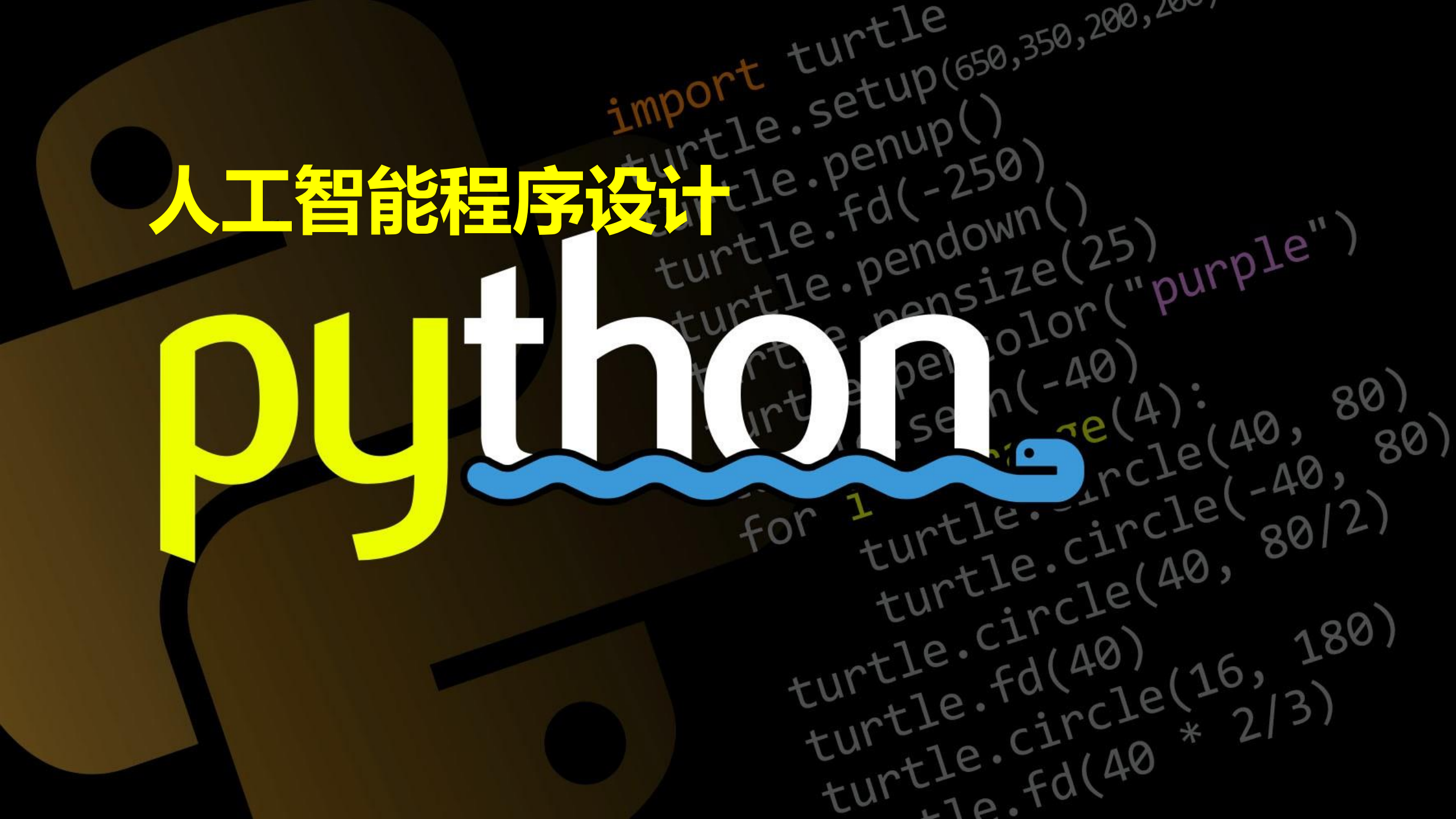
人工智能程序设计


python
import turtle
turtle.setup(650,350,200,200)
turtle.penup()
turtle.fd(-250)
turtle.pendown()
turtle.pensize(25)
turtle.pencolor("purple")
turtle.seth(-40)
for 1 in range(4):
turtle.circle(40, 80)
turtle.circle(-40, 80)
turtle.fd(40)
turtle.circle(40, 80/2)
turtle.fd(40)
turtle.circle(16, 180)
turtle.fd(40 * 2/3)

人工智能程序设计

14.1 自然语言处理基础

北京石油化工学院 人工智能研究院

刘 强

什么是自然语言处理

自然语言处理（NLP）是人工智能的重要分支，让计算机能够理解、分析和生成人类语言。

核心目标：

- 让机器理解人类语言的含义
- 从文本中提取有价值的信息
- 生成自然流畅的文本内容

NLP发展历程

NLP经历了四个重要发展阶段：

阶段	时间	特点
早期阶段	1950s-1980s	基于规则和符号推理
统计方法时代	1980s-2000s	概率模型和统计学习
机器学习时代	2000s-2010s	SVM、CRF等算法
深度学习时代	2010年至今	RNN、Transformer架构

早期阶段（1950s-1980s）

技术特点：

- 基于规则和符号推理方法
- 人工编写语法规则和词典
- 只能处理简单的语法分析

局限性：

- 规则难以覆盖语言的复杂性
- 扩展性差，维护成本高
- 无法处理歧义和上下文

统计方法时代（1980s-2000s）

技术特点：

- 引入概率模型和统计学习
- N-gram语言模型
- 隐马尔可夫模型（HMM）

突破：

- 词性标注准确率提升
- 语言建模取得进展
- 开始处理大规模文本数据

机器学习时代（2000s-2010s）

技术特点：

- 支持向量机（SVM）
- 条件随机场（CRF）
- 特征工程驱动

成果：

- 命名实体识别性能提升
- 情感分析应用落地
- 文本分类广泛应用

深度学习时代（2010年至今）

技术特点：

- 循环神经网络（RNN/LSTM）
- 注意力机制
- Transformer架构

革命性提升：

- BERT、GPT等预训练模型
- 端到端学习，无需人工特征
- 多任务学习和迁移学习

NLP核心技术领域

现代NLP包含五个核心技术领域：

1. **文本预处理** - 分词、词性标注、命名实体识别
2. **语义理解** - 词向量、语言模型
3. **文本分类** - 情感分析、垃圾邮件过滤
4. **信息抽取** - 实体关系抽取、事件检测
5. **文本生成** - 机器翻译、对话系统

文本预处理

基础任务：

- **分词**：将文本切分为词语单元
- **词性标注**：标注每个词的词性（名词、动词等）
- **命名实体识别**：识别人名、地名、机构名等

重要性：

- 为后续分析奠定基础
- 影响下游任务的性能
- 中文和英文处理方式不同

语义理解

核心技术：

- **词向量**：Word2Vec、GloVe将词语映射为向量
- **语言模型**：理解词语在上下文中的含义
- **句子表示**：BERT等模型捕获句子级语义

应用：

- 相似度计算
- 语义搜索
- 问答系统

文本分类

任务定义：

根据文本内容将其归类到预定义的类别中

典型应用：

- 垃圾邮件过滤（垃圾/正常）
- 情感分析（正面/负面/中性）
- 新闻分类（体育/财经/科技等）
- 意图识别（查询/购买/投诉等）

信息抽取

任务定义：

从非结构化文本中提取结构化信息

核心任务：

- **实体抽取**：识别文本中的实体
- **关系抽取**：识别实体之间的关系
- **事件检测**：识别文本描述的事件

应用场景：

知识图谱构建、舆情监控、金融分析

文本生成

任务定义：

基于给定条件生成自然流畅的文本

典型应用：

- **机器翻译**：将文本从一种语言翻译到另一种
- **对话系统**：生成对话回复
- **文本摘要**：生成文本的简短摘要
- **内容创作**：自动写作、文案生成

NLP典型应用领域

NLP技术在多个领域得到广泛应用：

- 搜索与推荐
- 智能客服系统
- 内容审核与监控
- 机器翻译服务
- 金融科技应用
- 教育技术应用

搜索与推荐

应用案例：

- Google搜索：语义理解提升检索准确率
- 今日头条推荐：基于内容理解的个性化推荐

核心技术：

- 查询理解和扩展
- 文档语义匹配
- 用户意图识别

智能客服系统

应用案例：

- 阿里小蜜：电商客服自动问答
- 微软小冰：情感对话机器人

核心功能：

- 意图识别
- 知识库问答
- 多轮对话管理
- 情感分析

内容审核与监控

应用场景：

- 违规内容识别（涉黄、涉政等）
- 垃圾信息过滤
- 舆情监控和分析

技术要点：

- 文本分类模型
- 敏感词检测
- 情感倾向分析

机器翻译服务

应用案例：

- Google翻译
- 百度翻译
- 有道翻译

技术演进：

- 规则翻译 → 统计翻译 → 神经机器翻译
- Transformer架构带来质的飞跃

金融科技应用

应用场景：

- **智能投顾**：分析财经新闻和研报
- **风险评估**：分析企业公告和舆情
- **舆情监控**：实时监控市场情绪

技术要点：

- 金融领域命名实体识别
- 事件抽取和因果分析
- 情感分析和趋势预测

教育技术应用

应用场景：

- **智能批改**：作文自动评分和批改
- **个性化学习**：学习内容推荐
- **知识图谱**：构建学科知识体系

技术要点：

- 文本质量评估
- 知识点抽取
- 语法错误检测

实践练习

练习 14.1.1：NLP发展理解

分析NLP四个发展阶段的特点和局限性，说明深度学习为什么能够带来突破性进展。

实践练习

练习 14.1.2：核心技术对比分析

比较文本预处理、语义理解、文本分类、信息抽取、文本生成五个核心技术的特点和应用场景。

实践练习

练习 14.1.3：应用场景调研

选择一个NLP应用领域，调研其技术实现方案和面临的挑战。

实践练习

练习 14.1.4：Python环境准备

安装NLP相关的Python库，为后续实践做准备：

- NLTK：经典NLP工具库
- spaCy：现代化NLP库
- jieba：中文分词库
- transformers：预训练模型库