Scrapy: Mining Data from a Pokemon Website

Entendendo Web Scraping

O que é Web Scraping?

Web Scraping refere-se à técnica de extração de dados da web de forma automatizada. No contexto da apresentação sobre o uso do Scrapy em um site de informações Pokémon, este slide destaca a importância de coletar dados específicos do site para análise.

Pokemon Website

Escolhendo o Website de Pokemon

Para esse projeto, foi escolhido um website de Pokemon que contém informações detalhadas sobre cada Pokemon.

https://pokemondb.net/pokedex/all

Setting Up Scrapy

Criar um Novo Projeto:

- No terminal, navegue até o diretório onde deseja criar seu projeto Scrapy Isso criará uma estrutura básica de diretórios e arquivos para o seu projeto;
- Criar um arquivo de Spider em que você pode definir como o Scrapy deve extrair os dados do site específico;
- Editar o arquivo do Spider (geralmente localizado em nome_do_projeto/spiders/nome_do_spider.py);
- Definir as regras de extração de dados usando seletores XPath ou CSS.

Building the Spider

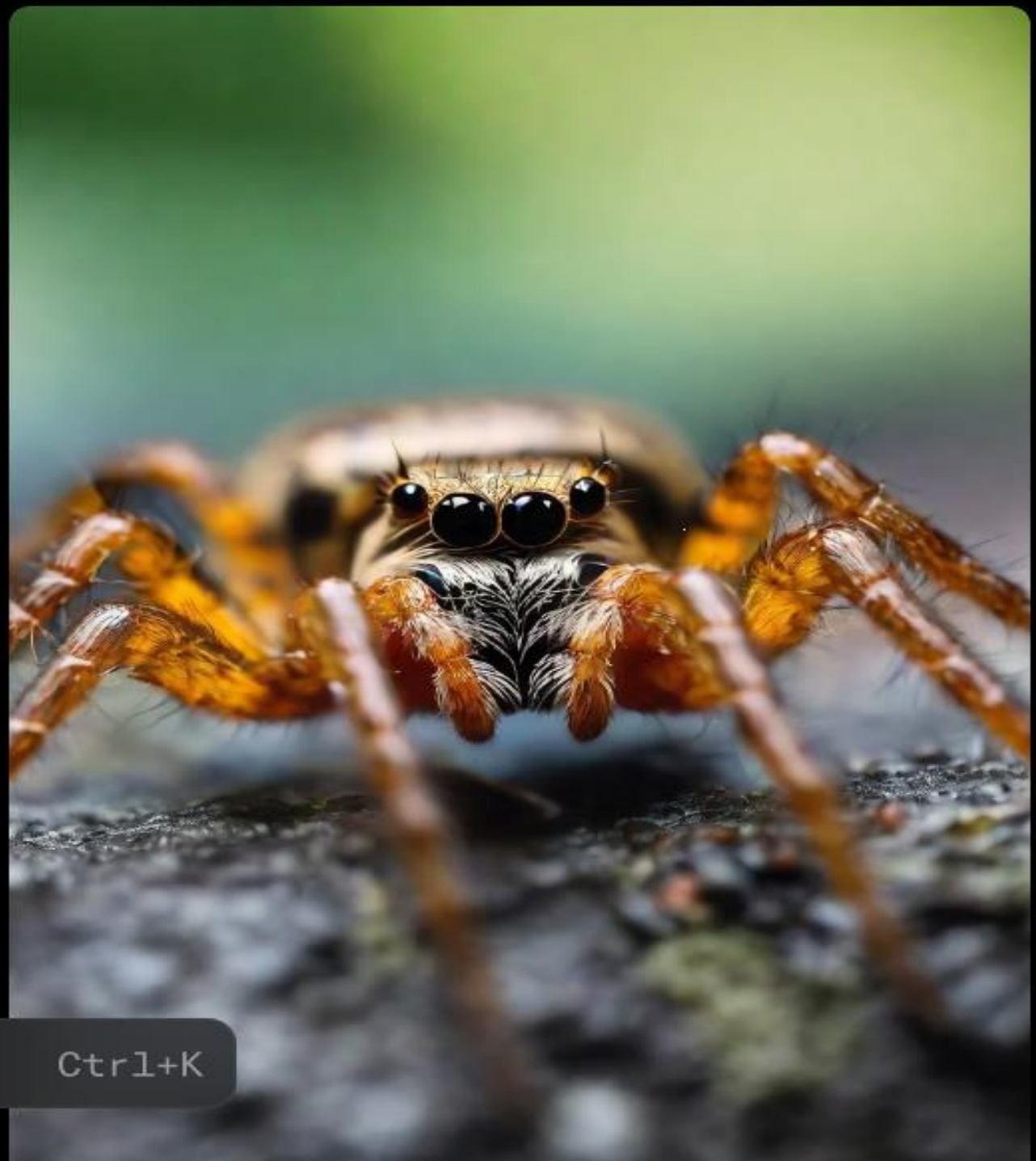
A Aranha (Spider) é o componente principal do Scrapy que define como extrair dados de um site específico. Uma Aranha personalizada precisa ser criada para o site escolhido sobre Pokémon.

Passos para Construir a Aranha:

- Identificar a estrutura do site-alvo e os dados a serem extraídos.
- Criar um novo arquivo em Python e importar as bibliotecas Scrapy necessárias.
- Definir a classe da Aranha e configurar seus atributos, incluindo o nome, URLs iniciais e domínios permitidos.
- Implementar o método 'parse' para lidar com a resposta e extrair os dados desejados usando seletores XPath ou CSS.

Executar a Aranha usando a ferramenta de lin
comando do Scrapy e especificar o formato d

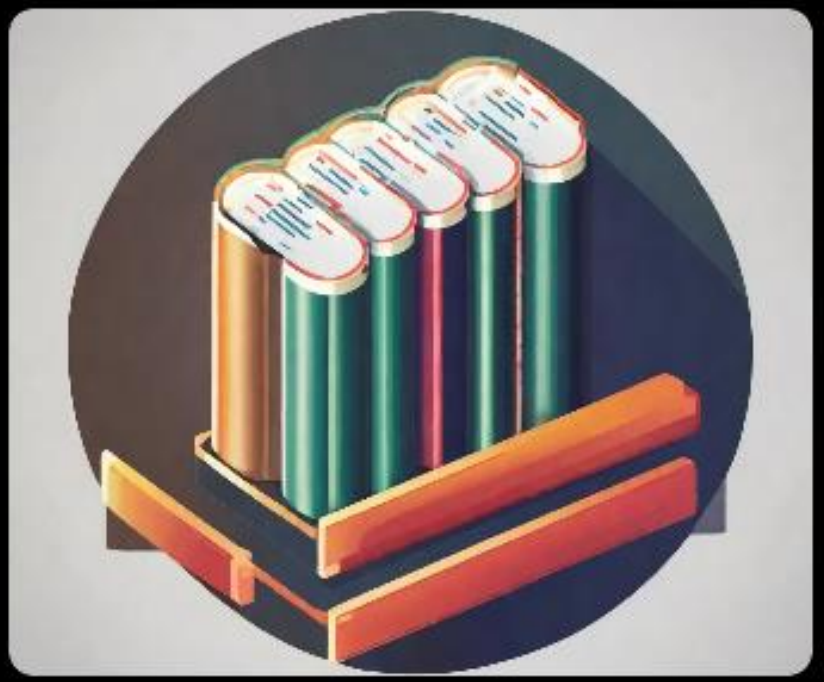
Ctrl+K



Storing the Data

Databases

Uma opção para armazenar os dados extraídos é utilizar bancos de dados. Eles oferecem uma maneira estruturada e eficiente de armazenar e gerenciar grandes quantidades de dados.



CSV Files

Outra opção é armazenar os dados em arquivos CSV. Esse formato é amplamente suportado e pode ser facilmente acessado e manipulado usando várias ferramentas.

Outras Opções

Também existem outras opções de armazenamento disponíveis, como serviços de armazenamento em nuvem, que proporcionam escalabilidade e acessibilidade para os dados armazenados.

Analyzing the Data

O que deveria ser buscado de cada Pokémon obrigatoriamente:

- Número
- URL da página
- Nome
- Próximas evoluções do Pokémon se houver (PokéNum, nome e URL)
- Tamanho
- Peso
- Tipos (água, veneno, ...)
- Habilidades (link para outra página)
  - URL da página
  - Nome
  - Descrição do efeito

Desafio e Considerações

Estrutura Dinâmica do Site:

- Desafio: Sites podem ter estruturas dinâmicas que mudam ao longo do tempo, o que pode exigir ajust
  frequentes no código do Scrapy.

Identificação de Elementos na Página:

- Identificar corretamente os elementos na página, como nome, tipo e estatísticas dos Pokémon, pode
  desafiador em um site complexo.