“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，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。

对于问题 1，通过数据观测，先观察一级标签这一列的详细情况，并通过画图展示。利用 jieba 中文分词工具对留言详情进行分词和剔除停用词，并在分词后的结果的基础上利用词云图画出每一类的 TOP100 高频词。再利用 TF-IDF 方法进行加权操作，用 sklearn 中的 TfidfVectorizer 方法来抽取特征值，然后利用卡方检验找出每个分类中关联度高的词语，并用朴素贝叶斯分类器分类。选择四种机器学习模型（逻辑回归，多项式朴素贝叶斯，线性支持向量机，随机森林），评估它们的分类效果，通过箱线图进行展示。针对效果好的 LinearSVC 模型，通过混淆矩阵，来查看效果。最后查看 F1 分数，来评估模型。

对于问题 2，采用 sklearn 中的 TfidfVectorizer 方法，对留言主题这一列计算相似度，初步得到了热点问题的一些关键点。在此基础上，借用索引的方式找出每个热点问题的时间，进行排序查看，同时根据每个热点问题的文本数来确定其热度，用于热点问题排序，得到热点问题表。最后通过 iloc 方法找出热点问题对应的行，然后通过 concat 函数进行合并，得到热点问题留言明细表。

对于问题 3，先对留言主题和答复意见进行分词和去停用词，但我们针对每一行数据我们都进行了特殊的去停用词，我们根据每行数据的特点，区别对待，最终得到适用于每行自己的去停用词。采用的是对留言主题的第一行的每个分词在答复详情相对应的行进行遍历，最终，通过每行的匹配次数除以答复详情相对应行的分词数确定精确度。采用上四分位数，中位数，下四分位数作为分界点，表示答复详情的好坏，进行最终的数据分析。

关键词：数据挖掘；TF-IDF 法；LinearSVC 模型；数据采集

目录

1.问题重述

在当前大数据时代之下,信息高度发达已经成为重要特征。通过对那些具有一定价值的信息群来挖掘与整合的工作,就是在大数据时代之下,有效的探究与应用数据的挖掘技术。合理的借助数据挖掘技术,能够更好的满足当前日常信息管理对于数据的要求,也能更好的对信息互扰等麻烦进行管理。信息时代最大的特点就是浩如烟海的数据每时每刻都在被产生。这些数据不是杂乱无章,毫无意义,反而蕴藏着巨大的信息价值。数据挖掘技术可以帮助人们从海量的数据中,发现具有创新性、不可见性、多元性的信息,而这些信息对某个领域的发展将起着至关重要的作用,这在智慧政务系统中也不例外。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。因此，我们需要一个更智能，更便捷个性化的系统，帮助政府高效准确的收集民众留言和信息，使内容的产生更有目的性和精确性，以及对提升政府的管理水平和施政效率具有极大的推动作用，从而实现价值的最大化。

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，但是对于所研究对象不同，数据集不同，所表现的算法也会存在差异。由于目前，大部分电子政务系统还是依靠人工根据经验处理，留言信息特征较难处理，非结构化的数据处理比较困难，数据容量大，且相似度量不好定义。此外，留言信息冗多，热点问题及时提取收集成为大部分电子政务系统的难题。

本研究的主要内容为：根据来自互联网公开来源的群众问政留言记录，及相关部门对部分群众留言的答复意见,利用数据挖掘方法解决以下两方面问题。

1.群众留言高效分类。根据理网络问政平台的群众留言，采用数据挖掘，TF-IDF 方法
方法，对留言内容按照一定体系分类，并做出了效果评估。

2.热点问题挖掘。通过对特定时间，地点以及人群的留言进行分类，定义合理地热度评价指标，并给出评价结果。

2. 模型假设

5. 假设所有群众在给出模型后的一段时间内正常在网络问政平台上留言；
6. 假设留言详情的长度和它的类别之间没有相关性。

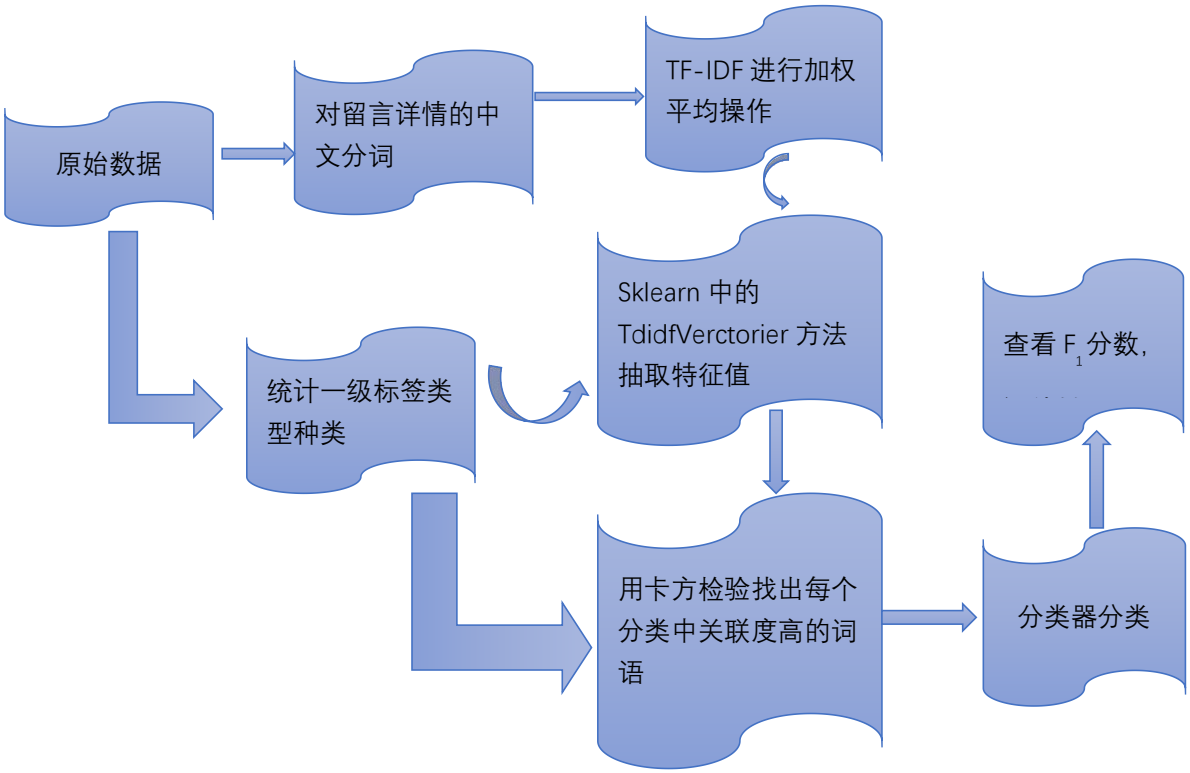
针对实际的群众问政留言记录而言，为了提升政府的管理水平和施政效率，我们必须做出如下假设：

1. 假设所给的数据集是完备可靠的；
2. 假设网络爬虫提取的数据是完全正确的；
3. 假设答复意见的相关性、完整性和可解释性与答复意见的质量成正比；
4. 假设所有问题的类型均属于我们的分类三级标签体系；

3. 分析方法和过程

3.1 问题 1 的分析方法

3.1.1 流程图



3.1.2 数据清洗

在题目给出的数据中，为了防止出现重复的数据干扰计算，我们首先对数据进行了去重。在对留言详情进行挖掘分析之前，先要把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。在附件 2 的留言详情中，以中文文本的方式给出了数据。为了便于转换，先要对这些留言详情进行中文分词。这里采用 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 HM　M 模型，使得能更好的实现中文分词效果。

其结果如下：

Out[9]:

	review	cat	cat_id	clean_review
4865	抵制D7县这种以公补私行为，遏止以公补私现象泛滥	教育文体	3	抵制D7县这种以公补私行为遏止以公补私现象泛滥
7541	建议把K市旅游局的网站做好	商贸旅游	5	建议把K市旅游局的网站做好
7536	K8县九嶷新村小区电梯停运停运两天了	商贸旅游	5	K8县九嶷新村小区电梯停运停运两天了
6988	中度双相情感障碍能否享受大病医保?	劳动和社会保障	4	中度双相情感障碍能否享受大病医保
186	A7县世锦小学垃圾站规划在小区门口离最近住户不到5米	城乡建设	0	A7县世锦小学垃圾站规划在小区门口离最近住户不到5米
3582	西地省成考招生报名系统一直不能预约！！！	教育文体	3	西地省成考招生报名系统一直不能预约
3749	A7县的幼师真的命苦!	教育文体	3	A7县的幼师真的命苦
5462	L3县合仁坪矿业几十人得矽肺病获赔无门	劳动和社会保障	4	L3县合仁坪矿业几十人得矽肺病获赔无门
6735	请求K市人社局对吕保端信访复查事项进行复查	劳动和社会保障	4	请求K市人社局对吕保端信访复查事项进行复查
1253	L市公交何时能开通到站定位提醒功能?	城乡建设	0	L市公交何时能开通到站定位提醒功能

3.1.3 词云图

在分词后的结果的基础上，生成每个分类的词云，在每个分类中罗列前 100 个高频词，然后画出这些高频词的词云。其中两个分类的结果如下：

3.1.4TF-IDF 算法

在对职位描述信息分词后，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，把职位描述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

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:

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

或

$$\text{词频(TF)} = \frac{\text{某个词个词在文本中的次数}}{\text{该文本出现次数最多的词的出现次数}}$$

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。

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

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

TF-IDF=词频(TF)X 逆文档频率(IDF)

实际分析得出 TF-IDF 值与一个词在职位描述表中文本出现的次数成正比,某个词文本的重要性 TF-IDF 大。计算文本中每个词的 TF-IDF 值，进行排序，次数最多的即为要提取的职位描述表中文本关键词。

我们用 TF-IDF 来提取文本特征，可以用 python 中 TfidfVectorizer 这一工具，对所有样本分词，并且通过设置 N-gram 来获得特征，然后以这些词作为维度特征对每个样本向量化，最后到模型中训练。

3.1.5 朴素贝叶斯分类器

为了以比较效果，我们要用到分类器，，基于我们要分类的是基于词频的数据，在这里我们选择了朴素贝叶斯分类器。朴素贝叶斯分类算法是贝叶斯分类算法中最简单的一种，采用了“属性条件独立性假设”：对已知类别，假设所有属性互相独立，也就是在概率的计算时，可以将一个类别的每个属性直接相乘。这在概率论中应该学过，两个事件独立时，两个事件同时发生的概率等

于两个事件分别发生的乘积。 给定一个属性值，其属于某个类的概率叫做条件概率。对于一个给定的类值，将每个属性的条件概率相乘，便得到一个数据样本属于某个类的概率。我们可以通过计算样本归属于每个类的概率，然后选择具有最高概率的类来做预测。朴素贝叶斯分类器的原理如下：
我们由训练集可以计算出所有 p（c）类别概率和 p（x|c）以类别为条件下特征的概率，由

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

可以求出 p（x,c）联合分布概率，这种模型称为生成模型。所以我们说贝叶斯分类器是生成模型。
由于属性条件的独立性假设，各个特征的 p（x）的值应当是一个常量（虽然我们不知道具体是多少），而且 p（x）在训练集和测试集上应当一致，因此我们可以省略上式的分母。朴素贝叶斯分类器公式:

$$h_{nb}(x)=\arg\max P(c)\prod_{i=1}^{d}P(x_i|c)$$

贝叶斯分类器即用 p（c）和 p（x|c）来计算 p（c|x），计算时不除以分母，通过比较测试样本的每个类别的 p（c|x）大小，确定测试样本最有可能属于哪个类别。

3.1.6 机器学习模型

为了更好地评估它们的分类效果，我们分别尝试了逻辑回归，多项式朴素贝叶斯，线性支持向量机，随机森林这四种机器学习模型。

逻辑回归(Logistic Regression, LR)模型其实仅在线性回归的基础上，套用了一个逻辑函数，但也就由于这个逻辑函数，使得逻辑回归模型成为了机器学习领域一颗耀眼的明星，更是计算广告学的核心。它的直观描述是：

$$P（y=1|x;\theta)=\frac{1}{1+e^{-\theta^T x}} \qquad （1）$$

首先来解释一下 $P(y=1|x;\theta)$ 的表示的是啥？它表示的就是将因变量预测成 1（阳性）的概率，具体来说它所要表达的是在给定 x 条件下事件 y 发生的条件概率，而 θ 是该条件概率的参数。将它分解一下：

$$\theta=g(z)=\frac{1}{1+e^{-z}} \qquad （2）$$

（1）式就是我们介绍的线性回归的假设函数，那（2）式就是我们的 Sigmoid

函数啦。由于线性回归在整个实数域内敏感度一致，而分类范围，需要在[0,1]。逻辑回归就是一种减小预测范围，将预测值限定为[0,1]间的一种回归模型，其回归方程与回归曲线如下图所示。逻辑曲线在 z=0 时，十分敏感，在 z>>0 或 z<<0 处，都不敏感，将预测值限定为(0,1)。为什么会用 Sigmoid 函数？因为它引入了非线性映射，将线性回归(-∞,+∞)值域映射到 0-1 之间，有助于直观的做出预测类型的判断：大于等于 0.5 表示阳性，小于 0.5 表示阴性。其实，从本质来说：在分类情况下，经过学习后的 LR 分类器其实就是一组权值θ，当有测试样本输入时，这组权值与测试数据按照加权得到

$$h_\theta = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

这里的$x_1, x_2, x_3, \cdots\cdots x_n$就是每个测试样本的 n 个特征值。之后在按照 Sigmoid 函数的形式求出$P(y=1 \mid x;\theta)$，从而去判断每个测试样本所属的类别。

多项式朴素贝叶斯模型常用于文本分类，其基本原理如如下：

已知类别$C = \{C_1, C_2, C_3, \cdots, C_k\}$与文档集合$D = \{D_1, D_2, D_3, \cdots, D_n\}$，设某一文档$D_j$的词向量为$D_j = \{d_{j1}, d_{j2}, \cdots, djl_j\}$（可重复），设训练文档中出现的单词（单词出现多次,只算一次）即语料库为 V

对于待分类文档$A = \{A_1, A_2, \cdots, A_m\}$，则有：

第一步，计算文档类别的先验概率：

$$P(C_J) = \frac{\sum D_j \in C_j \mid D_j \mid}{\sum_{j=1}^{n} \mid D_J \mid}$$

P(C_i)则可以认为是类别C_i在整体上占多大比例(有多大可能性)。

第二步，某单词d_{jl_j}在类别C_i下的条件概率

$$P(d_{jl_j}) = \frac{\mid d_{jl_j} \mid +1}{\sum D_J \in \mid D_j \mid + \mid V \mid}$$

$P(d_{jl_j} \mid C_i)$可以看作是单词d_{jl_j}在证明D_j属于类C_i上提供了多大的证据。

第三步，对于待分类文档 A 被判为类C_i的概率，假设文档 A 中的词即OBJ相互独立，则有

$$P(C_i \mid A) = \frac{P(C_i \cap A)}{P(A)} = \frac{P(C_i)P(A \mid C_i)}{P(A)}$$

$$= \frac{P(C_i)P(A_1, A_2, \cdots, A_m \mid C_i)}{P(A)}$$

$$= \frac{P(C_i)P(A_1 \mid C_i)P(A_1 \mid C_i)\cdots P(A_m \mid C_i)}{P(A)}$$

对于同一文档 P(A)一定，因此只需计算分子的值。多项式模型基于以上三步，最终以第三步中计算出的后验概率最大者为文档 A 所属类别。

线性支持向量机处理的是线性不可分的数据集。对于线性支持向量机的优化问题，就是在线性可分支持向量机的基础上加了一个松弛变量。其学习的优化问题为：

$$\min_{w,b,\xi} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{N} \xi_i$$

$$s.t \quad \begin{cases} yi(wi+b) \geq 1-\xi i; i=1,2,\cdots,N \\ yi(wi+b) \geq 1-\xi i; i=1,2,\cdots,N \end{cases}$$

所求的的分类超平面和决策函数与线性可分支持向量机相同。

随机森林是一个包含多个决策树的分类器，并且其输出的类别是由个别树输出的类别的众数而定。随机森林可以处理大量的输入变数，可以计算各例中的亲似度，对于数据挖掘非常有用。

根据下列算法而建造每棵树：

第一步，用 N 来表示训练用例（样本）的个数，M 表示特征数目。

第二步，输入特征数目 m，用于确定决策树上一个节点的决策结果；其中 m 应远小于 M。

第三步，从 N 个训练用例（样本）中以有放回抽样的方式，取样 N 次，形成一个训练集（即 bootstrap 取样），并用未抽到的用例（样本）作预测，评估其误差。

第四步，对于每一个节点，随机选择 m 个特征，决策树上每个节点的决定都是基于这些特征确定的。根 据这 m 个特征，计算其最佳的分裂方式。

第五步，每棵树都会完整成长而不会剪枝，这有可能在建完一棵正常树状分类器后会被采用。

3.1.7 F1 分数进行评估

F1 分数（F1 Score），是统计学中用来衡量二分类模型精确度的一种指标。它同时兼顾了分类模型的精确率和召回率。F1 分数可以看作是模型精确率和召回率的一种调和平均，它的最大值是 1，最小值是 0。F1 分数，又称平衡 F 分数（balanced F Score），它被定义为精确率和召回率的调和平均数。

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

更一般的，我们定义 F_β 分数为

$$F_\beta = (1+\beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}$$

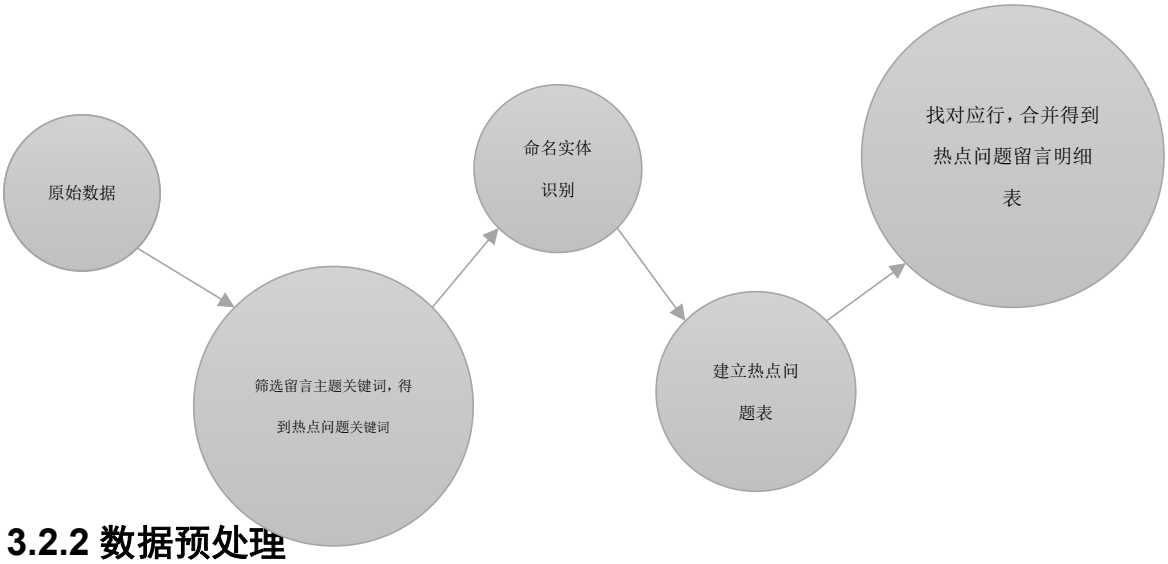
除了 F_1 分数之外，F_2 分数和 $F_{0.5}$ 分数在统计学中也得到大量的应用。其中，F_2 分数中，召回率的权重高于精确率，而 $F_{0.5}$ 分数中，精确率的权重高于召回率。

人们通常使用准确率和召回率这两个指标，来评价二分类模型的分析效果 F 分数被广泛应用在信息检索领域，用来衡量检索分类和文档分类的性能。但是当这两个指标发生冲突时，我们很难在模型之间进行比较。为了解决这个问

题，人们提出了 F_β 分数。F_β 的物理意义就是将准确率和召回率这两个分值合并为一个分值，在合并的过程中，召回率的权重是准确率的 β 倍。F_1 分数认为召回率和准确率同等重要，F_2 分数认为召回率的重要程度是准确率的 2 倍，而 $F_{0.5}$ 分数认为召回率的重要程度是准确率的一半。早期人们只关注 F_1 分数，但是随着谷歌、百度等大型搜索引擎的兴起，召回率和准确率对性能影响的权重开始变得不同，人们开始更关注其中的一种，所以 F_β 分数得到越来越广泛的应用。F 分数也被广泛应用在自然语言处理领域，比如命名实体识别、分词等，用来衡量算法或系统的性能。

3.2 问题 2 分析方法与过程

3.2.1 流程图



3.2.2 数据预处理

先将原始数据进行数据清洗，主要是要将按不同的，不兼容的规则所得到的各种数据集一致起来，筛选过滤的与留言主题无关的停用词数据，并且合理地处理缺失数据和异常数据。针对题目中所给的原始数据而言，各系统间的数据存在较大的不一致性，数据重复和信息冗余现象普遍以及由于数据丢失的不确定性造成的数据不完整，这将严重影响数据挖掘和建模的执行效率，可能甚至会导致数据结果存在偏差。因此，必须要对原始数据进行预处理，为数据挖掘内核算法提供干净，准确，更有针对性的数据，提供高质量的数据信息，提高挖掘效率。

3.2.3sklearn 中的 TfidfVectorizer 方法

为了计算留言主题的相似度，我们使用 TfidfVectorizer 生成 TF-IDF 向量，其步骤如下：
(1)使用 TF-IDF 算法，找出每个职位描述的关键词;
(2)对每个岗位描述提取的关键词，合并成一一个集合，计算每个岗位描述对于这个集合中词的词频，如果没有则记为 0;
(3)生成各个岗位描述的 TF-IDF 权重向量，计算公式如下:

TF-IDF=词频(TF) X 逆文档频率(IDF)

3.2.3HanLP 方法

HanLP 是由一系列模型与算法组成的 Java 工具包，目标是普及自然语言处理在生产环境中的应用。HanLP 具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。HanLP 主要功能包括分词、词性标注、关键词提取、自动摘要、依存句法分析、命名实体识别、短语提取、拼音转换、简繁转换等等。

本题采用 HanLP 方法（调用了 Java 库），基于感知机的中文命名实体识别，加载了一些与此文本有关的语料库，同时为了加强学习，为本题创建了感知机词法分析器，可进一步加强分词效果。

3.2.4 热点问题表

在前面的基础上，我们借用索引的方式找出了每个热点问题的时间，进行排序查看，同时根据每个热点问题的文本数来确定其热度，用于热点问题排序，然后得到热点问题表。下面是对应的时间的结果：

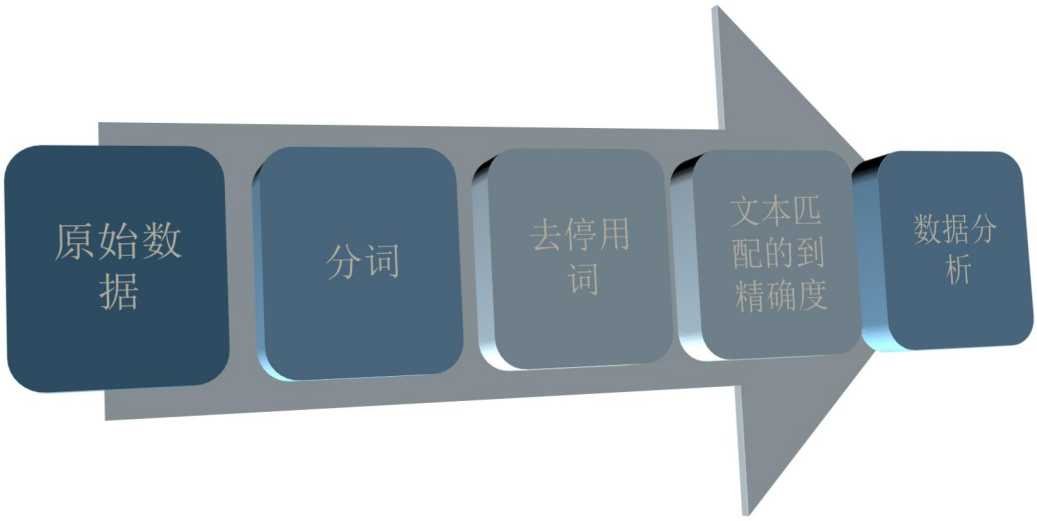
```
1617 2019-01-23 23:03:23
2281 2019-06-07 23:28:49
3535 2019-06-26 08:30:08
3349 2019-08-06 13:07:59
1604 2019-08-08 11:24:36
2787 2019-09-17 08:46:31
4054 2019-12-02 17:00:06
```

3.2.5 热点问题留言明细表

通过 iloc 函数找出热点问题对应的行，然后通过 concat 函数进行合并，得到表格。

3.3 问题 3 分析结果和方法

3.3.1 流程图



3.3.2 数据预处理

先要把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。在附件 4 的留言详情和留言主题中，以中文文本的方式给出了数据。为了便于转换，先要对这些留言详情和留言主题进行中文分词。这里采用 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 HM　M 模型，使得能更好的实现中文分词效果。为了得到适用于每行结果，针对每一行数据我们都进行了特殊的去停用词，我们根据每行数据的特点，区别对待。最终取得了适用于每行自己的去停用词。下面分别是留言主题和留言内容的分词结果：

Out[6]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
0	区景荟华苑	物业管理	有	问题	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	区潇楚	南路	洋湖	段	怎么	还	没	修好	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	请	加快	提高	市	民营	幼儿园	老师	的	待遇	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	在	市	买	公寓	能	享受	人才	新政	购房	补贴	吗	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	关于	市	公交站点	名称	变更	的	建议	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	区	含浦镇	马路	卫生	很差	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	区	教师	村	小区	盼望	早日	安装	电梯	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	反映	区	东澜湾	社区	居民	的	集体	民生	诉求	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
8	反映	市美麓	阳光	住宅楼	无故	停工	以及	质量	问题	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
9	反映	市洋湖	新城	和	顺路	洋湖	壹号	小区	路段	公共	绿化带	的	问题	NaN	NaN	NaN	NaN	NaN

Out[8]:

	0	1	2	3	4	5	6	7	8	9	...	3275	3276	3277	3278	3279	3280	3281	3282	3283	3284
0	现将	网友	在	平台	问政	西地省	栏目	向	胡华衡	书记	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	网友	您好	针对	您	反映	区潇楚	南路	洋湖	段	怎么	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	市民	同志	你好	您	反映	的	请	加快	提高	民营	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	网友	您好	您	在	平台	问政	西地省	上	的	留言	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	网友	您好	您	的	留言	已	收悉	现将	具体内容	答复	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	网友	您好	针对	您	反映	区	含浦镇	马路	卫生	很差	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	网友	您好	针对	您	反映	区	教师	村	小区	盼望	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7	网友	您好	您	的	留言	已	收悉	现将	有关	情况	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

3.3.3 文本匹配

对于文本的匹配，我们采用的是对留言主题的第一行的每个分词在答复详情相对应的行进行遍历，最终，精确度通过每行的匹配次数除以答复详情相对应行的分词数来确定。

3.3.4 数据分析

最后，进行数据分析，决定采用上四分位数，中位数，下四分位数作为分界点，来表示答复详情的好坏，并取了四个标题：优秀，不错，一般，差。

4. 结果分析

4.1 问题 1 的结果分析

4.1.1 一级标签分类模型

关于一级标签的分类模型，我们根据附件 2 给出的数据，在进行分词和提取关键词等操作之后，通用逻辑回归（Logistic Regression），多项式朴素贝叶斯(Multinomial NB)，线性支持向量机(Leaner SVC)，随机森林(Random Forest Classifier)这四种方法分别建立了一级标签分类模型并评估了它们的分类效果，其结果如下：

```
model_name
LinearSVC                0.813463
LogisticRegression       0.726385
MultinomialNB            0.651342
RandomForestClassifier   0.354950
Name: accuracy, dtype: float64
```

从上面的效果图我们可以发现，逻辑回归的分类效果最好。

4.1.2 对分类方法评价

为了衡量一级标签分类模型的精确度，我们用 F-Score 对该模型进行评价，计算得到了每个分类的准确率、召回率以及 F1-score 分数，其结果如下：

accuracy 0.8490131578947369

	precision	recall	f1-score	support
城乡建设	0.76	0.90	0.82	663
环境保护	0.91	0.81	0.86	310
交通运输	0.93	0.75	0.83	202
教育文体	0.90	0.89	0.89	525
劳动和社会保障	0.84	0.91	0.87	650
商贸旅游	0.87	0.76	0.81	401
卫生计生	0.90	0.78	0.83	289
avg / total	0.86	0.85	0.85	3040

从上图我们发现准确率和召回率呈负相关，一个高。的确，一般来说，精确度和召回率之间是矛盾的，所以我们无法通过准确率和召回率判断各个类别分类的效果。为了更好的评价一级标签分类模型的分类效果，我们选择通过 F1-Score 进行判断。我们发现，所有类别中 F1-Score 最高达到了 0.89，最低也有 0.81，从而我们可以认为我们建立的一级标签分类模型很好，能够起到提高分类效率和准确率的作用。

4.2 问题 2 的结果分析

4.2.1 热点问题表

本题研究主要的内容是将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，定义合理的热度评价指标，并给予出相应的评价结果，我们通过文

本相似度，命名体识别两部分入手，采用 sklearn 中的 TfidfVectorizer 方法，对留言主题这一列计算相似度，从而初步得到热点问题的一些关键点，进而采用的是 HanLP 方法（调用了 Java 库），基于感知机的中文命名实体识别，加载了一些与此文本有关的语料库，同时为了加强学习，从而创建了感知机词法分析器，可进一步加强分词效果，接着借用索引的方式找出每个热点问题的时间，进行排序查看，同时根据每个热点问题的文本数来确定其热度，最终我们得到以下的热点问题表：

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	53	2019-11-13到2020-1-5	A市万家丽南路丽发新城	附近搅拌站扰民，危害健康，影响人们的正常生活
2	2	44	2019-2-18到2019-12-3	A市	咨询各方面问题
3	3	42	2019-1-4到2020-1-6	A市	加快各方面建设
4	4	24	2019-7-7到2019-9-1	伊景园	捆绑销售车位
5	5	22	2019-1-3到2019-12-14	A市	公交路线，地铁，高速公路的建议
6	6	16	2019-2-15到2019-12-15	A7县，A5区，A1农业大学，A2区	麻将馆，广播，饭店扰民
7	7	16	2019-4-17到2019-12-6	A4区	夜间施工扰民
8	8	14	2019-1-21到2019-12-24	A3区	反映各方面问题
9	9	14	2019-1-18到2019-12-27	A市,A2区	公交车不准时
10	10	10	2019-1-11到2019-12-2	A市,A3区	请求解决各方面问题

从热点问题表我们可以看到，“附近搅拌站扰民，危害健康，影响人们的正常生活”这一问题热度高，
相关部门在处理群众反映的问题时，通过这一表格就能得到最高热度的问题，然后查看同一行的其他信息就能查看问题发生的地点和时间范围，这种有针对性的处理方法，极大的提升了服务效率。

4.2.2 热点问题留言明细表

热点问题的留言详情表，在得到热点问题表的基础上，通过 iloc 方法找出热点问题对应的行，然后通过 concat 函数进行合并，得到以下的表格:

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	188809	A909139	A市万家丽南路丽发新城居民区附近搅拌	2019/11/19 18:07:54	A市万家丽南路丽发新城居民区，开发商在小	0	1
1	189950	A909204	投诉A2区丽发新城附近建搅拌站噪音扰	2019/11/13 11:20	我是A2区丽发新城小区的一名业主，我要投诉	0	0
1	213464	A909233	投诉丽发新城小区附近违建搅拌站噪音	2019/12/10 12:34	我是暮云街道丽发新城小区的业主，我要投诉	0	0
1	214282	A909209	A市丽发新城小区附近搅拌站噪音扰民和	2020/1/25 9:07	你们管不管A2区丽发新城小区啊！这个附近建	0	0
1	231136	A909204	投诉A2区丽发新城附近建搅拌站噪音扰	2019/12/2 11:20	尊敬的领导，我是A2区丽发新城小区的一名业	0	0
1	239648	A909211	A市A2区丽发新城小区附近搅拌站明目张	2020/1/6 22:41	丽发新城小区附近近日突然建起了搅拌厂！将	0	0
1	243692	A909201	丽发新城小区附近的搅拌站噪音严重扰	2019/11/15 11:23	领导您好！我是暮云街道丽发新城小区的业主	0	2
1	253040	A909202	投诉A2区丽发新城附近建搅拌站噪音扰	2019/12/4 12:10	投诉A2区丽发新城小区附近违建搅拌站！该站	0	0
1	258378	A00084226	丽发新城社区附近搅拌站修建严重影响	2019/11/23 0:00	开发商在A市暮云街道丽发新城社区附近百米	0	0
1	259788	A909221	A市暮云街道丽发新城社区搅拌厂危害巨	2019/12/7 0:00	A市暮云街道丽发新城社区（丽发新城小区）	0	0
1	272361	A909242	丽发新城小区旁建搅拌厂严重扰民	2019/12/4 8:46	本人是丽发新城小区的居民，最近小区附近居	0	1
1	284485	A909222	黑心搅拌厂建在A市丽发新城小区附近	2019/12/18 0:00	A市A2区丽发新城小区近期百米范围内修建搅	0	0
1	190108	A909240	丽发新城小区旁边建搅拌站	2019/12/21 15:11	丽发新城小区旁边建的搅拌站几百米外就是小	0	1
1	213464	A909233	投诉丽发新城小区附近违建搅拌站噪音	2019/12/10 12:34	我是暮云街道丽发新城小区的业主，我要投诉	0	0
1	215563	A909231	A2区丽发新城小区旁边的搅拌厂是否合	2019/12/6 12:21	领导，您好！我相咨询A2区丽发新城小区旁边	0	0
1	217700	A909239	丽发新城小区旁的搅拌站严重影响生活	2019/12/21 2:33	开发商把特大型搅拌站，水泥厂从绿心范围内	0	1
1	231136	A909204	投诉A2区丽发新城附近建搅拌站噪音扰	2019/12/2 11:20	尊敬的领导，我是A2区丽发新城小区的一名业	0	0
1	233158	A909242	丽发新城小区旁建搅拌厂严重扰民!	2019/12/5 8:46	本人是丽发新城小区的居民，最近小区附近居	0	0

通过热点问题明细表，我们可以热点问题的详细信息，以便相关部门给出

针对性的解决方案，提高服务效率。

4.3 问题 3 结果分析

本题我们将从三个方面进行具体研究，分别从去停用词，文本匹配以及数据优劣分析。本题特色亮点在于去停用词，使用留言主题和答复详情这两列，但我们针对每一行数据都进行了特殊的去停用词，根据每行数据的特点，区别对待。最终取得了适用于每行自己的去停用词。进行数据分析，决定采用上四分位数，中位数，下四分位数作为分界点来评价答复意见的好坏，我们的得到了如下的表格：

	acc
count	2813.000000
mean	0.086709
std	0.063435
min	0.000000
25%	0.042553
50%	0.077778
75%	0.120000
max	0.444444

通过以上三方面的研究，对于本题的答复意见的质量的相关性，完整性和可解释性都得出了精准的描述，我们所研究的对于留言点问题的筛选，排列和保存均得出了效果最佳的研究方案，使智慧政务系统更加高效，智能。

5. 结论

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，这对于处理群众留言热点问题提供了更便捷，更高效的处理手段，这也对于政府了解民意有重大意义，但这同时也是文本分析的一个课题，一个难题。传统的文本解读已经不能满足数据庞大冗多的网络留言信息。本文采用 TF-IDF 方法，HanLP 方法等方法和模型统计归类留言热点问题，建立一级标签分类体系，深入分析和归类群众留言问题现状。统计新兴系统在服务方面不同领域的需求量，得出智慧政务系统更适合信息繁多和科技高度发达的当下社会，另外我们通过不断实验和评估以及对未来发展趋势的推断，可以看出智慧政务系统发展前景是更加出色的。

6. 参考文献

[1]张琳,李朝辉.文本分类中一种改进的特征项权重计算方法[J].福建师范大学学报(自然科学版),2020,36(02):49-54.
[2]黄春梅,王松磊.基于词袋模型和 TF-IDF 的短文本分类研究[J].软件工程,2020,23(03):1-3.
[3]宋欣燃.大数据时代的数据挖掘技术与应用探析[J].中国新通信,2020,22(05):109.
[4]叶章浩.基于决策树的数据挖掘技术在就业信息管理系统中的应用研究[J].科技传播,2019,11(16):132-134.