# 《Python中文自然语言处理基础与实战》课程大纲

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **课程编码** |  | | **课程类别** | 选修 |
| **课程名称** | Python中文自然语言处理基础与实战 | | | |
| **英文名称** | Fundamentals and practice of Chinese natural language processing in Python | | | |
| **学分** | **4** | | **建议修读学期** | 7 |
| **总学时数** | 64 | **其中：实践学时** | **实验学时** | **24** |
| **其他学时** |  |
| **预修课程** | 高等数学、线性代数、概率统计、Python程序设计、数据挖掘 | | | |
| **考核方式** | 考察 | | | |
| **适用专业** | 数据科学与大数据技术 | | | |
| **大纲编写组** |  | | **大纲审核人** |  |

## 一、课程目标

1. 知识与技能（支撑毕业要求：工程知识、问题分析）。通过学习，掌握自然语言处理的基本流程、基本知识和基本性质。通过案例实践掌握对完成自然语言处理任务所必须的各项技能。

2. 过程与方法（支撑毕业要求：设计/开发解决方案、使用现代工具）。通过学习，认识自然语言处理的建模和算法过程，并学会运用现代信息技术获取相关信息以及解决自然语言处理问题的方法等。

3. 情感·态度·价值观（支撑毕业要求：个人和团队、终身学习、职业道德）。通过学习，养成良好的自主学习习惯，具有吃苦耐劳的态度；在学习交流中能表现出团队协作能力。

## 二、课程目标与毕业要求的对应关系

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **毕业要求** | **指标点** | | **权重** | **课程目标** |
| 工程知识1 | 指标1-1 | 熟悉数学和数据科学基础知识，具有解决数据科学问题的基本能力。 | 0.1 | 知识与技能 |
| 指标1-2 | 具有解决大数据、人工智能问题的能力。 | 0.1 | 知识与技能 |
| 问题分析2 | 指标2-1 | 具有基本的研究能力，能够基于问题的性质进行科学的分析，并采用科学的方法对工程问题进行研究。 | 0.1 | 知识与技能 |
| 指标2-2 | 能够通过实验来验证工程问题，能够对实验数据进行分析并得出适当的结论。 | 0.1 | 知识与技能 |
| 设计/开发解决方案3 | 指标3-1 | 能够设计针对大数据工程问题的解决方案，设计满足特定需求的系统。 | 0.1 | 过程与方法 |
| 使用现代工具4 | 指标4-1 | 具有使用现代大数据工具和信息技术工具的技能。 | 0.1 | 过程与方法 |
| 指标4-2 | 具有对大数据工程问题进行预测与模拟的能力。 | 0.1 | 过程与方法 |
| 个人和团队5 | 指标5-1 | 具有较强的工程实践能力。 | 0.1 | 情感•态度•价值观 |
| 指标5-2 | 具有团队协作能力，能够就工程问题进行有效沟通和交流。 | 0.1 | 情感•态度•价值观 |
| 终身学习7 | 指标7-1 | 具有自主学习的意识和较强的终身学习的意识。 | 0.05 | 情感•态度•价值观 |
| 职业道德8 | 指标8-2 | 能够在工程实践中理解并遵守工程职业道德和规范，履行责任。 | 0.05 | 情感•态度•价值观 |

## 三、教学内容、重难点和课时安排

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 章 | 重难点 | 内容 | 课程目标 | 学时 |
| 第1章 绪论 |  | 1.1 NLP的基本概念 | 知识 | 2 |
|  | 1.2 NLP的发展历程 | 知识 |
|  | 1.3 NLP的研究内容 | 知识 |
|  | 1.4 NLP的几个应用场景 | 知识 |
|  | 1.5 NLP的研究现状 | 知识 |
|  | 1.6 学习NLP的困难 | 知识 |
|  | 1.7 NLP基本流程 | 知识 |
| ※ | 1.8 NLP的开发环境配置（Python） | 技能 |
| 第2章 语料库 |  | 2.1 语料库概述 | 知识 | 6 |
|  | 2.2 语料库种类与原则 | 知识 |
|  | 2.3 NLTK库 | 知识 |
| ※ | 2.4 获取语料库 | 技能 |
| ※ | 2.5 实战之构建语料库与语料操作 | 过程与方法 |
| 第3章 正则表达式 |  | 3.1 正则表达式 | 知识 | 6 |
| ※ | 3.2 正则表达式应用 | 过程与方法 |
| 第4章 中文分词技术 |  | 4.1 中文分词简介 | 知识 | 6 |
|  | 4.2 基于规则的分词 | 知识 |
| ※ | 4.3 基于统计的分词 | 技能 |
| ※ | 4.4 中文分词工具jieba库 | 技能 |
| ※ | 4.5 中文分词的应用 | 过程与方法 |
| 第5章 词性标注与命名实体识别 |  | 5.1 词性标注 | 知识 | 6 |
| ※ | 5.2 命名实体识别 | 技能 |
| ※ | 5.3 中文命名实体识别 | 过程与方法 |
| 第6章 关键词提取 |  | 6.1 关键词提取技术简介 | 知识 | 6 |
|  | 6.2 关键词提取算法 | 技能 |
| ※ | 6.3 实战之提取文本关键词 | 过程与方法 |
| 第7章 文本向量化 |  | 7.1 文本向量化概述 | 知识 | 6 |
|  | 7.2 文本离散表示 | 知识 |
|  | 7.3 分布式表示-Word2Vec | 技能 |
|  | 7.4 分布式表示-Doc2Vec | 技能 |
| ※ | 7.5 实战之论文相似度计算 | 过程与方法 |
| 第8章 文本分类与聚类 |  | 8.1文本挖掘简介 | 知识 | 8 |
|  | 8.2 分类算法 | 技能 |
|  | 8.3 聚类算法 | 技能 |
|  | 8.4 文本分类与聚类的步骤 | 知识 |
| ※ | 8.5 分类实战：垃圾短信分类 | 过程与方法 |
|  | 8.6 聚类实战：新闻文本聚类 | 过程与方法 |
| 第9章 文本情感分析 |  | 9.1 情感分析简介 | 知识 | 6 |
|  | 9.2 情感分析的常用方法 | 技能 |
| ※ | 9.3 基于词典的情感分析 | 过程与方法 |
| ※ | 9.4 基于文本分类的情感分析 | 过程与方法 |
| ※ | 9.5 基于LDA模型的情感分析 | 过程与方法 |
| 第10章 NLP中的深度学习技术 |  | 10.1循环神经网络概述 | 技能 | 6 |
|  | 10.2 RNN结构简介 | 技能 |
|  | 10.3 深度学习工具简介 | 知识 |
| ※ | 10.4基于LSTM的文本分类与情感分析 | 过程与方法 |
| ※ | 10.5基于Seq2Seq的机器翻译 | 过程与方法 |
| 第11章 智能问答系统 |  | 11.1 问答系统简介 | 知识 | 4 |
|  | 11.2 问答系统的主要组成 | 知识 |
| ※ | 11.3 基于Seq2Seq的智能问答系统 | 过程与方法 |
| 第12章 基于TipDM数据挖掘建模平台实现垃圾短信识别 |  | 12.1 实现垃圾短信识别 | 过程与方法 | 2 |

## 四、课程教学方法

教学方法：讲授法、案例教学法，实验法、探究法等，突出理论分析中学生的参与性，强调师生的共同探讨与互动。

## 五、实践教学安排

|  |  |  |
| --- | --- | --- |
| 章节 | 讲授 | 实践教学 |
| 第1章 绪论 | 2 | 0 |
| 第2章 语料库 | 4 | 2 |
| 第3章 正则表达式 | 4 | 2 |
| 第4章 中文分词技术 | 4 | 2 |
| 第5章 词性标注与命名实体识别 | 4 | 2 |
| 第6章 关键词提取算法 | 4 | 2 |
| 第7章 文本向量化 | 4 | 2 |
| 第8章 文本分类与聚类 | 4 | 4 |
| 第9章 文本情感分析 | 4 | 2 |
| 第10章 NLP中的深度学习技术 | 4 | 2 |
| 第11章 智能问答系统 | 2 | 2 |
| 第12章 基于TipDM数据挖掘建模平台实现垃圾短信识别 | 0 | 2 |
| 合计 | 40 | 24 |

## 六、课程教学评价

|  |  |  |
| --- | --- | --- |
| **课程教学目标** | **考核内容** | **评价依据** |
| **知识与技能** | 自然语言处理基础、语料库、词性标注与命名实体识别、 | 作业、实验 |
| **过程与方法** | 中文文本的预处理、中文分词技术、关键词提取算法、文本向量化、机器学习算法及其应用、文本情感分析、基于深度学习的NLP | 作业、实验、小论文 |
| **情感•态度•价值观** | 应用案例 | 作业、考勤 |

## 七、成绩评定方法

本课程平时成绩及期末成绩占比分别为30%，70%，平时考核形式为考勤、作业、实验，期末考核的形式为小论文。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **笔试相关试题占分比例%** | **小设计报告**  **占分比例%** | **平时作业相关内容占分比例%** | **课程分目标达成评价方法** |
| **知识与技能** | 10 | 60 | 30 | **分目标达成度**={0.15ｘ(分目标试题平均/分目标试题总分)+0.7ｘ(小设计分目标平均成绩/分目标总分) +0.15ｘ(分目标平时成绩)}/分目标总分) |
| **过程与方法** | 10 | 60 | 30 |
| **情感•态度•价值观** | 10 | 60 | 30 |

## 八、课程学习资源

1.选用教材

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 教材名称 | 编者 | 出版社 | 出版时间 | 是否马工程教材 | 是否国家、省规划教材 | 是否国外教材 |
| Python中文自然语言处理基础与实战 | 肖刚，张良均 | 人民邮电出版社 | 2021年12月 | 否 | 否 | 否 |

2.主要参考书目

[1] 宗成庆, 统计自然语言处理，清华大学出版社，北京， 2013年

[2] 李航，统计学习方法, 清华大学出版社，北京，2012年

[3] 张良均．Python与数据挖掘 [M]．北京：机械工业出版社．2016．

[4] 张良均，谭立云. Python数据分析与挖掘实战（第2版）[M]．北京：机械工业出版社.2019.

## 九、课程学习建议

理论是基础，python语言编程是技能，完成案例项目是目标。

## 十、评分标准

【请按照本门课程采用的课程考核方式选择下表之一填写】

表1：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **课程目标** | **评分标准** | | | | |
| **90-100** | **80-89** | **70-79** | **60-69** | **0-59** |
| **优** | **良** | **中** | **及格** | **不及格** |
| **知识与技能** | **完全掌握自然语言处理的基本流程、基本知识和基本性质。完全掌握自然语言处理任务所必须的各项技能。** | **较好掌握自然语言处理的基本流程、基本知识和基本性质。较好掌握自然语言处理任务所必须的各项技能。** | **基本掌握自然语言处理的基本流程、基本知识和基本性质。基本掌握自然语言处理任务所必须的各项技能。** | **一般自然语言处理的基本流程、基本知识和基本性质。一般掌握自然语言处理任务所必须的各项技能。** | **不能掌握自然语言处理的基本流程、基本知识和基本性质。不能掌握自然语言处理任务所必须的各项技能。** |
| **过程与方法** | **完全认识自然语言处理的建模和算法过程，并学会运用现代信息技术获取相关信息以及解决自然语言处理问题的方法** | **较好认识自然语言处理的建模和算法过程，并学会运用现代信息技术获取相关信息以及解决自然语言处理问题的方法** | **基本认识自然语言处理的建模和算法过程，并学会运用现代信息技术获取相关信息以及解决自然语言处理问题的方法** | **一般认识自然语言处理的建模和算法过程，并学会运用现代信息技术获取相关信息以及解决自然语言处理问题的方法** | **不能认识自然语言处理的建模和算法过程，并学会运用现代信息技术获取相关信息以及解决自然语言处理问题的方法** |
| **情感•态度•价值观** | **通过学习，完全养成良好的自主学习习惯，具有吃苦耐劳的态度；在学习交流中能表现出团队协作能力，并形成数据驱动的科学价值观。** | **通过学习，养成良好的自主学习习惯，具有吃苦耐劳的态度；在学习交流中能表现出团队协作能力，并形成数据驱动的科学价值观。** | **通过学习，基本养成良好的自主学习习惯，具有吃苦耐劳的态度；在学习交流中能表现出团队协作能力，并形成数据驱动的科学价值观。** | **通过学习，养成了自主学习习惯，具有吃苦耐劳的态度；在学习交流中能表现出团队协作能力，并形成数据驱动的科学价值观。** | **通过学习，尚未养成自主学习习惯，具有吃苦耐劳的态度；在学习交流中能表现出团队协作能力，并形成数据驱动的科学价值观。** |