“智慧政务”中的文本挖掘应用

摘要

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。因此，利用自然语言处理和文本挖掘技术对群众问政留言记录的研究具有重大意义。

对于问题 1，分别通过传统机器学习分类模型和谷歌开源的 Bert 模型进行群众留言分类。在传统学习模型中，把一级标签作为划分标签，利用 jieba 分词工具和哈工大停用词词库对留言详情进行数据预处理， 把再通过 TF-IDF 技术继续把留言详情表达成 TF-IDF 权重向量，分别来训练朴素贝叶斯分类器，SVM 分类器，逻辑回归分类器，随机森林分类器四种分类器，进行留言分类；谷歌开源 Bert 预训练语言模型具有强大的表义能力，使用该模型对预处理后的留言详情进行句子层面的特征表示，把一级标签作为划分标签，最后将获得的特征向量接入 softmax 分类器进行分类。

对于问题 2，先对留言详情命名实体识别，找出留言详情的实体类、时间类和数字类。又进行如同问题 1 的去停用词和 jieba 分词，并进行 TF-IDF 权重向量表示。把不同留言利用余弦相似度，分析其相似度。再利用 K-means 方法，把 TF-IDF 向量特征的留言进行聚类，并建立热度指标，结合其相似度，找出热度前五的留言。

关键词:传统机器学习模型　Bert 模型　命名实体识别　相似度分析　K-means 聚类　热度指标

Text mining application in "smart government"

Abstract

In recent years, with wechat, microblog, mayor's mailbox, sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand public opinion, gather people's wisdom and gather people's spirit. Therefore, it is of great significance to use natural language processing and text mining technology to study the records of political messages.

Aiming at the problem of the first , the traditional machine learning classification model and Google's open-source Bert model are used to classify the public comments. In the traditional learning model, the first level tag is used as the partition tag, and the message details are preprocessed by using the jieba word segmentation tool and the Harbin Institute of technology's stop words thesaurus, We will continue to express the message details into TF-IDF weight vector through TF-IDF technology, respectively to train naive Bayes classifier, SVM classifier, logical regression classifier and random forest classifier to classify the message; Google open source Bert pre training language model has strong semantic ability, using the model to carry out sentence level analysis on the message details after pre-processing In feature representation, the first level tag is used as the partition tag, and finally the feature vector obtained is connected to the softmax classifier for classification.

Aiming at the problem of the second, first name the entity recognition of message details, and find out the entity class, time class and number class of message details. We also use the de stop words and jieba segmentation as in question 1, and use TF IDF weight vector to represent them. We use cosine similarity to analyze the similarity of different messages. Then, k-means method is used to cluster messages with TF-IDF vector features, and the heat index is established. Combined with its similarity, the top five heat messages are found.

Keywords: Traditional machine learning model Bert model Named Entity Recognition Similarity analysis K-means clustering Heat index

目录

1.挖掘目标.....4
2 问题 1 分析方法与过程.....4
2.1 流程图.....4
2.2 传统机器学习模型.....5
2.2.1 数据分析.....5
2.2.2 去除停用词.....6
2.2.3 jieba 分词.....6
2.2.4 特征处理.....7
2.2.5 四种分类器原理.....8
2.2.6 构建分类器.....10
2.3 Bert 模型.....11
2.3.1 Bert 模型简介.....11
2.3.2 Bert 模型使用.....12
3.问题 2 分析方法与过程.....13
3.1 命名实体识别.....13
3.2 余弦相似性分析.....14
3.3 使用 K-means 方法进行聚类.....15
3.3.1 思想原理及算法步骤.....15
3.3.2 聚类步骤.....16
3.4 热度评价指标的建立.....17
4 结果分析.....19
4.1 问题 1 结果分析.....19
4.1.1 传机器学习模型.....19
4.1.2 Bert 模型.....20
4.1.3 问题 1 总结.....21
4.2 问题 2 热点问题表.....22
参考文献.....23

1.挖掘目标

本次建模的目标是利用互联网公开来源的群众问政留言记录，通过传统机器学习分类器和 Bert 模型、相似度计算、K-means 聚类以及　达到以下目标:

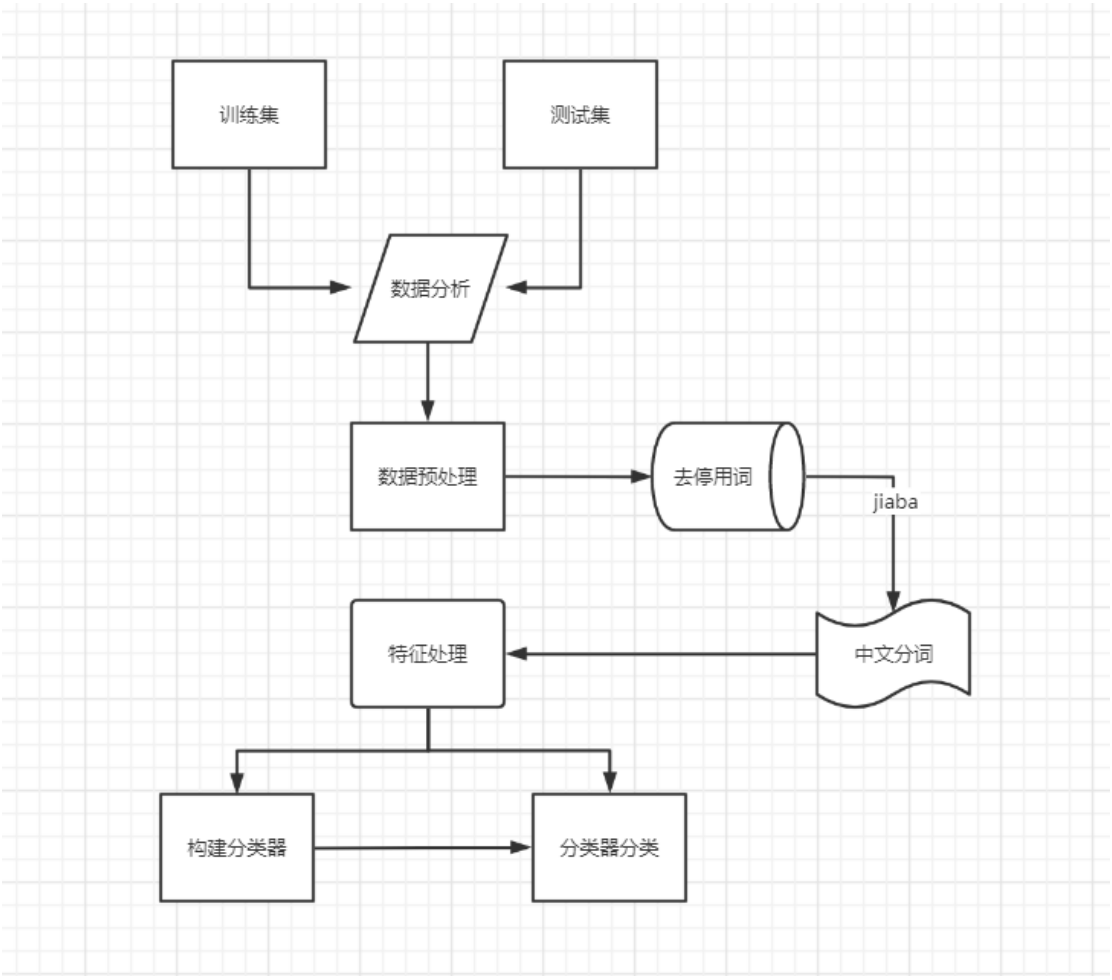
1)利用文本分词和文本分类的方法，对群众问政留言记录，建立关于留言内容的一级标签分类模型。

2)利用命名实体识别、相似度分析、文本聚类的方法，并建立热度指标，建立

2 问题 1 分析方法与过程

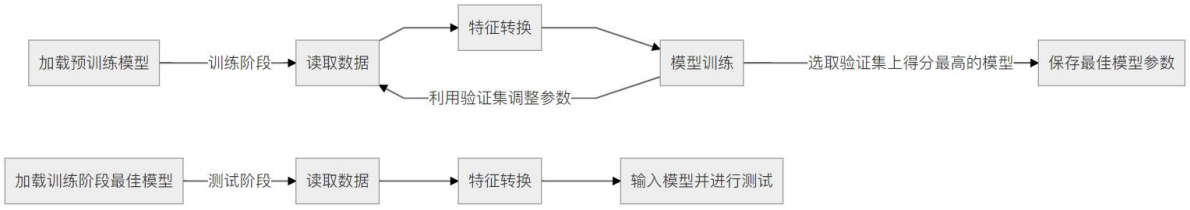
2.1 流程图

传统机器学习模型:



图一

Bert 模型：

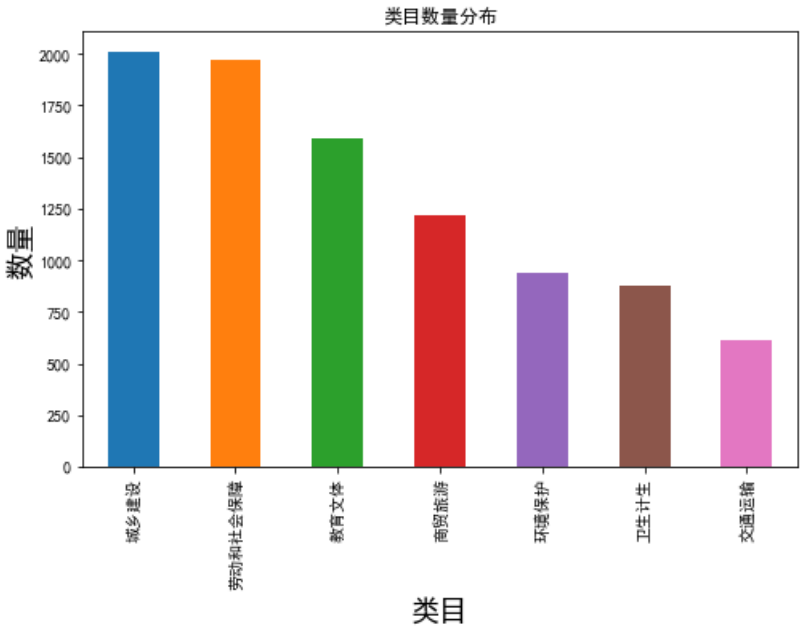


图二

2.2 传统机器学习模型

2.2.1 数据分析

先对本题中给出的全部数据进行分析，发现在 9210 条数据中未出现缺失值。分析已知人工已经处理好的一级标签数据，得到数据中的一级标签包含 7 个类别，分别为城乡建设、环境保护、交通运输、教育文体、劳动和社会保障、商贸旅游、卫生计生如下图所示：



图三

继续将类别转换成相应的 id，方便之后分类模型的训练。

label_id_df	一级标签	label_id
0	城乡建设	0
1	环境保护	1
2	交通运输	2
3	教育文体	3
4	劳动和社会保障	4
5	商贸旅游	5
6	卫生计生	6

图四

2.2.2 去除停用词

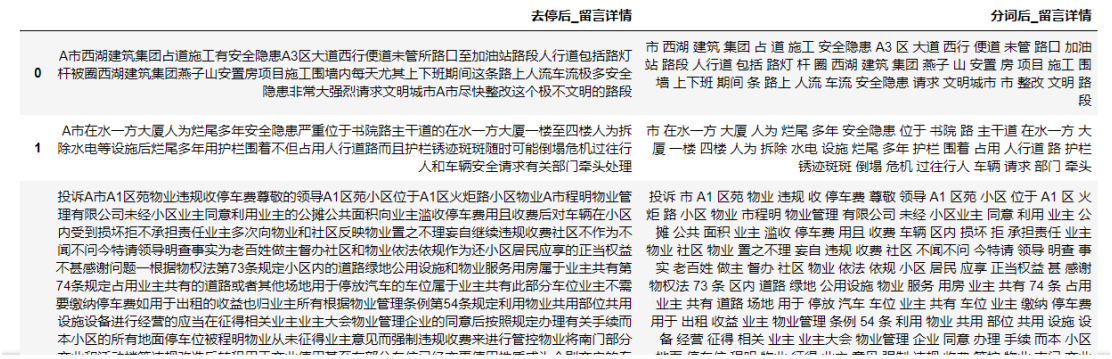
停用词是指在信息检索中，为节省存储空间和提高搜索效率，在处理自然语言数据(或文本)之前或之后会自动过滤掉某些字或词，这些字或词即被称为 Stop Words(停用词)。这些停用词都是人工输入、非自动化生成的，生成后的停用词会形成一个停用词表。这里我们采用的是哈工大去停用词表，将留言详情列除去停用词，生成去停后_留言详情.xlsx。

	留言详情	去停后_留言详情
0	A市西湖建筑集团占道施工有安全隐患\n\t\t\t\t\t\n\t\t\t\t\tA3区大道西行便道，未管所路口至加油站路段，人行道包括路灯杆，被圈西湖建筑集团燕子山安置房项目施工围墙内。每天尤其上下班期间这条路上人流车流极多，安全隐患非常大。强烈请求文明城市A市，尽快整改这个极不文明的路段。\n\t\t\t\t\t\n\t\t\t\t\t\n\t\t\t\t\t	A市西湖建筑集团占道施工有安全隐患A3区大道西行便道未管所路口至加油站路段人行道包括路灯杆被圈西湖建筑集团燕子山安置房项目施工围墙内每天尤其上下班期间这条路上人流车流极多安全隐患非常大强烈请求文明城市A市尽快整改这个极不文明的路段
1	A市在水一方大厦人为烂尾多年，安全隐患严重\n\t\t\t\t\t\n\t\t\t\t\t\t位于书院路主干道的在水一方大厦一楼至四楼人为拆除水、电等设施后，烂尾多年，用护栏围着，不但占用人行道路，而且护栏锈迹斑斑，随时可能倒塌，危机过往行人和车辆安全。请求有关部门牵头处理。\n\t\t\t\t\t\n\t\t\t\t\t\n\t\t\t\t\t	A市在水一方大厦人为烂尾多年安全隐患严重位于书院路主干道的在水一方大厦一楼至四楼人为拆除水电等设施后烂尾多年用护栏围着不但占用人行道路而且护栏锈迹斑斑随时可能倒塌危机过往行人和车辆安全请求有关部门牵头处理
	投诉A市A1区苑物业违规收停车费\n\t\t\t\t\n\t\t\t\t\t\t尊敬的领导：A1区苑小区位于A1区火炬路，小区物业A市程明物业管理有限公司，未经小区业主同意，利用业主的公摊公共面积向业主滥收停车费用，且收费后对车辆在小区内受到损坏拒不承担责任，业主多次向物业和社区反映，物业置之不理妄自继续违规收费，社区不作为不闻不问。今特请领导明查事实，为老百姓做主，督办社区和物业依法依规作为，还小区居民应享的正当权益，不甚感谢!问题一、根据《物权法》第73条规定，小区内的道路、绿地、公用设施和物业服务用房，属于业主共有。第74条规定，占用业主共有的道路或者其他场地用于停放汽车的车位，属于业主共有。此部分车位业主不需要缴纳停车费，如用于出租的	投诉A市A1区苑物业违规收停车费尊敬的领导A1区苑小区位于A1区火炬路小区物业A市程明物业管理有限公司未经小区业主同意利用业主的公摊公共面积向业主滥收停车费用且收费后对车辆在小区内受到损坏拒不承担责任业主多次向物业和社区反映物业置之不理妄自继续违规收费社区不作为不闻不问今特请领导明查事实为老百姓做主督办社区和物业依法依规作为还小区居民应享的正当权益不甚感谢!问题一根据物权法第73条规定小区内的道路绿地公用设施和物业服务用房属于业主共有第74条规定占用业主共有的道路或者其他场地用于停放汽车的车位属于业主共有此部分车位业主不需要缴纳停车费如用于出租的收益也归业主所有根据物业管理条例第54条规定利用物业共用部位共用设施设

图五（去除停用词后）

2.2.3 jieba 分词

去除停用词后，进行分词。在对热点问题留言明细进行挖掘分析之前，我们要做一些预处理操作以便后续将非结构化的文本信息转换为计算机能够识别的结构化信息。在留言明细表中，以中文文本的方式给出了数据，因此采用 python 的中文分词包 jieba 进行分词。jieba 采用了基于前缀词典实现的高效词图扫描，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），同时采用了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，对于未登录词，采用了基于汉字成词能力的 HMM 模型，使得能更好地实现中文分词效果。我们将去除停用词的留言详情进行 jieba 分词，生成分词后_留言详情.xlsx。



	去停后_留言详情	分词后_留言详情
0	A市西湖建筑集团占道施工有安全隐患A3区大道西行便道未管所路口至加油站路段人行道包括路灯杆被圈西湖建筑集团燕子山安置房项目施工围墙内每天尤其上下班期间这条路上人流车流极多安全隐患非常大强烈请求文明城市A市尽快整改这个极不文明的路段	市 西湖 建筑 集团 占 道 施工 安全隐患 A3 区 大道 西行 便道 未管 路口 加油 站 路段 人行道 包括 路灯 杆 圈 西湖 建筑 集团 燕子 山 安置 房 项目 施工 围墙 上下班 期间 条 路上 人流 车流 安全隐患 请求 文明城市 市 整改 文明 路段
1	A市在水一方大厦人为烂尾多年安全隐患严重位于书院路主干道的在水一方大厦一楼至四楼人为拆除水电等设施后烂尾多年用护栏围着不但占用人行道路而且护栏锈迹斑斑随时可能倒塌危机过往行人和车辆安全请求有关部门牵头处理	市 在水一方 大厦 人为 烂尾 多年 安全隐患 位于 书院 路 主干道 在水一方 大厦 一楼 四楼 人为 拆除 水电 设施 烂尾 多年 护栏 围着 占用 人行道 路 护栏 锈迹斑斑 倒塌 危机 过往行人 车辆 请求 部门 牵头
	投诉A市A1区苑物业违规收停车费尊敬的领导A1区苑小区位于A1区火炬路小区物业A市程明物业管理有限公司未经小区业主同意利用业主的公摊公共面积向业主滥收停车费用且收费后对车辆在小区内受到损坏拒不承担责任业主多次向物业和社区反映物业置之不理妄自继续违规收费社区不作为不闻不问今特请领导明查事实为老百姓做主督办社区和物业依法依规作为还小区居民应享的正当权益不甚感谢问题一根据物权法第73条规定小区内的道路绿地公用设施和物业服务用房属于业主共有第74条规定占用业主共有的道路或者其他场地用于停放汽车的车位属于业主共有此部分车位业主不需要缴纳停车费如用于出租的收益也归业主所有根据物业管理条例第54条规定利用物业共用部位共用设施设备进行经营的应当在征得相关业主业主大会物业管理企业的同意后按照规定办理有关手续而本小区的所有地面停车位被程明物业从未征得业主意见而强制违规收费来进行管控物业将南门部分	投诉 市 A1 区苑 物业 违规 收 停车费 尊敬 领导 A1 区苑 小区 位于 A1 区 火炬 路 小区 物业 市程明 物业管理 有限公司 未经 小区业主 同意 利用 业主 公摊 公共 面积 业主 滥收 停车费 用且 收费 车辆 区内 损坏 拒 承担责任 业主 物业 社区 物业 置之不理 妄自 违规 收费 社区 不闻不问 今特请 领导 明查 事实 老百姓 做主 督办 社区 物业 依法 依规 小区 居民 应享 正当权益 甚 感谢 物权法 73 条 区内 道路 绿地 公用设施 物业 服务 用房 业主 共有 74 条 占用 业主 共有 道路 场地 用于 停放 汽车 车位 业主 共有 车位 业主 缴纳 停车费 用于 出租 收益 业主 物业管理 条例 54 条 利用 物业 共用 部位 共用 设施 设备 经营 征得 相关 业主 业主大会 物业管理 企业 同意 办理 手续 而本 小区

图六（留言详情分词后）

2.2.4 特征处理

在预处理过后，需要把这些词语转换为向量，以供挖掘分析使用。这里采用 TF-IDF 算法，把分词后留言信息转换为权重向量。TF-IDF 算法的具体原理如下：

第一步，计算词频，即 TF 权重(Term Frequency)。

词频(TF)=某个词在文本中出现的次数 (1)

考虑到文章有长短之分，为了便于不同文章的比较，进行"词频"标准化，除以文本的总词数或者除以该文本中出现次数最多的词的出现次数即:

$$\text{词频(TF)} = \frac{\text{某个词在文本中的出现次数}}{\text{文本的总词数}} \quad (2)$$

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。

$$\text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库的文本总数}}{\text{包含该词的文本数+1}}\right) \quad (3)$$

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

TF-IDF =词频(TF) X 逆文档频率(IDF) (4)

第四步，生成 TF-IDF 向量

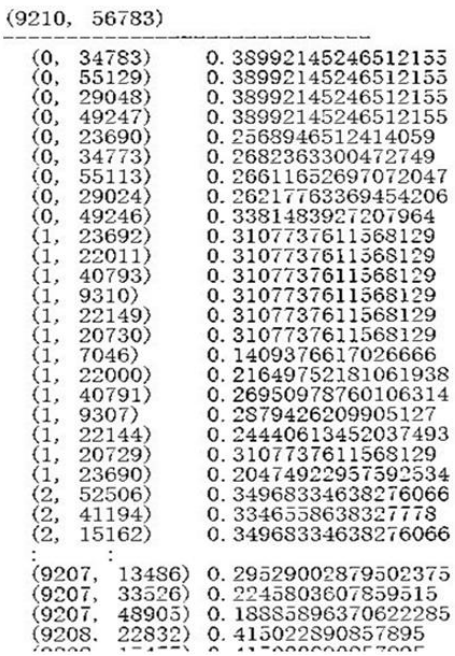
生成 TF-IDF 向量的具体步骤如下:

(1)使用 TF-IDF 算法

(2)合并成一个集合，计算每个热点问题描述对于这个集合中词的词频，如果没有则记为 0;

(3)生成各个词语描述的 TF-IDF 权重向量，计算公式如下:

TF-IDF=词频(TF) X 逆文档频率(IDF) (5)



图七(TF-IDF 的特征值)

2.2.5 四种分类器原理

2.2.5.1 朴素贝叶斯分类器

朴素贝叶斯的思想基础为：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就为此待分类项属于哪个类别。朴素贝叶斯分类的正式定义如下：设 x = { a1 ， a2 ， ... ， am}为一个待分类项，而每个 a 为 x 的一个特征属性。有类别集合 C = { y1 ， y2 ， ... ， yn} 计算每个 y 在 x 基础的概率分布，P(y1|x),P(y2|x), ...,P(yn|x) 如果 P(yk|x)= max { P(y1|x),P(y2|x), ...,P(yn|x)，则 x∈yk 。关键是如何计算每个条件概率。对于此我们可以统计训练集中的条件概率估计，并假设各个属性是条件独立地根据贝叶斯定理有

$$P(y_i|x) = \frac{P(x|y_i)\,P(y_i)}{P(x)}$$

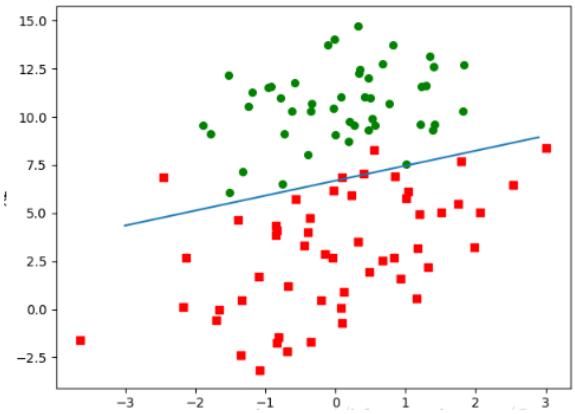
因为分母对于所有类别为常数所以将分子最大化即可，又因为各特征属性是条件独立的，所有

$$P(x|y_i)\,P(y_i) = p(y_i)\prod_{j=1}^{m} P(a_j|y_i)$$

2.2.5.2 逻辑回归分类器

直观来说，用一条直线对一些现有的数据点进行拟合的过程，就叫做回归。逻辑回归分类的主要思想：根据现有数据对分类边界建立回归公式，并以此分类。建立拟合参数的过程中用到最优化算法,这里用到的是常用的梯度上升法。

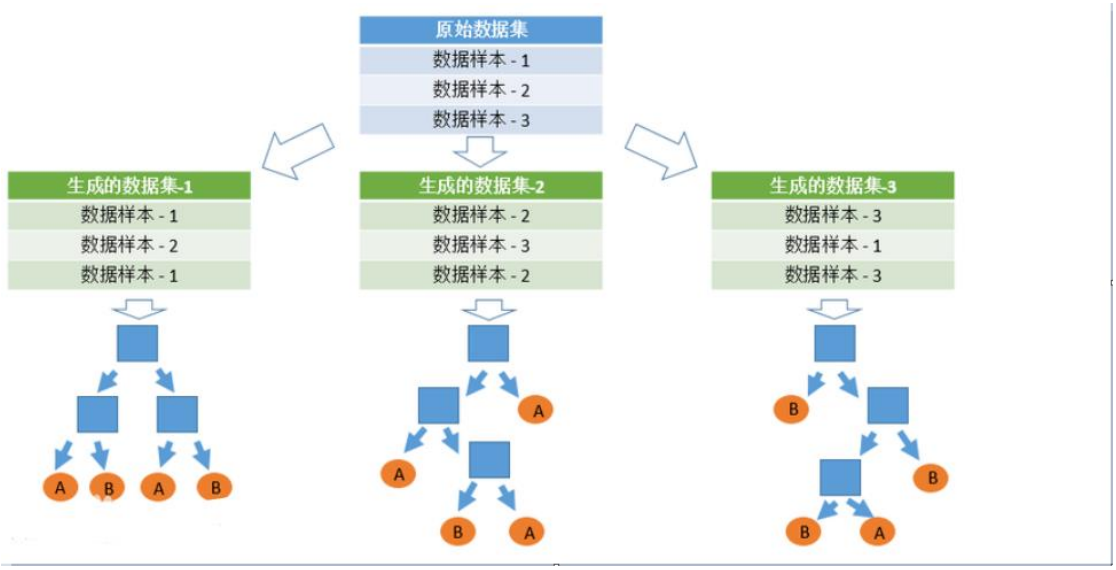
一个直观的图片：



图八

2.2.5.3 随机森林分类器

随机森林用于分类是，即采用多个个决策树分类。通过有放回的抽样从原始数据集中构建多个子数据集，然后利用每个子数据集构建一颗决策树，最终的分类效果有多颗决策树预测得到的众数决定，提高准确率。过程如下图所示

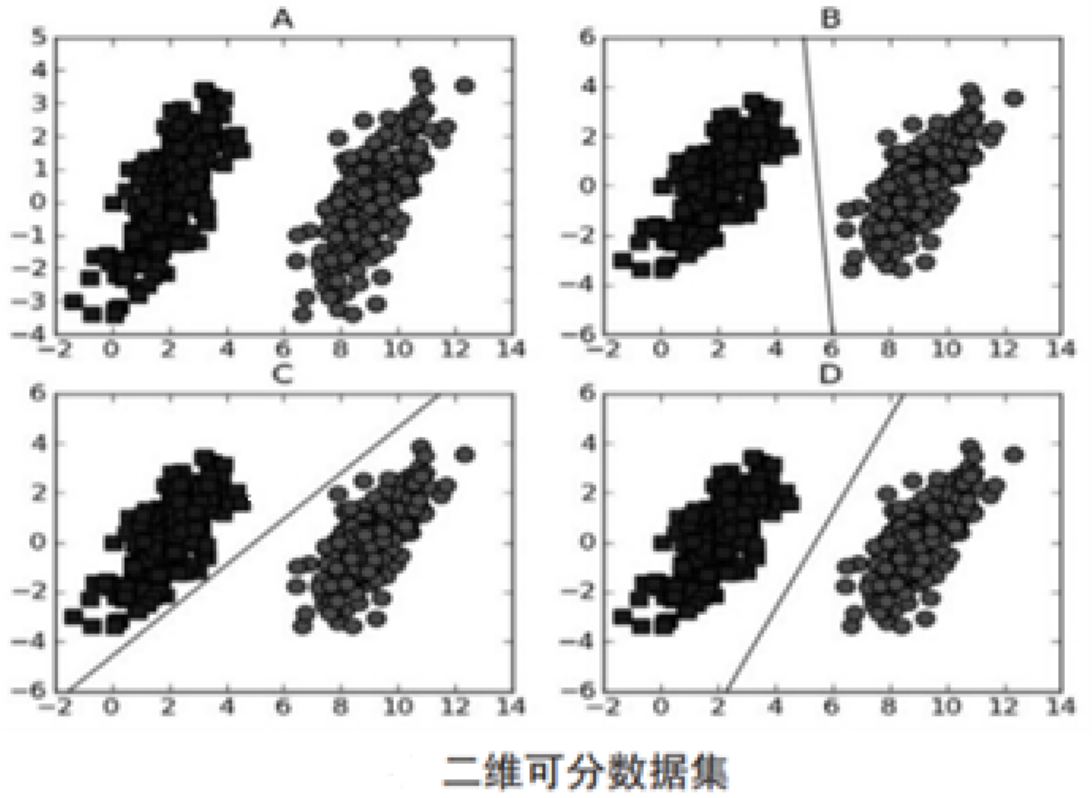


图九

2.2.5.4 SVM 分类器

SVM（Support Vector Machines）——支持向量机是在所有 知名的数据挖掘算法中最健壮，最准确的方法之一，它属于二分类算法，可以支持线性和非线性的分类。假设在一个二维线性可分的数据集中，图 A 所示，我们要找到一个超平面把两组数据分开，这时，我们认为线性回归的直线或逻辑回归的直线也能够做这个分类，这条直线可以是图 1B 中的直线，也可以是 图 C 中的直线，或者图 D 中的直线，但哪条直线才最好呢，也就是说哪条直线能够达到最好的泛化能力呢？那就是一个能使两类之间的空间大小最大的一个超平面。这个超平面在 二维平面上看到的就是一条直线，在三维空间中就是一个平 面，因此，我们把这个划分数据的决策边界统称为超平面。离这个超平面最近的点就叫做支持向量，点到超平面的距离叫间隔。支持向量机就是要使超平面和支

持向量之间的间隔尽可 能的大，这样超平面才可以将两类样本准确的分开，而保证间隔尽可能的大就是保证我们的分类器误差尽可能地小，尽可能地健壮。



图十

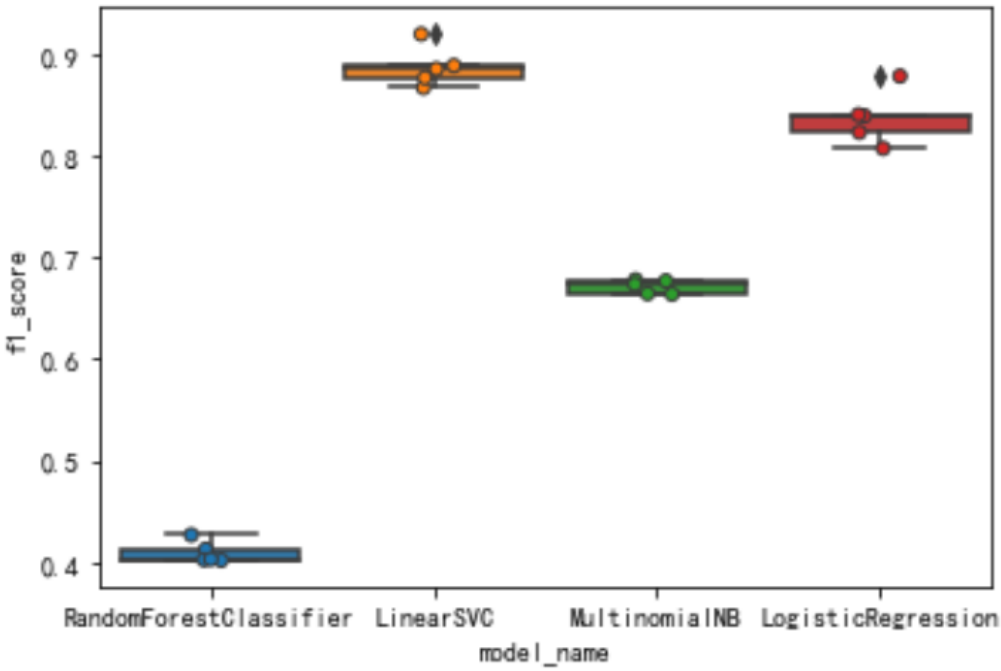
2.2.6 构建分类器

把全部数据分割成 80%作为训练集，20%作为测试集（以下训练都会按照此种分割）。分别尝试以上四种的分类器,并评估它们的 f1-score。

可以从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。

朴素贝叶斯分类器是通过找出特征之间的概率分布并利用贝叶斯算法计算概率。，算法较简单，发现该算法对缺失数据不太敏感。

逻辑回归分类器和 svm 分类器的 f1-score 超过 0.8,其中 svm 分类器 f1-score 分数最高。支持向量机算法是一种二分类模型，其主要思想是在 N 维空间中寻找出一个分类超平面以将低维的线性不可分变成高维的线性可分模型，且超平面将分割距离最大化。在高维空间中通过引入多种类核函数将计算复杂度降低，通过引 入惩罚因子和σ 参数调节校正分类拟合度与误差估计来进行分类。因此接下来实验结果分析主要针对 svm 分类器。

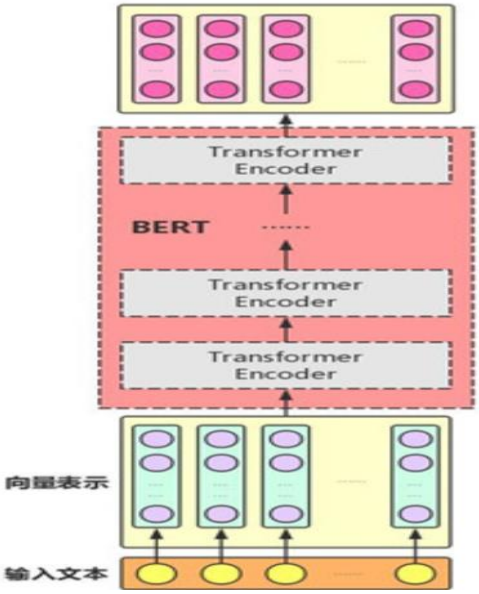


图十一（箱体图）

2.3 Bert 模型

2.3.1 Bert 模型简介

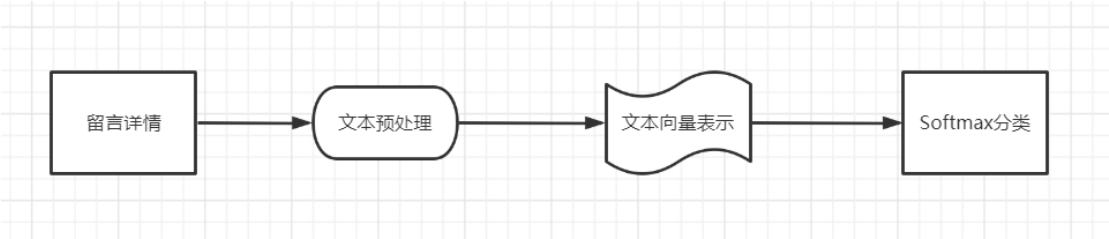
Bert 是一个自然语言处理的通用模型，其基础是 Attention 机制。Attention 机制的提出是为了解决一些深度学习模型无法并行等缺点。在此之后谷歌提出了完全建立在 Attention 之上的 Trans-former 模型 ，Bert 正是由多个 Transformer 模型的 Encoder 结构堆叠而成。Bert 具有预训练的特点，即先使用大规模的语料进行无监督学习得到预训练模型。使用者可以在具体的自然语言处理任务中直接使用此模型或者微调后使用。
过程如下图所示：



图十二

2.3.2 Bert 模型使用

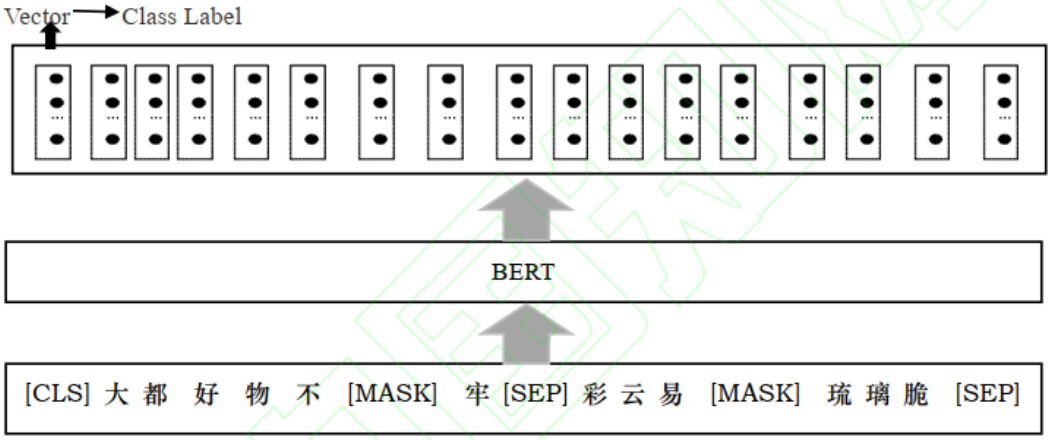
2.3.2.1 过程



图十三

2.3.2.2 生成句向量

直接利用谷歌开源的预训练模型，将留言详情利用 Bert 模型输出为句子级别的向量，即 Bert 模型输出最左边[CLS]特殊符号的向量，认为这个向量可以代表整个句子的语义，如下图所示



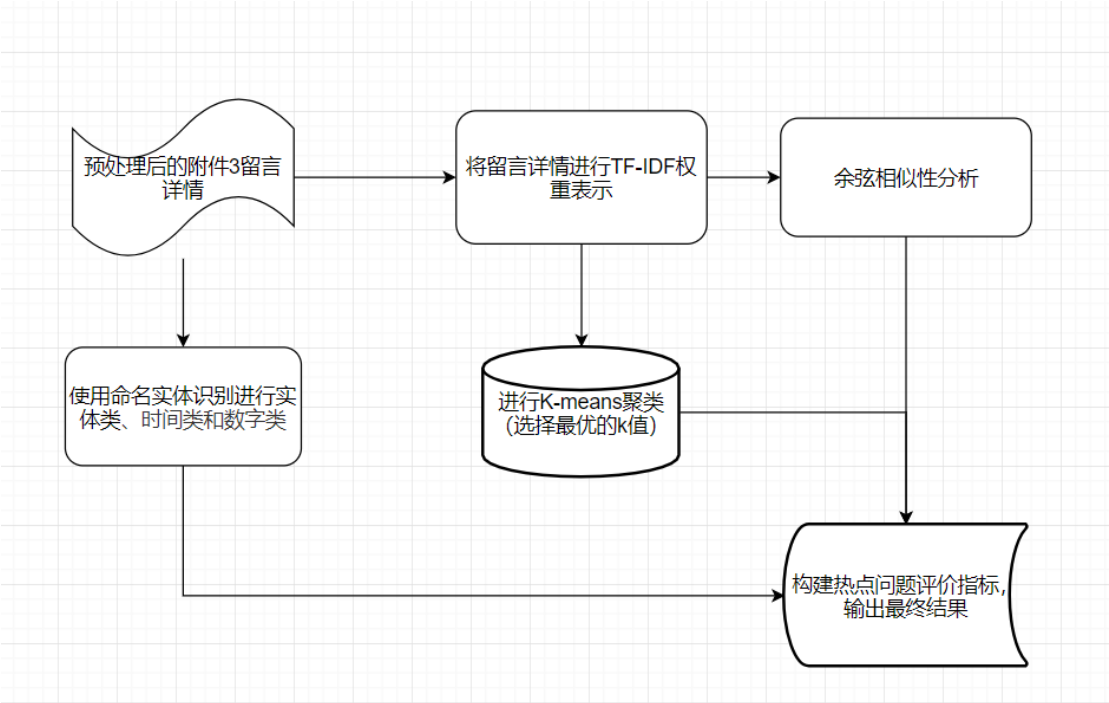
图十四

2.3.2.2 Softmax 回归模型

Softmax 回归分类算法是逻辑回归分类算法在多分 类问题上的推广。在多分类问题中，分类标签可以取两个以上的值。Softmax 回归分类算法对手写数字分类等问题有很好的效果，如 MNIST 手写数字的识别正确率可达 92%，适用于多分类问题。

通过 Bert 模型输出留言详情的向量表示，我们直接选用 Softmax 回归模型来进行训练，输出分类结果，并打印 f1_score。

3.问题 2 分析方法与过程

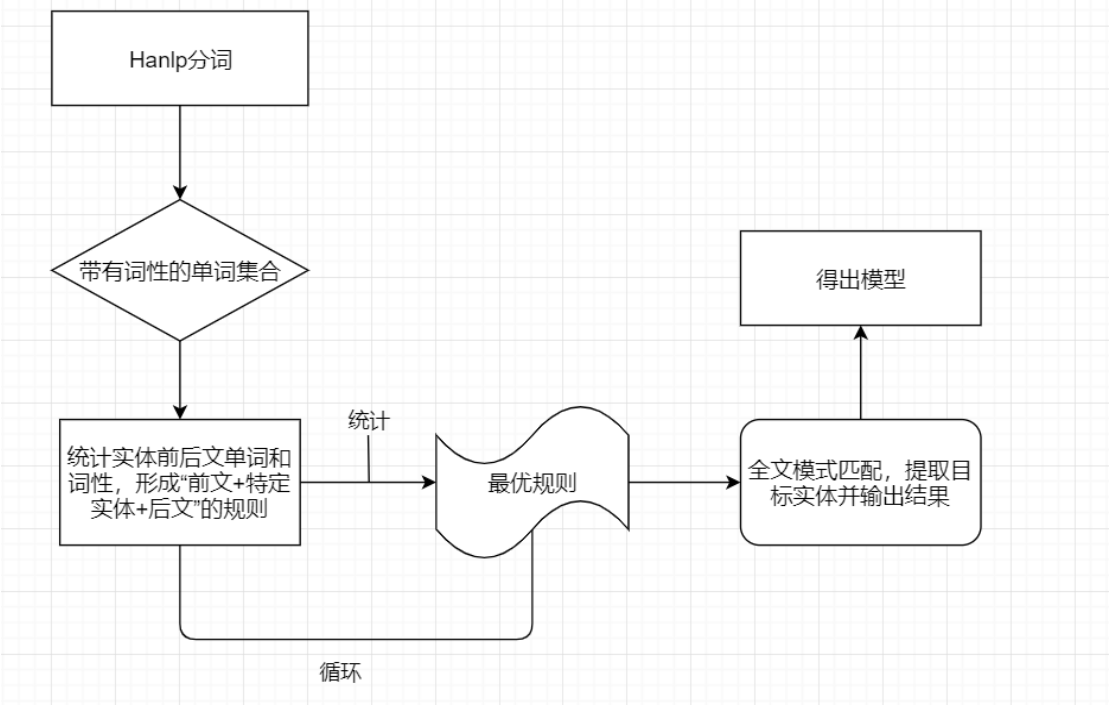


图十五（总体流程图）

3.1 命名实体识别

热点问题留言明细表经过如问题 1 的去停用词和分词后，我们通过命名实体识别来提取特定地点、特定人物等重要信息。命名实体识别(Named Entity Recognition 简称 NER)，又称作"专名识别"，是指识别文本中具有特定意义的实体，主要包括人名、地名、专有名词等。它是信息提取、问答系统等应用领域的重要基础工具，在自然语言处理技术走向实用化的过程中占有重要地位。

具体实现步骤如图 2 所示



图十六（命名实体识别流程图）

首先，统计这些实体出现的前后文单词和词性，并考虑它们之间的联系，概括出特定实体前后出现的高频词汇；最后，利用这一规则在全文中进行模式匹配。利用投票原理，对匹配度高的规则分配高分，相反，匹配度低的规则赋予低分。然后，对所有匹配的规则进行分数排序，得到投票分数最高的规则，并从规则中剥离出特定实体，这个实体即为我们的目标实体。

热点问题留言特定地点的提取，我们统计出该实体出现的前文频率较高的为：介词词性的词语等，后文为：小区、发生事件、物业等。通常出现这些词汇的前后就是特定地点。然后我们再根据这个词的词性，判断它是否属于地点名。这样，就可以获得我们需要的特定地点信息。

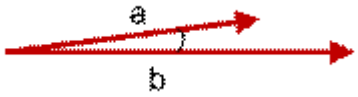
	特定地点	出现次数		关键词	次数
1	A7	682	**11**	街道	199
2	小区	549	**12**	A5	177
3	A3	439	**13**	西地省	160
4	A2	264	**14**	A6	144
5	A4	236	**15**	社区	132
6	A1	209			

3.2 余弦相似性分析

相似度度量（Similarity），即计算个体间的相似程度，相似度度量的值越小，说明个体间相似度越小，相似度的值越大说明个体差异越大。对于多个不同的文本或者短文本对话消息要来计算他们之间的相似度如何，一个好的做法就是将这些文本中词语，映射到向量空间，形成文本中文字和向量数据的映射关系，通过计算几个或者多个不同的向量的差异的大小，来计算文本的相似

度。

余弦相似度用向量空间中两个向量夹角的余弦值作为衡量两个个体间差异的大小。余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，这就叫"余弦相似性"。



假定 a 和 b 是两个 n 维向量，则 a 与 b 的夹角的余弦等于：

$$\cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}}$$

$$= \frac{a \bullet b}{||a|| \times ||b||}$$

余弦值越接近 1，就表明夹角越接近 0 度，也就是两个向量越相似，夹角等于 0，即两个向量相等，这就叫"余弦相似性"。

1）利用问题 1 TF-IDF 算法把留言详情向量化
2）利用余弦相似性分析，将各类别之间相似度记录下，保存文件为 相似度.xlsx

3.3 使用 K-means 方法进行聚类

3.3.1 思想原理及算法步骤

（1）算法原理：

K-means 算法是最为经典的基于划分的聚类方法，是十大经典数据挖掘算法之一，归类划分体系于聚类分析算法体系下分割算法类启发式。K-means 算法的基本思想是：以空间中 k 个点为中心进行聚类，对最靠近他们的对象归类。通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果。

（2）核心思想：以集群内资料平均值为集群的中心。

任务是把所有的实例分配到若干的簇，使得同一个簇的实例聚集在一个簇中心的周围，它们之间距离的比较近，而不同簇实例之间的距离比较远。

（3）算法步骤：

1） 任意选择 k 个对象作为初始的类的中心
2） 将每个实例分配到距它最近的簇中心，得到 k 个簇；
3） 根据类中实例的平均值,将每个实例(重新)赋给最相近的类
4） 更新类的平均值，
5） 直到不再发生变化,即没有对象进行被重新分配时过程结束 。

3.3.2 聚类步骤

3.2.2.1 质心计算

对于分类后的产生的 k 个簇，分别计算到簇内其他点距离均值最小的点作为质心（对于拥有坐标的簇可以计算每个簇坐标的均值作为质心）

3.2.2.2 距离度量

将对象点分到距离聚类中心最近的那个簇中需要最近邻的度量策略，在处理文本数据时常采用的是余弦相似度（余弦相似性）来计算文本相似度。

给定两个属性向量，A 和 B，其余弦相似性 θ 由点积和向量长度给出，如下所示：

$$similarity = cos(\theta) = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2 \times \sum_{i=1}^{n} (B_i)^2}}$$

这里的A_i, B_i分别代表向量 A 和 B 的各分量。

给出相似性范围从-1 到 1：-1 意味着两个向量指向的方向正好截然相反，1 表示它们的指向是完全相同的，0 通常表示它们之间是独立的，而在这之间的值则表示中间的相似性或相异性。在信息检索的情况下，由于一个词的频率（TF-IDF 权）不能为负数，所以这两个文档的余弦相似性范围从 0 到 1。并且，两个词的频率向量之间的角度不能大于 90°。

对于文本匹配，属性向量 A 和 B 通常是文档中的词频向量。余弦相似性，可以被看作是在比较过程中把文本长度正规化的方法。

将两个文本根据他们词，建立两个向量，计算这两个向量的余弦值，就可以知道两个文本在统计学方法中他们的相似度情况。

3.2.2.3 聚类效果评价

轮廓系数（Silhouette Coefficient）结合了聚类的凝聚度（Cohesion）和分离度（Separation）两种因素，是聚类效果好坏的一种评价方式。可以用来在相同原始数据的基础上用来评价不同算法、或者算法不同运行方式对聚类结果所产生的影响。该值范围位于[-1,1]，值越大，表示聚类效果越好。具体计算方法如下：

假设我们已经通过一定算法，将待分类数据进行了聚类。常用的比如使用 K-means ，将待分类数据分为了 k 个簇 。对于簇中的每个向量。分别计算它们的轮廓系数S_i。

对于其中的一个点 i 来说：计算点 i 与其同一个簇内的所有其他元素距离的平均值，记作a_i，用于量化簇内的凝聚度，表示 i 向量到同一簇内其他点不相似程度的平均值。选取 i 外的一个簇 b，计算 i 与 b 中所有点的平均距离，遍历所有其他簇，找到最近的这个平均距离,记作b_i，即为 i 的邻居类，用于量化簇之间分离度，表示 i 向量到其他簇的平均不相似程度的最小值。

$$a_i = average\left(\text{i 向量到所有它属于的簇中其它点的距离}\right)$$

$$b_i = \min\left(\text{i 向量到与它相邻最近的一簇内的所有点的平均距离}\right)$$

$$S_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

可见轮廓系数的值是介于 [-1,1] ，越趋近于 1 代表凝聚度和分离度都相对较优。将所有点的轮廓系数求平均，就是该聚类结果总的轮廓系数。

注意：上部分中所说的“距离”，指的是不相似度（区别于相似度）。“距离“值越大，代表不相似度程度越高。当簇内只有一点时，我们定义轮廓系数S_i为0。

从上面的公式，不难发现若S_i小于 0，说明 i 与其簇内元素的平均距离小于最近的其他簇，表示聚类效果不好。如果a_i趋于 0，或者b_i足够大，即a_i远远小于b_i，那么S_i趋近于 1，说明聚类效果比较好。

（1）在距离表 points 中查询距离 distance;

（2）根据中心点划分簇 cluster;

（3）编写 cal_k_means 函数用以进行计算 k-means 聚类过程;

（4）编写主函数 k_means 框架；

（5）用轮廓系数函数 slt_coefficient 评价聚类效果。

3.4 热度评价指标的建立

我们通过对关键特征词进行聚类分析来衡量热度，聚类结果是群众留言热度可参考因素之一，点赞数与反对数是热度可参考因素之二。这两个表现特征作为衡量留言问题热度的指标,借此构建一套衡量热度的综合评价指标体系,并运用因子分析方法对指标体系进行检验,期望能够为热点问题的界定和评价提供参考。

（1）指标体系的确立：

表 2 热度评价指标体系

指标	指标含义
聚类结果	热点问题反映的密集程度
点赞数和反对数	部分群众并无留言，通过赞同或反对方式来表达自己的态度，这也可以体现问题的热度

(2) 将点赞数和反对数提取出来，通过 spss 工具求得各条留言点赞数和反对数占比，将其与聚类度量值结合列表如下:

表 3 指标数据集（部分）

聚类度量值	点赞数	反对数
0.176	0.3	0.0
0.802	0.2	0.1
0.056	0.2	0.0
0.044	0.0	0.0
0.023	0.1	0.1
0.071	0.1	0.0
0.012	0.4	0.2

（3）运用 SPSS 工具，对收集到的数据进行 KMO 测度和巴特利特球体检验，其检验结果如表 3 所示：

表 3 KMO 测度和巴特利特球体检验结果

取样足够的 Kaiser-Meyer-Olkin 度量		0.713
Bartlett 的球形度检验	近似卡方	67.736
	df	35
	Sig.	0.001

表 3 中得到 KMO 值为 0.713，巴特利特球体检验的卡方统计值的显著性为 0.001，小于 1%，两条结果均说明了数据具有相关性，适合做因子分析。

接下来对这些热度评价指标进行因子分析中的降维处理，其结果如表 4 所示：

表 4 公共因子提取及解释信息比例情况

成分	初始特征值			旋转平方和载入		
	合计	方差(%)	累计(%)	合计	方差(%)	累计(%)
1	3.836	37.643	37.128	3.731	38.682	37.265
2	2.206	21.872	53.974	1.978	21.847	53.478
3	0.812	9.018	86.500	1.579	16.564	74.534

从表 4 中可以看出，按照特征值超过 1 和信息解释百分比超过 80%为标准，一共提取 3 个公共因子，这和本文提出的三个维度的指标体系刚好吻合。可以进一步发现三个公共因子能够分别解释热度相关信息的 43.72%、 35.4% 和 15.26% ，最后累计能够解释总体信息的 78.9%。这一结果显示了三个公共因子能够较好地反映问题热度的总体信息。

然后，再进一步根据因子分析反馈的成分矩阵对各个公共因子含义以及构成进行分析，结果如表 5 所示。

表 5 各成分得分系数矩阵

	公共因子		
	1	2	3
聚类度量值	0.657	0.732	0.614
点赞数	0.346	0.282	0.323
反对数	0.184	0.265	0.036

从表 5 中每一行的最高的相关系数可以判定该指标与哪个公共因子具有最高的相关性，例如第一行聚类度量值与第二公共因子的相关系数最高，达到 0.732，具有最高相关性。同理，可以总结出以下信息：第一公共因子与聚类度量值、点赞数有很高的相关关系；第二公共因子与聚类度量值有较高的相关关系；第三公共因子与聚类度量值和点赞数有较强的相关性。 将此处的分析结果和前面按照理论依据构建的指标体系相比较可以发现，3 个指标中所有指标的降维结果和上文构建的热度指标体系基本符合。

通过因子分析方法对热度指标进行检验，求得排行前五名的留言编号为 263672、208636、223297、268250、233743。

4 结果分析

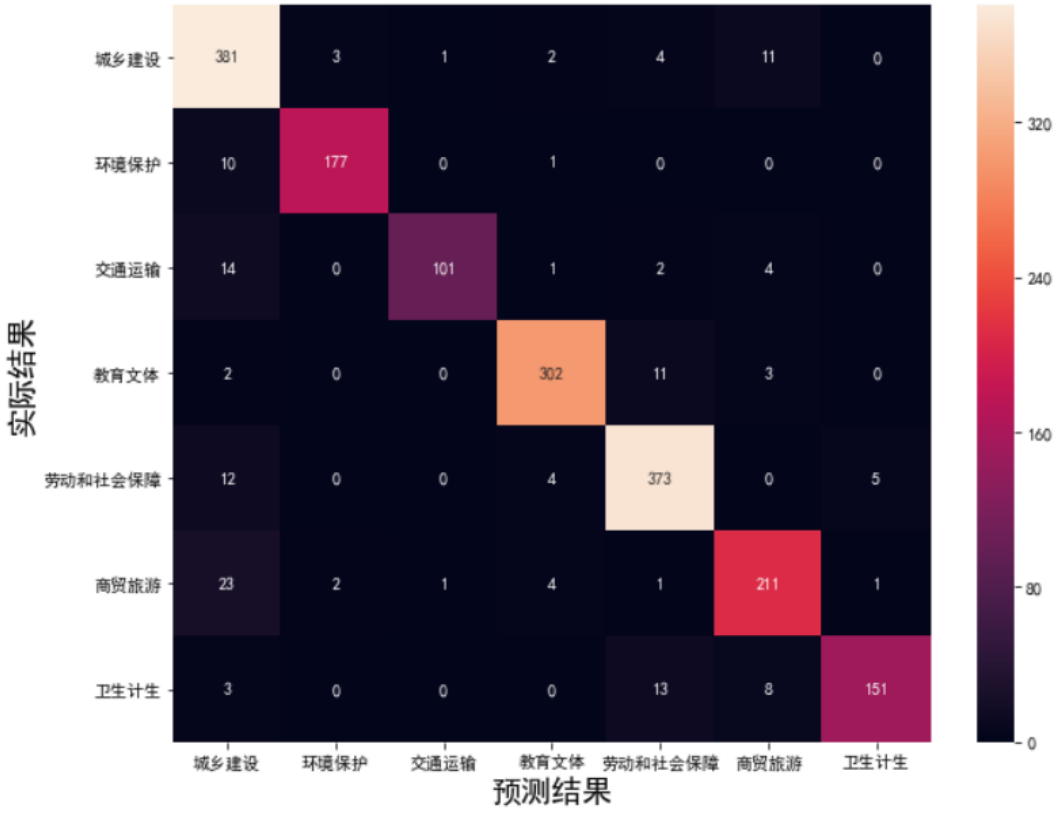
4.1 问题 1 结果分析

4.1.1 传机器学习模型

首先对文本数据进行预处理后，将去停用词并分词后的留言详情进行 TF-IDF 权重向量表示，把数据分割，80%的数据进行训练，20%的数据进行测试，通过选择分类器，采用 SVM 分类器，得出结果如下：

	精确率 precision	召回率 recall	F_1指标 f1-score	数量 Support
城乡建设	0.86	0.95	0.90	402
环境保护	0.97	0.94	0.96	188
交通运输	0.98	0.83	0.90	122
教育文体	0.96	0.95	0.96	318
劳动和社会保障	0.92	0.95	0.93	394
商贸旅游	0.89	0.87	0.88	243
卫生计生	0.96	0.86	0.91	175
Macro avg	0.94	0.91	0.92	1842
Weight avg	0.92	0.92	0.92	1842

图十七



图十八

由此可见，数据通过 SVM 分类器，每一类 F1-score 在 0.9 左右，平均 F1-score 达到 0.92，预测错误的主要是城乡建设与其他类，劳动和社会保障与其他类，但预测错误量较少。可见本次数据经预处理过后的留言详情，用 TF-IDF 权重表示，经 SVM 分类器后，测试效果较好。

4.1.2 Bert 模型

全部数据 80%的数据进行训练，20%的数据进行测试，利用谷歌开源的预训练模型，通过 Bert 模型把留言详情向量化，经 softmax 分类，并把模型训练参数设成如下图所示：

```
self.require_improvement = 1000                                 # 若超过1000batch效果还没提升，则提前结束训练
self.num_classes = len(self.class_list)                         # 类别数
self.num_epochs = 30                                            # epoch数
self.batch_size = 128                                           # mini-batch大小
self.pad_size = 150                                             # 每句话处理成的长度(短填长切)
self.learning_rate = 5e-5                                       # 学习率
self.tokenizer = BertTokenizer.from_pretrained(self.bert_path)  # bert 切词器
self.hidden_size = 768                                          # bert 隐藏层个数
```

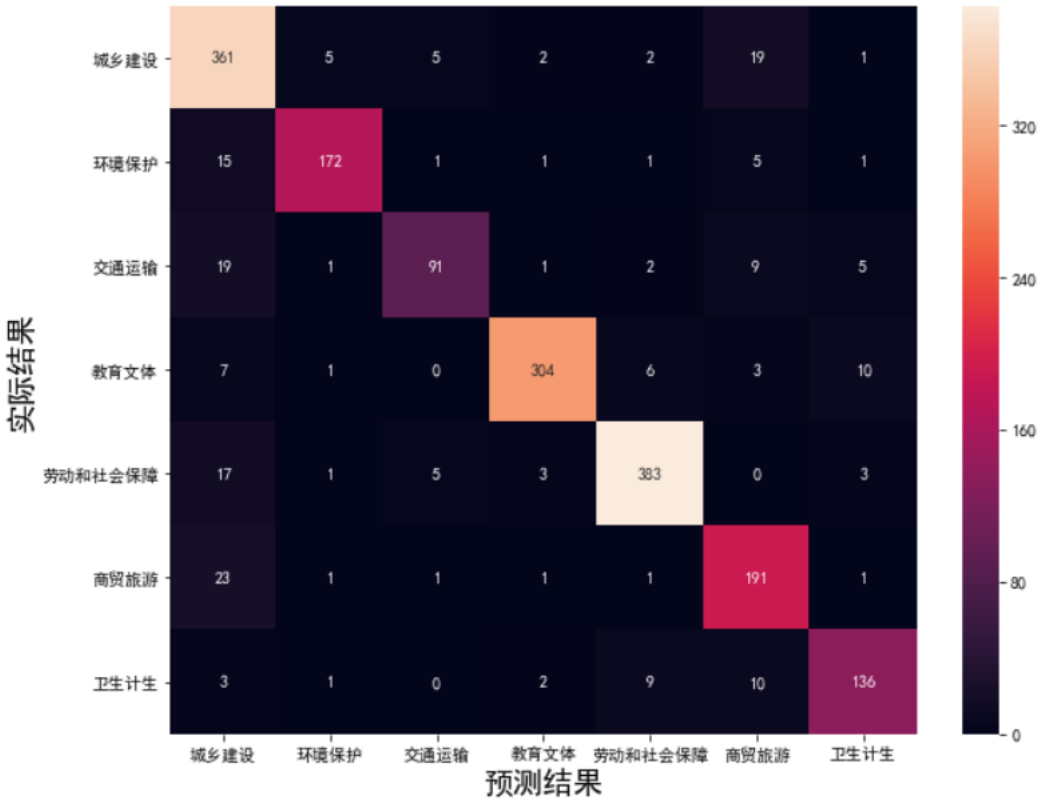
图十九

得出最好结果如下：

	精确率 precision	召回率 recall	F_1指标 f1-score	数量 support
城乡建设	0.82	0.95	0.89	402

环境保护	0.92	0.92	0.93	188
交通运输	0.96	0.83	0.90	122
教育文体	0.91	0.93	0.93	318
劳动和社会保障	0.94	0.91	0.93	394
商贸旅游	0.86	0.87	0.86	243
卫生计生	0.96	0.83	0.90	175
Macro avg	0.91	0.89	0.90	1842
Weight avg	0.91	0.91	0.89	1842

图二十



图二十一

由此可见，本次数据通过 Bert 模型处理， 经 softmax 分类，测试错误主要集中城乡建设与其他类，商贸旅游与其他类，卫生计生与其他类，错误数比 svm 模型高，F1-score 分数比 svm 模型低，但总体测试效果较好。

4.1.3 问题 1 总结

针对关于留言内容的一级标签分类模型，在尝试了传统机器学习模型与谷歌开源 Bert 模型，发现最终传统机器学习下预处理后的 svm 模型性能较好。

svm 模型适合于多标签的分类，理论依据强，比较成熟。且在本次数据集下，测试效果较好。因此笔者认为在数据集不大，不需要消耗大量资源的情况下可以选择传统学习 svm 模型对留言进行分类。

Bert 模型消耗资源较大，我们只尝试了 Bert 模型把文本转换为句向量后，

就经过 softmax 分类。还可把经 Bert 模型输出的句向量进入卷积神经网络再进行分类等，但 Bert 模型训练需要调节的参数的更多，可能性更多，因此笔者认为 Bert 模型在留言文本分类中的潜力更大。

4.2 问题 2 热点问题表

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	263672	72.3	2019/8/29 至 2019/9/10	A4 区绿地海外滩小区附近居民	A4 区绿地海外滩小区距长赣高铁最近只有 30 米，对生活造成严重影响。
2	208636	63.8	2019/8/18 至 2019/8/27	A 市 A5 区汇金路五矿万境 K9 县业主	A 市 A5 区汇金路五矿万境 K9 县存在一系列问题
3	223297	62.1	2019/4/5 至 2019/4/16	梅溪湖金毛湾业主	A 市金毛湾配套入学的问题
4	268250	53.8	2019/3/26 至 2019/4/6	A 市月亮岛路附近业主	A 市电力局沿月亮岛路施工新建一排谷山变电站 110kV 出线线路，采用高压铁塔线路方案，可能对附近群众造成健康影响。
5	233743	49.2	2019/3/12 至 2019/3/24	A 市 A3 区观沙岭银杉路附近	A3 区观沙岭郝家坪小学改扩建问题

参考文献

[1] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[2] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems. 2017: 5998-6008.

[3] Wilson L Taylor. 1953. cloze procedure: A new tool for measuring readability. Journalism Bulletin, 30(4):415－433.

[4]杨锋.基于线性支持向量机的文本分类应用研究[J].信息技术与信息化,2020(03):146-148.
[5]段丹丹,唐加山,温勇,袁克海.基于 BERT 的中文短文本分类算法的研究[J/OL].计算机工程:1-12[2020-05-07].https://doi.org/10.19678/j.issn.1000-3428.0056222.
[6]祁小军,兰海翔,卢涵宇,丁蕾锭,薛安琪.贝叶斯、KNN 和 SVM 算法在新闻文本分类中的对比研究[J].电脑知识与技术,2019,15(25):220-222.
[7]何跃,蔡博驰.基于因子分析法的微博热度评价模型[J].统计与决策,2016(18):52-54.
[8]邬启为. 基于向量空间的文本聚类方法与实现[D]. 北京交通大学, 2014.
[9] 胡春涛，秦锦康，陈静梅，等. 基于 BERT 模型的舆情分类应用研究［J］. 网络安全技术与应用，2019(11):41
[10] 李丹阳. 面向中文评论的情感分析方法研究[D]. 西安：西安工业大学，2019.