“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。

我们主要针对问题一（群众留言分类）进行了研究：由于长文本的无意义表达太多，我们通过导入分词库、sklearn 中的数据分割模块，调用停用词等来将长文本分割成一段段的词组，通过调用四种模型的图形化，进而得出所选择的最优模型。并并对问题一进行错误估计模型检验，尽量提高其精准度。

【关键词】 智慧政务、最优模型、精准度

Text mining application in "smart government"

[abstract]In recent years, as online platforms such as Wechat,Weibo,mayor's mailbox and sunshine hotline have gradually become an important channel for the government to understand public opinions,pool people's wisdom and gather people's morale,the amount of text data related to various In recent years,with wechat,microblog,mayor's mailbox,sunshine hotline and other online political platforms gradually becoming an important channel for the government to understand the public opinion, gather people's wisdom and gather people's spirit,the amount of text data related to various social situations and public opinions has been increasing,which has brought great challenges to the work of relevant departments that used to rely mainly on human to divide messages and sort out hot spots.At the same time,with the development of big data,cloud computing,artificial intelligence and other technologies,the establishment of smart government system based on natural language processing technology has become a new trend of social governance innovation and development,which has a great role in promoting the management level and efficiency of the government.

We mainly focus on problem one (classification of mass message): because there are too many meaningless expressions of long text,we divide the long text into phrases by importing thesaurus,data segmentation module in sklearn,calling stop words,etc.,and then get the optimal model by calling the four models.At the same time,we test the error estimation model to improve the accuracy as much as possible.

[Key words] smart government, optimal model, accuracy

目录

1 挖掘目标

由问题背景得知，近年来，随着网络问政平台逐步成为政府了解民意、民声的重要渠道。各类社情民意相关的文本数据量不断攀升，给工作人员带来了极大的不便。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。因此顺应时代发展的趋势我们根据附件给出了收集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。运用 python 语言中的 sklearn 库进行机械学习，处理得出最优的模型进行测试，并对测试集进行预测准确率分析。并根据网友提出的问题，进行统筹分类，做到热点问题及时发现，从而有针对性处理，最后根据网友留言进行答复。

2 总体步骤

2.1 总体流程

文本的总体构架及思路如下：

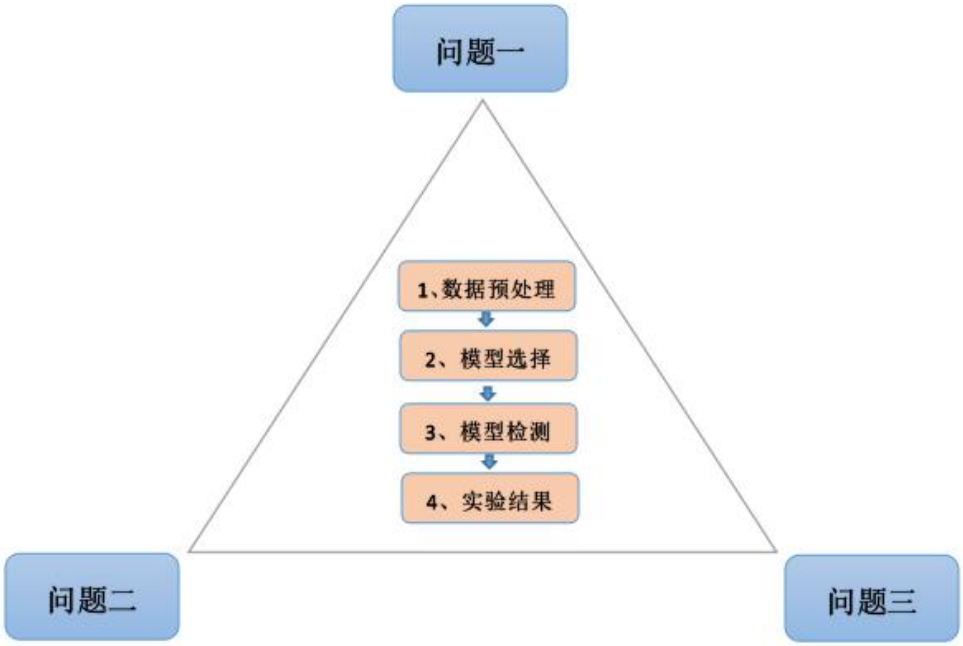

图 1 文本的总体流程

步骤一：数据预处理

根据题目所给的数据预测分析，首先确定实现问题的解决办法。数据和特征决定了机器学习的上线。特征处理是特征工程的核心部分，sklearn 提供了较为完整的特征处理方法，包括数据预处理、特征选择、降维等。

我们选取的是 Scikit-learn（图 2），这是一款基于 NumPy、SciPy 与 matplotlib 的机器学习 Python 模块，能够为数据挖掘（data mining）与数据分析提供简单有效的工具。SKLearn 面向所有人，让每个人能够在复杂环境中重复使用，并且建立 NumPy、Scipy、MatPlotLib 之上，因此我们用此来处理我们本次的数据。

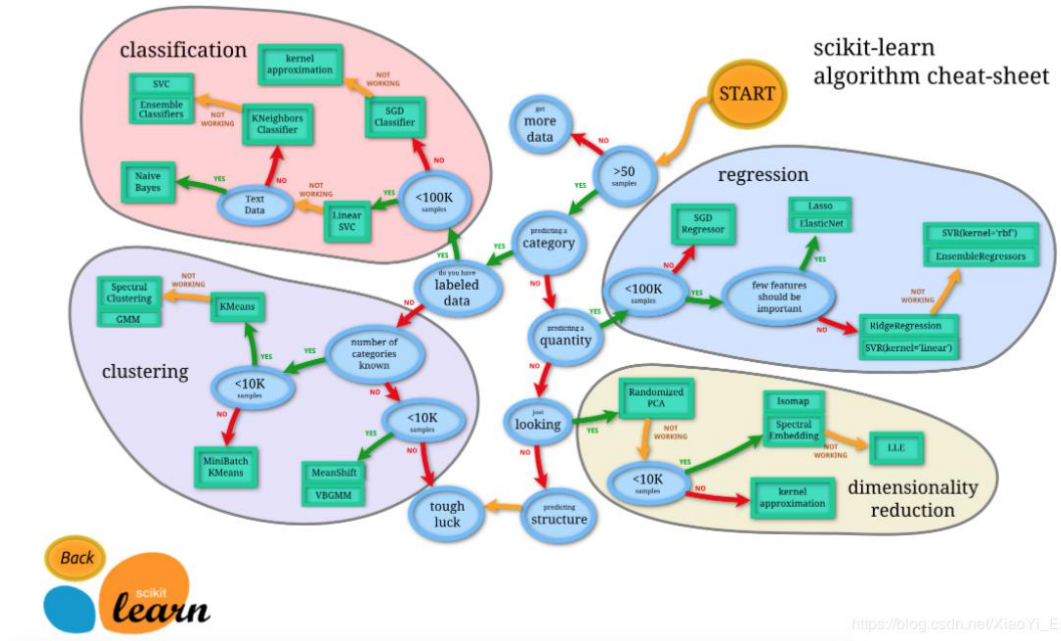

图 2 sklearn 的特点

步骤二：模型选择

由于数据为文本类，因此我们采用机器学习。机器学习通常包括分类（Classification）和回归（Regression），常用的分类器包括 SVM、KNN、贝叶斯、线性回归、逻辑回归、决策树、随机森林、xgboost、GBDT、boosting、神经网络 NN。我们从中选取贝叶斯、线性回归、随机森林、SVM 来进行初步的概率分析，从而选出准确率高的的模型来进行操作。

步骤三：模型检验

模型检验是确定模型的正确性、有效性和可信性的研究与测试过程。根据步骤二得出最优的数据模型 LinearSVC 模型，由于模型依旧存在检验预测标签和实际标签之间的差异，因此我们调用混淆矩阵，来进行模型检验。

步骤四：实验结果

该问题利用 Python 中的 sklearn 库来实现文本的多分类，从而建立一级标签的模型。

2.2 具体步骤

步骤一：数据预处理

根据题中所给的数据预测分析，首先对数据进行初步统计（如图 3），由柱状图的形式表达。我们将一级分类转换成了 Id(0 到 6),便于后续的处理。

由于我们的评价内容都是中文,所以要对中文进行一些预处理工作,这包括删除文本中的标点符号,特殊符号,还要删除一些无意义的常用词(stopword),因为这些词和符号对系统分析预测文本的内容没有任何帮助,反而会增加计算的复杂度和增加系统开销,所有在使用这些文本数据之前必须要将它们清理干净。

因此将附件 2 中的留言详情定义删除字母,数字，汉字以外的所有符号的函数，加载停用词，并且删除字母，数字，汉字以外的所有符号。

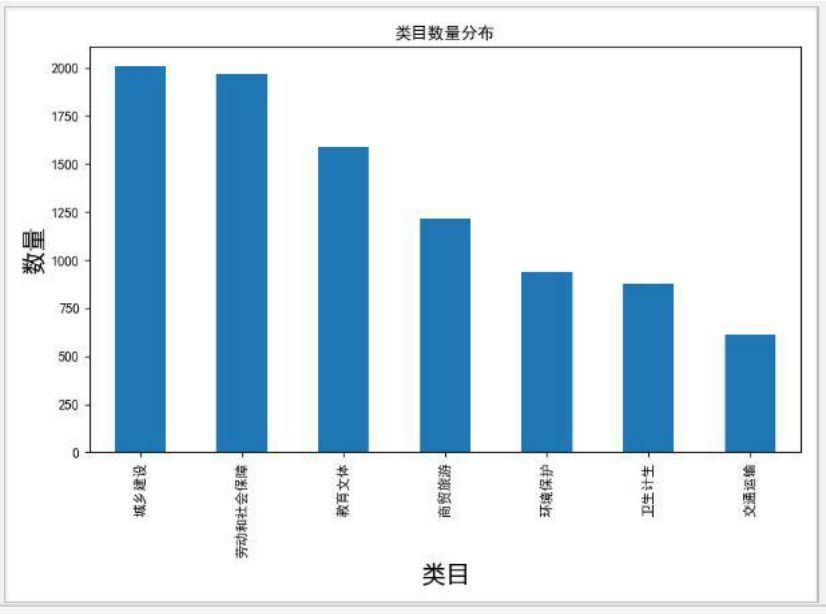

图 3 各大类的数据数目统计

步骤二：筛选最优模型 Linear Support Vector Machine

根据问题一的要求进行分析处理，我们运用逻辑回归（Logistic Regression）、多项式朴素贝叶斯（(Multinomial) Naive Bayes）、线性支持向量机（Linear Support Vector Machine）、随机森林（Random Forest）四个模型进行比较。

从箱体图（如图 4）上可以看出随机森林分类器的准确率是最低的，因为随机森林属于集成分类器(有若干个子分类器组合而成)，一般来说集成分类器不适合处理高维数据(如文本数据)。

文本数据有太多的特征值,使得集成分类器难以应付,另外上逻辑回归的准确率达到 80%，多项式朴素贝叶斯的概率达到 60%。其中线性支持向量机的准确率最高高达 90%。

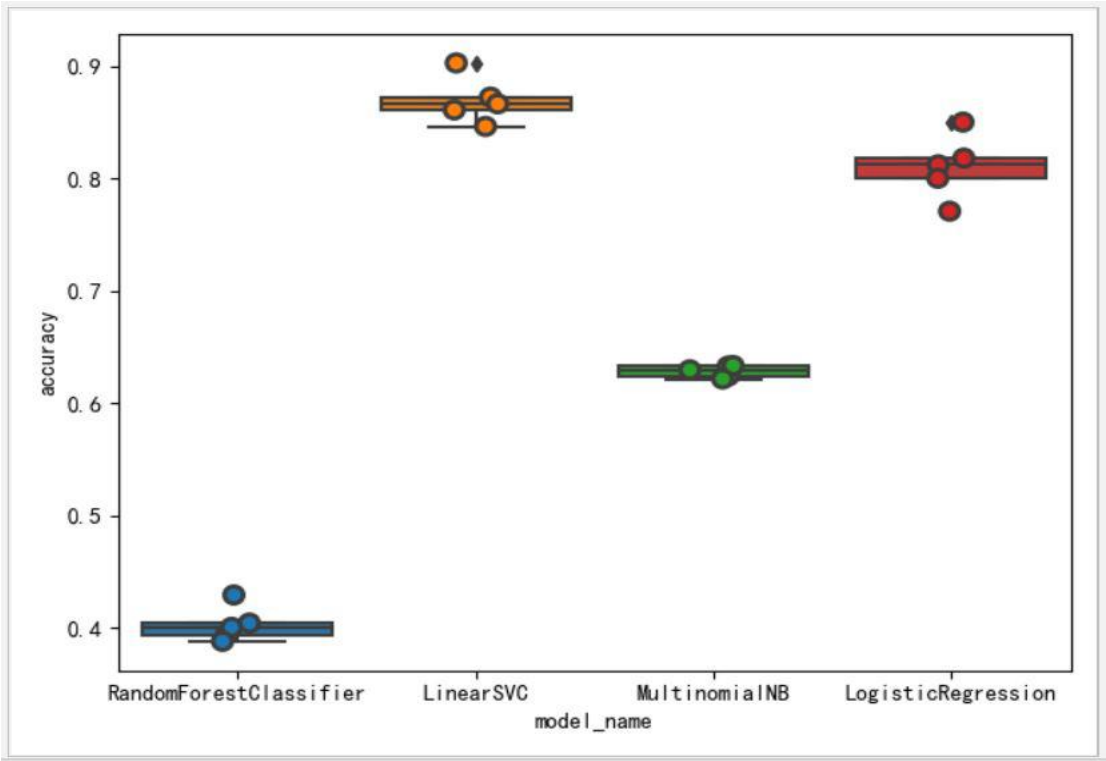

图 4 箱体图准确率统计

步骤三：模型检验

从平均准确率最高的 LinearSVC 模型中，我们将查看混淆矩阵（如图 5），检验预测标签和实际标签之间的差异。混淆矩阵的主对角线表示预测正确的数量，除主对角线外其余都是预测错误的数量，从上面的混淆矩阵可以看出"城乡建设"类预测最准确，只有两例预测错误。

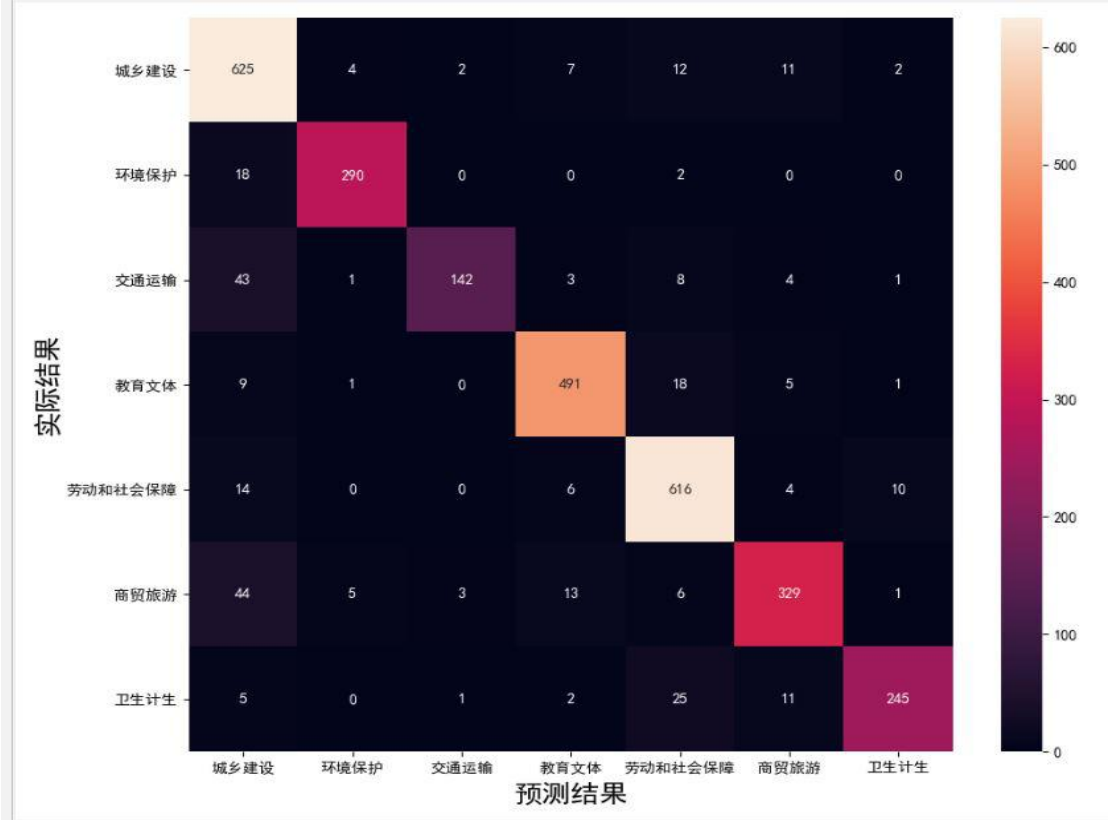

图 5 混淆矩阵

步骤四：提高预测精准度

为了提高预测的精确度，我们通过合并留言主题和留言详情数据两列数据使得其数据变得更加易于分辨，（如图 6）从而提高整个数据的预测准确率。

测试集预测结果：
[4 4 3 ... 5 6 6]
模型对于测试集预测的准确率:
0.9227095093356491
模型对整个数据集预测的准确率：
0.9792616720955483
权重为各类别数在y_true中所占比例为:
0.979266156263118

	precision	recall	f1-score	support
0	0.98	0.97	0.98	613
1	0.98	0.99	0.99	1969
2	1.00	0.97	0.98	877
3	0.98	0.97	0.97	1215
4	0.96	0.98	0.97	2009
5	0.98	0.98	0.98	1589
6	0.98	0.98	0.98	938
accuracy			0.98	9210
macro avg	0.98	0.98	0.98	9210
weighted avg	0.98	0.98	0.98	9210

图 6 合并后的数据处理

3、结论

根据题目所给的数据，建立了关于留言内容的一级标签分类模型，得到最优的模型——平均准确率最高的 LinearSVC 模型。混淆矩阵检验预测标签和实际标签之间的差异。

4、参考文献

[1]-派神-.使用 python 和 sklearn 的中文文本多分类实战开发[EB/OL].
https://blog.csdn.net/weixin_42608414/article/details/88046380.