# 《自然语言处理》教学大纲

## 一、课程基本信息

1. 课程名称：自然语言处理

2. 课程体系/类别：专业拓展课

3. 学分学时：4学分，90学时（含课堂教学42学时，实践教学48学时）

4. 适用专业：人工智能、数据科学与大数据技术

5.考核方式：考查

6. 先修课程：深度学习

7. 后续课程：知识图谱、大语言模型

8. 开课单位：人工智能学院

## 二、课程简介

本课程系统讲解了自然语言处理的基本概念、核心技术和主要应用，内容涵盖文本数据处理技术、规则自然语言处理、统计自然语言处理、语言模型与词向量表示、循环神经网络、Seq2Seq架构、注意力机制与Transformer、预训练模型的使用和模型的压缩与裁剪等内容以及在情感分析、机器翻译、智能问答、摘要生成和信息抽取等任务场景中的技术实践，旨在通过理论讲解和项目实践相结合的方式，帮助学生深入理解自然语言处理的基础原理和应用技术，了解国内外自然语言处理技术的发展概貌，具备运用基本原理和常用方法解决问题的能力，同时为以后从事自然语言处理与大模型相关的工作奠定基础。

## 三、课程思政

本课程通过自然语言处理技术的讲解和实践，强调自主创新和问题解决的重要性，激发学生的创新创造意识和爱国情怀。借助于科大讯飞人工智能实验平台，培育学生在掌握这一前沿技术的理论和方法的同时，具备社会责任意识和担当精神，将个人发展与国家发展相结合，使自然语言处理的应用场景拓展到更多社会公共服务中；同时引导学生深刻理解自然语言处理技术背后涉及的伦理和道德问题，加强学生在工程实践中的责任意识、道德意识、底线意识，更好地推动深度学习技术的发展和应用，从而为我国在人工智能领域取得领先优势，推动全球科技产业发展作出贡献。

## 四、课程目标

通过本课程学习，学生应达到的预期教学目标如下：

课程目标1：了解自然语言处理的基本概念、发展历程、应用场景以及研究方向与常规流程；

课程目标2：理解文本语料数据清洗与表示的操作方法以及K-means聚类、隐马尔科夫模型、条件随机场和朴素贝叶斯等常用的统计自然语言处理方法；

课程目标3：理解RNN循环神经网络、seq2seq网络架构、注意力机制、Transformer和模型的压缩与剪裁的基本概念、核心原理以及实现流程；

课程目标4：掌握预训练模型的调用与微调技术以及在情感分析、机器翻译、智能问答、摘要生成和信息抽取等经典任务场景中的实践技能；

课程目标5：能够按照指定的任务描述，利用所学知识和技术，选择合适的自然语言处理模型，根据性能需求、进行调优，建成模型，完成模型的部署与评估。

## 五、教学内容与课程目标的关系

**（1）理论教学**

本门课的主要教学内容与要求，以及学时分配如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **序号** | **教学内容** | **教学要求** | **学时** | **教学方式** | **课程思政**  **内容** | **对应**  **课程**  **目标** |
| 1 | 1、自然语言处理概述  （1）自然语言处理基本概念  （2）自然语言处理技术应用 | （1）了解自然语言处理的基本概念和发展历程  （2）了解自然语言处理的研究方向和应用场景  （3）理解自然语言处理的研究技术和常规流程 | 2 | 讲授 | 培养学生学习兴趣，推动技术向善的社会责任感 | 目标1  目标5 |
| 2 | 2、文本数据处理技术  （1）文本语料处理  （2）文本数据表示 | （1）理解常用文本数据处理工具的使用和处理方法  （2）掌握常用文本数据的表示方式 | 4 | 讲授 | 不同语言数据的不同处理与表示方式，培养严谨认真的专业态度 | 目标2 |
| 3 | 3、规则自然语言处理  （1）基于规则的词法分析  （2）基于规则的句法分析 | （1）理解基于规则的词法分析方法  （2）理解基于规则的句法分析方法 | 4 | 讲授 | 掌握规则自然语言处理的技术应用，认知专业知识的社会意义和价值 | 目标3 |
| 4 | 4、统计自然语言处理  （1）贝叶斯与文本分类  （2）k-means与文本聚类  （3）隐马尔科夫模型与序列标注  （4）条件随机场与序列标注 | （1）了解贝叶斯、隐马尔科夫模型、K-menas算法和条件随机场的基本思想和实现方式  （2）理解贝叶斯方法、隐马尔可夫模型、条件随机场和K-means聚类等机器学习方法应用于自然语言处理任务的基本概念和实现流程 | 6 | 讲授 | 掌握统计自然语言处理的技术应用，认知专业知识的社会意义和价值 | 目标2  目标5 |
| 5 | 5、深度学习自然语言处理   1. 语言模型与词向量表示   （2）卷积神经网络  （3）Seq2Seq架构  （4）Attention注意力机制  （5）Transformer模型  （6）预训练模型与微调  （7）模型压缩与剪裁 | （1）理解常见的语言模型以及词向量表示方法  （2）理解经典卷积神经网络的基本思想和核心原理  （3）理解Seq2Seq架构的基本概念和实现原理  （4）掌握Attention注意力机制的实现方式和Transformer模型的核心思想  （5）掌握预训练模型的下载、调用与微调方法  （6）掌握预训练模型的网络剪枝、参数量化和知识蒸馏的基本概念和实现方式 | 14 | 讲授 | 掌握深度自然语言处理的技术应用，认知专业知识的社会意义和价值 | 目标3  目标4 |
| 6 | 6、基础任务  （1）关键词提取  （2）词法分析  （3）句法分析 | （1）理解关键词提取、词法分析和句法分析的核心技术和应用场景  （2）掌握关键词提取、词法分析和句法分析的技术实践 | 4 | 讲授 | 识别基础任务场景的方法选择性，认知专业知识的应用价值 | 目标4  目标5 |
| 7 | 7、应用任务  （1）文本分类  （2）情感分析  （3）机器翻译  （4）智能问答  （5）摘要生成  （6）信息抽取 | （1）理解文本分类、情感分析、机器翻译、智能问答和摘要生成等自然语言处理应用场景的基本概念和核心实现技术  （2）掌握情感分析、机器翻译、智能问答和摘要生成等自然语言处理应用场景的模型选择思想和技术实践 | 6 | 讲授 | 识别具体应用场景的方法选择性，认知专业知识的应用价值 | 目标4  目标5 |
| 8 | 8、自然语言处理前沿方向  （1）前沿技术  （2）科研竞赛  （3）未来发展 | （1）了解大模型时代自然语言处理的核心任务以及基于大模型的自然语言处理新应用  （2）理解NLP常用的开源项目、开源数据集和最新研究论文  （3）掌握典型的NLP竞赛项目以及常见的NLP竞赛网站和网址 | 2 | 讲授 | 了解新技术对专业学科的发展作用，认知专业发展的意义 | 目标1 |

表1：教学内容与目标的关系

**（2）实践教学**

其中，实践环节主要包含以下项目：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **序号** | **实践名称** | **实践内容** | **学时** | **教学方式** | **课程思政**  **内容** |
| 1 | 基于NLP工具的中文分词 | （1）使用jieba工具实现中文分词 | 2 | 线上实验 | 要求学生处理实验数据必须坚持实事求实、严谨的科学态度。 |
| 2 | 基于TF-IDF的问答系统 | （1）使用TF-IDF算法实现文本向量化  （2）使用余弦相似度算法实现文本相似度计算  （3）基于本地知识库实现FAQ问答 | 2 | 线上实验 | 培养学生的批判性思维能力，鼓励学生在算法应用中保持理性、科学的态度 |
| 3 | 基于朴素贝叶斯的垃圾邮件分类 | （1）使用TF-IDF进行文本特征处理  （2）使用sklearn构建朴素贝叶斯分类模型  （3）基于分类模型评价指标进行模型评估 | 2 | 线上实验 | 增强学生的社会责任感，鼓励学生利用技术手段维护网络空间的清洁和安全 |
| 4 | 隐马科夫模型与中文分词 | （1）使用jieba进行中文分词  （2）隐马尔可夫模型的调用及训练  （3）维特比算法的实现  （4）使用模型完成中文分词 | 2 | 线上实验 | 培养学生将复杂问题进行拆解和细化的问题解决能力和严谨细致的科学态度 |
| 5 | 基于CRF模型的命名实体识别 | （1）定义模型评价指标  （2）进行文本特征提取  （3）调用sklearn-crfsuite库构建命名实体识别模型  （4）模型测试与效果评估 | 2 | 线上实验 | 培养学生将问题进行分门别类、系统分析的能力，提升其科技创新素养和实践能力 |
| 6 | 基于Kmeans的短文本聚类 | （1）使用TF-IDF进行文本特征处理  （2）使用Sklearn构建Kmeans模型  （3）Kmeans模型的训练和预测  （4）Kmeans聚类结果的词云图绘制 | 2 | 线上实验 | 培养学生的数据分析与处理能力，提升科学素养；强调算法在信息分类中的实际应用 |
| 7 | Gensim入门学习 | （1）使用Jieba工具进行文本语料预处理  （2）使用Gensiim进行文本向量化  （3）通过Gensim工具包调用TF-IDF模型进行文档相似度计算 | 2 | 线上实验 | 培养学生对中文语言的重视与理解，增强对数据处理的责任感和对技术应用伦理的认识 |
| 8 | 基于Word2Vec的词向量生成 | （1）使用Gensim训练Word2Vec词向量模型  （2）计算多种形式的词语相似度 | 2 | 线上实验 | 培养学生对语言模型的理解与数据处理能力，提升对数据伦理和技术应用的责任感 |
| 9 | 基于PyTorch构建循环神经网络RNN实现文本生成 | （1）基于PyTorch框架构建RNN模型  （2）RNN模型的训练、测试与评估  （3）利用RNN模型进行文本生成 | 2 | 线上实验 | 强调科技创新对社会发展的影响，培养创新精神和解决实际问题的能力 |
| 10 | 基于Seq2Seq+Attention的翻译系统 | （1）使用Pandas进行数据预处理  （2）基于Tensorflow框架构建Seq2Seq+CNN模型  （3）机器翻译模型的训练、保存、测试与评估 | 2 | 线上实验 | 培养严谨科学态度和创新精神，提升科技文化素养和跨文化交流能力 |
| 11 | 基于Transformer的翻译系统 | （1）使用Pandas进行数据预处理  （2）基于Tensorflow框架构建Transformer模型  （3）机器翻译模型的训练、保存、测试与评估 | 2 | 线上实验 | 培养严谨科学态度和创新精神，提升科技文化素养和跨文化交流能力 |
| 12 | 预训练模型的调用与调试 | （1）下载Bert等预训练模型  （2）查看模型的网络结构  （3）利用预训练模型进行文本特征提取  （4）文本向量化可视化展示 | 2 | 线上实验 | 培养学生学会知识迁移与举一反三的应用能力 |
| 13 | 使用TextPruner对Bert模型进行模型裁剪 | （1）预训练模型的词典剪裁  （2）预训练模型的Transformer剪裁  （3）预训练模型的Pipeline剪裁 | 2 | 线上实验 | 培养学生提升资源利用效率和模型性能，推动绿色计算理念，减少资源浪费的意识 |
| 14 | 使用TextBrewer对Bert模型进行知识蒸馏 | （1）定义教师模型与学生模型  （2）构造知识蒸馏所需数据集  （3）使用TextBrewer进行知识蒸馏 | 2 | 线上实验 | 培养学生资源节约和环境保护意识，理解知识传递与共享的重要性。 |
| 15 | 基于情感词典的文本情感分析 | （1）读取情感词典语料  （2）使用情感词典进行情感分类  （3）评估情感分类方法的性能 | 2 | 线上实验 | 促进学生理解情感分析在社会舆情管理中的作用，培养对社会问题的敏感度和解决能力 |
| 16 | 基于开放平台的情感分析 | （1）讯飞开放平台情感分析能力调用  （2）Jinjia2前端模版设计  （3）使用Flask框架调用情感分析模块  （4）情感分析结果前端演示 | 2 | 线上实验 | 强调科技服务社会的应用价值，鼓励学生关注技术对用户体验的提升 |
| 17 | 基于开放平台的机器翻译系统 | （1）讯飞开发平台机器翻译能力调用  （2）Jinjia2前端模版设计  （3）使用Flask框架调用机器翻译模块  （4）机器翻译结果前端演示 | 2 | 线上实验 | 强调科技在跨文化交流中的重要作用，鼓励学生关注翻译技术对全球化的促进 |
| 18 | 新冠病毒的FAQ问答系统 | （1）FAQ数据集预处理  （2）使用TF-IDF进行文本向量化  （3）使用余弦相似度计算问题相似度 | 2 | 线上实验 | 引导学生关注科技在公共卫生领域的应用，提升应对危机的能力，培养社会责任感和服务意识 |
| 19 | 基于Bert预训练模型的阅读理解 | （1）领域数据集预处理  （2）使用Bert构建阅读理解模型  （3）阅读理解模型的训练、测试与评估 | 2 | 线上实验 | 培养学生对人工智能在教育和知识普及中的应用的关注，增强科技服务社会的意识 |
| 20 | 基于Bart的开放域问答系统 | （1）领域数据集预处理  （2）使用Bart构建开放域问答模型  （3）开放域问答系统的训练、预测与评估 | 2 | 线上实验 | 掌握开放域问答的技术应用，培养对开放域问答系统公正性和透明性的思考 |
| 21 | 基于Bert的关系抽取任务实战 | （1）数据集预处理  （2）基于Bert设计关系抽取框架并构建相应模型  （3）关系抽取模型的调用与微调  （4）关系抽取结果测试 | 2 | 线上实验 | 培养学生从纷繁复杂的问题中理清关系、实事求是的能力和科学严谨的态度 |
| 22 | 基于Bert的事件抽取任务实战 | （1）数据集预处理  （2）基于Bert设计事件抽取框架并构建相应模型  （3）事件抽取模型的调用与微调  （4）事件抽取结果测试与效果评测 | 2 | 线上实验 | 培养学生透过现象看本质、抽丝剥茧、把握重点的问题解决能力。 |
| 23 | 基于LSTM的中文古诗词生成 | （1）定义数据分词器与数据生成器  （2）定义随机写诗、续写古诗和藏头诗函数  （3）使用Tensorflow框架构建LSTM网络  （4）中国古诗词模型的训练与预测 | 2 | 线上实验 | 强调中文古诗词的文化传承与语言美，增强学生对传统文化的认同感，展示技术与文化结合的可能性，鼓励学生在创新中保护和弘扬文化遗产 |
| 24 | 基于Bart的摘要生成 | （1）使用Sentence-BERT实现相似度计算  （2）利用Python实现TextRank算法  （3）使用Pytorch构建Bart摘要生成模型  （4）调用模型进行摘要生成 | 2 | 线上实验 | 强调精准提取信息的重要性，培养对信息真实性和可靠性的重视 |

表2：实践内容与学时分配

## 六、课程目标对毕业要求的支撑

课程目标对毕业要求的支撑关系如下：

| **毕业要求** | **毕业要求分解指标点** | **课程目标** |
| --- | --- | --- |
| 毕业要求6：专业知识 | 6-1：掌握扎实的人工智能基础知识、核心技术、主流算法，熟练掌握人工智能主流编程语言、框架工具以及核心模块库的使用。 | 目标2  目标3 |
| 毕业要求8：专业能力 | 8-2：掌握使用人工智能基础理论和专业知识解决实践问题的能力，了解前沿发展现状和趋势。 | 目标4  目标5 |
| 毕业要求9：创新能力 | 9-2：具备人工智能理论、技术、应用及交叉学科融合的科学研究能力以及对新知识、新技术的敏锐性。 | 目标1 |

表3：课程目标对毕业要求的支撑

## 七、课程教学方法

本课程侧重学生个人学习能力的培养，老师主讲理论基础知识，实验环节演变为技术的指导者，帮助学生解决实践操作中的问题，学生自主上机操作完成实操案例。学习模式采取“教师主讲理论知识，人工智能认知教育平台支撑，助教辅助，梯形团队管理”，主讲老师大班理论、重难点讲解，助教小班作业辅导。

1. 讲师课堂讲授

(1) 采用启发式教学，激发学生主动学习的兴趣，培养学生对专业的认识，引导学生明确学习目标。

(2) 采用案例式教学。理论教学与工程实践相结合，增强学生对专业的学习兴趣。

(3) 采用互动式教学。课内讨论和课外答疑相结合。

2. 助教辅导答疑

(1) 跟班辅导。做好考勤工作，随时关注学生学习状态，维持课堂记录，了解和记录学生知识掌握情况。

(2) 独自辅导。了解和记录学生任务和作业完成情况，主动巡查并进行答疑。

(3) 协助批改作业、试卷。

## 八、课程考核要求

1. 课程目标达成度

课程目标达成度评价包括课程分目标达成度评价，以及课程总目标达成度评价，具体计算方法如下：





根据本课程的教学内容、知识掌握的要求、以及前面几年的课程考核等综合因素考虑，本课程总目标达成度设置为0.6。

2. 课程考核细则

检验课程目标达成度，评价学生学习成果达成度。本课程成绩以百分计，满分100分。考核环节包含平时成绩、期末考核（平台）和期末考核（学校）等。各考核环节所占分值比例可根据具体情况微调，建议值及考核细则如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **考核**  **环节** | | **成绩** | **考核/评价范围与标准** | **考核方式** | **映射课程目标** |
| 平时成绩 | 出勤 | 10 | 根据学生出勤情况打分，三次缺勤，该出勤考核为0分；五次缺勤，平时成绩为0分。 | 过程评分 | 目标1  目标2 |
| 平台 | 10 | 根据平台课程学习情况（如进度、时长、正确率）进行考核，如果进度低于80%，则此项考核成绩为0分。 | 过程评分 | 目标4  目标5 |
| 作业 | 20 | （1）教学过程中包含课堂讨论及小组学习等环节，课堂表现考核学生的课堂参与程度；  （2）作业项考核学生的作业是否及时、独立地完成。 | 过程评分 | 目标1  目标2  目标3  目标4 |
| 期末考核  （平台） | | 20 | 考核范围为课程大纲考核范围内的所有内容。  注：允许平台考核不高于5%的超纲范围 | 平台考核 | 目标3  目标4 |
| 期末考核  （学校） | | 40 | （1）采用笔试，按照学校教学计划进行，考核范围为课程大纲包含的全部内容知识点和能力要求。  （2）期末考试具有一票否决权，如果期末笔试的卷面成绩在50分以下，则该课程直接不及格，课程成绩为期末笔试卷面成绩；如果期末笔试卷面成绩等于或高于50分，按照分值比例归于课程成绩。 | 笔试 | 目标1  目标2  目标3  目标5 |

表4：课程考核环节及评分细则

## 九、教材与参考书

教材：

[1] 《深度学习进阶：自然语言处理》.[日] 斋藤康毅 等著.陆宇杰 译.人民邮电出版社.2020

参考资料：

1. 人工智能实验平台中的《自然语言处理》课程；
2. 周涛,李艳凤,刘留,等.“深度学习”课程思政建设的实践探索[J]. 工业和信息化教育, 2024, (03):60-64.