**人工智能训练师（三级）操作技能考核**

**试题单**

准考证号：

试题代码：

试题名称：爬虫培训大纲编写

考核时间：10min

**1.**场地设备要求

（1）人工智能训练师主机；   
（2）Office软件。

**2.**工作任务

A企业是一家互联网金融机构，需要利用人工智能技术获取大量上市公司的财务报告。为了提高获取效率，现计划对新进技术人员进行网页爬虫工具的使用培训。通过这次培训，将会使新进技术人员掌握常用网页爬虫工具的使用，能对网络公开的上市公司财务报告进行获取，达到人工智能训练师四级/中级工的技能水平。  
请你根据要求补全素材4.1.2.docx中的培训大纲。

**3.**技能要求

（1）能编写初级培训讲义；   
（2）能对初级工、中级工开展知识和技术培训；

**4.**质量指标

（1）培训讲义编写逻辑合理；   
（2）培训资料正确有效；   
（3）语句组织通顺。

**人工智能训练师（三级）操作技能考核**

**试题评分表**

准考证号：

试题代码：

试题名称：爬虫培训大纲编写

考核时间：10min

测量分评分表

| 细则编号 | 配分 | 评分细则描述 | 规定或  标称值 | 结果或  实际值 | 得分 |
| --- | --- | --- | --- | --- | --- |
| M1 | 5 | 回答内容与答案类似得1分；最多5分； | 根据数据 |  |  |
| 合计配分 | 5 | 合计得分 | |  |  |

**参考答案**（尽量将细则内容写在上面的表格内，写不下可另写，但要具体可评判）

**在下划线上填写答案。**

# 根据学习大纲补充学习目标

**引言**

学习目标：了解网页爬虫的重要性和应用场景。

内容：

介绍人工智能在金融数据采集中的应用，网页爬虫在获取上市公司财务报告中的作用。

**网页爬虫基础理论**

学习目标：**理解网页爬虫的基本原理与核心流程，掌握机器人协议（robots.txt）的规范遵循方法**。

内容：

什么是网页爬虫？

网页爬虫的工作流程和基本结构。

机器人协议（robots.txt）的作用和遵守方法。

**常用网页爬虫工具简介**

学习目标： **熟悉主流爬虫工具的核心功能与适用场景，能够根据需求选择合适工具**。

内容：

BeautifulSoup：用于解析HTML和XML文档。

Scrapy：一个快速、高效的网页爬虫框架。

Selenium：用于模拟浏览器操作，处理动态网页。

**环境搭建与工具安装**

学习目标：掌握网页爬虫工具的安装和环境配置。

内容：

安装Python和pip包管理工具。

安装并配置BeautifulSoup、Scrapy、Selenium。

**网页数据解析**

学习目标： **能够使用多种技术高效提取网页结构化数据，并完成实际解析任务**

内容：

使用BeautifulSoup进行HTML解析。

使用XPath和CSS选择器提取数据。

实践操作：解析一个简单的网页数据。

**动态网页处理**

学习目标：掌握处理动态网页的方法。

内容：

使用Selenium模拟浏览器操作。

处理动态加载的数据，如Ajax请求。

实践操作：爬取一个动态加载的网页数据。

**Scrapy框架实战**

学习目标： **能够独立构建Scrapy爬虫项目，实现复杂数据抓取与自动化存储**。

内容：

Scrapy项目的创建和基本配置。

编写爬虫脚本，设置抓取规则。

数据存储与导出：将数据保存为CSV、JSON等格式。

实践操作：构建一个Scrapy爬虫项目，获取上市公司财务报告。

**数据清洗与处理**

学习目标： **掌握数据清洗的核心方法，能够规范化处理爬取后的原始数据**。

内容：

数据去重、缺失值处理。

数据格式转换与规范化。

实践操作：对获取的财务报告数据进行清洗和处理。

**综合项目实践**

学习目标：综合运用所学知识进行网页爬虫项目。

内容：

项目介绍：构建一个完整的财务报告爬虫项目。

数据采集：获取多个网站的财务报告数据。

数据处理与分析：清洗并分析获取的数据。

项目评审：展示爬虫成果，讲解实现思路和遇到的问题。

**总结与答疑**

学习目标：巩固所学内容，解决学员疑问。

内容：培训内容总结，关键点回顾，学员提问与解答。

**评估测试**

学习目标：评估学员掌握的网页爬虫技能。

内容：

理论测试：关于网页爬虫基本概念和工具使用的笔试。

实操测试：要求学员完成指定的网页爬虫任务。