Assistant

`crawl_proxies.py` 使用说明书

概述
`crawl_proxies.py` 是一个用于从多个网站爬取代理服务器（Proxy）的 Python 脚本。它支持从指定的网站列表中提取代理 IP，并提供了多种功能，如代理 IP 的验证、去重、保存、以及通宵挂机模式等。该脚本适用于需要大量代理 IP 的场景，如网络爬虫、数据采集、匿名访问等。

适用人群
- **开发者**：需要代理 IP 进行网络爬虫、数据采集等开发工作。
- **网络安全研究人员**：需要测试代理 IP 的可用性和匿名性。
- **数据分析师**：需要代理 IP 来绕过网站的反爬虫机制，获取数据。
- **普通用户**：需要代理 IP 来访问被限制的网站或服务。

功能说明

1. 爬取代理 IP
- **功能描述**：从指定的网站列表中爬取代理 IP。
- **使用方法**：通过命令行参数指定代理网站列表文件（默认文件名为 `proxy_sites.txt`），脚本会自动从这些网站中提取代理 IP。
- **示例**：
  ```bash
  python crawl_proxies.py
  ```

2. 代理 IP 验证
- **功能描述**：验证爬取的代理 IP 是否可用。
- **使用方法**：通过命令行参数 `--validate` 启用代理 IP 验证功能，脚本会使用指定的验证 URL（默认是 `https://www.google.com`）来测试代理 IP 的可用性。
- **示例**：
  ```bash
  python crawl_proxies.py --validate
  ```

3. 代理 IP 去重
- **功能描述**：对爬取到的代理 IP 进行去重操作。
- **使用方法**：通过命令行参数 `--deduplicate` 启用去重功能，脚本会去除重复的代理 IP。
- **示例**：
  ```bash
  python crawl_proxies.py --deduplicate
  ```

4. 保存代理 IP
- **功能描述**：将爬取到的代理 IP 保存到文件中。
- **使用方法**：默认情况下，代理 IP 会保存到 `proxy_list.txt` 文件中。可以通过 `--timestamp` 参数在文件名中添加时间戳。
- **示例**：
  ```bash
  python crawl_proxies.py --timestamp
  ```

5. 通宵挂机模式
- **功能描述**：脚本可以持续运行，定期爬取代理 IP。
-**使用方法**：通过命令行参数 `--overnight` 启用通宵挂机模式，脚本会每隔指定的时间间隔（默认 2 秒）重新爬取代理 IP。
- **示例**：
  ```bash
  python crawl_proxies.py --overnight
  ```

6. 显示详细调试信息
- **功能描述**：显示详细的调试信息，包括每个代理 IP 的爬取和验证过程。
- **使用方法**：通过命令行参数 `--verbose` 启用详细调试信息。
- **示例**：
  ```bash
  python crawl_proxies.py --verbose
  ```

7. 显示无效代理
- **功能描述**：显示无效代理的警告信息。
- **使用方法**：通过命令行参数 `--show-invalid` 启用显示无效代理功能。
- **示例**：
  ```bash
  python crawl_proxies.py --show-invalid
  ```

8. 简单报告
- **功能描述**：显示简单的代理获取报告。
- **使用方法**：通过命令行参数 `--simple-report` 启用简单报告功能。
- **示例**：
  ```bash
  python crawl_proxies.py --simple-report
  ```

命令行参数说明

参数	描述
`--proxy`	指定代理服务器（例如：`http://127.0.0.1:8080` 或 `socks5://127.0.0.1:1080`）。
`--validate`	在保存前验证代理 IP 是否可用。
`--show`	实时显示爬取到的代理 IP。
`--verify-url`	指定用于验证代理的目标网站（默认：`https://www.google.com`）。
`--verify-ssl`	启用 SSL 验证（默认禁用）。
`--add-prefix`	保存代理 IP 时添加前缀（`http://`、`https://` 或 `socks5://`）。
`--timestamp`	保存代理 IP 时在文件名中添加时间戳(默认保存到 `proxy_list.txt`)。
`--deduplicate`	对文件中的代理 IP 进行去重。
`--deduplicate-file`	指定需要去重的文件（默认：`proxy_list.txt`）。
`--deduplicate-after-save`	保存代理 IP 后对文件内容进行去重。
`--show-invalid`	显示无效代理的警告信息。
`--simple-report`	显示简单的代理获取报告。
`--verbose`	显示详细的调试信息。
`--overnight`	启用通宵挂机无人值守模式。
`--interval`	自定义每次爬取的间隔时间（单位为秒，默认：2 秒）。

示例用法

示例 1：爬取代理 IP 并保存到文件
```bash
python crawl_proxies.py --timestamp
```
该命令会从 `proxy_sites.txt` 文件中列出的网站爬取代理 IP，并将结果保存到带有时间戳的文件中。

示例 2：爬取并验证代理 IP
```bash
python crawl_proxies.py --validate --verify-url "https://www.example.com"
```
该命令会爬取代理 IP，并使用 `https://www.example.com` 作为验证 URL 来测试代理 IP 的可用性。

示例 3：通宵挂机模式
```bash
python crawl_proxies.py --overnight --interval 5
```
该命令会启用通宵挂机模式，每隔 5 秒重新爬取代理 IP。

注意事项
1. **代理网站列表**：确保 `proxy_sites.txt` 文件中列出的网站是有效的，并且包含代理 IP 的表格。
2. **网络环境**：确保运行脚本的网络环境可以访问指定的代理网站和验证 URL。
3. **合法性**：在使用代理 IP 时，请遵守相关法律法规，避免用于非法用途。

结语
`crawl_proxies.py` 是一个功能强大的代理 IP 爬取工具，适用于各种需要代理 IP 的场景。通过灵活的命令行参数，用户可以根据自己的需求定制脚本的行为。希望本说明书能帮助您更好地理解和使用该脚本。