

پروژه عملی Web Scraping

🎯 هدف: در این پروژه شما یاد می‌گیرید که داده‌های یک وبسایت (مثل محصولات یک فروشگاه یا اخبار یک سایت خبری) را استخراج کنید و در قالب فایل CSV یا پایگاه داده ذخیره کنید.

🕒 بخش ۱ – آماده‌سازی

✔ محیط پایتون را آماده کنید (می‌توانید از PyCharm یا Jupyter Notebook استفاده کنید).

✔ کتابخانه‌های موردنیاز را نصب کنید:

```
pip install requests beautifulsoup4 pandas
```

✔ وبسایتی که قرار است از آن داده جمع کنید بررسی کنید. به ساختار صفحه نگاه کنید و تصمیم بگیرید چه اطلاعاتی می‌خواهید (مثلا: عنوان محصول، قیمت، لینک).

🕒 بخش ۲ – استخراج داده از یک صفحه

✔ با استفاده از requests محتوای HTML یک صفحه را دریافت کنید.

✔ با BeautifulSoup تگ‌هایی که داده در آن قرار دارد را پیدا کنید.

✔ داده‌ها (مثلا نام و قیمت) را استخراج کنید و در یک لیست یا دیکشنری ذخیره کنید.

✔ خروجی را در ترمینال چاپ کنید.

🕒 بخش ۳ – استخراج از چند صفحه (Pagination)

✔ بررسی کنید آیا سایت صفحه‌بندی دارد (مثلا صفحه ۱، ۲، ۳ ...).

✔ یک حلقه بنویسید که همه صفحات را واکشی کند.

✔ داده‌های همه صفحات را به لیست اصلی اضافه کنید.

✓ داده‌ها را در یک فایل CSV ذخیره کنید.

④ بخش ۴ - بهبود و پاکسازی داده‌ها

✓ داده‌های استخراج‌شده را تمیز کنید (مثلا حذف فاصله‌ها یا تبدیل قیمت به عدد).

✓ داده‌ها را با pandas در فایل Excel یا دیتابیس SQLite ذخیره کنید.

✓ اگر خطایی رخ داد (مثلا Timeout)، آن را مدیریت کنید.

⑤ بخش ۵ - جمع‌بندی و ارائه

✓ اسکرپر خود را یک‌بار کامل اجرا کنید.

✓ نتایج نهایی را در CSV یا دیتابیس بررسی کنید.

✓ یک تحلیل ساده انجام دهید (مثلا میانگین قیمت‌ها یا بیشترین تکرار یک کلمه در عنوان‌ها).

✓ خروجی پروژه را تحویل دهید.

✦ خروجی نهایی پروژه

یک اسکرپر که داده‌های چند صفحه را جمع‌آوری می‌کند.

فایل CSV یا دیتابیس شامل داده‌ها.

یک تحلیل ساده روی داده‌های جمع‌شده.