基于自然语言处理技术“智慧政务”中的文本挖掘应用

摘要

随着 QQ、微信、微博、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，在以前主要依靠人工来完成那些对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，随着碎片化的时代不断扩大，信息的时效不断缩短，好在有大数据、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。

针对问题一（群众留言分类问题）：建立多个文本技术相关的模型，然后进行多次对比其准确率。在建立模型之前，首先对数据设置自定词和停止词然后才分词处理，再把得到的文班进行转向量操作，选择 KNN 算法模型、SVM 模型等进行训练，最后将这些经过训练的模型通过测试，并对测试结果进行评估，对比不同模型测试后的准确率，取用准确率最高的算法模型。

针对问题二（热点问题挖掘）：数据挖掘问题，往往是从数据开始。首先对数据进行去重，中文分词，然后再将词语转换成向量。对数据处理完后，还需要将所得的向量化数据进行聚类，用 K-meas 确认阈值来进行聚类，或者用 DBSCAN 算法进行聚类，然后清除离群点，即可获得我们所需要的结果。

针对问题三（答复意见的评价）：开放性题目，我们自定义相应的量化评价指标，然后再针对题目要求来完成。分解数据后整合，运用不变的方法，应用指标要求，进行数据处理，然后经过挖掘，最后挖掘出所需要的地点、人物、事件，对用户给予最准确的答复。

关键词：中文分词 KNN 算法 TF-IDF 算法　量化评价指标

目录

一、问题分析

随着大数据时代的不断壮大，数据不仅仅流动快，传播广，而且数据量越来越大，数据的类型也不断增多，同时也使得数据的价值密度低，还有时效性要求高。在这种情况下，我们需要对数据的处理要更快，更好，更精准。从而能够为

政府或单位提供更快，更精准得了解民情，并给予解决方案，得以提升政府的管理水平和施政效率。

问题中一共给出了 4 个附件数据，附件一是附件二中的关于留言内容的三个等级的标签，附件二是带有一级标签的留言用户所写的留言主题和留言内容。附件三与附件二相差不大，不同点是附件三把附件二中的一级标签换为点赞书和反对数统计。附件四是对留言内容进行的答复，以及附加有答复时间。

问题中的问题一，在给定了附件二留言内容进行数据挖掘，建立关于留言内容的一级标签分类模型，然后运用算法模型进行预测，就是给留言内容再贴上附件一中的符合留言内容的一级标签，最后进行模型评估。在对不同的模型得出的准确率不同，进行对比，取用准确率最好的一个模型，并保留。

问题二

二、数据准备

2.1 节选所需数据项

根据题目分析，我们知道，在题目一中我们只需要用到附件二的留言内容和一级标题里的内容，附件三中留言内容和留言时间，附件四留言内容和答复。其他数据对问题的研究并不参与，故可以忽略。

2.2 确认指标

1）准确率

在已给出数据后，只需要建立不同算法的模型，并进行预测，所预测的结果，在通过 F-Score 对分类方法来进行评估，即可确认其准确率。对比不同模型得出的准确率去最高值即可。

2）量化评价指标

确认指标往往是开始的第一步，它为我们的运行提供了一个方向，由此开展我们的编写。

2.3 数据预处理

1）留言信息的分词处理

在对数据进行挖掘分析之前，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。在附件中的留言内容和答复信息中，以中文文本的方式给出

了数据，因为计算机并不能识别其中的语意，需要我们人工使用自然语言处理进行转换。我们先要对这些留言内容进行中文分词。使用 python 的中文分词包 jieba 进行分词。在采用 jieba 自带的语义库后，并引入我们自制的停用词：stopwords.txt 文件，这样就可以使一些没有用的词语进行剔除，留下最关键，我们最需要的词语，确认留言内容的语意。

2）TF_IDF 进行文本转向量处理

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 (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:

词频(TP) = 某个词在文本中的出现次数/文本的总词数 (2)
或
词频(TF) =某个词在文本中的出现次数/该文本出现次数最多的词的出现次数 (3)

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。

逆文档频率(IDF) = log(语料库的文本总数/包含该词的文本数+1 (4)

完成数据转向量处理后，即可进入我们数据挖掘的正题了。

三、问题一：群众留言分类

在完成了数据预处理后，后面的事情就简单许多了。我们首先是要建立模型，对数据进行训练，然后预测，并评比准确率即可。

3.1 建立朴素贝叶斯模型
3.2 建立 KNN 算法模型
3.3 建立 SVM 模型
3.4 对比模型准确率

朴素贝叶斯准确率： 0.8619192357794182
支持向量机准确率： 0.8897090751194094
knn准确率： 0.583792601735223

四、问题二：热点问题挖掘

4.1DBSCAN 聚类

五、问题三：答复意见的评价

六、总结

这次比赛，对于我们来说比较困难，完成所以问题还需一些时间。我作为队长也感到非常愧疚。