中文文本分类系统

方案设计

系统设计目的

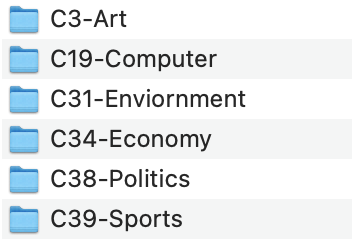
中文文本分类是自然语言处理中的一项重要任务，其目的是将给定的中文文本自动分类到不同的预定义类别中。中文文本分类可以应用于很多领域，如情感分析、舆情监测、文本过滤、垃圾邮件检测等。

中文文本分类系统的研究目的和背景是为了提高中文文本分类的准确率和效率，使其在实际应用中更加可靠和有效。当前，随着互联网的快速发展和人们对信息获取和处理的需求日益增加，中文文本分类系统得到了广泛的应用和研究。然而，由于中文的语言特点和文本结构的复杂性，中文文本分类面临着许多挑战，如词汇歧义、语义消歧、文本长度差异等。

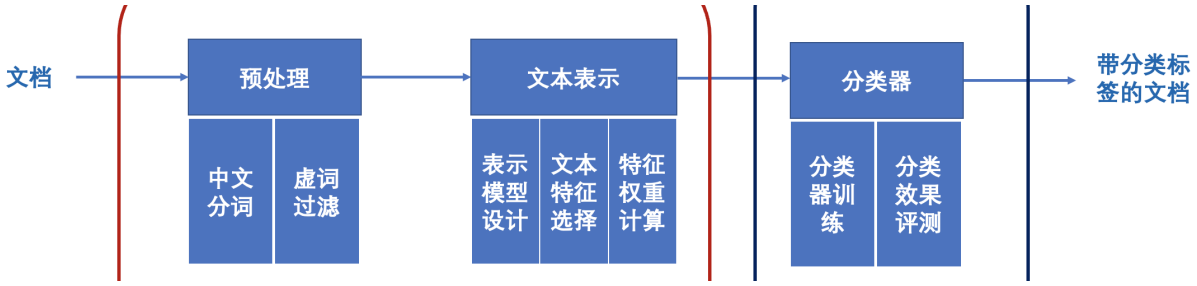
因此，中文文本分类系统的研究旨在通过构建更加准确和高效的中文文本分类模型，解决上述问题，提高中文文本分类的性能和实用性。此外，随着机器学习和深度学习技术的不断发展和应用，中文文本分类系统的研究也包括探索和应用新的算法和模型，以提高分类性能和效率。

数据集

数据集中包含训练集和测试集，包含艺术、计算机、环境、经济、政治和体育6个分类类别，训练集每个类别含有200个数据，测试集每个类别含有100个数据，数据分布均匀，每个数据都是一段中文文本报道，存储在txt文件中。



系统流程



预训练

预训练中主要包括中文分词和虚词过滤

中文分词是将一段中文文本分成一组有意义的词语序列，是中文文本处理的基础，常用的分词工具有jieba、pkuseg、ICTCLAS等。

虚词过滤是去除一些无意义的词语，例如介词、连词、副词等。这样可以减少特征数量，提高模型的训练效率。

经过预处理后，原始文本数据经过清洗、过滤、转换等操作，使得文本可以被计算机更好地理解和处理，提高最后分类的效率和准确性。

文本表示

文本表示将经过预处理后的文本表示成计算机可以理解的形式，即将文本转化为数值向量。主要包括：表示模型设计、文本特征选择、特征权重计算。

表示模型设计：文本表示模型的设计是文本分类的关键。常用的文本表示模型有词袋模型、TF-IDF模型、Word2Vec模型和BERT模型等。不同的模型有不同的优缺点，需要根据具体任务来选择。

文本特征选择：文本特征选择是在表示模型的基础上，根据特征选择算法选取有用的特征。常用的特征选择算法有卡方检验、互信息、信息增益等。

特征权重计算：对于文本特征选择后的词语，需要计算它们的权重。常用的权重计算方法有TF-IDF方法和词频方法等。

分类器

分类器是将经过文本表示后的文本分到不同的类别中。常用的分类器有朴素贝叶斯分类器、支持向量机分类器、随机森林分类器、深度学习分类器等。分类器训练的过程是通过训练数据集，学习分类器的参数，然后利用测试数据集来测试分类器的性能。

结果分析

选择不同的分类器，通过一些指标评判模型的分类效果，如Accuracy、Precision、Recall、F-measure等。

输出错误案例，分析可能的分类错误原因。

方案实现说明

中文分词

在本实验中，使用jieba库实现了中文分词。它采用了基于字典匹配和规则模板的方法实现中文分词，根据语料库和词频构建字典树，使用前向最大匹配分词。

虚词过滤

在本实验中，使用了停用词表对文本进行虚词过滤。其中停用词表是根据中文停用词表 、哈工大停用词表、百度停用词表和四川大学机器智能实验室停用词库构建的。

据此可以得到一个简单的预处理案例：

对中文分词和虚词过滤的简单测试

```
def preprocess(text):
    # 分词
    words = jieba.lcut(text)
    # 去除虚词
    words = [word for word in words if word not in stopwords]
    f_word = " ".join(words)
    # 返回处理后的词语列表
    return f_word, words
```

```
text = '这书是我昨天在杭电图书馆借的的。'
f_word, words = preprocess(text)
print(words)
```

```
['这书', '昨天', '杭电', '图书馆']
```

而综合所有的训练集数据，在经过中文分词和虚词过滤之后，可以得到如下的词云图：

文本表示

本实验中，使用了TF-IDF值作为特征选择的依据，这么做的原因是TF-IDF（Term Frequency-Inverse Document Frequency）是一种常用的文本表示方法，用于评估一个词语在文本中的重要程度。其中，TF表示词项频率（Term Frequency），用于衡量一个词在当前文本中出现的频率，TF越高，代表该词在文档中越重要。IDF表示逆文档频率（Inverse Document Frequency），用于衡量一个词的普遍重要性，IDF值越大，代表该词语越重要。具体来说，TF-IDF的计算公式如下：

$$tf(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

$$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$tf\text{-}idf(t,d,D) = tf(t,d) * idf(t,D)$$

随后进行特征选择，本实验选择了1000个特征，将每个文本表示成一个1000维的向量，其中每一维表示一个特征的TF-IDF值

最后是计算特征权重，使用**selector.scores**返回每个特征的得分，并排序，得到右图的特征权重表（仅展示前31个）

	Feature Words	Weights
0	政治	231.753898
1	体育	171.011286
2	艺术	151.928276
3	经济	110.958617
4	民主	67.613340
5	运动员	59.365736
6	文艺	57.517213
7	土壤	55.195053
8	干部	46.643702
9	社会主义	45.823928
10	文学	44.354118
11	作品	41.489863
12	训练	40.776333
13	运动	40.639840
14	企业	40.627949
15	知识经济	40.571082
16	投资	36.411964
17	增长	36.378907
18	全球化	35.960258
19	算法	35.136509
20	创作	31.513614
21	数据库	31.507000
22	agent	30.552870
23	奥运会	29.875805
24	小说	29.388399
25	比赛	29.312757
26	健身	28.586544
27	竞技	28.583580
28	数据	27.686066
29	服务器	27.338746
30	权力	27.218063

分类器训练：

分类器	原理	优点	缺点
朴素贝叶斯分类器	基于贝叶斯定理，假设每个特征相互独立	算法简单，计算速度快；对于大规模数据集和高维数据有较好的分类效果	对于特征之间存在较强关联或者样本数据分布不满足独立同分布假设的情况，效果可能不好
决策树分类器	通过构建树形结构进行分类	易于理解和解释，具有可视化的优势；不需要对数据进行归一化处理；能够处理具有缺失值的数据；适用于离散型和连续型特征数据	容易出现过拟合的情况，需要采取剪枝等方法进行处理；在处理特征关联性较强的数据时，分类效果可能不佳
支持向量机分类器	通过映射到低维空间，找到最优分界面进行分类	能够处理高维度数据集，分类效果较好；泛化能力强，对于小样本学习问题适用性较高	对于数据量过大和非线性问题计算复杂度较高；对于样本噪声和非均衡数据较为敏感
k近邻分类器	找到距离测试样本最近的k个训练样本进行分类	不需要事先对数据进行建模，分类效果较好；对于异常值不敏感；适用于非线性问题	需要保存所有的训练数据，对于数据量较大的情况，计算开销较大；对于高维数据和稀疏数据分类效果较差
随机森林分类器	通过多棵决策树的投票结果得到分类结果	能够处理高维度数据集，且对于缺失值和异常值的处理较为鲁棒；能够有效地避免过拟合问题	对于一些较为复杂的数据集和问题，分类效果可能不如其他算法；算法可解释性较差

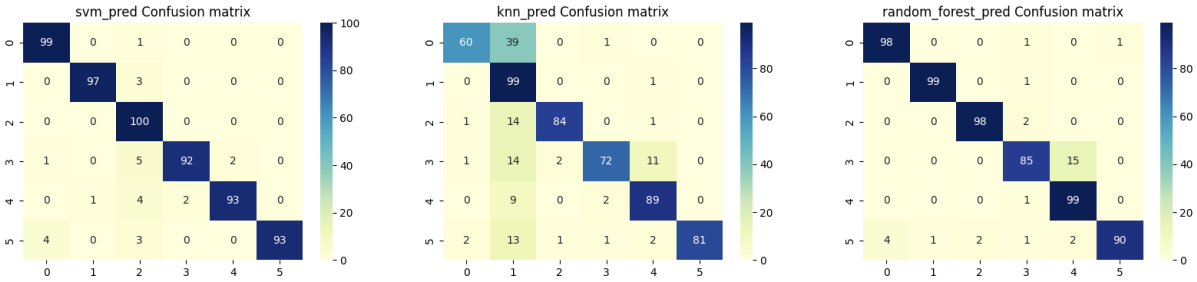
结果评测

经过实验，我们得到了如下结果：

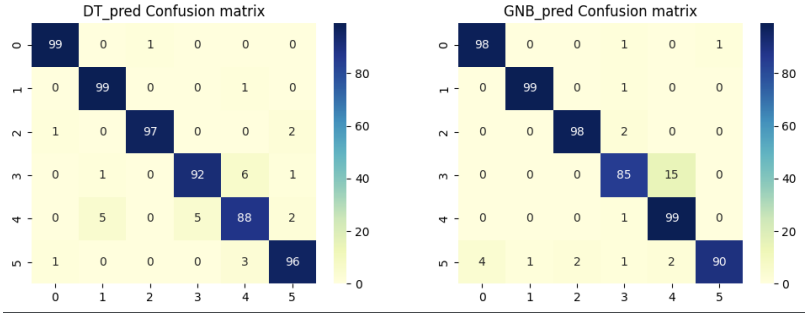
	Accuracy	Precision	Recall	F1-score	running time
朴素贝叶斯分类器	94.83%	0.95	0.95	0.95	9.1s
决策树分类器	95.17%	0.95	0.95	0.95	9.0s
支持向量机分类器	95.67%	0.96	0.96	0.96	9.5s
k近邻分类器	80.83%	0.87	0.81	0.82	10.0s
随机森林分类器	96.50%	0.97	0.96	0.96	15.4s

混淆矩阵

支持向量机、K近邻、随机森林：



决策树、朴素贝叶斯：



以上指标表明，我们的中文文本分类系统在数据集上表现良好，在表现最好的随机森林分类器中，准确率、召回率和F1值均达到了0.9以上，说明系统可以有效地对中文文本进行分类。

错误案例

输出随机森林分类器的一个错误分类：

```
【 文献号 】2-3871
【原文出处】新闻与传播研究
【原刊地名】京
【原刊期号】199601
【原刊页号】11-17
【分 类 号】G6
【分 类 名】新闻学
【 作　者 】张西明
【复印期号】199603
【 标　题 】中国电视业的现状和问题
【 正　文 】
　　截至1995年年中，中国的电视机拥有量达2.5亿台，　电视观众突破8亿，无线电视台为970余座，有线电视台则多达1200 多家，　另外还有1000家左右的教育台。电视行业的人通常用“三个一千”的形象说法来描绘这种格局，即“无线一千，有线一千，教育台一千”。〔1 〕从这组数字看，中国是无可争议的“世界上最大的视听市场”，而且是最繁忙的市场。可以这样说，每天我们的视野中都有各式各样新的建筑拔地而起，与此同时，我们身边看不见的空间里也每时每刻都在涌现着新的视听信号，渐渐地把电视机的频道全部占满。另一方面，犹如一处处欣欣向荣的建筑工地背后隐藏着这样那样的无序和混乱一样，透过电视的卫星信号、微波信号和有线电缆展现出的勃勃生机，也能明显感觉到夹杂其中的纷乱、混乱和浮躁。浮躁中透着勃勃生机，可以说是今天中国电视业的写照，而且也是其可以预见的明天。
　　本文意在总结当前我国电视业的主要发展趋势和特点，特别是电视行业内部对现状和发展的看法，笔者希望通过这篇文章与有关专家学者讨论这个领域的管理和发展问题。
　　一、有线电视的发展及遇到的问题
　　我国新兴的电视业市场中，发展情况并不是铁板一块。一些新的部门、部类正在崛起，一些传统部门则不得不挖潜、革新。这里只能择要对一些方面进行描述，破题之处是近年来电视业中发展较快、问题也较多的方面军——有线电视。
　　我国的有线电视，自80年代中期开始出现，最近几年迅猛发展。有线电视的优点，是能使观众同时收看到十几个乃至几十个频道，这在较大程度上丰富了电视屏幕和观众的文化生活。截至1995年5月份，　我国有线电视用户终端数为3000万，经广播电影电视部批准的有线台目前已达1200座。〔2 〕实际的用户终端数和有线电视台数量远远超过了前两个数。由于经济利益的驱动，全国各地发展有线电视的积极性都非常高
...
```

rf_pred is C19-Computer, while real label is C39-Sports

在test中，这篇文本属于体育板块，但是查看原文后发现文本并不太像体育报道，因此，错误分类的原因可能是因为部分数据的标注出现问题。

小结

本实验设计了一个中文文本分类系统，该系统分为预训练、文本表示和分类器三个步骤。通过使用jieba分词库进行中文分词、使用停用词表进行虚词过滤、使用TF-IDF算法进行特征选择和特征权重计算、使用朴素贝叶斯算法进行分类器训练，我们成功地对数据集进行了文本分类，并取得了不错的分类效果。

对比中文和英文的文本分类系统，主要的差别如下：

1. 分词：中文文本需要进行分词，将一个连续的字符串划分成有意义的词语，而英文文本则不需要进行分词。在中文文本分类中，分词的质量对文本特征的选择和分类器的准确性都有很大的影响。
2. 特征选择：中文和英文文本的特征不同，需要采用不同的特征选择方法。中文文本特征通常采用TF-IDF等算法，而英文文本则采用词袋模型或者N-gram模型等算法。这是因为中文文本特征通常需要考虑词语的词序信息，而英文文本词汇通常以空格或标点符号分隔，因此英文文本的特征主要是词汇。
3. 分类器：中文和英文文本的分类器也有所不同。中文文本分类器通常采用SVM、朴素贝叶斯等方法，而英文文本分类器则通常采用朴素贝叶斯、决策树等方法。此外，中文文本的分类器还需要考虑汉字的复杂性和歧义性等问题。