# 任务考核题

**代码题需提交py文件，放在一个文件夹中压缩发送  
命名格式：XXX-考核A**

**一、单选题（一个2分，共20分）**

1. **在进行复杂网页爬取时，面对动态加载的内容（如通过AJAX请求加载），以下哪个方法通常不是首选？（A）**

A. 使用requests库直接发送AJAX请求  
B. 分析JavaScript代码，找出数据接口并直接请求  
C. 使用selenium或Playwright等浏览器自动化工具  
D. 分析网络请求，使用mitmproxy等代理工具拦截并解析请求

### 2. 在处理网页中的验证码时，哪种方法通常不是自动化解决验证码的可行方案？（D）

A. 使用OCR技术识别验证码  
B. 分析验证码生成逻辑并编写算法绕过  
C. 训练深度学习模型来识别验证码  
D. 尝试不同的验证码输入

### 3. 在构建大规模的爬虫系统时，以下哪个问题通常不是首要考虑的？（D）

A. 高效地管理URL队列并避免重复爬取  
B. 处理网站的robots.txt文件以遵守爬虫协议  
C. 确保爬虫在高并发下的稳定性和性能  
D. 将爬取的数据直接存储到数据库中

### 4. 在面对具有反爬虫机制的网站时，以下哪种策略可能不是有效的应对方法？（A）

A. 通过分析网站日志来识别爬虫行为并进行优化  
B. 使用多个不同的用户代理来模拟不同的浏览器和设备  
C. 添加随机的延迟和间隔，以模拟人类用户的浏览行为  
D. 更换IP地址或使用代理服务来避免被网站封禁

### 5. 在处理网页的编码问题时，以下哪个描述可能是不准确的？（D）

A. 可以通过HTTP响应头中的Content-Type字段来确定网页的编码  
B. 网页中的<meta charset="...">标签也可以用来指定编码  
C. 如果网页没有明确指定编码，requests库会自动检测并解码内容  
D. 网页可以假设其编码为UTF-8无需进行额外处理

### 6.面对需要登录的网页时，以下哪种方法不是常用的登录策略？（B）

A. 使用表单数据进行登录  
B. 在URL中拼接用户名和密码进行GET请求登录  
C. 通过模拟浏览器行为进行登录  
D. 分析登录接口的API文档并直接调用

### 7. 在进行网页数据抓取时，如何避免被网站识别为爬虫并限制访问？（D）

A. 频繁更换IP地址  
B. 使用常见的用户代理字符串  
C. 减少请求频率并模拟人类用户行为  
D. 以上都是有效的方法

**8.怎么处理因网站结构变化而导致的爬取失败（B）**

A. 捕获异常并直接忽略错误页面  
B. 更新爬虫逻辑以匹配网站新结构  
C. 使用正则表达式匹配页面内容  
D. 放弃对该网站的爬取

**9.如何确保爬取的数据不违反目标网站的隐私政策和法律法规（C）**

A. 使用代理IP隐藏身份  
B. 使用自动化工具快速爬取数据  
C. 阅读遵守网站的robots.txt文件和隐私政策  
D. 对网站的robots.txt文件视而不见

**10.构建Python爬虫时，如何确保代码的可维护性和可扩展性（B）**

A. 使用全局变量和硬编码的URL  
B. 将爬虫逻辑与数据存储、日志记录等独立的模块  
C. 将代码写在一个脚本文件中以便于管理  
D. 使用复杂的嵌套循环和条件语句来处理情况

1. **多选（一个3分，共15分）**
2. **影响爬虫性能和效率的因素包括哪些（ABCD）**

A. 网络延迟  
B. 目标网站的反爬虫策略  
C. 爬虫代码的优化程度  
D. 使用的数据处理和存储方法

1. **处理分页抓取时，以下哪些方法和策略是有效的（BCD）**

A.分析页面结构，找到分页控件并手动点击  
B. 使用正则表达式或XPath提取下一页链接  
C. 循环抓取，直到没有新页面为止  
D. 根据URL模式生成分页链接列表

1. **以下哪些技术和工具可以用于模拟用户登录（ABD）**

A. 分析登录表单，提交正确的POST请求  
B. 利用Selenium模拟用户输入和点击登录按钮  
C. 在URL中拼接登录凭证  
D. 使用requests.Session对象保持会话状态

1. **在爬虫中，处理大规模数据推荐以下哪些技术和工具（ACD）**

A. 使用pandas进行数据处理和分析  
B. 将所有数据加载到单个Python进程中  
C. 利用redis缓存服务器，可以缓存页面内容  
D. 使用分布式存储和计算框架

1. **使用Srapy时应该注意（BD）**

A.在 items.py 中定义 Item 类后，数据存储必须使用 Item 对象，不能再用普通字典

B.解析后的数据可以直接在 Spider 中处理（如保存到文件），不一定非要通过 pipeline

C.Scrapy 的 Response 对象不支持 css() 和 xpath() 方法提取数据

D.pipeline 的主要设计用途是数据持久化（如存储到数据库或文件），但也可用于数据清洗或验证

**三、判断题（一个2分，共10分）**

**1.requests库是Python中最常用的HTTP请求库之一，支持自动处理Cookie和会话。（对）**

**2.使用自动化爬虫可以模拟浏览器行为，因此可以轻松绕过大多数反爬虫机制。（对）**

**3.在处理动态网页时，可以从HTML中提取所需数据，因为动态内容已经加载到页面中。（错）**

**4.Scrapy框架中的Spider类负责定义爬虫的初始URL、解析响应以及生成新的请求。（对）**

**5.Scrapy是一个功能强大的Python爬虫框架，它提供了请求调度、解析、存储等完整的功能。（对）**

**四，代码题（55分）**

1.（15分）使用异步爬虫获取豆瓣电影排行榜的电影名称和评分信息（1-10页），其网址为：<https://movie.douban.com/top250>

2.（20分）爬取猫眼电影今日票房（今日大盘-->查看更多里面）的数据【综合票房：影片，综合票房，票房占比，排片场次】四个数据

目标url：<https://www.maoyan.com/>

①.获取当页的类利用python爬取猫眼电影今日票房（今日大盘-->查看更多里面）的数据【综合票房：影片，综合票房，票房占比，排片场次】四个数据



②.实时爬取综合票房中的 影片、综合票房、票房占比、排片场次四个数据【获取当前调用的实时数据】

③.将获取到的数据写入Excel表格中或者数据库中保存

3.（20分）使用爬取某政策大数据平台（最少10页）

①.获取文章标题、内容、发布机构、技术领域、以及发布时间  
②.可自定义标题搜索，例如：人工智能  
③.数据存储到excel表格里面或者入库  
目标url：<https://www.spolicy.com/>