C 题：“智慧政务”中的文本挖掘应用

摘要

本文研究文本挖掘和自然语言处理在“智慧政务”中的实际应用问题。

近年来，随之智慧政务践行，影响力剧增，越来越多市民参与到智慧政务中来，这样就产生了大量的信息数据，给我们原先依靠人工来进行留言分类和热点整理的工作带来极大挑战。处理数据效率、准确性、群众的满意度都成为了智慧政务能否进一步发展的制约因素。

总体上是数据挖掘中的文本数据挖掘，要用到 NLP 自然语言处理技术，来达到使用计算科学高效、准确处理一些信息化时代人工所不能信息快速处理。同时根据题目中相关小题具体分为三个方向:

第一小题，群众留言分类问题:本文将建立基于机器学习、逻辑回归以及朴素贝叶斯的一级标签分类模型和分类 F-Score 评价模型。

第二小题，热点问题挖掘问题:本文将建立有关基于文本数据挖掘的时间地点提取和 K-means 文本聚类定义热点挖掘模型。

第三小题，答复意见评价问题:本文将针对文本数据相关性、完整性、可解释性三个评价指标分别建立基于机器学习数据相关系数计算的文本数据相关性系数模型，基于机器学习的文本数据结构完整性评价模型，基于机器学习可解释性工具和深度学习的文本数据可解释性评价模型。

同时我们提取时间比较得出有关时间差和评分的函数，作为时效性评价模型，数据并根据四个模型、主管逻辑分析和数据的定义，使用了组合赋权法给予三个模型得到附件一定的权重，共同构建了答复意见评价模型。

我们使用 Excel、Python 以及 Python 的拓展库、第三方库，以及 C 程序 R 程序拓展包完成模型的构建。

关键词：自然语言处理、文本分类、K-means 文本聚类、VSM 向量积、逻辑回归模型、朴素贝叶斯

目录

1 问题的重述

1.1 问题背景:

近年来，互联网技术渗透到生活的方方面面。“互联网+”、智慧化成为了普遍的事物发展方向，政府部门为刚好收集民情，了解民意，汇聚民智，凝聚民气，提高服务人民的工作效率与解决问题的效率，走上智慧政务发展道路，充分利用数据挖掘、大数据分析、人工智能、物联网、云计算等新一代信息技术，以用户创新、大众创新、开放创新、共同创新为特征，实现政府和公民的双向互动，成为社会治理创新发展新趋势，营造更好更透明更公平的政治环境。

同时随之智慧政务践行，影响力剧增，越来越多市民参与到智慧政务中来，这样就产生了大量的信息数据，给我们原先依靠人工来进行留言分类和热点整理的工作带来极大挑战。处理数据的效率、准确性，以及群众的满意度都成为了智慧政务能否进一步发展的制约因素。

1.2 需要解决的问题：

第一小题：使数据结构化表示（数据处理）、特征提取、分类标准和方法、模型构建。

第二小题：使数据结构化表示（数据处理）、特征提取、数据类型转换、命名实体识别、聚类中心选取、模型构建。

第三小题：使数据结构化表示（数据处理）、特征提取、数据类型转换、命名实体识别、词性标注、语料库提取、模型构建。

2 问题的分析

2.1 问题一分析：

属于文本数据挖掘中的文本分类问题，对于解决此类问题我们一般采取机器学习或者深度学习的方法对经过预处理和特征提取的数据进行分类和模型构建。

观察问题一所需的附件一、附件二的数据。

首先我们可以由附件一得知分类所需一级标签应有 12 种分类,我们需要由附件 2 中”留言主题”和”留言详情”与 12 种一级标签其中的七种的对应关系建立分类模型，并对这个模型进行评价。

附件 2 中留言编号、留言用户和留言时间对于我们数据的分析并没有实质性的帮助，我们可以直接剔除。而以“留言主题”和“留言详情”作为分类指标的两列数据。其中，“留言主题”内容较为精简，对于分类的指向性也更加明确，但也正是因为“留言主题”数据内容较少，在模型建立个过程中，若仅通过“留言主题”，对整条留言进行分类建立的话，会出现可提取到的文本数据特征比较少而导致分类指标不明确、分类结果不准确的情况。而“留言详情”可以为我们提供较多的文本特征选择，但是同样存在的冗余、不相关特征，也就是噪音数据比例也会上升，从而影响分类。此时若直接使用 TF-IDF 权

值向量标识，取得实际效果不明显。考虑到这两组数据关联性较强同时也彼此独立存在，可单独调用，所以我们在提取文本数据特征时，联合两组数据提取共同特征，并采取了主观赋权的方法给予“留言主题”列特征较高权重。

这样我们的模型就要输入一组意见数据，经过预处理、特征提取、与机器学习模型进行比较，得到返回的结果为其对应的一级标签。

再使用已知分类结果的一部分数据作为测试数据，通过朴素贝叶斯滑稽矩阵得出这个模型的 f 分数指标。

2.2 问题二分析：

属于文本数据挖掘中的热点挖掘问题，需解决此类问题，我们一般采取文本聚类方法对经过预处理的半结构化数据，进行文本聚类。我们想采用的是文本聚类中的层次凝聚法。

观察问题二所需的附件三的数据。

附件 3 的数据与附件 2 的数据总体上相似，主要依靠刘云主题和“留言详情”中提取的特征来进行“物以类聚”的文本聚类的分类。

我们考虑到点赞数和反对数均可作为反映热点的一种指标，但是对文本聚类没有直接的贡献，所以我们在现阶段的模型建立中，把反对数和点赞数这两个指标先剔除。

我们会与问题一的方法相同，将“留言主题”列和“留言详情”列数据结合赋权后合并。先使每一行数据形成一个聚类，根据聚类中簇的相似性，进行层次文本聚类，同时得到我们所需的模型。这样我们可以在放进去一系列意见后，聚类为一些反应相同问题或相似问题的新聚类(相似问题聚类)。新聚类中较大，意见数较多的问题，就是我们所需要挖掘的热点问题。

2.3 问题三分析：

我们需要建立一个回复意见评价体系，来为政府部门所做出的回复进行评价。这个评价体系我们采取了答复意见的相关性、完整性、可解释性作为三个评价指标。

我们将答复意见的相关性与数据相关性系数计算相对应；我们使用了词性标注和命名实体识别的方法，分析答复意见成分近似给出一个完整性评价分数；最后利用机械学习中的可解释性工具分析答复意见的可解释性。

观察问题三所需附件四的数据

人与附件 2 中数据相类似，相较多出了答复意见和答复时间两块内容，我们希望通过命名实体识别一起出留言时间和答复时间相较，计算时间差，给出答复效率作为新的评价指标。

将经过预处理、赋权、特征表示后的答复意见与“留言主题”和“留言详情”结合后的数据，建立语料库和字典后，做相似度计算并得出相关系数(我们所求为皮尔森相关系数)。

再利用上述数据，通过机器学习的可解释性工具，得到评价数据。

利用哈工大自然语言处理平台，重新对数据预处理得到又词性标注后的数据进行分析，同时提取提问时间和答复时间做比较。得到完整性和时效性相关评价。

3 模型基本假设

1、假设题目所给数据真实可靠（包括文本数据内容、时间、点赞数等）；
2、假设噪音数据在合理范围内，比例不影响模型建立；
3、假设数据 2 中所给分类合理准确、且对应数据的分类真实；
4、假设数据 3 所涉及时间、地点、人物，真实且有区分度；
5、假设同行、同列的数据具有足够局部相关性，支持各类模型建立；
6、假设全部数据可获取足够文本数据特征；

4 方法分析和理论基础

4.1 文本预处理

由于切分歧义是汉语分词所面临的最大难题,其中能用语法知识消解的就约占 90%以上,而涉及语义和语用知识的切分歧义则很少,因此本文有机地将分词过程和词性标注过程融合在一起,采用了动态规划解决这一问题,有利于切分歧义的消解。对于粗切分后的碎片,本文根据重叠词的模式进行了重叠词识别,并采用了一定的规则来识别碎片中的未登陆词。在信息检索的向量空间模型中,文本被形式化地表示为由词项及其权重组成的向量。因此如何使这个向量尽可能准确有效地表示出文本内容同时又要尽量地减少向量空间的维数一直是该模型的基础性问题。针对这个问题,本文提出了一个标引词选择的算法,该算法充分考虑了词项的词频、位置以及它与其它词项、重要语句间的关系,根据实例,证明了该算法的有效性。最后,本文利用这些算法设计了一个基于信息检索的文本预处理系统。该系统首先利用句末标点将文本进行断句,根据各个句子的不同位置为其设定不同的句子权重;然后依次处理各个句子,根据句子中的其它标点把这个句子打散成短句子组,对于其中的一些特殊标点,进行了特殊的处理;接着对每个短句子进行分词和词性标注;对文本的碎片进行重叠词和未登陆词的识别;最后,采用了本文提出的标引词算法对已经完成前期处理的文本进行标引词的选择【1】。

4.2 逻辑回归模型

对于保理公司来说,授信额度的计算与客户的违约可能性显得尤为重要。但在现实生活中,往往会出现客户资料缺失的情况发生。因此,本文就此类问题进行了分析并建立相关模型。采用 KNN 填补法填充缺失数据,并利用主成分分析法、逻辑回归构建违约率模型。采用向前逐步回归法、多元线性回归,得到授信额度估算模型。最后利用神经网络进行二分类回归,提供保守型公司和风险型公司的两个不同标准来判断客户是否违约。实证结果表明,模型可以作为较为理想的预测工具【2】。

4.3 文本聚类

聚类算法作为发现数据内在结构与分布特征的无监督学习方法,被广泛应用于各个领域。伴随着互联网的高速发展和在线文档数量的大幅增加,文本聚类已成为一项重要任务。讨论文本聚类算法的基本概念与应用场景,对文本聚类算法及评价方法进行综述【3】【4】。

4.4 KNN 算法

当今大数据时代,文本数据占相当大的比重,作为有效管理和组织文本数据的方法,分类逐渐成为关注的热点。KNN 是一种经典的分类算法,针对其分类速度和分类精度无法同时兼顾的不足,采用改进的 K-Medoids 聚类算法裁剪对 KNN 分类贡献小的训练样本,从而减少 KNN 相似度的计算量,并定义代表度函数有差别地处理测试文本的 K 个最近邻文本,以提高 KNN 的分类精度。实验结果表明,改进后的方法在分类速度上和分类精度上均有明显地提高【5】【6】。

4.5 朴素贝叶斯

对基于朴素贝叶斯的文本自动分类研究进行了系统的综述。探讨了多项式模型和多元伯努利模型等经典的朴素贝叶斯分类方法。重点分析了经典的特征选择方法以及包括 ALOFT 等在内的多种改进的特征选择方法。论文还对从加权、避免平滑等视角的 NB 改进算法进行了梳理。最后,提出了进一步改进 NB 的主要思路。 在信息化时代,数据已经成为一种宝贵的资产。如何高效准确地从这些数据中挖掘出有价值的信息和知识,一直备受学术界和工业界的关注。贝叶斯分类方法是机器学习和数据挖掘研究领域的重要数据处理方法之一。朴素贝叶斯分类方法具有简单、高效、分类效果稳定的优点,同时还具有坚实的理论基础,因此在实际应用中得到广泛的重视。另一方面,朴素贝叶斯为了简化分类模型,而假定分类数据各个属性间是相互独立的,这在实际应用中通常很难完全满足,如此就使得朴素贝叶斯方法在处理复杂问题时受到一定的限制。为此很多研究人员通过放宽属性独立性假设这个条件来提高朴素贝叶斯分类方法的分类性能【7】【8】【9】。

5 模型的建立和求解

5.1 第一部分：准备工作（数据预处理）

计算机是无法直接读懂人类语言的，也就是说我们在做自然语言处理的时候，先要对文本数据进行一定的改变，将中文变为计算机所可以读懂的数据，将其转化为计算机能够自动处理的数据；另外文本数据中存在一部分的噪声数据，他影响了我们文本分析的效率以及精确性。预处理工作对于降低文本噪声有显著的影响，其最终结果直接决定了能否选择到合适的文本特征进行文本表示以及能否提高文本分析的效率。由于文本数据和传统的结构化数据有很大差异，同时不同语言的文本形式和语法都不相同，因此对不同语言的数据进行预处理的方法也存在较大的差别。对于中文来说，文本预处理主要包括分词、剔除停用词、词性标注和特征提取等步骤。我们的准备工作以文本数据预处理为主。

由于不同模型使用数据类型不同，且数据类型间的转换容易造成误差，甚至是报错，所以对不同模型采取不同的中文预处理模式。但是大致的方法是一样的：

方法一：

其中最主要的是对第一小题的一级标签分类模型和分类 F-Score 评价模型，第三小题文本数据相关系数模型、文本数据时效性评价模型、文本数据可解释性评价模型。

这几个模型运用的模型基本为机器学习中 sklearn 和 gensim 这两个库中的函数或者

model 来制作，可以使用我们使用 pandas 导入的 Excel 中的数据，或者是 numpy 简单转化过的数据，所以我们对这几个模型所需要用到的数据就使用了 re 和 jieba 两个库以及一部分 python 自带的函数做了预处理。具体过程会在下一部分介绍。

方法二：

虽然赛题给出的数据是已经经过一定处理的，如：地点 A 市、B13 区，标准的时间表示，但是具体到小区等更有对热点指向性的数据才是对我们有意义的数据，需要做到对数据分词之后，进行词性标注和命名实体识别，所以我们使用了哈工大 ltp 的 Python 扩展包 pyltp 以及 NLTK 完成相应预处理操作。具体过程我会在下一部分介绍。

5.2 模型一的建立：

（1）使用 Excel。观察数据亦当中一直标签中的分类，剔除附件 2 中留言编号、留言用户和留言时间等无用数据。

（2）用 pandas 从附件 2 中导入剩下的三列数据作为 data

（3）考虑到题目数据读取进来可能会有重复数据。影响整体的特征频率。我们先使用 drup 对整个 data 剔除重复数据。同时我们做了一个测试。”留言主题”列和”留言详情”列单独做剔除时。会剔除掉更多重复数据。但是我们考虑到这样会改变数据的类型。这样剔除重复数据的同时，也会剔除一部分有用的数据，并且这一类数据理论上来说并不是我们在爬取数据或导出数据时系统出错而产生的重复，这样的数据往往能反应重要问题(会重复提到而产生重合)，在这个模型实际运用过程中会有较大的意义，所以我们不予剔除。

（4）接下来由于导入格式和数据本身的问题，我们需要使用 apply 函数、sub 函数和 re 包遍历”留言主题”列和”留言详情”列每一行，去敏如:’\n’、’\t’、’’等特殊字符,以及”留言详情”列’\u3000’、’\xa0’等干扰字符，。这一步放的较为前面，因为后续一些函数不支持对带有这些特殊字符的数据进行处理，从而产生报错。（去敏结果见附件）

（5）考虑到文本分类时，时间地点人物往往带来的误差高于他们带来的特征，所以我们在分词前也可以先剔除大小写字母和数字。

（6）这样我们得到了去敏和去重后较为干净的数据我们就可以应用 jieba 库进行分词了。打印后观察分词后的结果，可知:分词结果并不理想，有产生类似“省县”、“市公品”这样处理不精确产生的有歧义的词，以及对一些不该切分的词，切分错误的词“供应商”和“库”，我们挑取了其中一部分建立了词典，导入到 jieba 库，提高分词的准确性。（分词结果见附件）

（7）同时我们也观察得到，长度为 1 的单个字对于文本分类贡献不大(实际测试后得到的结果也是这样，即使这样也确实剔除了一部分能表示特征的词)。我们使用循环的方式删除所有长度为 1 的单字。

（8）我们导入停用词库(合并百度停用词库，nlp 停用词库，去重后得到 2600 词，再去除特殊符号(我们的数据中并没有特殊符号))，使用循环的方法剔除停用词。

（9）接下来是转接步骤，为了下面特征提取和模型建立，我们使用 join 函数先将数据改成每个元素以空格链接的字符串数组。同时以 loc 函数建立他们的索引所对应的标签。

（10）对得到的数据用 sklearn 中 train_test_split 进行切分分为训练集和测试集，经过测试 test_size 参数设置为 0.2(参考了简单随机抽样最大样本公式，避免过抽样或欠抽样所带来的误差)较优。

（11）接下来我们做特征向量的提取方法有 tf 词频统计、TF-IDF 权重等等，但是经

过我们测试，备受推崇的 TF-IDF 并不能取得较好分类结果(多次尝试结果准确率在66.8%-68.6%之间，原因我们估计是总样本容量比较小，对 idf 权重影响比较大)。我们则使用了词频作为我们的特征提取标准，还考虑到前面提及到的”留言主题”列和”留言详情”列对分类贡献程度不一的问题，同时借鉴了 TF-IDF 中共享维度的方法，和组合赋权方法(主观赋权和客观赋权相结合)，对我们训练集 66017 词和测试集 11623 词，提取到了我们的文本数据特征 top6000(词频较高 6000 词(测试得到 5000-12000 词特征之间区别不明显，为节省纬度资源选取 6000 作为参数))

（12）根据选定的 top6000 使用循环加入列表，再使用 numpy 中的 array 函数转换为数组，也是特征向量作为模型可以训练和测试的数据。

（13）最后是模型的选取。我们选取了 KNN（最邻近分类算法)、logisticregression(逻辑斯蒂回归分类器)、GaussianNB 和 MultionmialNB 两种朴素贝叶斯算法，最后逻辑回归算法明显优于朴素贝叶斯优于附近分类算法(原因我们估计可能还是总样本容量比较小)。最后我们选择了逻辑回归分类器，作为我们的分类模型。

（14）另外我们直接利用 sklearn 中的函数，和 F-Score 公式，计算得到我们模型的评价，经过一系列参数调整后，得到的 F-Score 的值为 0.898 附近。（详情见附件 Q1.py）

5.3 模型二的建立：

（1）pandas 导入整个附件 3 作为 data，观察由于导入格式和数据本身的问题，我们先使用 apply 函数、sub 函数和 re 包遍历留言主题”列和“留言详情”列每一行，去敏如:’\n’、’\t’、’’等特殊字符,以及”留言详情”列’\u3000’、’\xa0’等干扰字符。这一步放的较为前面，因为后续一些函数不支持对带有这些特殊字符的数据进行处理，从而产生报错。

（2）此题对热点聚类时，需要用到时间、地点、人群、对于我们已经进过一定处理的数据（地点以 A 市、A2 区、A6 县），我们不需要剔除时大小写字母和数字。

（3）此模型数据的预处理，基本为 nltk 库中的函数或者 model 来完成（自带分词库和停用词库）所以数据先按列转为 txt（列表不能转为 txt，失去数据索引，但选取聚类中心时，无需索引）文档便于保存、导出和重新调用。

（4）使用 scikit-learn 中的 TF-IDF 模型、VSM 向量空间模型来制作每一簇元素的特征向量。

（5）对“留言主题”列和“留言详情”列进行文本聚类。使用聚类算法进行聚类（K-means，DBSCAN，BIRCH 等，我们使用 K-means 算法）得到的初步聚类中心，可以认为剔除一些不合理中心，如:“政府”“街道”等。

（6）因为预处理的时候没有带上“留言编号”，所以最终我们还需要按顺序把“留言编号”与“留言主题”对应起来，所以要先将“留言编号”和“留言主题”按照 df 格式读取，再单独将“留言编号”按行保存为 txt。

（7）然后通过 VSM 向量空间模型来存储每个文档的词频和权重，特征抽取完后，因为每个词语对实体的贡献度不同，所以需要对这些词语赋予不同的权重。计算特征词在向量中的权重方法——TF-IDF。(tf-idf 向量矩阵表见附件)

（8）所以接下来使用 scikit-learn 工具函数计算 TF-IDF 值，生成 tf-idf 向量矩阵表，并将矩阵保存为 csv 类型。最后进行 K-means 聚类，完成了文本归类，将数量最多的前五类数据的索引提取出来，结合 Excel 的功能完成了表 2 热点问题明细表。

（9）再进一步对表 2 热点问题明细表进行文本处理，运用基于 pyltp 库的命名体识别和 time 库，再次预处理，提取出前五类热点问题中里面的时间、地点、人群，根据信

息量、点赞数、问题反应持续时间，判断热度指数，并结合 Excel 的操作完成表 1 热点问题表。（详情见附件 Q2.py）

5.4 模型三的建立：

先建立四个分模型

相关性：根据模型一的方式（参考模型一的建立（1）-（9））对附件 4 的“留言详情”列、“留言详情”列和“答复意见”列数据做分词、去敏、去重、去停用词等预处理，处理好的分词数据，进行多重数据类型转换（先用遍历和循环的方式转变为列表类型、再用 numpy 中的 array 函数转变为数组类型、再用 astype 函数调参，使得普通数组转变为 unicode 标记数组，才可投入相似度计算模型）。将转变好类型的所有数据转变为字典类型，并分别用于提取语料库，我们以“留言主题”“留言详情”的合并数据为本征数据（以下称本征数据），计算“答复意见”的相似的和相关性系数，提取本征意见 TF-IDF 权值向量和总特征数以及本征数据语料库作为模型参数，遍历“答复意见”计算相关性系数。（字典、语料库结果见附件）

完整性：根据模型一的方式（参考模型一的建立（1）-（9））单独对“答复意见”列去敏、去重，再使用 pyltp 库，数据做分词、去停用词等预处理，处理好的分词数据，也运用 pyltp 库做词性标注和命名实体识别，分析语句成分，我们在假设大部分答复意见完整性较好情况下，默认语句成分相同较多的为较高完整度，设每个成分分值相同，发现得到完整系数普遍较高，且两极分化明显，使用函数拟合方法进行评分函数优化。

可解释性：在对答复意见进行数据挖掘分析之前，需要先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，以中文文本的方式给出数据，为了便于转换，先对答复意见中的信息进行中文分词，这里采用了 python 的中文分词包 jieba 进行分词，并提取出高频词，生成句子中汉字所有可能成词的情况所构成的有向无环图 DAG，对于其他未登录词，采用了基于汉字成词能力的 HMM 模型。这里采用 jiebe 自带的语义库。数据的可解释性！也称为可读性，是指数据被人理解的难易程度，如果数据具有解释性或包含注释性信息，而且数据书写规范，则数据的可解释性越高。相反如果数据隐晦难懂就根本不具备分析的条件。观察表中数据，“留言详情”“答复意见”，“答复时间”等，结合可解释性的定义和理解，我们先导入数据库，引用 matplotlib.pyplot 库，从 sklearn.tree 中引入 DecisionTree 库，查看数据内容，根据一系列查阅，我们可以假定，如果其中的一些顾客的答复意见可以转换为可被人理解的程度，那么这些因素都会有各自的影响因子，对于分类变量，我们需要创建虚拟变量，并且丢掉每类的描述特征，更方便运用。也可以输入 dot-Tpngtree.则会生成 tree.png 参考 dot 解决，这提供给我们一个解释性工具，然后继续分析评估这个模型。也可以运用 LDA 主题模型，以挖掘出更多信息，在判断文档相似性时，应进行语义挖掘，而语义挖掘的有效工具即为主题模型。LDA 模型采用词袋模型，将每一篇答复意见视为一个词频向量，从而将文本信息转换为易于建模的数字信息。

时效性：导入意见时间和答复时间为两列数组，同时建立每一组数据所对应的相同的索引。使用 pyltp 库，依次切词、词性标注、命名实体识别得到两组时间数据，根据遍历每一索引值，得到每一行数据的“提议-答复时差”，根据这一数据，以“越快，效率越好，评价越高”原则得到简单的线性函数评分（权重平均分配到 20 天内），建立时效性评分函数模型。

总模型建立，默认四个评分权重一致，相加后得到关于答复意见总评分，同时考虑到对于整个数据列表用遍历方式得到每一行方式的总评分较为不合理，所以以相关性系

数评分为例，改写一份用输入行索引方式得到某一行的总评分模型。（详情见附件 Q3.py 和 Q3-tmp.py）

6 结果分析

6.1 模型一

因为调用 sklearn 库的同时已经建立了 F-Score 模型，已经可以直接对模型进行相应客观评价，我们一直以 F-Score 分数作评价和结果分析：

1、我们使用 TF-IDF 权值向量时所得结果分数在 66.8%-68.6%之间（适应以下各种情况，所以这个分数是一个范围，下述均为使用 TF 词频权重得到结果）。

2、仅使用“留言主题”列数据或“留言详情”列数据，得到结果分数分别为 69.0%和 75.1%（朴素贝叶斯模型，0.2 切分度，top6000 特征下），而使用两个数据相结合有明显提升，且通过客观赋权计算，接近“留言主题”接近两倍于“留言详情”权重是最为合理的（在提取特征前，现将”留言主题”列特征元素，多提取一次，可以视为更高词频也就是更高权重）。

3、为避免过抽样或欠抽样所带来的误差，通过简单随机抽样最大样本公式，对于去重后 9051 条数据，在允许抽样误差为 1%，结果误差为 5%的情况下，所抽取测试集不少于 1100（比例在 0.1 到 0.15 之间），所以我们使用 0.15、0.2、0.25、0.3 分别作为切分参数，对训练集抽样结果来看相差不大，且 0.2 的情况下略优。

4、对于总词库，词频特征维度取值，没有较好理论支持，实验得出 5000-12000 词特征之间区别不明显，为节省纬度资源选取 6000 作为参数。

5、确定了特征词频和维度以及训练集、测试集切分度后，经过对数据分析观察，我们采取了机械学习方法对我们的模型进行构建，尝试了 KNN 模型（结果分数为 72.8%）、GaussianNB 朴素贝叶斯模型（结果分数为 86.0%）、MultionmialNB 朴素贝叶斯模型（结果分数为 85.5%）最终我们替换为 LogisticRegression 逻辑回归模型，结果分数进一步提升为 89.8%。

6.2 模型二：

基于与模型一所使用相同库和函数处理，去除停用词表中的代表时间地点的部分，得到的预处理数据进行操作。

1、经过测试，模型二选取 TF-IDF 模型，观察聚类结果有较大改观（经过分析应该是聚类对于文本的分类相比一般的分类更加细致，或者 KNN 算法运用于聚类时与 TF-IDF 向量更加契合）。

2、同样，对于这一题相应提高“留言主题”列权重，有助于聚类结果优化。

3、根据所取得特征，利用 KNN 算法，得到聚类中心，依照“少数服从多数”原则，提取出“扰民”“实习”。

4、计算对聚类中心距离时，余弦值为[-1,1],通过 sim=0.5+0.5*cosn 归一化为[0,1],值越大相似度越大；sim=1-sim 将其转化为值越小距离越近。

5、计算聚类间的最大距离、最小距离、平均距离做出误差分析，结果显示前五聚类中心，得到结果误差较小。

6、通过层次聚类算法，得到聚类后热点问题索引，用 Excel 筛选出这些数据，得到

热点问题留言明细表，再通过主观分析得到热点问题表（结合明细表与聚类中心）。

7、定义热度指标时使用点赞数添加权值时应当设定一定的阈值。

6.3 模型三：

其中关于相关系数计算系数模型和可解释性评价模型，我们依旧使用模型一所使用相同库和函数处理，去除停用词表中的代表时间地点的部分，得到的预处理数据进行操作。

1、其中第一步计算（0，1）之间的数值观察发现普遍低于 0.1，所以我们将计算结果加入到最后计算前全部*5，以平衡数据大小不一指标不同，避免降低这一指标的权重，而产生的大误差。

2、词袋、字典和语料库建立的时候一些引用模型不便于修改，所以观察所得数据，认为去除一些噪音。

对于另外两个模型：数据完整性评价模型和时效性评价模型。

3、制作完整性使用 pyltp 库，pyltp 库比较强大，预处理结果明显优于前两者，我们在假设大部分答复意见完整性较好情况下，默认语句成分相同较多的为较高完整度，设每个成分分值相同，发现得到完整系数普遍较高，且两极分化明显，使用函数拟合方法进行评分函数优化。

4、对时效性性模型，我们并未对时间这样的指标作出具体分析，所以提取到时间后，我们针对最大时长 30 天，最短时长 10 天，根据“越快，效率越好，评价越高”原则得到简单的线性函数评分（权重平均分配到 20 天内）

7 模型的评价与改进（推广）

7.1 模型一：

1、优点：运用到的逻辑回归模型较为简单有效；在大部分参数定义上有较好理论支撑，便于根据实际需要进行参数修正；得到 F-Score 分数较高有使用价值；如果有更多数据进行训练提升空间较大。

2、缺点：目前建立的模型是对数据 2 进行的分类训练后的结果以及评价，缺少输入外部数据的设置，利用价值还可以提升；预处理结巴分词结果仍有错误，停用词导入有错误。

3、改进方法：把整个模型包装起来，可供调用，需要做好调试；对结巴库参数进行修正、同时根据预处理出来数据修正停用词表内容，同时指定引擎避免导入时出现警告；使用深度学习方法可能还要进步空间。

7.2 模型二：

1、优点：TF-IDF 权重向量使用，数据可信度更高；NLTK 的中文预处理更加强大，数据清洗与结构化更加完善，便于分析；KNN 算法与聚类模型较为契合，测试结果好。

2、缺点：有多次数据类型间的转换，容易产生错误；NLTK 的 model 较为复杂；命名实体识别仍有不足之处，希望地点数据可以导入。

3、改进方法：若是在提高准确率后的模型一基础上做出聚类，得到的热点问题更加精确，反映的问题也更加统一和突出；导入地点数据或做有关地名识别深度学习。

7.3 模型三：

1、优点：增加了时效性作为新的评价指标，使评价更加有可信度；对于四个模型得到四样数据，合理处理数据间的关系，拟合产生结果可靠。

2、缺点：相关性计算得到结果偏低，可能因为程序逻辑问题，简单的放大数据可能并不可靠；完整性语句成分权重相同；对于我们评价系统得不到一个有效评估；指标之间过于独立；遍历方法计算相似度，字典和语料库过小。

3、改进方法：题目较为开放，我们希望可以找出更多的参考指标，对文本进行二次挖掘（如情感分析等等）；对于数据大小不一的问题应该通过客观计算，得到放大、统一数据的有理论支撑的方式；结合主观判断对，不同成分进行赋权；可以让时效因素影响别的指标（不同时效情况下，别的数据权重产生相应变化），并且提高相关性权重（问答对应是评价的基础）。

8 结论

看似简单的数据、文本，其实暗藏许多信息。走向大数据时代，为“更好服务于人民”的政务处理进步，提供了新的发展。根据文本中某些特征词词频，我们可以将文本转化为结构化、可视化信息，再利用计算机，寻找文本数据相似度、分类聚类。可以更好处理一类数据或者更加重要的数据，提高服务效率。同样对于政府部门的答复意见，对这类文本数据，加以结构化，运用到以各项指标模型为基础的评价模型，可以提高政府部门答复水平和效率，体现服务为民的工作态度。

参考文献:

【1】. 何金凤. 基于中文信息检索的文本预处理研究[D].电子科技大学,2008.

【2】. 杨睿哲,王智敏.客户信息不完全下的授信评估问题——基于逻辑回归、神经网络等模型[J].现代商业,2019(36):91-92.

【3】. 史梦洁.文本聚类算法综述[J].现代计算机(专业版),2014(03):3-6+25.

【4】. 吴启明,易云飞.文本聚类综述[J].河池学院学报,2008(02):86-91.

【5】. 樊存佳,汪友生,边航.一种改进的 KNN 文本分类算法[J].国外电子测量技术,2015,34(12):39-43.

【6】. 潘登. KNN 算法的相似度研究[D].东北师范大学,2014.

【7】. 贺鸣,孙建军,成颖.基于朴素贝叶斯的文本分类研究综述[J].情报科学,2016,34(07):147-154.

【8】. 阿曼. 朴素贝叶斯分类算法的研究与应用[D].大连理工大学,2014.

【9】. 朱晓丹. 朴素贝叶斯分类模型的改进研究[D].厦门大学,2014.