南京農業大學

NANJING AGRICULTURAL UNIVERSITY

数字人文下的文本分类

目录
CONTENT

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

文本分类基本知识

1、文本分类

◆文本分类就是使用计算机模型自动为文本划分类别，是信息检索和自然语言处理中的核心任务。从数字人文的角度看，文本分类是大规模数据资源组织和整理的必要技术，小到特殊语句的自动识别，大到文本风格的自动发现，都可以使用文本分类的思路和模型来实现。

2、分类和聚类

◆文本分类一般和聚类一起探讨，其区别在于有没有预先划分好的类别。文本分类任务需要预先定义好具体类别，再接着将待分类文本划分入这些类别中；而文本聚类任务则不需要预先定义具体类别，只需提前规定好划分的类别数量，其根据文本内部的特征以及类别数量自动将待分类文本聚合成若干个类别。因此从机器学习的角度来看，文本分类是有监督学习，需要大量预先标注类别的训练语料；而文本聚类则是无监督学习，无需训练语料。

3、二分类和多分类

◆文本分类任务从模式上可以分为二分类和多分类。对于一篇文本来说，二分类就是判断其是不是属于某一个类别的；而多分类是判断其属于哪一个类别。对于包含N个类别的文本多分类任务，可以分解成N个二分类任务来解决。

4、空间向量模型

◆空间向量模型是文本分类任务的核心，是使用机器学习模型完成文本分类任务的前提。该模型将语料库看作一个多维空间，这样语料库中的文本就可以用空间中的向量来表示。文本向量的表示有很多种，传统的表示中向量的维度与空间的维度一致，每一个维度对应文本的一个特征，特征一般是重要的词语，特征值一般是加权后的词频如TFIDF等。空间向量模型下，使用向量之间的夹角余弦值来表示相似程度，进而用于文本分类。

5、特征与降维

◆高维度是空间向量模型的一大特点，也是影响该模型性能的主要桎梏。不经任何调整的空间维度与整个语料中词语的数量一致，因此维度可能非常大，造成所谓的维度灾难现象。因此一般需要对空间维度进行降维，减少计算的复杂度。降维的方法有卡方、信息增益、TFIDF等，主要思想都是通过高效的计算方法找出对于文本分类最有用的少量词语作为特征，重新构建向量空间，从而降低空间维度。

5、特征与降维

◆深度学习的表示学习思路为向量空间的降维带来了不一样的解决思路。结合语言模型的表示和神经网络结构的设计，可以从大规模语料中学习得到较低维度的稠密向量，用于表示词语、句子或者文本，从而大大降低了空间维度，能够有效提高文本分类的效果。

6、TFIDF

◆TFIDF是一种简单有效地特征加权方法，用于找出对文本分类任务重要的特征词语。TFIDF可以分成两个部分，TF表示词语频率，IDF表示倒文档频率，TFIDF是两者的乘积，如果将TFIDF值看作特征词语对文本分类任务的重要程度，那么权值越高越好。一方面，重要特征词的TF值和IDF值都应该越高越好；

6、TFIDF

◆另一方面，在文本分类任务中，一些常见词语并不能帮助提高文本分类的效果，如计算机类文本中的“硬件”、“系统”等词语，这时即使它们的TF值很高，也无益于文本分类任务，因此需要限制它们的权值，IDF发挥的就是这样的作用。那些出现在多数文本中的词语的IDF值会很低，这样即使它们的TF值很高，TFIDF值也不会太高，从而整体的权重得到了控制。

7、文本分类器

◆在空间向量模型的基础上，使用机器学习模型来最终实现文本分类的任务，这类模型一般称作文本分类器。传统的分类器主要有朴素贝叶斯、K最近邻和支持向量机等，深度学习兴起之后，分类器的设计变得更加多样，结合低维稠密向量的表示学习结果，通过对神经网络输出层的softmax等设计，或者再结合传统的支持向量机等模型，能够达到更好的文本分类效果。

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

文本分类在数字人文研究中的应用

提取文献中具有特定价值的文本

◆虽然古籍文本中包含大量具有分析价值的文字，但这些高价值的文字分布散乱。传统的依赖人工标注的方法不仅费时费力，而且效率低下。一些学者采用NLP技术来代替人工进行抽取，例如: 黄水清[1]等基于CRF,BiLSTM, Bi-LSTM-CRF三种序列标注模型抽取部分《十三经注疏》中抽取引书文献，并基于文献计量学中的引文分析的方法分析著述者的引用行为。

提取文献中具有特定价值的文本

◆在另一份相关研究中，周好[2]等基于文本分类的思路以SVM,Bi-LSTM,BERT等模型抽取古籍引书句，BERT模型在实验中取得了较为良好的效果。鲁国轩[3]等设计了一种识别数字人文相关研究的机器学习分类算法，对图情领域的数字人文文献有较好的识别效果。梁媛[4]等利用文本分类的思想从《春秋三传》中抽取描述同一事件的不同文本，证实深度学习算法可用于古汉语平行语料库的构建。赵建明[5]采用机器学习方法识别《史记》中的伪作，筛选出《史记》中语言风格明显和其他文本有差别的文章。

为已有文本构建自动分类体系

◆人文学科的研究者在从事文献整理工作时的重要环节之一就是正确地给文本内容分类，如果以人工方法执行这一环节，则不可避免地需要对所分类文本进行通读，且分类粒度较难控制。数字人文研究者们试图利用文本分类技术来应对这一问题，相关研究如秦贺然[1]等基于利用sklearn工具包地特征提取方法将命名实体特征加入分类器用于典籍文本地自动分类，取得了良好效果。胡昊天[2]等基于sikuBERT和sikuRoberta预训练模型对《四库全书》子部14个类别地古籍文本进行分类，最高取得了90.39%的整体分类F值。

分析特定文本的情感、意境等信息

◆王东波[1]等为争对先秦典籍的问句设计了分类体系，分别使用统计学习模型和深度学习模型开展分类研究。胡韧奋[2]等以空间向量模型将唐诗文本转化为文本向量，使用NB算法和SVM算法构建文本分类器，实现对唐诗题材的自动分类。蒋俊成[3]将多种深度学习模型应用于古代诗歌的意境识别和情感分析，并以此为基础设计了诗歌自动推荐系统推荐具有相似意境的诗歌。张馨怡[4]等为古典诗歌训练对应词向量，利用TextCNN模型筛选出具有爱国情怀的古诗词并进行用词分析。

南京農業大學
NANJING AGRICULTURAL UNIVERSITY

非物质文化遗产的文本分类

文本预处理

◆预处理工作是分类的第一步也是关键一步，主要包括分词、去停用词。

a）分词

对于现代汉语分词算法一般选用jieba库。同时，由于非遗语料中含有大量专有名词，比如“文水鈲子”、“阿诗玛”及“汗青格勒”等，这些语词或代表一种音乐艺术、一个民间故事，对于非遗文本的分类有着重要的区分作用。我们需要将这些专有名字作为用户自定义词库，加载入jieba。

b）去除停用词

中文停用词典常用的有哈工大停用词表、百度停用词表等，可以从https://github.com/goto456 /stopwords获取。

划分训练集、测试集

◆一般而言，我们将数据集分为两部分：训练集和测试集。其中训练集用来做特征工程、构建分类模型，而测试集用来对模型的效果进行评测。此外，验证集来自对训练集的再划分，目的是为了模型的选择和调参。可使用sklearn工具包的train_test_split方法将数据集划分。

特征提取

◆常用的特征提取方法有TF-IDF、Word2vec，而TF-IDF在实现上较容易，但仅仅利用词频这一静态信息，词的位置和词间关系等相应的动态信息没有被使用。

◆word2vec采用CBOW或skip-gram两种模型实现词向量的预测工作，考虑语词的上下文信息，其调用也非常简单。

◆应当注意的是，到此我们只是求出了一个样本中每个语词的向量化表示，而对于完成的样本表示而言，一般选择把所有的词向量相加，再求平均值。当然，如果利用深度学习模型，特征的学习过程可以由模型自己完成（比如pytorch的nn.Embedding）。

构造分类器

◆除了sklearn库，nltk库同样功能强大，其中提供了NaiveBayesClassifier、DecisionTreeClassifier、MaxentClassifier三种类型的分类器。分类器都提供了类方法可以训练出一个分类器实例，有了这个实例，便能对新的样本进行分类预测，以及其进行准确度评测。

构造分类器

◆更复杂的分类器如TextRNN、TextRCNN及TextCNN需要借助于pytorch、tensorflow等深度学习框架进行搭建，模型结构一般分为embedding层、隐层（lstm/conv+maxpooling）和输出层（linear+softmax），有兴趣可参照我们组的另一个github项目（https://github.com/veigaran/ NLP_ROAD），该仓库包含了常见的机器学习分类算法，也实现了TextCNN、TextRNN、BERT等深度学习模型的分类算法，可自行探索。

模型评估

◆一般而言，我们使用准确率、召回率和F值来评价模型效果，对于多分类任务而言，可以选择宏平均或微平均。