“智慧政务”中的文本挖掘应用

摘　要：

随着信息时代的发展，各种问政渠也迅速开通以便利人民市场问题反馈。各类与社情民意相关的文本数据量不断攀升,同时不可避免地对相关部门的工作带来了极多的困扰。在相关问题的处理中，人工处理会花费大量时间，造成资源的浪费。一次有一套完整的面多问政平台事件处理的算法，则是可以节约人力、物力的有利条件之一。

本文完成了问政事件类别的基本分类，以及高热度事件的提取并制定了一套关于留言回复的完整评价系统。经过长时间的修改和调试，最终取得了较为准确的分类效果。本文第一部分简单介绍了论文的整体框架并介绍本题背景，说明挖掘目标并列举了当前国内外在这些方面的研究现况；第二部分详细的描述了本次语言处理的大致流程以及对附件中数据进行简单基本分析；第三部分详细描述了针对留言分类，热点问题提取，留言回复等算法的具体实施步骤，以及其中运用的算法的详细介绍。在 3.2.1 中，我们对数据在分词、去停用词、去重等方面进行了预处理并对停词表进行了分析和修改以增加分类精确度。

在 3.2.2 中我们还通过绘制词云图等观察数据以精确模型。在本题中，我们针对留言分类采用了 TF-IDF 权重策略以及朴素贝叶斯模型，因为相对其他算法而言 TF-IDF 模型实现简单，结果也比较符合实际情况。在 3.3.3 中，我们详细描述了在本题中才用的热度计算方法，并根据实际数据对方法进行了改善。最后使用 gensim 库中的方法计算稀疏矩阵的余弦值得出留言事件和留言回复的相似度，来确定回复的完整性，可解释性，相关性。

本文所设计的算法较为准确地完成了留言分类和留言归类等要求，其中涉及的方法是在对比了其他模型后确定的。停词表也进行了大量的修改，最后得出的分类结果和归类结果精确度也较好。

关键词：TF-IDF 模型 朴素贝叶斯模型 稀疏矩阵 停词表

目录

1. 挖掘目标

1.1 挖掘背景

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，各类与社情民意相关的文本数据量不断攀升，同时不可避免地对相关部门的工作带来了极多地困扰。对相关问题的处理时根据以往的手工操作，不仅花费的时间多，也会出现各种各样的问题，难以准确把握民意导向。因此题需要根据网络公共资源中的群众问政留言记录，及相关部门对部分留言的答复意见的文本数据中建立有效的数学模型对数据进行留言的分类，热点问题的挖掘和对于其中答复意见的评价。

成功的执政者，无一不把民心作为治理的重要基础，把社情民意作为重要的决策依据，采取各种方式及时获取信息，使政策措施与民心民意相一致。如今，大数据互联网的不断发展，使国家获取社情民意有了不一样的途径，也面临新的挑战。准确预测、预警、科学预测民意走向是时代的需要。因此，基于各种网络政治平台的数据，建立一个综合的模型，以信息分类、热点问题挖掘和相关回复评价为基础，对一个国家获取民意走向具有重要意义。

1.2 挖掘目标

根据题目中给出的附件 1 给出的问政平台划分体系和附件 2 中的群众留言详情进行深入分析梳理。挖掘出附件 2 中群众留言的重点内容和核心问题，研究留言内容中关键词与问政平台划分体系的关系模型，并检测模型的可靠性，计算模型的准确率。其次根据给出的群众留言详情，定义热度评价指标并统计出在某段特定时间内群众反映的热点问题。最后根据附件 4 中相关部门对群众留言的答复意见，设计一套可行的模型。并对其在相关性、完整性、可解释性等方面进行评价。

1.3 研究现状

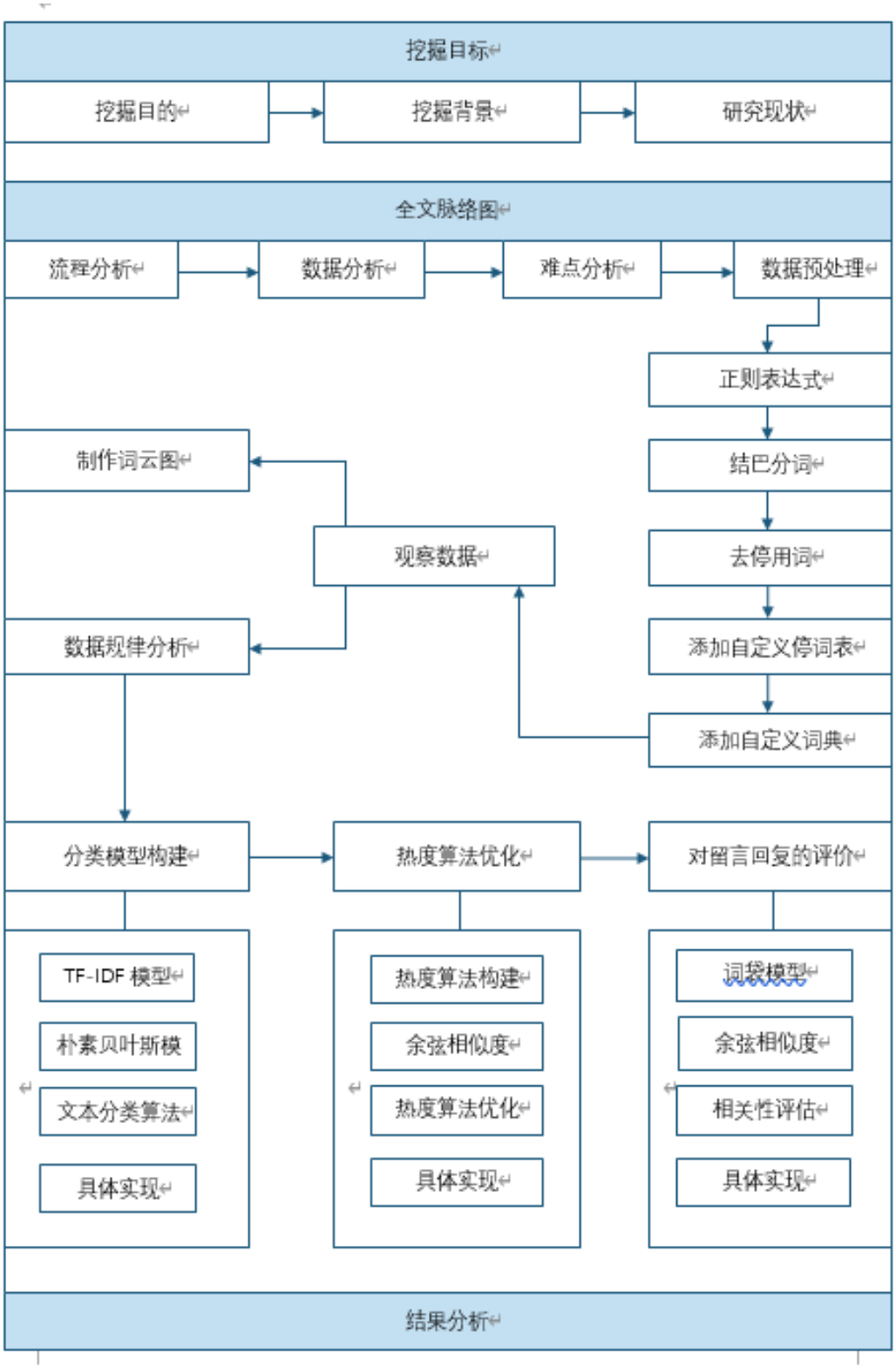
国内对于文本分类的研究比较晚。很大一部分原因是因为中文文本的分类相对于英

文更加的困难。

1981 年，侯汉清教授首先进行了文本分类的研究，之后也有很多国内的研究者对于文本分类进行了研究与实践，例如 1981 年侯汉清对文本分类工作进行了深入的探讨，2001 年,卜东波、白硕比较和分析文本分类系统的三种关键技术[1],2012 年平源对于支持向量机分类器的研究与应用[2],2019 年郭超磊进行了基于 SA-SVM 的中文文本分类研究等等。因为中文文本的特殊性,在做分类处理都过程时一般需要满足某种形式的分词。分词工具的选择尤为重要，快速发展的分词技术让我们有比较多的选择,如开源的 HanLp、jieba、FudanNLP、LTP、THULAC、NLPI 等。商业用途的阿里云 NLP、腾讯文智、百度 NLP、BosinNLP 等都是技术较为成熟的分词工具,而在中文文本分类也应用了很多较为成熟的算法。如决策树、K 近邻分类算法、随机森林、朴素贝叶斯分类、神经网络分类算法、支持向量机算法等。国外对于文本分类的研究开始于 1950 年,通过人工去定义标准确而对文本进行分类,是早期文本分类的方法,这样的方法既耗费时间,也需要专家对某一领域的知识有充分的了解,才可以定义出合适的规则。随后 H.R.Luhn 将词频计算这一思想引入文本分类中,给予了各个研究者一定基础，一时间内很多经典的文本分类数学模型被提出，如 Salton 提出了利用空间向量模型对文本进行描述，H.Borko 等人提出了利用因子分析法对文本进行自动分类等。[3]

2. 全文脉络图

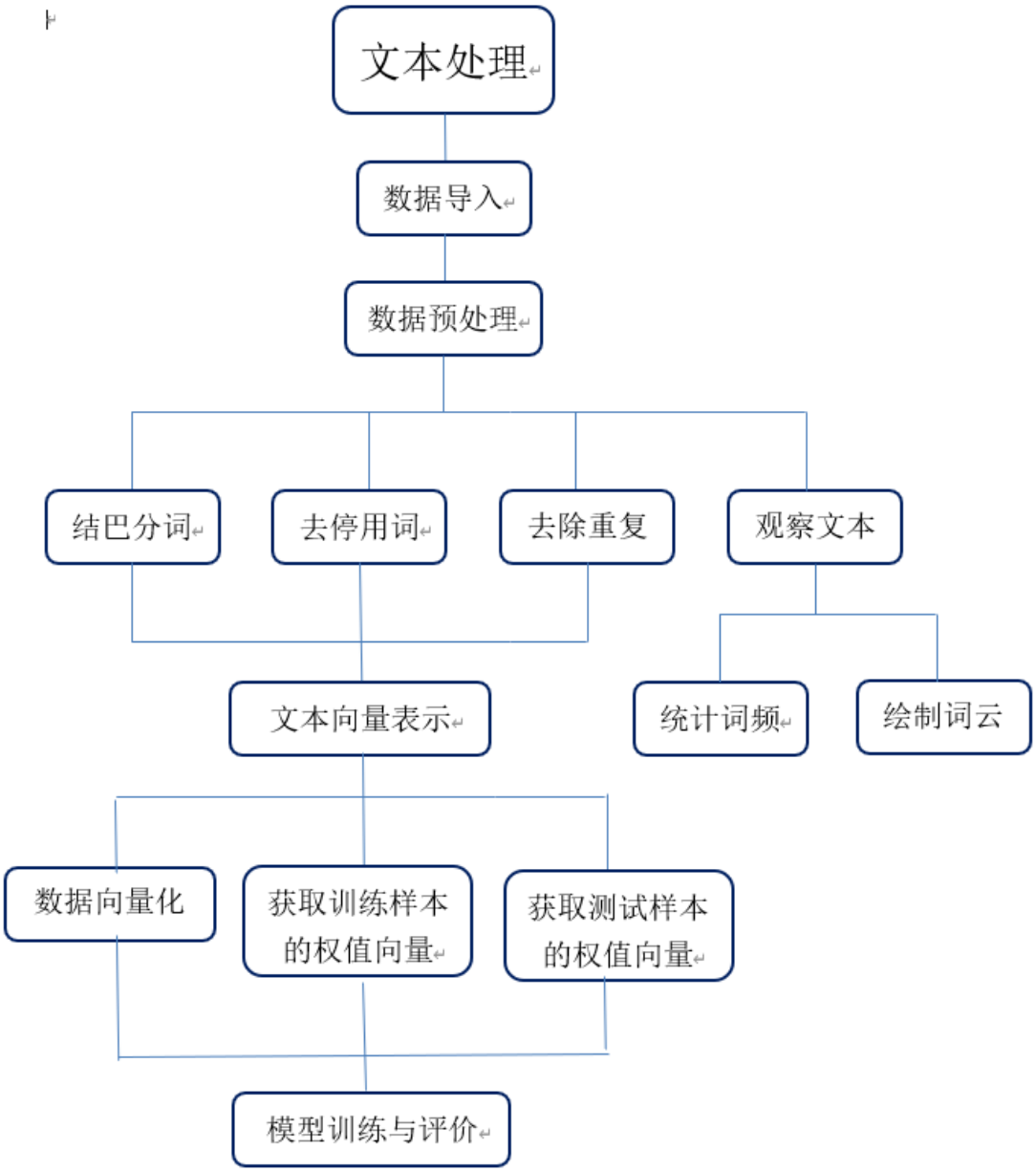
本次文本处理主要流程如下：



3. 分析方法与过程

3.1. 分析流程

3.1.1 总体流程分析



3.1.2 数据分析

获得的数据共有 4 个附件。附件 1 给我们提供了一种内容分类三级标签体系，附件 2 的数据提供了用户留言时的编号、ID、主题、时间、详情和该留言所属的一级标签，附件 3、附件 4 则和附件 2 类似，不同点在于附件 3 的数据于附件 2 除去一级标签内容后添加了该条留言获得的点赞数和反对数，而附件 4 的数据则是于附件 2 除去一级标签内容后添加了对于每条留言内容有关人员相应的答复意见及答复时间。在所给与的附件中我们可以了解到三级标签体系中各二三级标签所属的一级标签，利于更加精确的分类。此外用户所留言的内容中涉及到关于城乡建设、党务政务、劳动和社会保障等 15 类的内容,在留言消息获得的点赞数和反对数中一定程度的了解该条留言内容反应的问题受众大小及人们所持的意见，在留言答复意见的内容和时间中反映相关部门面对此类问题所解决的方式，是否能够解决问题及与留言内容是否有关系。

3.1.3 难点分析

1.题目的要求是对于群众留言的分类，寻找热点问题，对答复进行评价等，而在题目给予的数据内容里包含着部分对于题目解决没有明显作用的字符序列，如类似特殊符号和电话、空格、时间等格式的数据，因此在将数据内容进行后续的处理前如何将数据进行一定程度的清洗，获得较为干净的数据，使后续对数据的分析的结果更为准确是开始处理数据后的第一个难题。

2.由于题中的数据均为群众对政府进行问政的内容，因此在对数据进行分词操作时，如何在词典比较完整的添加一些与政府日常政务有关的出现频率较高的词汇，是精确关键词选取时的准确率提升模型的泛化能力过程中不可避免地一个点。

3.在对数据内容进行分词处理后，观察数据时可以明显发现分词过后的数据中含有很多的功能词既停用词，这些词往往应用的范围非常广泛或自身没有明确的意义，并且很少单独表达文档相关程度的信息。例如“的”、“里面”、“在”等。因此如何去建立一个比较完整并且适用本题目的停词表，来保证文本中关键词的密度，使关键词更为突出是去停用词中一个难点。

4.在对比赛数据的特征进行深入的了解后，结合数据特征选择模型是较为关键也是较为困难的一步，因为模型选择也说明如何将文本数据进行预处理并且向量化，应该选择什

么类型的模型，配置什么样的参数。多样的解决方案为问题不仅扩大了解决问题的范围还增加了大量的复杂度。因为这并不是简单的哪个模型复杂、准确率高而选择哪个的问题。因此如何在各类数学模型中找到对于问题解决较优的解，以找到匹配数据集的模型配置是问题中间较难的一个点。 5.因题目中要求对群众留言内容进行分类，而题目给予的数据集并不是理想中的那般均匀分布，数据不平衡对于训练出来的模型有着一定程度上的影响，例如有两个类===x 类、y 类，分别占数据集的 80%、20%。而在训练集上都是 x 类的实例，这样每次预测时 x 类的准确度可以到达 80%，但是这很可能是一个没有用的分类器，因此我们如何根据数据分布提取提取训练集使各类样本在训练的过程中有着相同的话语权，保证模型的学习达到更好的效果是解决问题不可避免地一个难点。

3.2. 具体步骤

3.2.1 数据预处理

分词是自然语言处理的基础，分词的准确性直接影响着后续的文本分析的质量。与英文不相同的是，中文的词与词之间没有分隔符，需要我们自行分句、分词。因此在进行中文自然语言处理时，我们需要先进行分词。目前的分词方法有 Jieba 分词、SnowNLP 分词、PkuSeg 分词、THULAC 分词、HanLP 分词。其中我们预选 thulac 分词和 jieba 分词这两个工具。经过对测试结果的对比，我们发现结巴分词的颗粒度较细，比较适合我们本次文本处理。而 thulac 分词对一些人名的识别要强于 jiaba 分词。同时为了优化分词效果，我们还加入了自定义的词典，以精准分词。精准的分词将会对后续的文本表征以及聚类等产生更好的作用，从而提高精确度。

停用词可以节省存储空间并提高搜索效率，某些单词或符号在处理自然语言数据之前会自动过滤掉。这些符号或词语称为停用词。停用词不会自动生成，生成的停用词将于用构建停用词表。本次文本处理的停用词分为两部分，一部分去次网络上的停用词表，里面都是一些普遍的中文词语及符号，另一部分则有较高针对性，需要通过对文本进行观察统计并经过人工筛选后填入停用词表。

经过对数据的观察，我们发现在问政平台有部分数据出现重合现象，这会影响我们的文本分类精确度。因此我们应该对数据进行去重。在本次实验中，我们采用的去重方法是 pandas 的 drop_duplicates 方法，此方法可以将 dataframe 中存在在重复的行或者

几行中的某几列的重复值去掉。因为我们需要将重复的数据保留一行，所以讲 drop_duplicates 中的 keep 参数设为 first。

3.2.2 数据规律分析

在正式开始文本处理前我们对文本进行了词频统计，绘制词云图，可以更加直接的观察到数据的词频，以下是我们在没有把人工筛选的数据加入停词表时的词频情况：

从中可以看出如果只使用网上获取的停词表，会出现许多不重要的词语没有去除的情况。因此我们需要反复测试数据来筛选出一些对实验不重要的词语，以增加精确度。并且，我们可以在停词表中看到，有许多语句的分词并不准确。并且，由于文件中的文本数字均来自问政平台，文本词语较有针对性，因此应针对的去除某些政务问题中的提及频率高但对文本分析并无实际意义甚至会产生干扰的词语进行去除。经过了这些改善，会大大精确我们的分词效果，并提高文本处理效率。

3.2.3 分类模型构建

一、使用到的方法

（一）TF-IDF 权重策略

TF- IDF（Term Frequency- Inverse Document Frequency）即“词频-逆文本频率”是文本分类中一种经典计算特征权重的方法，由 TF 和 IDF 两部分组成。

TF 表示词频，统计文本中每个词出现的频率。其计算公式为如下公式（1）。

$$\mathbf{TF(x)} = \frac{\mathbf{n(x)}}{\sum \mathbf{n}} \quad (1)$$

▲ n(x)表示特征词 x 在文本中出现的次数

▲ ∑n表示文本中所有特征词的个数

IDF 表示逆文本频率，用来衡量某个词在语料库中所有文本中的频率，反映词语在文本中的重要性。对于 IDF 值，如果一个词在多个文本中出现，它的 IDF 值越低；反之，它的 IDF 值更高，这说明这个词会更重要性。其计算公式如下公式（2）。

$$\mathbf{IDF(x) = \log\frac{N}{N(x)}} \quad \mathbf{(2)}$$

▲ N 表示语料库中文本数量

▲ N(x) 表示含有特征词 x 的文本数

为防止某个词语在语料库中不存在即分母为 0，可以采取平滑，使语料库没有出现的词也能得到一个合理的 IDF 值，需要对公式（2）进行改进得到如下公式（3）。

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

再根据公式（1）、公式（3），得到 TF-IDF 值的计算公式如下公式（4）。

$$\mathbf{TF - IDF = TF\,(x) \times IDF\,(x)} \quad \mathbf{(4)}$$

▲ TF (x) 表示特征词 x 在当前文本中的词频

从上面公式（3）可以看出，TF-IDF 的值是由词频 TF 和逆向文本词频 IDF 的乘积得到的。TF-IDF 值越大，则表示该特征词对这个文本的重要性越大。

（二）朴素贝叶斯模型

朴素贝叶斯分类算法是贝叶斯分类算法的一种，它基于在给定类别值的条件下，各特征属性值间是相互独立的，利用类别的先验概率和样本信息通过贝叶斯公式计算未知文本属于某一类别的后验概率，最大的后验概率即为文本分类的判别结果。朴素贝叶斯分类的原理如下。[4]

如果 X 和 Y 相互独立，则有：

$$\mathbf{P(X,Y) = P(X)P(Y)}$$

因为全概率公式：

$$\mathbf{P(X) = \sum_k P(X|Y = Y_k)P(Y_k)}$$

▲ P(X) 条件 X 出现的概率，属先验概率

▲ $\mathbf{P(Y_k)}$ 在不考虑相关因素的情况下，随机事件出现 Y 情况的概率，属先验概率

▲$\mathbf{\sum_k P(Y_k) = 1}$

可得到贝叶斯公式：

$$\mathbf{P(Y_k|X)} = \frac{\mathbf{P(X|Y_k)P(Y_k)}}{\sum_{\mathbf{k}} \mathbf{P(X|Y = Y_k)P(Y_k)}}$$

▲ $P(X|Y_k)$ 已知事件出现 Y 情况的条件下，条件 X 出现的概率，属后验概率

▲ $P(Y_k|X)$ 在条件 X 下的概率，随机事件出现 Y 情况的概率，属后验概率

（三）模型优缺点

1. TF-IDF 模型

优点：实现简单，相对容易理解，结果符合实际情况。

缺点：高度依赖语料库，需要选择质量较高的，并与需处理的文本匹配处理后的语料库进行训练。并且仅仅通过每个词语的频率来衡量每个词的重要性还不够全面，有时候重要的词语不会经常出现，并且没有设置词语出现的条件，因此出现在前面位置的单词与出现在后面位置的单词一样重要，权重可以较为重点放在全文的第一段或每一段的第一句。

2. 朴素贝叶斯模型

优点：朴素贝叶斯算法假设了数据集属性之间是相互独立的，因此算法的逻辑性十分简单，并且算法较为稳定，当数据呈现不同的特点时，朴素贝叶斯的分类性能不会有太大的差异。换句话说就是朴素贝叶斯算法的健壮性比较好，对于不同类型的数据集不会呈现出太大的差异性。当数据集属性之间的关系相对比较独立时，朴素贝叶斯分类算法会有较好的效果。

缺点：属性独立性的条件同时也是朴素贝叶斯分类器的不足之处。数据集属性的独立性在很多情况下是很难满足的，因为数据集的属性之间往往都存在着相互关联，如果在分类过程中出现这种问题，会导致分类的效果大大降低。[5]

二、进行分类前的数据处理

（一）文本特征选取及向量表示

1. 特征选取及向量表示

特征选择指对现有的特征空间筛选重要的特征重新组成新的特征集，能有效提高文本分类的准确率。特征抽取指对当前特征空间进行变换压缩生成新的语义空间，可以一定程度上解决词语歧义问题，降低维度。[6]文本分类的流程：文本预处理，抽取文本特征，构造分类器。其中最关键部分为文本特征的抽取。为了降低文本表示的维度，我们要对特征词进行选取，选取出重要程度高的关键词，为了让我们训练数据质量提高，并使得最终训练的效果更好。目前常见的基础的文本特征算法有信息增益、开方拟合检验、潜在语义分析、期望交叉算熵以及计算文档频率等。由于在文本处理中，每个词在不同文本中的出现的次数是不一样

的，即每个特征在文本中的重要程度是不一样的，因此我们需要考虑每一个特征项的重要程度，根据这种情况，我们引进了特征词权重算法。常见的特征词权重算法有 TF 的改进、信息熵的引用、TF-IDF ，根据比较和测试，我们选用了 TF-IDF 权重策略进行本次的特征词权重计算。对于文本中的每一个特征即每一个词，它都有一个 TF-IDF 权值，计算得到的权值越大则这个词的越重要，反之，这个词的重要性就小，而对于权值较小的词，我们将剔除掉。而对于一条文本而言，它会有一个 TF-IDF 权值向量，我们需要把文本数据用数值型的值表达出来，然后进行 TF-IDF 权值向量表示及计算，并最终得到我们所需要的样本数据。

2. 具体实现：

①加载得到的文本数据集，并将其进行数据集划分。

②实例化 CountVectorizer 类，通过词频向量转换器，创建词袋数据结构。

③将文本数据转换为词向量即词频矩阵，计算词汇在该训练文本数据中出现的频率。

④将上面得到的结果稀疏矩阵，通过 TF-IDF 模型进行权值计算，并得到关于 TF-IDF 值的稀疏矩阵对象，并最终转成训练集的样本数据。

⑤维度共享，共享词库，使得训练样本和测试样本的维度保持一致，指定词典为创建的词库，将文本数据转换为词频矩阵，进行 TF-IDF 模型计算，并将测试样本数据转换成向量矩阵。

（二）文本分类算法

1. 分类算法与模型选择

在文本分类方面，目前实现的方法大致分为两类，一类是基于传统机器学习的文本分类，另一类是基于深度学习的文本分类，不同的分类选择，对数据的处理和实现不同，对数据处理的结果也有不同，因此在分类算法的选择上面，我们需要对不同文本分类算法进行大致的了解及分析，最终得到适合我们文本数据的算法和模型。其中基于传统机器学习的文本分类有朴素贝叶斯模型、随机森林模型、SVM 分类模型、KNN 分类模型、神经网络等模型，而基于深度学习的文本分类的算法模型有本卷积神经网络、文本循环神经网络、循环卷积神经网络、动态记忆网络等文本分类算法模型。经过分析和比对，我们选择了基于传统机器学习的文本分类算法模型——朴素贝叶斯模型。

朴素贝叶斯分类器（NBC），是以贝叶斯定理为基础的简单概率分类器，其通常有三种实现方式，一是基于多项式模型实现，二是基于伯努利模型实现，三是基于高斯模型实现。其中，由于多项式模型会考虑词语在文档中出现的次数，考虑到词语的词频问题，因此多项式模型主要用于文本的主题分类，而伯努利模型不会考虑词频，只考虑这个词有没有出现，

因此其主要用于文本情绪分析。[5]区别于多项式模型适合用于特征属性离散的情况和伯努利模型适合用于缺失值较多的情况，高斯模型可以解决特征为连续型变量的情况，在此模型中将假设每一维特征的所有属于某类别的观测值都服从高斯分布。经过测试和对比，由于文本中很多特征为连续值型数据，因此我们选择了高斯朴素贝叶斯模型。

2. 具体实现

①确定特征属性后，由人工实现对每个特征属性进行适当的划分，得到训练样本集合和测试样本集合，即得到有待进行分类的文本数据和进行测试验证的文本数据。

②训练，生成分类器即在此程序中的实例化高斯朴素贝叶斯模型，由程序实现利用分类器计算各不同类别在训练样本集合中出现的频率和各不同特征属性划分对不同类别的条件概率估计。

③测试，使用分类器对有待分类数据进行分类，由程序内部实现并最终得到的后验概率即为文本分类的判别结果。

3.2.4 热度算法优化

符号说明

Count:对某件事情的群众反应条数

Comment：回帖数，即点赞数+反对数

Time_sub:事件的讨论起始时间到结束时间的天数

G: 重力加速度,它的数值大小决定了排名随时间下降的速度快慢

$$score = \frac{(count + comment)}{time\_sub^{G}}$$

从公式中我们可以获知影响热度排名的因素总共有 4 个。首先是 count，即同一件时间的反应次数，count 越大，热度指标越高。其次还有 comment，回帖数，人们的回帖数越多则证明人们对该事件的关注度越高。由此讨论次数，回帖数与热度排名为正相关。第三个因素，反应时间段，反应时间段在其他条件不变的情况下，讨论时间越长，排名越低。或者说，一个事件的排名，会随着时间不断下降。最后一个因素是 G，它的数值大小决定了排名随时间下降的速度。

3.2.5 对留言回复的评价

按照题目要求，对附件 4 中每条留言信息的“留言详情”与“答复意见”的具体内容转化为计算机可以识别的向量化数据，通过提取的特征向量计算两个文本向量之间的余弦相似度获得其每一条留言信息中”留言详情”与”答复意见”的相关性大小，依据余弦相似性的能够体现在方向上的差异的特点，可以相信余弦相似度大的留言信息其答复意见也相应的具有一定的完整性和可解释性。因此通过余弦相似度对相关部门答复意见的质量做出评价。

一、概念

（1）词袋模型(Bag of words)

词袋模型是能够有效的描述文本特征的数学模型，它的本质是将文本向量化，每个词都与文本中的其他词都互不相关，忽略文本中的语序和语法，并将出现的词的频率作为权重。

优点：词袋模型能够从文档中提取特征词，并使用特征项矩阵对它们进行建模。这允许将每个文档描述为一个单词包。而且只需要记录单词的数量，语法和单词的顺序就可以忽略不计。

缺点：词袋模型对相似词之间的表达有着较为严重的缺陷，例如“他不喜欢哈尔滨”和“他喜欢哈尔滨”这两个文本表达的意思是完全不一样的，但在词袋模型之中会被认为高度相似。

（2）余弦相似度

在进行文本相似度分析时，通常把文本特征转化为向量的形式。两个向量的夹角越接近 0，则两个向量在同一方向上的距离就越近余弦值越接近 1。两个向量的夹角越接近 180 度，则两者相反余弦值为负 1。如果两个向量之间的夹角是 90 度，它们就不相关。对于基于词袋模型构建的向量，词项出现的频率不小于 0，因此两向量之间的余弦值范围在 0 到 1 之间。所以可以通过计算文本特征向量之间的余弦值来表示这两文本之间的相似度[7]。

二、具体实现

（1）对文本集进行分词，删除停用词（哈工大停用词表），并将文本集生成分词列表。

（2）（建立词袋模型）基于“留言详情”的每条具体内容建立词典，获取词频向量并提取词典特征数后，利用词典将分词列表集转换为稀疏向量集生成语料库。

（3）使用“TF-TDF 模型”处理语料库。

（4）对稀疏向量建立索引，同时将“答复意见”的每条内容也转换为稀疏向量。

（5）计算两个向量之间的余弦相似度。

三、自定义答复评价指标

根据计算所给数据中的答复意见与留言详情之间的相似度，得到留言答复质量的评估指标。

①相关性：通过取整个答复系统与留言间的相似度的平均值，分析得到相关值已达 50%。

②完整性：统计计算出来的相似度为 0 的留言的答复，通过观察和计算得到可以得到，对于已经解决的问题，和分配给相关部门解决的问题，答复系统中都能做到较完整的回复。

结论：对于所给留言的答复中，答复系统拥有较高质量。对于一些未能很快解决是事件留言，也给出了相关的答复。

4.结论

本次比赛，我们根据附件 1 提供的内容分类三级标签体系和附件 2 收集自互联网公开来源的群众问政留言记录的数据特点，针对数据中的留言内容，利用 TF-IDF 权重模型结合朴素贝叶斯中的高斯朴素贝叶斯模型进行多分类问题的分类模型构建，并通过 F-Score 对此次程序的分类方法进行评价；通过分析附件 3 关于留言主题数据和留言详情数据的特点，经过数据预处理等过程，利用 TF-IDF 权重策略、Bow 模型和 LDA 主题模型等方法对数据进行分析、挖掘，结合 Gensim 文本摘要自动生成算法对数据进行相似度计算，统计分析得到关于某一时间段内反应特定地点或特定人群问题的留言归类，并根据自定义热度评价指标对数据进行分析和评估得到经过排序后的排名前五的热点问题及其分别得到的热度指数，最终得到表 1 -热点问题表和表 2 －热点问题留言明细表；根据附件 4 留言详情和相关部门对群众留言的答复意见数据的特点，结合相似度算法，通过自定义指标对数据中的答复意见进行评价，分析得到留言记录中留言回复对留言的相关性以及完整性，并从中评估获得对于留言答复意见质量。实现本次的挖掘目标：建立基于自然语言处理技术的智慧政务系统，提升政府的管理水平和施政效率，并

对相关部门对群众留言的答复意见的质量提供一套评价方案。

通过这套文本挖掘算法，可以通过对留言进行分类，以便后续将群众留言分派至相关的部门处理，解决人工处理时遇到的工作量大、效率低、差错率高等的问题，并通过本套算法，及时得到热点问题信息，及时发现热点问题，有助于相关部门进行有针对性的处理，提升服务效率，制定有效的答复质量的评价方案，可以提升政府部门对相关留言问题的处理效率和施政效率。总而言之，就是让网络问政平台真正成为政府了解民意、汇聚民智、凝聚民气的重要渠道，最终达到真正的权为民所用，利为民所谋，通过自然语言处理和文本挖掘技术对留言进行划分和对热点问题进行整理，让相关部门整理和处理群众留言事件得到便利，提升政府的管理水平和施政效率。

通过这次的比赛，让我们认识到了团队的重要性，并通过大家的讨论和学习让我们对关于 Python 语言知识的理解加深，让我们对 Python 语言更加感兴趣。本次比赛锻炼了自己的动手能力，增长我们的实践能力，并对之前所学的理论知识有了一定质量上认识和理解。在这次比赛中，我们查阅大量资料，学习理论，实现代码，编写论文，让我们开拓眼界，锻炼自己的思维能力。世上无难事只怕有心人。我们三位成员付出了许多心血和时间，只为在过程的最后能给自己交上一个满意的答卷。坚信未来的我们会越来越好。

参考文献

[1] 庞剑锋，卜东波，白硕. 基于向量空间模型的文本自动分类系统的研究与实现[J]. 计算机应用研究, 2001, (09): 23-26.

[2] 平源. 基于支持向量机的聚类及文本分类研究[D]. 北京邮电大学, 2012.

[3] 朱梦. 基于机器学习的中文文本分类算法的研究与实现[D]. 北京邮电大学, 2019.

[4] 张聪慧. 朴素贝叶斯分类算法在提升电信客户满意度方面的研究应用[J]. 科技视界, 2019, (02): 122-123.

[5] 马刚. 朴素贝叶斯算法的改进与应用[D]. 安徽大学, 2018.

[6] 石凤贵. 基于 TF-IDF 中文文本分类实现[J]. 现代计算机, 2020, (06): 51-54.

[7] 王志刚，谢恺，朱慧. 降成本政策的文本分析——基于文本相似度计算原理[J]. 地方财政研究, 2020, (03): 90-97.