西北師範大學

2021年学生“创新创业能力提升计划”

课题研究项目申请表

项目名称：新闻文本分类算法

申 请 者：周学铭

指导教师：代祖华

学　　院：计算机科学与工程学院

西北师范大学创新创业学院制

填 表 须 知

1、主要目的：推动创新创业教育与思想政治教育紧密结合、与专业教育深度融合，促进学生全面发展。倡导学生开展研究性学习，支持学生在科研中学习，使学生早进课题、早进实验室、早进团队，培养学生的科学研究能力；倡导学生理性创业，鼓励学生将创业实践与专业学习相结合，将创新项目与创业实训相结合，激发学生的创造力。

2、实施原则：面向全体，分类施教；引导激励，项目带动；自我实践，动态管理；尊重创造，宽容失败；师生共创，教学相长。

3、课题研究项目组成员在指导教师指导下开展研究，指导教师应在学术思想、研究技术手段与研究方法、研究工作成果分析总结方面给予指导，引导学生自主学习、自主完成研究计划；及时指导和跟踪研究活动，对研究活动作出评价，并给出书面意见和建议；应注重学生创新思想的激发，培养团队的协作精神；应培养学生科学研究的兴趣、实事求是的科学精神、坚韧顽强的意志品格。

4、立项按照学生自主申报、学院初审推荐、学校专家委员会评审、公示评审结果的程序进行，经主管副校长批准后立项，学校发文公布，根据评审结果确定立项课题的资助级别。

5、项目申请人原则上为学有余力，身体健康，对科学研究有浓厚兴趣，能够保证开展课题研究所需时间和精力，具备初步科研和动手能力的各年级学生。鼓励学生组成项目组合作申请项目，原则上每组 3－5 人为宜，最多不超过 5 人。

6、资助项目立项年限（指迄止时间）原则上应在 1 年内完成；个别项目由于课题研究的实际需要也可 2 年时间内完成结题、答辩和研究成果汇编工作。

7、资助项目立项申请表要求逐项认真仔细填写，内容应言简意赅，思路清晰，论证充分，表述明确。

8、每年 5 月校院组织资助课题中期检查，每年 11 月组织资助课题结题、答辩。

9、如填表有不明事宜，请与学校学生“创新创业能力提升计划”指导委员会办公室（联系电话：7971943）或学院“学生创新创业能力提升计划指导小组”（学院团委）咨询。

一、项目申请简表（10 分）

<table>
<tr><td colspan="2">项目名称</td><td colspan="8"></td></tr>
<tr><td colspan="2">项目类别
标记（√）</td><td colspan="8">（　）一般项目
（√）竞赛专项项目
（　）“国家级大学生创新创业训练计划”专项培育项目</td></tr>
<tr><td colspan="2">申请金额</td><td colspan="2">2000</td><td>起止年月</td><td colspan="5">2021.4.30—2021.7.15</td></tr>
<tr><td rowspan="4">申请人</td><td>负责人姓名</td><td>周学铭</td><td>性　别</td><td>男</td><td>学号</td><td>201871010130</td><td>民族</td><td>汉</td></tr>
<tr><td>专　　业</td><td colspan="3">计算机科学与技术</td><td>班级</td><td colspan="3">18 卓越班</td></tr>
<tr><td>所在学院</td><td colspan="2">计算机科学与工程学院</td><td colspan="2">联系电话</td><td colspan="3">15557883208</td></tr>
<tr><td colspan="8"><table>
<tr><td>成员（不超过四人）</td><td>姓　名</td><td>学号</td><td>专业</td><td>项目中的分工</td><td>签 字</td></tr>
<tr><td></td><td>常雅伦</td><td>201871030107</td><td>计算机科学与技术</td><td>文档</td><td></td></tr>
<tr><td></td><td>何飞</td><td>201871030110</td><td>计算机科学与技术</td><td>测试</td><td></td></tr>
<tr><td></td><td>谢林江</td><td>201871030131</td><td>计算机科学与技术</td><td>UI 设计</td><td></td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td></tr>
</table></td></tr>
<tr><td>项目内容概述</td><td colspan="8">随着互联网技术和移动通信技术的飞速发展，网络新闻成为人们日常生活、学习和工作的重要信息来源之一。相比于其他传统媒体，网络新闻具有内容丰富、形式多样，可以实现实时更新，用户获取与阅览网络新闻不受时间和空间的限制，能够实现随时获取，用户可以根据自己的需求从互联网中获取满足自身需求的新闻等优点。网络新闻能够给用户带来更好的立体、全方位的信息接收体验。但面对海量且混乱无序的网络新闻信息，用户所面临的难题不再是如何找到新闻资源，而是如何从海量的新闻资源中高效准确的获取能够满足自身需求的信息。因此，为满足网络新闻用户在大数据时代背景下的多样化和个性化需求，对网络新闻进行有效的信息组织与管理已成为亟待需要解决的问题。
二十世纪九十年代以后，随着互联网上文本信息资源的飞速增长，文本分类方法得到了空前的关注。而基于知识工程的文本分类技术已经完全不能满足需求，从而被新兴的基于统计的机器学习文本分类方法逐渐替代，并很快成为文本分类领域新的主流方法，至今仍是众多学者研究的重点。基于统计的机器学习的文本分类方法通过对已知类别的样本数据进行学习，利用学习到的类别特征构建分类器，然后利用分类器对待分类的文本信息进行分类，最终获取文本信息的类型。与基于知识工程的方法相比较，这种方法中分类器的构造不需要人工的参与，大大减少了人力物力，可以处理大量的文本信息，并且由于这类分类器的训练是基于某些算法而非特定领域的专业知识，因此具有更好的通用性和很好的适应力，在文本的分类效率和准确率上都有非常显著的提升。正是由于具有较可靠的理论基础和更好的分类结果，基于机器学习的文本分类方法得到了学者们的广泛关注，至今仍是研究人员应用与研究的重点和主流，应用领域广泛，如文本挖掘、模式识别、信息检索、数据挖掘、学习系统等领域。机器学习方法中目前比较常用的</td></tr>
</table>

	有类中心向量法、K 最近邻法、支持向量机法等等。 本项目以深度学习方法为理论基础，运用深度学习的相关理论和模型来重构网络新闻文本分类过程，达到解决传统文本分类存在的问题，提高文本分类效果的目的。

二、项目背景及可行性分析（20 分）

1.项目背景

随着互联网技术和移动通信技术的飞速发展，网络新媒体已经成为信息交互的有效平台。其中非结构化的新闻文本作为信息的一种重要承载形式呈爆炸式增长。相比于其他传统媒体，网络新闻具有内容丰富、形式多样，可以实现实时更新，用户获取与阅览网络新闻不受时间和空间的限制，能够实现随时获取，用户可以根据自己的需求从互联网中获取满足自身需求的新闻等优点。网络新闻能够给用户带来更好的立体、全方位的信息接收体验。但同时如何能高效准确地对海量新闻文本进行分类,从而提取能够满足自身需求的信息成为用户所面临的难题。并且由于其内容简短,表达方式多样化和语法结构不规范,增加了分类的难度。因此，为满足网络新闻用户在大数据时代背景下的多样化和个性化需求,迫切需要一种有效的文本分类算法对文本语义进行更好地提取,从海量的新闻文本中挖掘出有价值的信息。

2.现状分析

二十世纪九十年代以后，随着互联网上文本信息资源的飞速增长，文本分类方法得到了空前的关注。而基于知识工程的文本分类技术已经完全不能满足需求，自从深度学习思想被提出以来,已经在图像识别、机器翻译和语音识别等领域中取得了出色的表现。和传统机器学习算法相比,深度学习模型通过多层非线性空间的变换,能够刻画出数据的本质特征,为提高新闻文本分类模型的准确性提供了良好的基础。深度学习模型中的卷积神经网络（Convolutional Neural Network,CNN）已成为一种主流的文本分类模型。

3.创新点与项目特色

本项目基于卷积神经网络的新闻文本分类框架,对文本分类中的特征表示、特征提取和分类器构造等关键环节进行了不同程度地改进。以便于高效准确地对海量新闻文本进行分类,提取所需信息，用于解决目前在信息时代之下用户最关心的问题，从而满足用户大数据时代背景下的多样化和个性化需求。

本项目的创新点有：

（1）与基于知识工程的方法相比较，这种方法中分类器的构造不需要人工的参与，大大减少了人力物力，可以处理大量的文本信息，并且由于这类分类器的训练是基于某些算法而非特定领域的专业知识，因此具有更好的通用性和很好的适应力，在文本的分类效率和准确率上都有非常显著的提升。

（2）利用 Python 爬取相关的测试数据；利用神经网络等机器学习算法进行开发；

同时熟练利用 Java 开发较为完善的可视化界面。

（3）用户的使用成本基本为零。

（4）目前社会上该类产品较少，竞争不太激烈。主要运用的技术都是传统的数据挖掘算法，例如：朴素贝叶斯等。对于神经网络在文本分类算法上的开发较为少见。

4.可行性分析

技术的可行性：Java 目前已存在相当成熟的前后端框架，开发可视化界面较为方便。Python 爬取数据的能力卓越。目前市面上主流的分类算法还是较为传统的数据挖掘算法。

法律的可行性：Java 的框架大多是开源的，没有知识产权相关条件的约束。

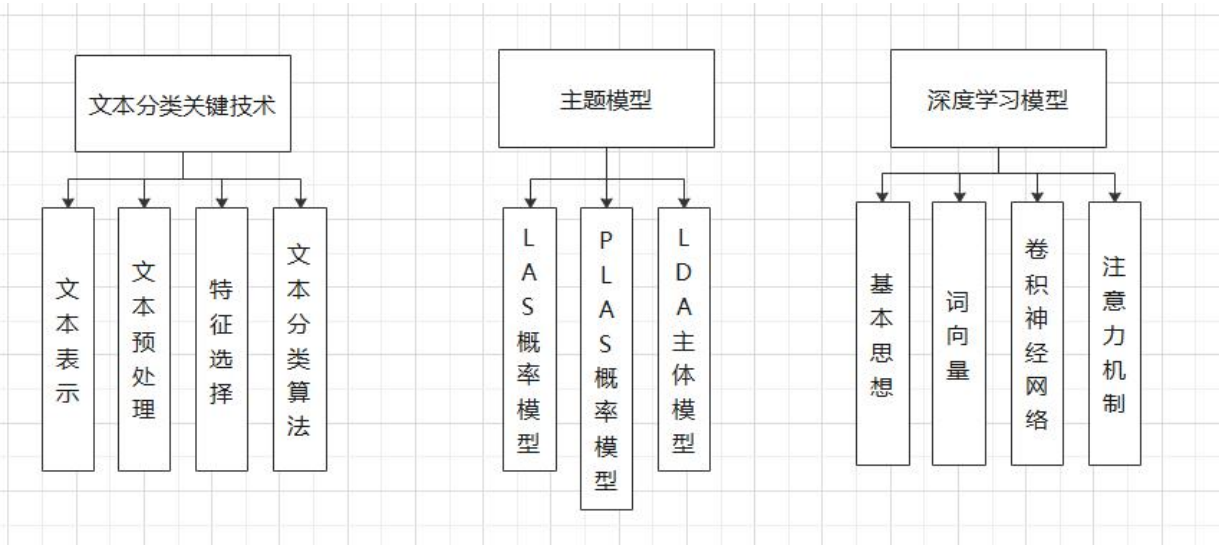
三、项目实施方案（20 分）

1 研究目标

本项目致力于设计一种能够输出分类的准确率不低于 80%，且可以提供简单的可视化界面，能够输入单条新闻，输出新闻的分类，或者支持本地上传 csv/xlsx 文件，批量输入新闻，并输出新闻分类的算法平台。

2 研究内容

① 算法理论基础



② 卷积神经网络(CNN):卷积神经网络仿造生物的视知觉（visual perception）机制构建，可以进行监督学习和非监督学习，其隐含层内的卷积核参数共享和层间连接的稀疏性使得卷积神经网络能够以较小的计算量对格点化（grid-like topology）特征，例如像素和音频进行学习、有稳定的效果且对数据没有额外的特征工程（feature engineering）要求。

③ 注意力机制（Attention Mechanism）源于对人类视觉的研究。在认知科学中，由于信息处理的瓶颈，人类会选择性地关注所有信息的一部分，同时忽略其他可见的信息。上述机制通常被称为注意力机制。人类视网膜不同的部位具有不同程度的信息处理能力，即敏锐度（Acuity），只有视网膜中央凹部位具有最强的敏锐度。为了合理利用有限的视觉信息处理资源，人类需要选择视觉区域中的特定部分，然后集中关注它。例

如，人们在阅读时，通常只有少量要被读取的词会被关注和处理。综上，注意力机制主要有两个方面：决定需要关注输入的哪部分；分配有限的信息处理资源给重要的部分。

④ TF-IDF（term frequency－inverse document frequency）是一种用于信息检索与数据挖掘的常用加权技术。TF 是词频(Term Frequency)，IDF 是逆文本频率指数(Inverse Document Frequency)。

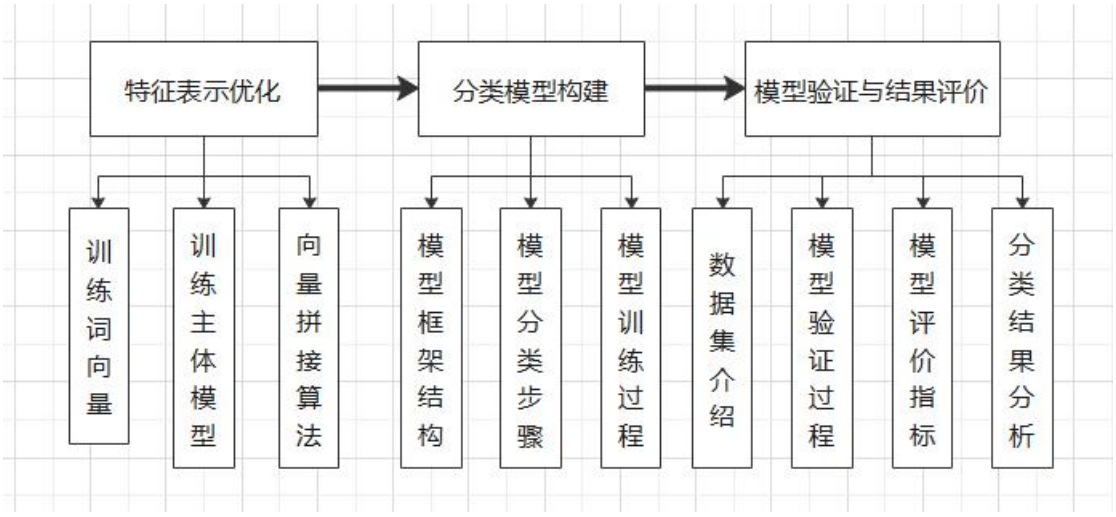
⑤ Word2vec，是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络，用来训练以重新建构语言学之词文本。网络以词表现，并且需猜测相邻位置的输入词，在 word2vec 中词袋模型假设下，词的顺序是不重要的。训练完成之后，word2vec 模型可用来映射每个词到一个向量，可用来表示词对词之间的关系，该向量为神经网络之隐藏层。

⑥ Layui（谐音：类 UI）：是一套开源的 Web UI 解决方案，采用自身经典的模块化规范，并遵循原生 HTML/CSS/JS 的开发方式，极易上手，拿来即用。其风格简约轻盈，而组件优雅丰盈，从源代码到使用方法的每一处细节都经过精心雕琢，非常适合网页界面的快速开发。layui 区别于那些基于 MVVM 底层的前端框架，却并非逆道而行，而是信奉返璞归真之道。

⑦ Spring Boot：SpringBoot 是由 Pivotal 团队在 2013 年开始研发、2014 年 4 月发布第一个版本的全新开源的轻量级框架。它基于 Spring4.0 设计，不仅继承了 Spring 框架原有的优秀特性，而且还通过简化配置来进一步简化了 Spring 应用的整个搭建和开发过程。另外 SpringBoot 通过集成大量的框架使得依赖包的版本冲突，以及引用的不稳定性等问题得到了很好的解决。

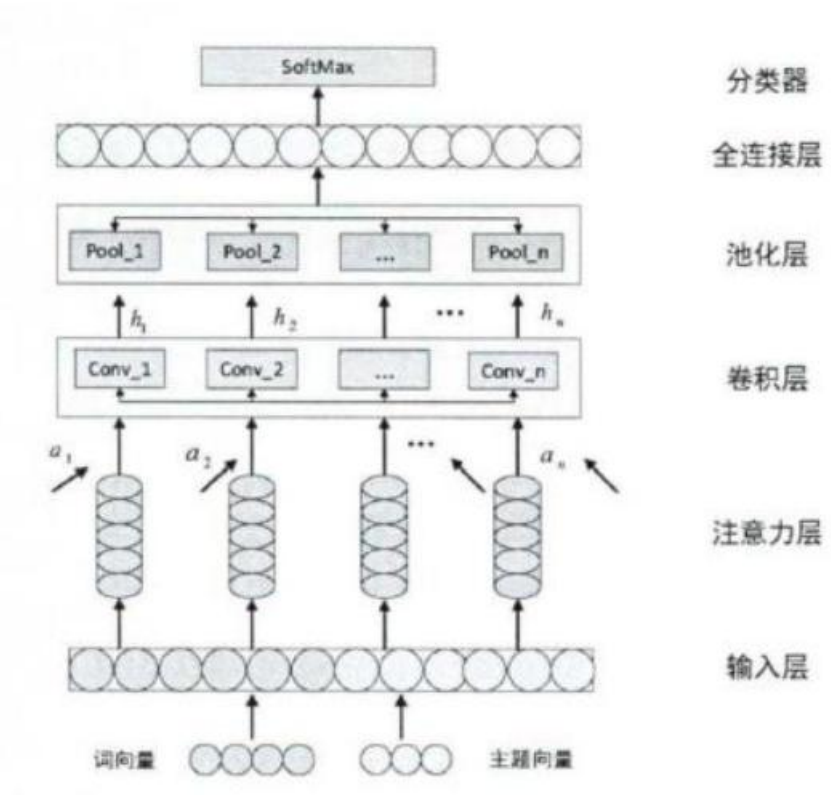
3 拟采取的研究方法及技术路线

① 新闻文本分类过程



1. 训练词向量：词向量的可以通过训练学习到词的分布式表示，给定 n 篇文章，就可以出词向量。关于词向量原理，先从神经网络语言模型(Neural Network Language mNNLM）谈起，NNLM 直接通过一个神经网络结构对 n 元条件概率进行评估,元语言模型进行建模，通过词的 n-1 个历史词，来预测当前词。

2. 模型结构设计：基于 CNN 的新闻文本分类模型主要由输入层、注意力层、卷积层、池全连接层和 softmax 分类器六层组成,模型结构图如图所示;



3. 数据集：使用的新闻文本语料集为 THU 公开新闻数据集，数据内容和类样，适合于本文的文本分类任务。本文使用其中的 10 大类别，每个类别选择样本数量为 8000 条。

② 系统设计流程

4 项目进度时间安排

（文献调查）：2021 年 4 月 21 日到 2021 年 4 月 30 日
（社会调查）：2021 年 5 月 1 日到 2021 年 5 月 7 日
（可行性研究）：2021 年 4 月 21 日到 2021 年 5 月 10 日
（需求分析）：2021 年 4 月 21 日到 2021 年 5 月 10 日
（方案设计）：2021 年 4 月 23 日到 2021 年 4 月 30 日
（实验研究）：2021 年 5 月 8 日到 2021 年 6 月 1 日
（数据处理）：2021 年 5 月 1 日到 2021 年 5 月 15 日
（研制开发）：2021 年 5 月 25 日到 2021 年 6 月 30 日
（单元测试）：2021 年 6 月 1 日到 2021 年 6 月 30 日
（综合测试）：2021 年 7 月 1 日到 2021 年 7 月 15 日

四、项目预期成果（5 分）

该网站初步上线，通过新闻数据，使用新闻文本分类算法，根据新闻标题和内容，进行分类，同时用户可通过输入单条新闻，输出新闻的分类，同时支持本地文件的批量上传，单条新闻从上传到解析完成，时间不会超过 5s 钟，且同时具有异步传输的效果，可最大化程度上提供网站的效率和实用性，最大程度上为用户提供良好的交互，供用户快速有效的获取新闻信息。本项目作为一种核心算法，可作为知识产权出售获取收入。

五、经费预算

资料费	实验费	打印费	交通费	其他	合计
200	1300	200	100	200	2000
指导教师签字			负责人签字		

六、指导教师审查推荐意见

指导教师签字：

年　　月　　日

七、学院学生“创新创业能力提升计划”项目评审小组审查推荐意见

（1）是否同意予以立项：_____

（2）建议资助金额：_____元

（3）本学院所具备的保证申请者开展此项研究所必须的基本条件：

负责人签字：

年　　月　　日（公章）

八、学校“创新创业能力提升计划”评审委员会审核意见

（1）是否予以立项：_____

（2）资助金额：_____元

（3）项目执行时间：_____年_____月至_____年_____月

年　　月　　日（公章）