Introdução ao Scrapy e Extração de Dados da Web

O que é Web Scraping?

Definição: processo de coletar informações de sites de forma automatizada.

Exemplo: pegar os nomes e preços de livros de uma loja online, sem precisar copiar manualmente.

O que é o Scrapy?

Um framework em Python que facilita a coleta de dados de páginas da web.

Diferente de simplesmente usar requests + BeautifulSoup, o Scrapy é:

- Mais rápido (assíncrono).
- Estruturado (usa "spiders" para organizar o código).
- Já tem recursos prontos para navegar, coletar e exportar os dados.

O mínimo de HTML que precisamos

HTML é como um esqueleto do site.

Cada elemento tem uma "caixa" chamada tag.

- <h1> → título principal
- <p> → parágrafo
- <a> → link
- → caixa genérica

Cada "caixa" pode ter:

class → apelido que o site dá para organizar estilos.

id → identificador único.

Exemplo:

```
<div class="produto">
    <h2 class="nome">Camiseta Azul</h2>
    <span class="preco">R$ 59,90</span>
</div>
```

Nome do produto: está dentro da tag <h2 class="nome">.

Preço: está dentro da tag .

Primeiros Passos com o Scrapy

https://docs.scrapy.org/en/latest/intro/tutorial.html

O que é .css no Scrapy?

No Scrapy, .css é um método para selecionar partes do HTML usando seletores CSS.

Ele não significa que estamos mexendo no estilo (cores, fontes etc.), mas sim que estamos usando a mesma sintaxe do CSS para encontrar elementos.

O resultado é um SelectorList (uma lista de trechos do HTML) que você pode manipular para extrair dados.

```
response.css("SELETOR")
```

Principais complementos:

::text → pega apenas o texto dentro da tag.

::attr(atributo) → pega o valor de um atributo da tag (ex: link, título, imagem).

.get() → retorna o primeiro resultado encontrado.

.getall() → retorna todos os resultados encontrados em uma lista.

response.css("SELETOR")

Limitações e Ética

Nem todo site permite scraping (verificar o robots.txt).

Não usar scraping para sobrecarregar sites ou roubar dados pessoais.

O que é o robots.txt?

É um arquivo de texto que os sites colocam na raiz do domínio para dizer o que robôs (crawlers) podem ou não acessar.

Exemplo: quando você entra em https://example.com/robots.txt, esse arquivo mostra as regras de permissão.

Criando um Projeto Scrapy

scrapy startproject loja

cd loja

scrapy genspider produtos exemplo.com