**《自然语言处理》课程实验教学大纲**

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

1. **实验教学任务与目的**

本课程实验以自然语言处理程序综合设计为目的，以 Pycham 的 自然语言处理库的应用为目标完成相关实验项目的设计与实验。

**二、实验项目内容、基本要求与学时分配**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **序号** | **实验项目名称** | **实验学时** | **实验类别** | **实验类型** | **实验要求** | **实验内容提要** | **思政元素** |
| 1 | 基于NLP工具的中文分词,基于NLP工具的词性标注 | 2 | 专业 | 设计 | 必做 | 1、了解中文分词的概念及工具；  2、了解jieba分词工具中的分词方法的调用。  3、了解词性标注的概念及工具；  4、了解jieba工具的词性标注方法的调用。 | 良好的开端的重要性 |
| 2 | 汉字笔画的语音处理 | 2 | 专业 | 设计 | 必做 | 1．通过读取文件，取得数据  2．利用生成算法生成相关汉字  3．对汉字进行笔画判定 | 理论联系实际 |
| 3 | 基于朴素贝叶斯的文本分类方法 | 2 | 专业 | 设计 | 必做 | 本实验旨在通过实现一个基于朴素贝叶斯分类器的垃圾邮件检测系统，理解并掌握文本分类的基本流程和方法。具体目标包括：  文本预处理：学习如何对文本数据进行预处理，包括去除无效字符、分词、过滤掉长度较短的词汇等。  特征提取：从预处理后的文本中提取有效的特征，即构建词向量，表示每封邮件中特定词汇的出现情况。 | 理论联系实际 |
| 4 | 基于pytorch框架的对话系统实战 | 2 | 专业 | 设计 | 必做 | 本实验的主要目的是通过实现一个简单的聊天机器人来熟悉和理解序列到序列（Seq2Seq）模型及其关键组件，包括编码器、解码器以及注意力机制。通过使用PyTorch框架和TorchScript，学生将能够了解如何构建、训练和部署一个基于神经网络的聊天机器人。此外，本实验还将帮助学生掌握自然语言处理（NLP）中的一些基本技术，如词汇表构建、文本规范化以及模型评估或文件存储的过程，掌握客户端发送数据的类封装技术 | 紧跟AI时代发展，服务社会各领域 |

注：1、实验类别：基础、专业基础、专业、其它。

2、实验类型：演示、验证、综合、设计研究、其它。

3、实验要求：必做、选做。

**三、实验项目的目的及要求**

**实验一 基于NLP工具的中文分词和词性标注（2课时）**

（一）实验目的和要求

1、了解中文分词的概念及工具；

2、了解jieba分词工具中的分词方法的调用。

3、了解词性标注的概念及工具；

4、了解jieba工具的词性标注方法的调用。

**（二）实验设备及材料**

1．硬件：多媒体计算机

2．软件：中文分词工具——jieba分词工具

**（三）思政元素**

使用国产华为设备，增强民族自信。

**（四）实验内容**

1、了解jieba分词工具中的分词方法的调用。

2、了解jieba工具的词性标注方法的调用。

**实验二 汉字笔画的语音处理（2课时）**

**（一）实验目的和要求**

1． 掌握文件服务器的使用

2． 掌握汉字生成算法

3． 汉字书写与文本生成

4． 文本的语音处理

**（二）实验设备及材料**

1．硬件：多媒体计算机

2．软件：中文分词工具——jieba分词工具

**（三）思政元素**

理论联系实际、实事求是。

**（四）实验内容**

1．通过读取文件，取得数据

2．利用生成算法生成相关汉字

3．对汉字进行笔画判定

**实验三 基于朴素贝叶斯的文本分类方法（2 课时）**

**（一）实验目的和要求**

本实验旨在通过实现一个基于朴素贝叶斯分类器的垃圾邮件检测系统，理解并掌握文本分类的基本流程和方法。具体目标包括：

文本预处理：学习如何对文本数据进行预处理，包括去除无效字符、分词、过滤掉长度较短的词汇等。

特征提取：从预处理后的文本中提取有效的特征，即构建词向量，表示每封邮件中特定词汇的出现情况。

模型训练：使用朴素贝叶斯分类器对提取的特征进行训练，学习邮件是否为垃圾邮件的分类规则。

模型评估：通过测试集验证模型的分类效果，了解模型的准确性和泛化能力。

**（二）实验设备及材料**

1．硬件：多媒体计算机

2、Pycharm 配置好的虚拟环境和相关文件

**（三）思政元素**

对比国内外AI模型，开好新场景，实现新功能。

**（四）实验内容**

1．客户端数据采集

2．数据集创建

3．模型训练

**实验四 基于pytorch框架的对话系统实战（2课时）**

**（一）实验目的和要求**

本实验的主要目的是通过实现一个简单的聊天机器人来熟悉和理解序列到序列（Seq2Seq）模型及其关键组件，包括编码器、解码器以及注意力机制。通过使用PyTorch框架和TorchScript，学生将能够了解如何构建、训练和部署一个基于神经网络的聊天机器人。此外，本实验还将帮助学生掌握自然语言处理（NLP）中的一些基本技术，如词汇表构建、文本规范化以及模型评估

**（二）实验设备及材料**

1. 学生需要确保计算机上安装了Python和PyTorch库。
2. 了解并配置好必要的依赖项，如Ptorch, numpy等。。

**（三）思政元素**

AI时代技术更新的非常快，课程内容要紧跟时代的发展，争取学有所用，用就能用好。

**（四）实验内容**

本实验基于序列到序列（Seq2Seq）模型，通过构建一个基于PyTorch框架的聊天机器人来深入理解该模型的工作原理及其在自然语言处理（NLP）中的应用。Seq2Seq模型是一种典型的编码器-解码器架构，其中编码器负责将输入序列转换成一个固定长度的内部表示（上下文向量），而解码器则利用这个内部表示来逐步生成输出序列。为了提升模型性能，本实验还引入了注意力机制，使得解码器在生成每个输出词时能够关注输入序列的不同部分。

**四、实验考核方式与标准**

1、每次实验课后认真完成实验报告并上传到到自己的学习主页。

2、评分办法：实验纪律20%，实验完成情况30%，实验报告50%。

3、实验成绩占本课程总分的15%计入期末总成绩。

**五、实验教材（实验指导书）及教学参考书**

实验指导书为自编教材

执笔人：栗青生

编制日期: 2024年9月