Relatório 8 - Web Scraping com Python p/ Ciência de Dados

Edryck Freitas Nascimento

1. Introdução

O objetivo desta aula foi aprender a realizar a coleta de dados da web, ou raspagem de dados (*web scraping*) utilizando a linguagem Python junto das bibliotecas *BeautifulSoup* e *Requests*. A tarefa consistia em assistir a vídeo aula "*Web Scraping with Python - Beautiful Soup Crash Course*" e fazer um programa simples aplicando os conceitos aprendidos.

2. Desenvolvimento

Inicialmente, foi realizada a busca de um *site* para realizar a coleta de dados. Encontrando o *site* eu iniciei a construção de um programa simples para armazenar os episódios de cada anime que foi lançado, em seguida, são armazenados em um arquivo CSV.

O programa segue o seguinte fluxo:

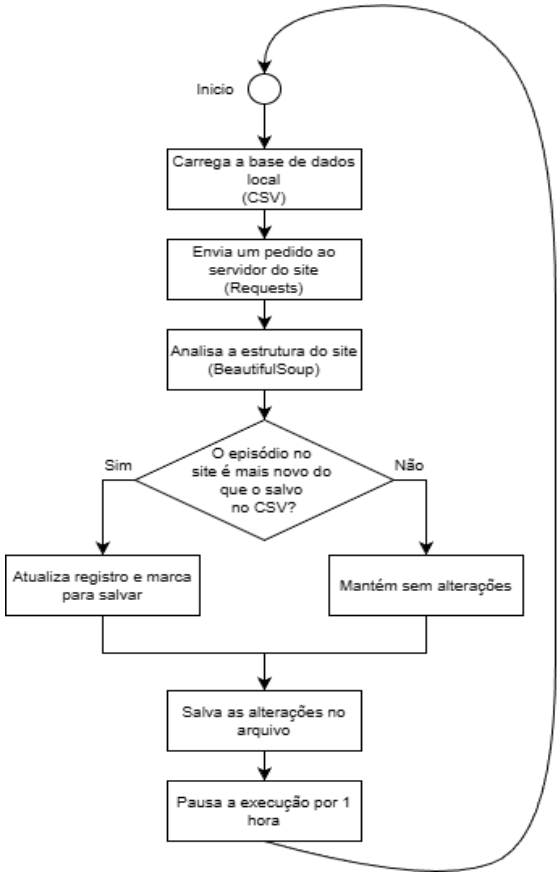

Figura 1. Fluxograma do programa (autoria própria).

O *loop* do programa funciona da seguinte maneira:

1. Ele carrega a base de dados local, que é o arquivo CSV.
2. Envia um pedido ao servidor do site, usando a biblioteca *Requests*. O retorno desse pedido é salvo em uma variável no formato de texto.
3. Em seguida, analisa a estrutura do site para encontrar as *tags* que tenho interesse, essa parte é utilizando a biblioteca *BeautifulSoup*.
4. Para cada *tag* encontrada, é comparado se o episódio é mais novo que o que estava salvo ou se o anime estava salvo, isso porque eu só realizo essa extração na primeira página.
    - Se sim, atualiza o registro do arquivo, no caso de o anime estar na lista, atualiza só o episódio, senão adiciona o anime, episódio e o link no arquivo.
    - Senão, não realiza nenhuma mudança no arquivo.
5. Salva o arquivo com os novos dados, no caso que houve alguma mudança.
6. E por último, pausa a execução por uma hora antes de voltar ao passo um, para isso utilizei a biblioteca *time*.

Porém, o maior desafio da coleta é a volatilidade do *site*. Se o *site* mudar o nome de alguma classe CSS, o programa quebra. Por isso foi adicionado a lógica de verificar se houve mudanças antes de salvar, ela é importante para evitar corromper o arquivo local com dados vazios.

Além da função principal (procurar novos episódios), ele conta com uma função para carregar o arquivo CSV e ler o conteúdo dele, e outra para escrever no final arquivo CSV os novos episódios ou anime encontrado.

3. Conclusão

Web scraping é uma técnica que transforma uma grande quantidade de informações não estruturadas da internet em conjuntos de dados organizados e pronto para análise. O projeto mostra que a eficiência de um coletor de dados depende de robustez da sua lógica de comparação e persistência, desta forma garante que o fluxo de informações seja atualizado continuamente, mas ainda tem que estar atento a mudanças do ambiente que está sendo coletado os dados, algumas mudanças no HTML do site é o bastante para fazer esse programa não funcionar, tendo que ficar atento a mudança e sempre atualizando.

Referências:

Vídeo aula **Web Scraping with Python - Beautiful Soup Crash Course**: https://www.youtube.com/watch?v=XVv6mJpFOb0.

Site **Animes Online**:
https://animesonlinecc.to/.