“智慧政务”中的文本挖掘应用

摘要

本文通过对群众留言数据进行分类、对热点问题进行数据挖掘，并给出一套评价方案。为智慧政务平台处理留言信息提供科学的依据。

针对问题 1 ：通过 python 的数据清洗操作对留言详情进行去重，得到不同留言详情内容。利用 jieba 中文分词工具对留言详情内容进行分词，并去除停用词，再通过 TF-IDF 算法利用 TF-IDF 算法把留言详情内容转换为权重向量，得到每个留言详情的 TF-IDF 权重向量，采用分类模型对训练集的文本进行学习，然后使用测试集文本去测试分类效果，最后评价分类模型的性能。

针对问题 2：对附件 3 的数据进行数据清洗、数据处理和数据分析，定义热度评价指标，再根据 TF-IDF 权值向量特征构造聚类器，利用 K-means 算法实现文本聚类，对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，并从中得出排名前五的热点问题，以及对应的热点问题留言明细。

针对问题 3：利用 1-9 标度法构建答复意见质量评价指标，利用 matlab 工具计算权重，得出对答复留言的质量评价方案。

关键词：文本分类 数据挖掘 热点问题 模型评价 机器学习 层次分析法

目录

一、挖掘目标

本次数据挖掘是通过研究群众留言，利用 python 等软件以及利用层次分析法达到以下三个目标：

1、对群众留言进行文本分类，构建一级分类标签模型，对该模型进行评价，得出分类的情况，并利用 python 画出相应的图形。

2、利用 K-means 算法对留言进行聚类，定义相应的热度指标，得出排名前五的热点问题表以及相对应的留言明细表。

3、采用层次分析法对评价指标进行分析，得出相应的留言质量评价方案。

二、分析方法与过程

2.1 问题 1 分析方法与过程

2.1.1 问题 1 流程图

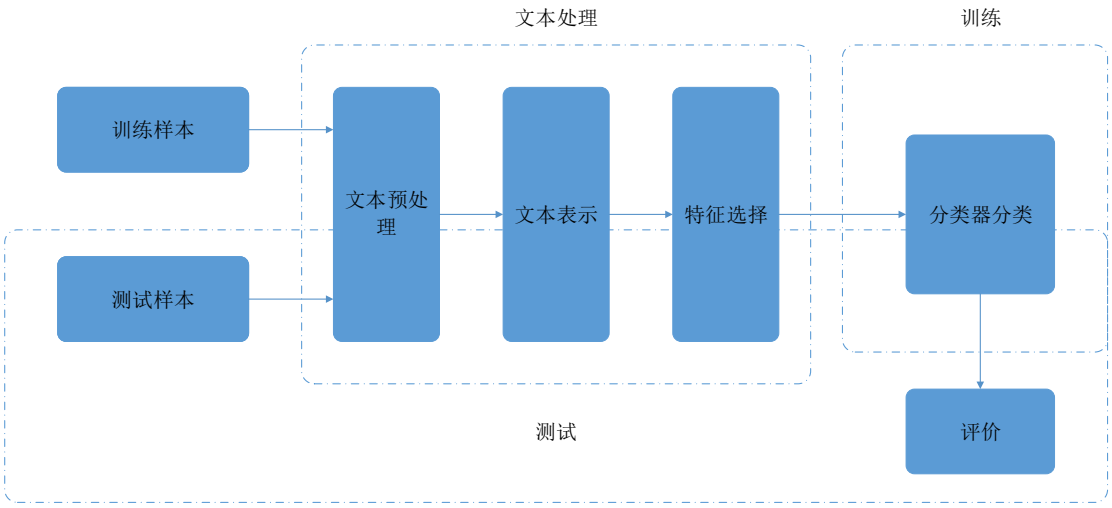

图 2-1 问题 1 流程图

文本分类数据一般由训练集和测试集两部分组成，如图 2-1 所示。两个集合的数据都需要进行文本处理中所包含的步骤。最后我们采用机器学习中的分类器通过训练样本学习，完成之后对测试样本输入分类器中进行分类预测，最后我们对分类的效果进行评价。

2.1.2 文本预处理

由于中文文本词与词之间没有间隔，因此需要使用分词手段将词区分出来，其中区分出来的词中有些对分写没有帮助的信息也要去除，也就是出停用词，最终把能够表达文本关键信息的关键词从文本中提取出来，然后文本表示为这些关键词的集合。因此，在对数据进行预处理的过程涉及到数据清洗、文本分词、以及去除停用词等。

1、留言详情的去重

在题目给出的数据中，有出现留言详情内容相同的数据。为了确保在分类模型中准确度的可靠性和稳定性，我们因此将留言详情内容相同的数据进行去重，得到最近的数据进行接下来的操作。

2、对留言详情进行中文分词

分词是中文文本预处理中不可缺少的步骤，因为中文文档不同于英文，在分类操作中使用单词表示文本时必须先进行分词处理。目前的分词技术已经逐步完善，本文将选择 python 中的 jieba 分词工具进行分词操作。

3、去停用词

对于文本分类而言，有时候并不能以某些词在文本中出现的频率高低来判断该词在文本中的重要程度。比如“一二三四”、“你我他”、“这个”、“的”，这些没有特殊语义并且出现频繁词语，也就所谓的停用词，我们应该将这些停用词从文本中清楚掉，避免它们对后续分类产生干扰。

2.1.3 特征选择指标

在对留言详情内容进行预处理值后，需要把这些词语转换为向量，以供数据

的挖掘分析使用。本文将采用 TF-IDF 算法，把留言详情内容转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重（Term Frequency）。

$$\text{词频 TF} = \frac{\text{词在文章中出现的次数}}{\text{文章总的词汇数}} \qquad (1)$$

第二步，计算 IDF 权重，即逆文档频率（Inverse Document Frequency），需要建立一个语料库（corpus），用来模拟语言的使用环境。IDF 越大，此特征性在文本中的分布越集中，说明该分词在区分该文本内容属性能力越强。

$$\text{逆文档频率 IDF} = \log\frac{\text{语料库的总文档数}}{\text{包含该词条的文档数+1}} \qquad (2)$$

第三步，计算 TF-IDF 值（Term Frequency Document Frequency）。

$$\text{TF - IDF = 词频（TF）} \times \text{逆文档频率（IDF）} \qquad (3)$$

实际分析得出 TF-IDF 值与一个词在留言详情内容中文本出现的次数成正比，某个词文本的重要性越高，TF-IDF 值越大。

2.1.4 文本分类

经过文本的预处理阶段后我们得到一个用于文本分类的特征集合，生成留言内容的 TF-IDF 权重向量后，根据每个留言的 TF-IDF 权重向量，选择一个合适的文本分类器实现文本的分类。本文首先采用逻辑回归(Logistic Regression)、多项式朴素贝叶斯（Multinomial）、线性支持向量机（Linear Support Vector Machine）和随机森林（Random Forest）四种机器学习模型对留言进行分类，并估算他们的准确值，然后选取准确值最高的模型对训练集进行文本分类，然后使用测试集文本去测试分类模型的性能。

通过对四种机器学习模型都进行 5 次的留言分类，得到每种模型 5 次对应的准确率如表 2-2 所示，并绘制如图 2-3 箱型图；

	model_name	fold_idx	accuracy
0	RandomForestClassifier	0	0.400772
1	RandomForestClassifier	1	0.412252
2	RandomForestClassifier	2	0.395580
3	RandomForestClassifier	3	0.445796
4	RandomForestClassifier	4	0.399336
5	LinearSVC	0	0.847850
6	LinearSVC	1	0.874172
7	LinearSVC	2	0.861878
8	LinearSVC	3	0.900996
9	LinearSVC	4	0.867257
10	MultinomialNB	0	0.630099
11	MultinomialNB	1	0.653974
12	MultinomialNB	2	0.644199
13	MultinomialNB	3	0.647124
14	MultinomialNB	4	0.637721
15	LogisticRegression	0	0.772326
16	LogisticRegression	1	0.810155
17	LogisticRegression	2	0.808287
18	LogisticRegression	3	0.839602
19	LogisticRegression	4	0.785951

图 2-2 四种分类模型分类准确率

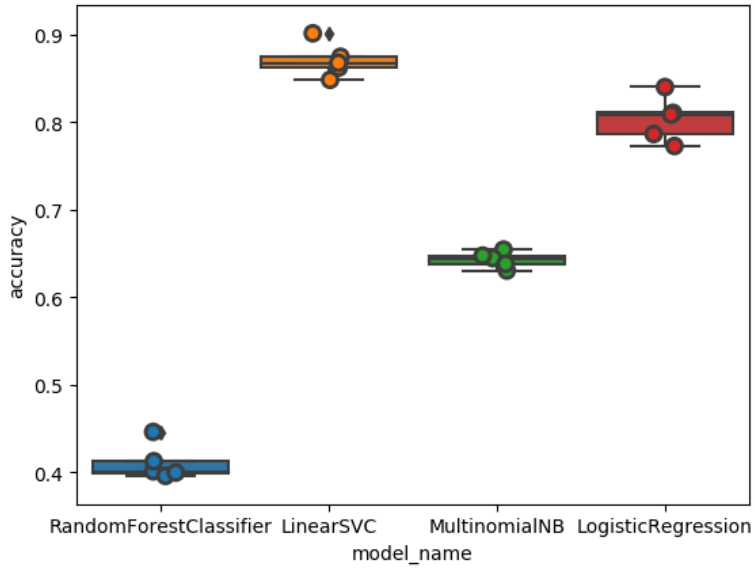

图 2-3 分类模型准确率

从可以箱体图上可以看出随机森林分类器的准确率是最低的，因为随机森林属于集成分类器(有若干个子分类器组合而成)，一般来说集成分类器不适合处理高维数据(如文本数据)，因为文本数据有太多的特征值，使得集成分类器难以应付，另外三个分类器的平均准确率都在 60%以上。其中线性支持向量机的准确率最高。

通过计算，得到四种模型的平均准确率，如表 2-1 所示：

表 2-1 分析模型平均准确率表

model_name	accuracy
LinearSVC	0.870431
LogisticRegression	0.803264
MultinomialNB	0.642623
RandomForestClassifier	0.410747

通过上述箱型图以及分类模型平均准确率表可以看出线性支持向量机的平均准确率达到了 87%。因此，针对问题 1 的分类模型的选取我们采用线性支持向量机。

2.2 问题 2 的分析方法与过程

2.2.1 问题 2 流程图

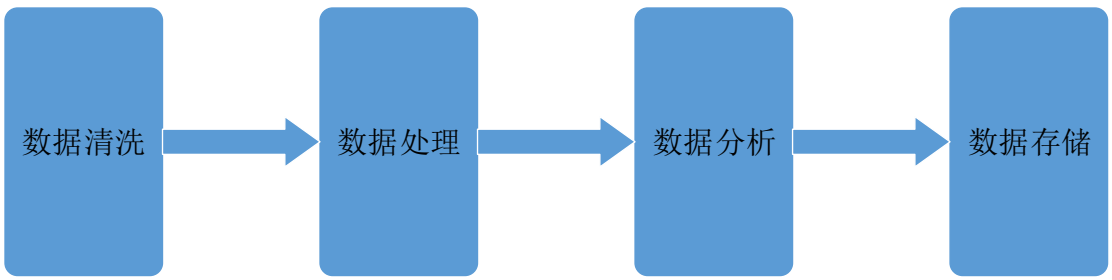

图 2-4 问题 2 流程图

2.2.2 K-means 文本聚类

在进行数据分析阶段前，同样需要对附件 3 所给定的留言信息进行数据的清洗工作，保证数据的合理性及后期操作的可靠性，还需运用 jieba 分词工具对留

言详情进行分析，去除停用词。将数据清洗和数据处理工作完成后，我们接下来进行数据的分析过程，该过程分为两个步骤：

步骤一：计算特征值。我们首先将经过数据处理获得的特征词集合转换成词频向量,然后将词频向量再转换成 TF-IDF 向量。

步骤二：对热点问题进行分类。针对该问题的求解我们通过 K-means 文本聚类模型来实现热点问题的挖掘。

1、特征降维

在使用向量空间模拟表示文本的时候,集合中的每个特征词都对应这向量空间的一个向量。通常情况下，特征词集合中的特征词数量十分庞大，这会导致构成的向量空间的维度相当高，会产生分类器过拟合的问题，因此，针对该问题，我们运用特征选择的特征降维方法对特征词向量进行降维操作。

2、构造聚类模型

Kmeans 算法是属于最为典型的分割式分群算法,而主要应用是要在大量高维度的数据点,找出最具代表性的数据点。这些所谓的数据点可以称为群聚中心,并根据这些群聚中心来进行资料分类,用少数的代表点来代表特定的类别资料,可以大幅度降低系统的计算量和数据量。此分群算法可以将数据依据分群的目标分成许多群聚,而被划分为同一群资料会有某些特性极为相似,相反的被分为不同群的数据就会有某些特性会有明显的不同。此方法预先预定群聚的数目,经由反复迭代运算,逐次降低每一个目标函数的误差值,直到目标函数不再产生变化,就能产生分群的效果。

在使用分割式分群演算时,是希望使每一个群聚中心,每一个数据点,与群聚中心的距离都能有最小的平方误差。如果我们默认资料内含有 h 个群聚中心,其中第 k 个群聚可以用集合G_k来表示,假设G_k含有N_k个样本$\{X_{1k}, X_{2k}, X_{3k} \ldots X_{nk}\}$,此群聚的平方误差$e_k$则可定义为：

$$e_k = \sum_i |x_{ik} - y_i|^2 \qquad (4)$$

其中x_{ik}属于群集G_k。

则 h 个群聚数的总和平方误差 E 即是每个群聚旷方的总和,称为分群的「误差函数」:

$$E = \sum_{k=1}^{k} e_k^2 \qquad (5)$$

Kmeans 就是一种可以找出 h 以及相关群聚中心使得 E 值为最小的算法。

3、Kmeans 算法的流程

步骤 1:设定分群群聚数目为 h,并选取 h 个样本作为 h 个群聚的群聚中心。

步骤 2:输入全部的样本,计算每一笔样本到各个默认群集中心之间的距离,然后再比较该笔的样本距离哪一个群集中心最为接近,这笔数据记录就会被指派到最为接近的那个群集中心,此时就会产生了初始群集的成员集合

步骤 3:再根据群内的每一个样本重新计算出该群集的质量中心,利用新的质量中心来做为该群新的群集中心。指定完新的群中心之后,再一次比较每一笔样本与新的群集中心之间的距离,然后根据距离,再度重新分配每一个资料所属的群集。

步骤 4:持续反复步骤 3,一直执行到群集成员不再变动为止。

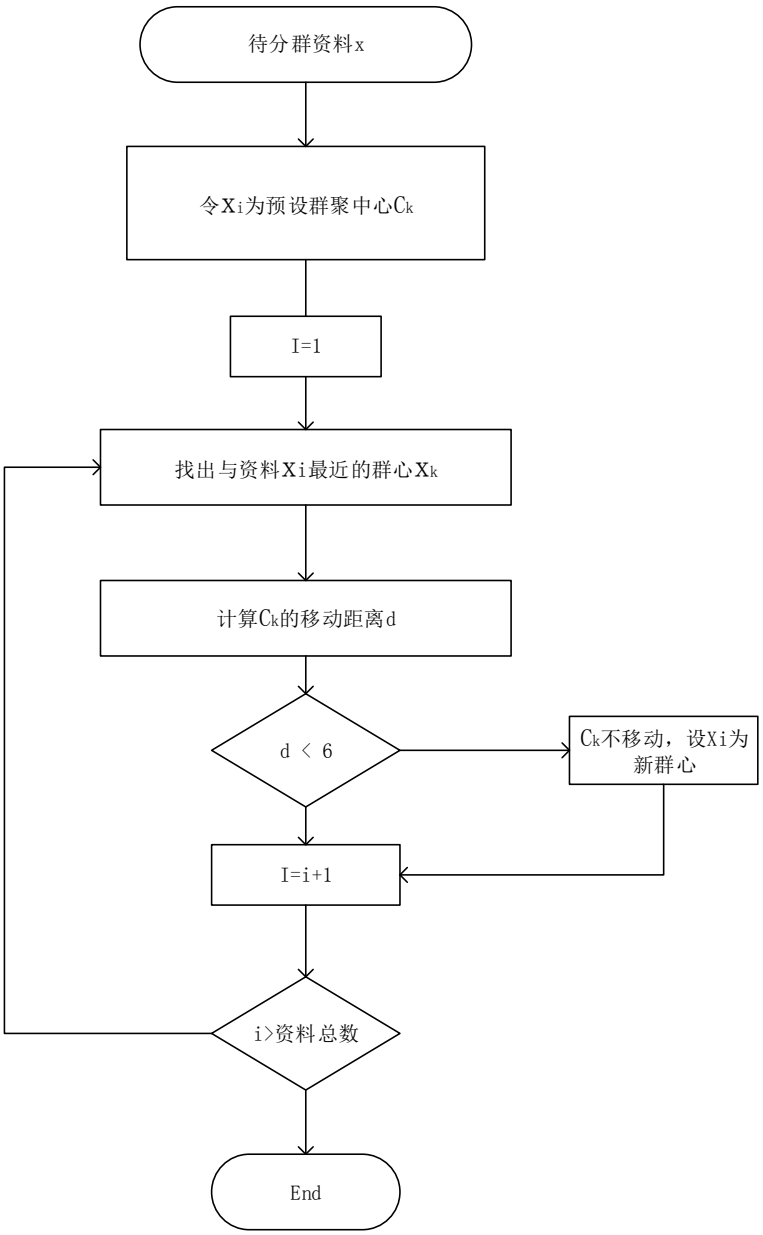

图 2-5　Kmeans 主题聚类模型构建流程图

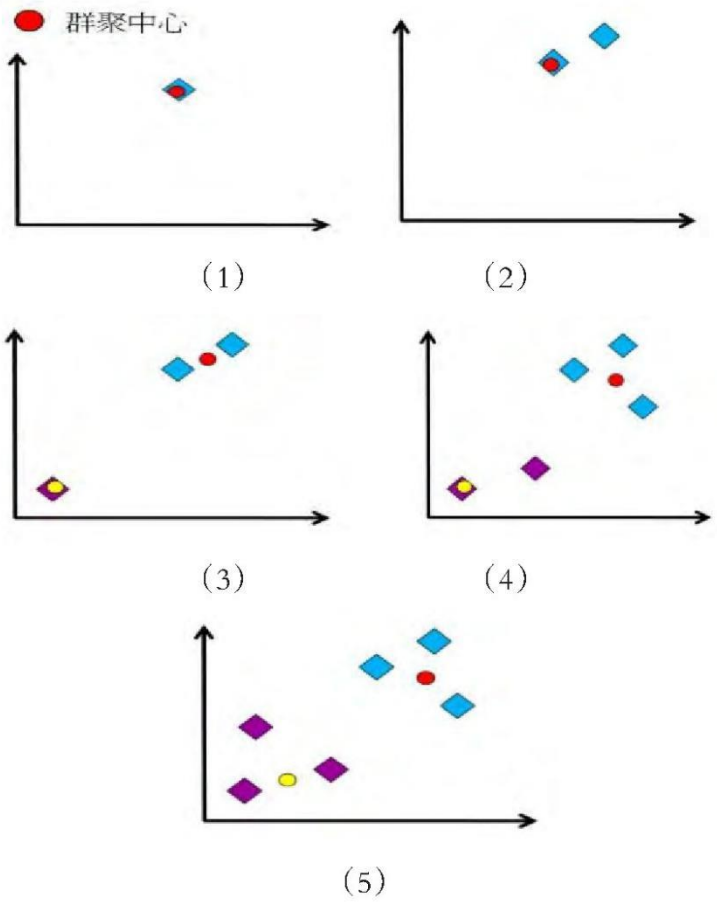

图 2-6 聚类主题模型过程图

由于附件 3 给出了条 4326 记录，无重复记录，把所有的留言详情都用来挖掘分析，将留言详情进行分词、求 TF-IDF 向量，特征降维、并利用 K-mean 聚类，将文本记录分类，程序见附件 km_model.py，K-mean 聚类出来的相关热点问题的留言次数作为热热度评价指标，以得出来的排名前 5 的热点问题，保存在附件“热点问题表.xls”，同时给出相应热点问题对应的留言信息，并保存在附件“热点问题留言明细表.xls”。

2.3 问题 3 的分析方法与过程

本文通过使用层次分析法 AHP 来判断矩阵并求权重：

AHP 的特点是在对复杂决策问题的本质、影响因素、内在关系等进行深入分析的基础上，利用较少的定量信息使决策的思维过程数学化，从而为多目标、多

准则或无结构特性的复杂问题提供简便的决策方法，是对难于完全定量的复杂系统作出决策的模型和方法。这是一种实用的多准则决策方法，能够统一处理决策中的定性和定量因素，具有高度的逻辑性、系统性 、简洁性和实用性等优点。

2.3.1 问题 3 流程图

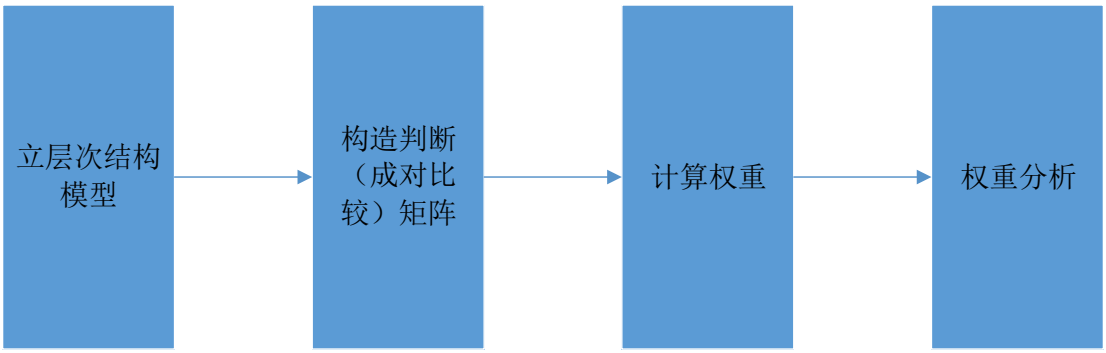

图 2-7 问题 3 步骤流程图

2.3.2 答复意见质量评价指标建立

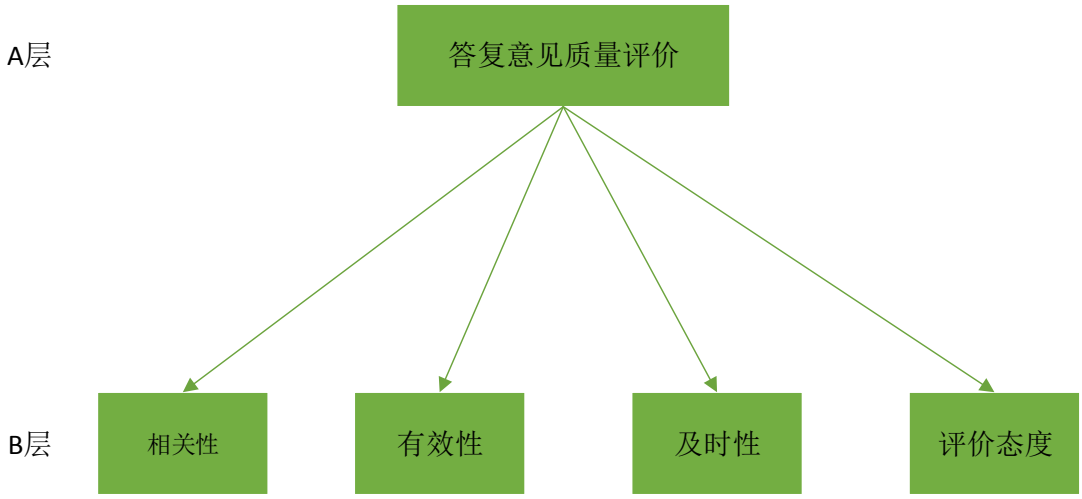

图 2-8 答复意见质量评价指标

将问题分解成两个层次，最上面成为目标层，及对答复意见质量的评价，第二层为准则层，有相关性，有效性，及时性，评价态度四个准则，两层之间的联系用直线表示，如图 2-8 所示。

2.3.3 构造判断（成对比较）矩阵

表 2-2　1-9 标度法

标度	含义
1	表示两个因素相比，具有同等重要性
3	表示两个因素相比，前者比后者稍重要
5	表示两个因素相比，前者比后者明显重要
7	表示两个因素相比，前者比后者强烈重要
9	表示两个因素相比，前者比后者极端重要
2,4,6,8	表示上述判断的中间值
倒数	若因素 i 与因素 j 的重要性之比为a_{ij}，则因素 j 与因素 i 的重要性之比为a_{ij}/a_{ji}

利用 1-9 标度法，通过对评价指标的重要程度相互比较，构造出相应的比较矩阵，得到答复评价质量指标的判断矩阵为：

$$A=\begin{pmatrix} 1 & 1/2 & 5 & 1/7 \\ 2 & 1 & 7 & 1/2 \\ 1/5 & 1/7 & 1 & 1/7 \\ 7 & 2 & 7 & 1 \end{pmatrix} \qquad (6)$$

2.3.4 一致性检验

对构造的判断矩阵进行一致性检验，用来确定权重分配是否合理，计算一致性比例 CR，并对值进行判断：

$$\mathrm{CR}=\frac{CI}{RI} \qquad (7)$$

$$\mathrm{CI}=\frac{\lambda_{\max}-\mathrm{n}}{\mathrm{n}-1} \qquad (8)$$

表 2-3 平均随机一致性指标 RI 值

阶数	1	2	3	4	5	6	7	8	9
RI	0	0	0.58	0.9	1.12	1.24	1.32	1.41	1.45

当 CR＜0.10 时，认为判断矩阵的一致性是可以接受的，否则应对判断矩阵作适当修正。

由公式可算出：

$$\lambda_{max} = 4.22$$

可得 CI=0.073、CR=0.08.。一致性比例小于 0.1，因此权重值合理。

三、问题分析

3.1 问题 1 结果分析

针对问题 1 我们采用平均准确率较高的线性支持向量机对留言进行分类，我们将所给定的留言详情划分训练集和测试集，其中测试集占总留言详情的 33%。因此，我们首先采用线性支持向量机对样本集学习，完成之后对测试集输入模型中进行分类预测，通过查看混淆矩阵的方式，来显示预测标签和实际标签之间的差异。

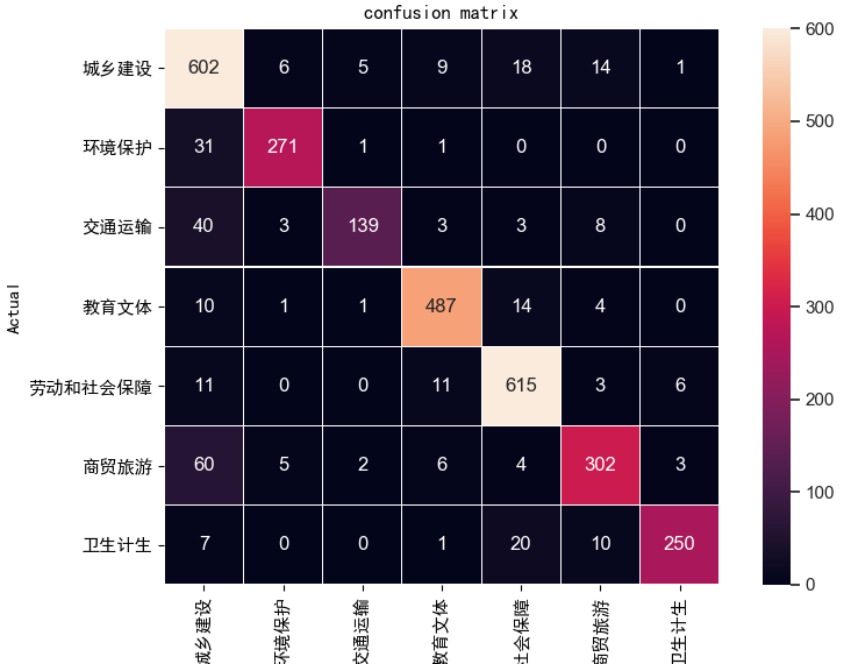

图 3-1 分类预测混淆矩阵

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.“城乡建设”和“劳动和社会保障”预测的错误数量较多。

由于问题 1 所涉及到的是多分类的情况，而多分类模型一般不使用准确率(accuracy)来评估模型的质量，因为 accuracy 不能反应出每一个分类的准确性，

因为当训练数据不平衡(有的类数据很多,有的类数据很少)时，accuracy 不能反映出模型的实际预测精度，这时候我们就需要借助于 F1 分数、ROC 等指标来评估模型。本文我们选取 F-Score 对分类方法进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_i R_i}{P_i + R_i} \qquad (9)$$

其中 Pi 为第 i 类的查准率，Ri 为第 i 类的查全率。

下面我们将查看各个类的 F1 分数，如表 3-1 所示：

表 3-1 各个类的指标值

	precision	recall	f-score	support
城乡建设	0.79	0.92	0.85	655
环境保护	0.95	0.89	0.92	304
交通运输	0.94	0.71	0.81	196
教育文体	0.94	0.94	0.94	517
劳动和社会保障	0.91	0.95	0.93	646
商贸旅游	0.89	0.79	0.94	382
卫生计生	0.96	0.87	0.91	288

从以上表中 F1 分数上看，"教育文体"类和“贸易旅游”类的 F1 分数最大，“交通运输”类 F1 分数最差只有 81%，究其原因可能是因为“交通运输”分类的训练数据和测试数据最少，使得模型学习的不够充分,导致预测失误较多。

通过上述表的各个类的 F1 分数，我们可以计算出平均 F1 值为 0.89。因此，说明我们选取的分类模型具有一定的可取性。

3.2 问题 2 结果分析

表 3-2 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	30	2019-01-09 至 2019-12-24	A 市居民	A 市有些地区施工产生噪音，严重扰民，影响居民休息
2	2	26	2019-01-06 至	A 市物业公司	A 市物流公司存在私吞公共收入，乱收取公共停车费用等问题，通过危害社会秩序和侵犯他人权益来谋求自己

			2019-12-30		的利益
3	3	23	2019-01-05 至 2019-12-16	A 市房产业主	A 市房产业主存在长时间拖欠居民房产证，严重侵犯住户的合法权益
4	4	21	2019-01-02 至 2019-12-25	A 市买房公民	A 市存在公积金相关问题，买房公民迫切了解公积金政策以此解答自己的疑惑，进而在这方面维护自己的合法权益
5	5	20	2019-03-05 至 2019-12-30	A 市某些区楼层电梯	A 市有些区存在电梯安全隐患以及信号覆盖问题，严重影响人们的生命安全

表 3-3 部分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	188059	A00028571	A 市 A3 区中海国际社区三期与四期中间空地夜间施工噪音扰民	2019/11/22 16:54:42	A 市 A3 区中海国际社区三期四期中间，即蓝天璞和洲幼儿园旁边那块空地一直处于三不管状态…	0	0
1	188399	A00097934	A 市利保壹号公馆项目夜间噪声扰民	2019/7/3 6:23:25	您好，我想举报 A 市利保壹号公馆项目 夜间噪声扰民 A 市利保壹号公馆项目：位于 A 市 A3 区咸嘉湖路熊家湾巷 1 号…	0	0
1	193665	A000100428	A 市人民西路违建严重影响后面居民夜间休息	2019/3/28 23:40:37	该房屋原是一层平房，2018 年年底开始施工，夜间偷偷动用挖掘机拆除原有房屋，严重影响后面居民夜间休息…	0	0

1	197319	A00033665	A 市电力局强占小区公摊范围强制施工，影响人身安全	2019/4/22 11:04:29	尊敬的胡书记：您好！我是 A 市 A4 区陡岭路佳阳悦景馨都的业主，向您反映一个情况，寻求您的帮助。事情发生在 2019 年 4 月 22 日…	2	0
1	198779	A00058878	A 市五一路袁家岭佳兆业广场施工跟放炮一样	2019/6/19 1:01:15	五一路袁家岭佳兆业广场施工，最近一个多月来，从早到晚，不晓得做什么施工，跟放炮一样…	0	0
1	205478	A909124	A5 区万科金域华府别墅区夜间 10-凌晨 4 点施工扰民	2019/12/20 1:10:42	作为一个上班族，每天白天很忙，甚至于晚上还要加班，尤其做电商的，晚上休息时间本来就不够，近日来每天晚上都在不停的施工扰民…	0	0
1	209549	A0007057	A4 区天健盛世 A1 区工地长期深夜施工扰民	2019/7/15 19:01:20	位于 A4 区新河街道的天健盛世 A1 区工地，长期深夜或通宵施工扰民。多次投诉，毫无结果…	0	0
1	210725	A00013709	A 市地铁窑岭站半夜施工持续扰民	2019/11/21 1:24:26	自 2019 年 7 月住过来至今(2019 年 11 月 21 日)，A 市地铁窑岭站半夜施工持续扰民…	0	0
1	211060	A00018706	反映 A1 区火炬路电缆铺设影响交通的问题	2019/10/12 21:44:42	火炬路 A1 区苑南门往西，右侧道路铺设电缆，从今年 6 月下旬开始施工，开挖、铺管…	1	0

从上述表、表中可以看出，当前人民群众关心的热点难点问题众多，比如：生命安全问题、买房住房问题、日常生活保障问题。这些都是我国发展所带来的这样那样的矛盾和问题。这些发展中遇到的与人民群众利益息息相关的民生问题，如果解决不好，就极有可能酿成大事，影响社会稳定。因此，高度关注民生着力解决好人民群众关心放映的热点难点问题，是对各级部门的一项重大考验。

3.3 问题 3 结果分析

针对问题三利用 Matlab 工具进行权重计算，可得答复评价质量权重如下表所示：

表 3-4 答复意见评价质量权重表

	相关性	及时性	答复态度	有效性	权值
相关性	1	1/2	5	1/7	14.13%
及时性	2	1	7	1/2	27.51%
答复态度	1/5	1/7	1	1/7	4.72%
有效性	7	2	7	1	53.63%

由上表体现出来的权重大小可以看出，在答复意见评价质量方面，有效性，及时性这两个指标对答复质量的影响程度较大。因此政府在答复留言是应着重考虑这两方面的因素进行回答。

四、总结

文本分类技术作为数据挖掘的基础技术，广泛地被应用于自然语言处理和文本挖掘等领域。群众问政留言记录数据获取成本低且数据量大，通过对留言信息的分类，热点问题等操作，有效地帮助和推动政府在互联网、大数据、信息化的时代管理水平和施政效率的提高。随着深度学习的发展，打破了文本信息处理技术的瓶颈，使文本分类成为新的研究热点。本文就是将深度学习模型和深度学习方法引入文本分类中，主要研究内容包括

（1）概述了文本分类的基本步骤，详细介绍了应用在文本分类过程中所实施的具体操作，其中包括，数据清洗、文本分词、去除停用词、生成特征向量和分类模型的学习和测试。

(2)介绍和采用了集中流行的深度学习模型，例如线性支持向量机、K-means

算法，并且将几种算法进行比较，突出各自的优势。

但由于所给的数据量比较偏少的原因，我们在训练和测试的过程中产生的结果有可能会出现个别类别相对于其他类别来说，准确率比较低，或者说所产生的误差有可能比较大，因此在以后的测试中，需满足足够的数据量，使得训练和测试过程和实验结果更加优化。同时，针对留言信息的应用场景，其具有时效性，应当对深度学习模型进行更进一步的优化，使其更有利于实际的应用。

五、参考文献

[1]李曼.自然语言处理在网站分类中的应用[J].电信网技术,2018(05):81-84.
[2]何伟. 基于朴素贝叶斯的文本分类算法研究[D].南京邮电大学,2018.
[3]朱文峰. 基于支持向量机与神经网络的文本分类算法研究[D].南京邮电大学,2019.
[4]董杰盛.自然语言处理在新闻分类中的应用[J].科学咨询(教育科研),2019(11):12-14.
[5]张迪. 基于深度学习的中文文本分类算法研究[D].西安科技大学,2019.
[6]任世超. 基于机器学习的文本分类算法研究[D].成都信息工程大学,2019.
[7]冀先朋. 多标签文本分类算法的研究与应用[D].山东大学,2019.
[8]张燕,杜红乐.基于层次分析法的乡村旅游影响因素研究[J].微型电脑应用,2020,36(04):54-56.
[9]续拓. 基于聚类算法的深度学习训练改进研究[D].西安电子科技大学,2018.
[10]马婵媛. 基于 K-Means 的分布式文本聚类系统的设计与实现[D].西安电子科技大学,2018.
[11]刘江华.一种基于 kmeans 聚类算法和 LDA 主题模型的文本检索方法及有效性验证[J].情报科学,2017,35(02):16-21+26.