵。此后几经协	0	0
3	191777	A00097615	实际工资交五	19/11/7 0:57	是不肯交公积金。	0	0
3	202249	A00092267	房公积金无法	9/5/10 15:25	金商贷组合贷款，	0	0
...
3	280743	A000106808	九富达集团	19/9/8 11:57	代还买个车搞得	0	0
4	188451	A00013004	石塘铺村有党	9/4/11 17:54	关门了。但是石塘	0	2
4	191153	A909097	泉塘小区麻将	9/12/15 18:43	没有什么实质性的	0	0
...
4	288526	A00088728	A7县集镇麻将	9/5/13 11:28	社会风气，麻烦您	0	2
5	190522	A00056153	又有一所民办	9/10/9 16:27	的全日制公办高级	0	6
5	206039	A000113118	职业学院要求	9/12/6 13:47	个合理的说法，什	0	0
...
5	287492	A0003274	星沙实验小学	19/9/3 11:06	的车辆还要在本已	0	0

7. 答复意见评价

政府部门对留言进行整理后会由相关部门进行答复，而政府作为官方部门，其答复往往会作为政府工作能力的表现之一，同时也会作为评价政府的标准。

本研究将使用相关性、规范性和时效性三个角度并建立合适的评价指标对政府的答复意见做出评价。

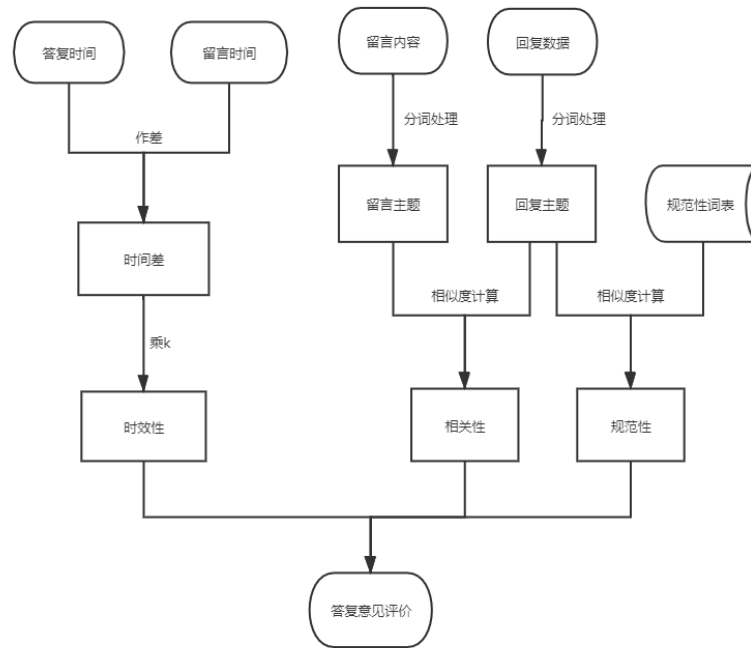


图 7-1 答复意见评价指标制作流程图

政府的留言回复也是用于监督政府工作的衡量指标，获取了使用政府问政平台上的留言数据，得到下表所示的政府留言及回复数据说明表 7-1：

表 7-1 政府留言及回复数据说明表

变量名	详细说明	取值范围
留言编号	连续变量	2549-185986
留言用户	文本数据	/
留言主题	文本数据	/
留言时间	单位：年/月/日	2011/10/3-2020/1/8
留言详情	文本数据	/
答复意见	文本数据	/
答复时间	单位：年/月/日	2011/11/14-2020/1/9

相关性：留言相关性便能看出政府工作人员在阅读留言的严谨性与工作的态度。本研究将留言详情与答复意见使用 Python 编程语言中的 jieba 进行分词，分词后进行文本相似度计算，得到的相似度数即为答复相关性值。

规范性：政府回复需要具有一些固定格式显示其规范性，由于数据中未给出标准格式，故而我们自己建立规范词表，其中包括您好、留言、收到、调查核实、收悉等词。建立规范词表后将其与每一条答复意见中的答复内容进行相似度计算。

由于规范词典的词数较少，而答复意见中文本较多，计算出的文本相似度即规范性数值较小，故我们对文本规范性数值乘以一个 k 值以进行后续计算。在此研究中取 k=10。

时效性：时效性是判断政府是否及时解决问题的一个重要指标。本研究使用答复时间-留言时间计算其时效性。由于计算结果的数值较大，故对其乘以一个 K 值，将其数值调小。其公式表示如下：

$$o=(T_1-T_2)*k \quad \dots\dots\dots (7-1)$$

其中 T_1 表示答复时间， T_2 表示留言时间。用文字表示其公式如下：

$$\text{时效性}=(\text{答复时间}-\text{留言时间})*k \quad \dots\dots\dots (7-2)$$

由于计算所得的时效性越高，时间间隔越小；时效性越低，其时间间隔越大，表示其时效性与时间间隔呈负相关。故 k 的取值为负数。且由于有大量数据时间间隔较大，故 $|k|<1$ ，在本研究中，取 $k=(-0.001)$ 。

答复意见质量计算：

使用均值表示答复意见质量，其公式表示如下：

$$N=\frac{\chi+K+O}{3} \quad \dots\dots\dots (7-3)$$

其中，N 表示答复意见质量， χ 表示答复意见与留言详情的相关性，K 表示答复意见的规范性，O 表示答复的时效性。用文字表示其公式如下：

$$\text{答复意见质量}=\frac{\text{相关性}+\text{规范性}+\text{时效性}}{3} \quad \dots\dots\dots (7-4)$$

使用该方法对每一条留言详情与答复意见进行遍历计算，最终得到对每一条答复意见的评分如图 7-2 所示：

	答复意见	评分
0	现将网友在平台《问政西地省》栏目向胡华衡书记留言反映“A2区景蓉花苑物业管理有问题”的调查核...	0.377980
1	网友“A00023583”: 您好! 针对您反映A3区潇楚南路洋湖段怎么还没修好的问题,A3区洋...	0.094674
2	市民同志: 你好! 您反映的“请加快提高民营幼儿园教师的待遇”的来信已收悉。现回复如下: 为了改善...	0.159287
3	网友“A000110735”: 您好! 您在平台《问政西地省》上的留言已收悉, 市住建局及时将您反...	0.250337
4	网友“A0009233”, 您好, 您的留言已收悉, 现将具体内容答复如下: 关于来信人建议“白竹坡...	0.290829
5	网友“A00077538”: 您好! 针对您反映A3区含浦镇马路卫生很差的问题,A3区学士街道、...	0.147322
6	网友“A000100804”: 您好! 针对您反映A3区教师村小区盼望早日安装电梯的问题,A3区...	0.167191
7	网友“UU00812”您好! 您的留言已收悉。现将有关情况回复如下: 一、关于小区附近幼儿园的问...	0.148391
8	网友“UU008792”您好! 您的留言已收悉。现将有关情况回复如下: 据查, 美麓阳光项目位于A...	0.150293
9	网友“UU008687”您好! 您的留言已收悉。现将有关情况回复如下: 您所反映的地点为洋湖新城...	0.228879
10	网友“UU0082204”您好! 您的留言已收悉。现将有关情况回复如下: 经查, 该处属原大托村四...	0.094821
11	网友“UU008829”您好! 您的留言已收悉。现将有关情况回复如下: 经A5区人防办、洞井街道...	0.233083
12	网友“UU00877”您好! 您的留言已收悉。现将有关情况回复如下: 经查, A4区政府已责成区市...	0.328690
13	网友“UU0081480”您好! 您的留言已收悉。现将有关情况回复如下: 经查, 您所反映的相关警...	0.320201
14	网友“UU0081227”您好! 您的留言已收悉。现将有关情况回复如下: 261路公交车全程24...	0.230894
15	网友“UU008444”您好! 您的留言已收悉。现将有关情况回复如下: 经查, 新开铺路(绕城高速...	0.197906

图 7-2 答复意见评分示意图

8. 模型评价

优点: 本文的模型均基于政府问政平台上的留言数据及回复数据进行求解, 因此模型具有很好的应用参考价值。并且本文引用了朴素贝叶斯、线性支持向量机模型与 K-means 聚类, 为文本分类与文本匹配建立了模型, 可以大力提高政府的工作效率。最后对政府回复从多角度进行了评价并给出评价指标。

缺点: 本文的模型在考虑影响因素方面尚有欠缺, 因此模型的效果会有所影响。除此之外, 模型要可行, 必须进行不断地试验修正, 而本文只用了一次的实验, 因此模型准确度会有所影响。

9. 参考文献

- [1] 张振华, 许柏鸣. 基于在线评论文本挖掘的商业竞争情报分析模型构建及应用[J]. 情报科学, 2019, 37(02):151-155+162.
- [2] 祝永志, 荆静. Chinese Word Segmentation Technology based on Python Language% 基于 Python 语言的中文分词技术的研究[J]. 通信技术, 2019, 052(007):1612-1619.
- [3] 钟熙, 孙祥娥. 基于 Kmeans++聚类的朴素贝叶斯集成方法研究[J]. 计算机科学, 2019(s1).
- [4] 杨立才, 李金亮, 姚玉翠, et al. 基于 F-score 特征选择和支持向量机的 P300 识别算法[J]. 生物医学工程学杂志, 2008, 025(001):23-26, 52.
- [5] 冀先朋. 多标签文本分类算法的研究与应用[D]. 2019.
- [6] 陈慧, 田大钢, 冯成刚. 多种算法对不同中文文本分类效果比较研究[J]. 软件导刊, 2019, 18(05):79-84.
- [7] 顾穗珊, 孙山山. Research on the Supply of Competitive Intelligence Services for Small and Medium-sized Enterprises Dominated by the Smart Government in Big Data Era%大数据时代智慧政府主导的中小企业竞争情报服务供给研究[J]. 图书情报工作, 2014, 058(005):64-68.
- [8] 张锦, 李光, 曹伍, 等. 基于主成分分析的自动文本分类模型[J]. 北京邮电大学学报, 2006, 029(0z2):136-138, 143.
- [9] 薛彬, 陶海军, 王加强. 针对民生热线文本的热点挖掘系统设计[J]. 中国计量大学学报, 2017(3).
- [10] 段尧清, 何思奇, 林平. 基于新闻文本挖掘的政府态度识别实证研究[J]. 情报理论与实践, 2019, 42(9).