Relatório 8 - Prática: Web Scraping com Python p/ Ciência de Dados (II)

João Pedro Gomes

1. Introdução

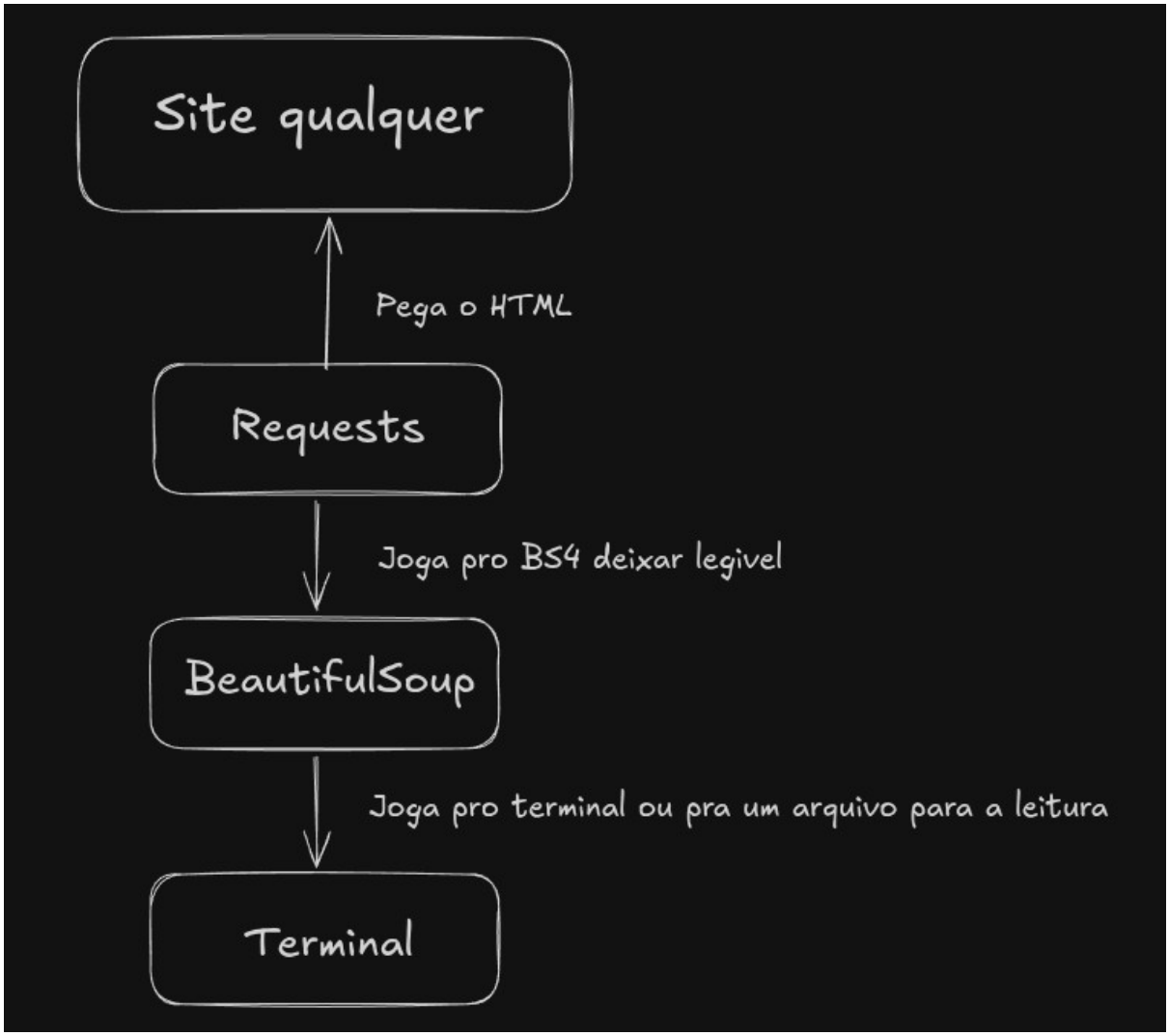
O card manda assistir o vídeo “Web Scapring with Python – Beautiful Soup Crash Course, freeCodeCamp.org”, e depois praticarmos o que aprendemos e fazer esse relatório.

2. Desenvolvimento

Esse vídeo ensina a fazer o web scraping que é puxar o HTML de qualquer site puxando qualquer tag html que ele tem, como divs, buttons e por classes também.

Pra isso ele começa introduzindo a biblioteca BeautifulSoup que organiza o texto puxado pra deixar ele legivel no console e não só um monte de texto, aqui ele só utiliza ela sozinha pois o arquivo que puxamos o código está baixado no computador e não está online, junto com o BS ele usa o lxml que é um parser do BS que é mais rápido pra organizar o HTML.

Para pegaro HTML de um site que está online é preciso usar a biblioteca requests que faz a requisição e pega o html do site e depois passamos o beutifulsoup no html que ele pegou pra deixar legivel.



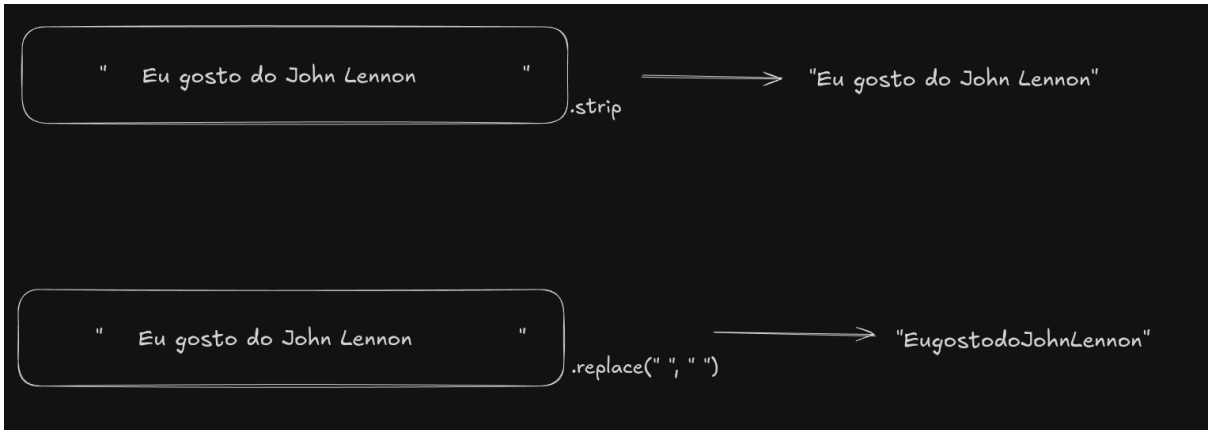
Na primeira parte ele ensina alguns comandos pra pegar as tags do arquivo baixado e eles são:

find(tag html): Pega a primeira tag da pagina que bate com a tag do parametro e retorna ela.

find_all(tag html): Mesma coisa que o de cima só que retorna todas as tags que batem com a tag do paraemetro

E no find_all() também da pra colocar a tag e depois de uma virgula colocar a classe que você quer, ai ele pega todas as tags com aquela classe, usando como parametro o nome class_ = “”.

Na segunda parte ele usa algumas fuções pra manipular o texto como o .text que pega apenas o texto da tag mas remove a tag em si e o .strip remove espaçoes desnecessarios em um texto mas no video ele usa o .replace com espaçoes vazios também que faz a mesma coisa que o .strip porem ele tira todos os espaços e junta o texto todo.



E também ensina como salvar arquivos com texto em cada chamada do código quando ele abre e executa o código, no meu caso ele não funciona pois o site mudou de tags e não deu pra puxar os dados e salvar em arquivos.

Prática:

Na minha prática eu peguei o site taylorswift.com que é o site da prórpria Taylor, e fiquei puxando links, textos dele, ele tem bem pouca coisa aproveitavel pra puxar porque a maioria é imagem mas ainda deu pra puxar algumas coisas legais e praticar o conteudo.

3. Conclusão

Eu gostei muito do conteúdo, foi uma coisa nova pra mim foi muito fácil de entender o video e a lógica por tras das bibliotecas, é muito util pra pegar dados de forma rápida, monitorar algo em algum site pra ver a mudança, etc…, o vídeo foi muito bem explicado com um inglês fácil de entender e é algo que eu espero ver mais pois gostei muito de fazer esse card.