“智慧政务”中的文本挖掘应用

摘要

该次试题中，我们进行了数据读取，数据探索分析，文本预处理，分词以及停用词处理，预测模型的构建，文本分类以及情感分析等操作，对关于群众留言的数据集进行分类，对按时间分类的动态主观模型进行了留言热点话题的分析排序，以及对关于政府答复内容建立了简单的评价系统。

关键字：文本分类，情感分析，主观模型。

一、问题背景

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。

二、实验原理

F-Score 模型评价方法：

通常使用 F-Score 对分类方法进行评价：

$$F_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{2P_iR_i}{P_i+R_i}$$

其中P_i为第i类的查准率，R_i为第i类的查全率。

朴素贝叶斯，它是一种简单但极为强大的预测建模算法。贝叶斯原理是最大的概念，它解决了概率论中“逆向概率”的问题，在这个理论基础上，人们设计出了贝叶斯分类器，朴素贝叶斯分类是贝叶斯分类器中的一种，也是最简单，最常用的分类器。朴素贝叶斯之所以朴素是因为它假设属性是相互独立的，因此对实际情况有所约束，朴素贝叶斯模型由两种类型的概率组成：每个类别的概率 P(Cj)和每个属性的条件概率 P(Ai|Cj)。

朴素贝叶斯分类主要分三个阶段，第一阶段是准备阶段，在这个阶段我们需要确定特征属性，同时明确预测值是什么。并对每个特征属性进行适当划分，然后由人工对一部分数据进行分类，形成训练样本。第二阶段是训练阶段这个阶段就是生成分类器，主要工作是计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率。第三阶段是应用阶段，这个阶段是使用分类器对新数据进行分类。

k-近邻（kNN,k-NearestNeighbor）算法是一种基本分类与回归方法，在分类问题中的 k-近邻算法的输入为实例的特征向量，对应于特征空间的点；输出为实例的类别，可以取多类。k-邻算法假设给定一个训练数据集，其中的实例类别已定。分类时，对新的实例，根据其 k 个最近邻的训练实例的类别，通过

多数表决等方式进行预测。因此，k 近邻算法不具有显式的学习过程即属于有监督学习范畴。k 近邻算法实际上利用训练数据集对特征向量空间进行划分，并作为其分类的“模型”。k 值的选择、距离度量以及分类决策规则是 k 近邻算法的三个基本要素。

三、实验过程

每一个问题都可以用以下流程图来进行表示：

数据读取 → 数据探索分析 → 数据抽取 → 文本预处理 → 文本向量化 → 代码实现

问题一：

使用 python 的 pandas 库读取表格，因为测试数据总体较少所以以全部数据作为导入数据。再根据需要将导入的数据按照一级标签顺序分入多个数组，将这些数组组合作为简化后的导入数据。

对中文语言进行文本预处理，使用中文汉字集合^\u4E00-\u9FD5 去除多余数据，再使用 jieba 库，导入分词词典去除停用词，将留言详情转为分词形式。

由于无法直接对中文文本进行模型训练，因此将转为分词后的中文文本进行向量化表示，使用 sklearn 的 CountVrctorizer 模板，将分词模型变成了二进制型的矩阵模型。

之后构造分类器，本次代码我们分别构造了朴素贝叶斯模型，knn 模型以及 svc 模型进行模型训练，训练时我们将导入的整体数据按照 2:8 的比例分为测试数据以及训练数据。最后使用 sklearn 的 classification 板块进行模型评估。

问题二：

数据读取与问题一相同，问题二中我们首先对导入的数据按照留言时间制作了布朗运动模型，以留言时间的密集分布情况对留言内容按照时间划分，由此制作一个以时间为索引的字典，再对字典里面的元素按照如同问题一的方式进行文本预处理，即调用 jieba 库加载停用词表。

之后进行情感分析，调用 gensim 库，对留言内容进行主题模型构建，去除重复词，并且多次训练找到最适合的 K 值，即平均每条留言关键字的个数，最后将每条留言转化为由关键字构成的字符串。

之后对字典里每一数组的关键字进行统计，得出在该时间段内的主观情感主题。

最后我们需要规定一个热度指标用于对热点话题进行排序，我们规定，在单位时间内，按照某主题的留言次数为该话题在该单位时间的热度，以此对各个话题进行热度排序，得出以话题热度为索引的导出表，再另按照索引对具体留言整合导出，得到结果。

问题三：

问题三同样需要进行情感分析，我们在问题三种，对政府答复以及其对应的用户留言进行情感分析，对这两个模型得出的主题模型进行比较，比较其关键字的权值。其关键字之间越是相同，则判断该政府答复越是准确。

四、实验数据：

见附件。