自然语言处理课程教学大纲

英文名称：Natural Language Processing 课程编码：AA35450

学 时：32（理论授课24+实验8） 学 分： 2

课程性质：专业必修课 课程类别：理论课

先修课程：高等数学，概率论与数理统计，数据结构与算法分析，机器学习

开课学期：第五学期

适用专业：人工智能专业

一、课程教学目标

通过本课程的理论教学、实验和项目训练，使学生具备下列能力：

1、了解自然语言处理的概念和发展历程，了解其发展现状和发展趋势，掌握自然语言处理必备的知识和技术基础。

2、掌握自然语言处理常见任务的相关理论方法和编程技术，包括中文分词算法，关键词提取算法，词向量技术，文本分类算法，文本信息抽取算法，机器阅读理解算法，文本生成和摘要算法。了解对话系统和当前的相关方法。

3、针对复杂工程问题，能够综合运用所学的自然语言处理算法和实践技能进行问题分析、模型选择与优化，并设计具体方案解决实际问题，对方案进行测试评估。

4、在课程讲解过程中，引入思政环节，提升学生的思政水平，培养学生正确的马列主义观、爱国精神、科学探索精神与工匠精神等。

5、在课程实施过程中，充分锻炼学生的创新精神、创业意识和创新创业能力。

二、课程基本内容

**2.1 理论教学（支撑教学目标1、2、4）**

**1、****自然语言处理概述（支撑教学目标1、4）**

**[教学目的与要求]：**

使学生了解自然语言处理的概念和发展历程，了解其发展现状和发展趋势。

**[本章主要内容]：**

（1） 自然语言处理的定义

（2） 自然语言处理的应用领域

（3） 自然语言处理的发展历程

（4） 自然语言处理的研究现状和发展趋势

（5） 自然语言处理的知识和技术储备

**[本章重点]：**

（1）对自然语言处理的定义和研究内容的理解

（2）自然语言发展历程中的技术发展综述

**[本章难点]：**

（1）自然语言处理的技术更迭和趋势

**[本章思政教学内容]：**

**（1）中文自然语言处理难题与成果**

介绍中文自然语言处理的特点，遇到的难题，以及在中文自然语言处理研究中涌现出来的杰出人物，启发学生对母语的热爱和自豪感，激发学生向前辈学习，积极投身到中文自然语言处理的学习和研究中。

**2、Python语言基础和机器学习常见算法（该内容为选讲，如果学生已经学习过相关内容，可以略过）（支撑教学目标1）**

**[教学目的与要求]：**

掌握自然语言处理的编程和算法基础。

**[本章主要内容]：**

（1）Python编程基础

（2）机器学习算法基础

**[本章重点]：**

（1）机器学习算法

**[本章难点]：**

（1）结合Python语言的机器学习算法编程技术

**3、中文分词算法（支撑教学目标2，4）**

**[教学目的与要求]：**

掌握常见的中文分词算法。

**[本章主要内容]：**

（1）基于词表的分词算法

（2）基于统计模型的分词算法

（3）基于序列标注的分词算法

（4）中文分词工具

**[本章重点]：**

1. 基于统计模型的分词算法
2. 基于序列标注的分词算法

**[本章难点]：**

（1）基于序列标注的分词算法

**[本章思政教学内容]：介绍中文分词工具并讲解国内研究机构在中文分词研究上的贡献**

通过介绍并扩展哈工大LTP、中科院计算所NLPIR、清华大学THULAC和jieba等优秀的中文分词工具，进一步增强学生的民族语言和文化的自信，激发学生技术报国的热情。

**4、关键词提取（支撑教学目标2）**

**[教学目的与要求]：**

 掌握常见的关键词提取算法。

**[本章主要内容]：**

1. TextRank 关键词提取算法
2. TF-IDF 关键词提取算法

**[本章重点]：**

（1）TextRank 关键词提取算法

（2）TextRank 关键词提取算法

**[本章难点]**

（1）不同关键词提取算法的原理解析

**5、词向量技术（支撑教学目标2）**

**[教学目的与要求]：**

了解词向量技术的发展历程，掌握Word2vec、BERT等词向量模型。

**[本章主要内容]：**

（1）词向量技术发展历程

（2）Word2vec

（3）注意力机制

（4）深度学习概述

（5）BERT预训练模型

**[本章重点]：**

（1）Word2vec

（2）BERT预训练模型

**[本章难点]**

（1）注意力机制

（2）BERT预训练模型

**6、文本分类（支撑教学目标2）**

**[教学目的与要求]：**

了解文本分类任务，掌握常见的文本分类算法。

**[本章主要内容]：**

（1）基于统计方法的文本分类

（2）基于深度学习的文本分类

（3）开放领域文本分类

**[本章重点]：**

（1）基于统计方法的文本分类

（2）基于深度学习的文本分类

**[本章难点]**

（1）算法的理论理解

（2）算法的编程实践

**7、文本信息抽取（支撑教学目标2）**

**[教学目的与要求]：**

了解文本信息抽取任务，掌握命名实体识别、关系抽取和事件抽取的常见模型。

**[本章主要内容]：**

（1）命名实体识别

（2）实体关系抽取

（3）事件抽取

**[本章重点]：**

（1）基于BiLSTM+CRF的命名实体识别方法

（2）基于PCNN的关系抽取方法

（3）基于动态多池化卷积神经网络的事件检测模型

**[本章难点]**

（1）算法原理理解

（2）算法编程实践

**8、机器阅读理解（支撑教学目标2）**

**[教学目的与要求]：**

了解机器阅读理解任务，掌握机器阅读理解的常见模型。

**[本章主要内容]：**

（1）机器阅读理解发展历程

（2）抽取式阅读理解

（3）选择式阅读理解

**[本章重点]：**

（1）抽取式BiDAF模型

（2）选择题式CoMatch模型

**[本章难点]**

（1）算法原理理解

（2）算法编程实践

**9、文本生成与文本摘要（支撑教学目标2）**

**[教学目的与要求]：**

了解文本生成与摘要任务，掌握文本摘要的常见模型。

**[本章主要内容]：**

（1）文本生成与文本摘要概述

（2）抽取式文本摘要

（3）生成式文本摘要

（4）文本摘要案例

**[本章重点]：**

（1）抽取式文本摘要相关模型

（2）生成式文本摘要相关模型

**[本章难点]**

（1）算法原理理解

（2）算法编程实践

**10、对话系统（支撑教学目标2）**

**[教学目的与要求]：**

了解对话系统任务和常见方法。

**[本章主要内容]：**

（1）任务型对话系统原理与方法

（2）闲聊式对话系统原理与方法

（3）基于Pytorch框架的对话系统实战

**[本章重点]：**

（1）任务型对话系统原理与方法

（2）闲聊式对话系统原理与方法

**[本章难点]**

（1）算法原理理解

（2）算法编程实践

**2.2 实验内容（支撑教学目标2、3、5）**

**1、教学目标**

通过综合型实验，使学生进一步理解并掌握自然语言处理算法；能够基于具体实例要求，选择和设计相关自然语言处理算法并编程实现算法，测试评估算法性能。

**2、主要内容**

综合型实验由学生根据实验要求独立完成。

实验共8个实践主题，每个主题2学时，共计16学时。

（1）中文分词实验

（2）关键词提取实验

（3）文本分类实验

（4）命名实体识别实验

（5）关系抽取实验

（6）机器阅读理解实验

（7）文本摘要实验

（8）对话系统实验

三、教学安排

本课程由课堂教学和实验两部分组成。

课堂教学：占24学时，围绕本课程的基本概念、基本使用方法进行授课。

实验：占16学时，分别针对本课程每章的重点内容设置。

本课程的学时安排如下表：

|  |  |
| --- | --- |
| **讲 课 内 容** | **学 时** |
| **讲课** | **实验** | **上机** |
| 1  | 自然语言处理概述 | 2 |  |  |
| 2  | Python语言基础和机器学习算法概述 | 2 |  |  |
| 3 | 中文分词 | 2 | 1 |  |
| 4  | 关键词提取 | 2 | 1 |  |
| 5  |  词向量技术 | 2 |  |  |
| 6 |  文本分类 | 2 | 1 |  |
| 7 | 文本信息抽取 | 4 | 2 |  |
| 8 | 机器阅读理解 | 4 | 1 |  |
| 9 | 文本生成和文本摘要 | 2 | 1 |  |
| 10 | 对话系统 | 2 | 1 |  |
| 合 计 | 32 |

四、教学方法

本课程由理论教学和实验共两部分组成。

1、理论教学：围绕本门课程的基本概念、基本原理和基本技术，采用案例式教学、情景式教学、互动式教学等教学方法，并辅以网络教学资源等进行授课。

2、实验：用情境式教学法设计各种实验资源和项目，学生根据教师给定的实验指导书、实验讲解视频以及实操例题等各种学习资源，加强对于各种机器学习算法的理解与掌握，然后通过实操练习项目使学生能够把所学的知识和技能举一反三，融会贯通。同时，通过实验报告的撰写，培养学生文档撰写能力、问题分析与表述能力。

五、教学目标达成度评价

（1）教学目标1的达成度通过闭卷笔试综合考评；

（2）教学目标2的达成度通过闭卷笔试和实验进行综合考评；

（3）教学目标3的达成度实验进行综合考评；

（4）教学目标4的达成度通过闭卷考试进行综合考评；

（5）教学目标5的达成度通过实验进行综合考评。

六、课程考核方式与成绩评定

**考核方式：**闭卷笔试、实验。

**成绩评定：**闭卷笔试成绩50%，平时及期中测试20%，实验成绩20%，工程项目实践成绩10%。

七、课程教材及主要参考书

**[主要参考教材]：**

1.冯建周等.自然语言处理[M].北京:中国水利水电出版社,2022.

2.Zhiyuan Liu etc. Representation Learning for Natural Language Processing，springer.2021

3.宗成庆等.文本数据挖掘[M].北京:清华大学出版社，2023

制定人： 栗青生

审定人：

批准人：

 2024年6月