Curso Webscraping em R – Programa de Transformação Digital

Autor: Michel Alves

|  |  |
| --- | --- |
| **Conteúdo** | **Carga Horária** |
| Unidade 1   * 1. O que é webscrapping?   2. Quando é necessário realizar webscrapping?   3. Estrutura de uma página da WEB      1. HTML/XML      2. CSS   1.4 Ferramentas de desenvolvimento dos navegadores (que permitem visualizar a estrutura de uma página)  1.5 Principais bibliotecas no R para webscrapping | 2 Horas |
| Unidade 2  2.1 Extraindo dados: 2.1.1 Texto 2.1.2 Atributos 2.1.3 Tabelas 2.1.4 Importando arquivos | 2 Horas |
| Unidade 3  3.1 Tornando o webscrapping mais robusto: utilizando o xpath | 2 Horas |
| Unidade 4  4.1 Webscrapping usando o RSelenium | 2 Horas |

Pré-requisitos: Conhecimentos básicos em R

O curso será realizado em modalidade virtual com encontros semanais às sextas-feiras.

Data propostas: 11/03, 18/03, 25/03 e 01/04

# Unidade 1

## O que é Webscrapping?

O web scraping (raspagem de rede, em tradução livre), também conhecido como extração de dados da web, é o nome dado ao processo de coleta de dados estruturados da web de maneira automatizada. Os dados são então exportados para um formato mais útil ao usuário.

Em geral, esse método é usado por pessoas, empresas e, preocupantemente, criminosos, que desejam usar a vasta quantidade de dados da web disponíveis publicamente para tomar decisões mais inteligentes ou cometer crimes. Quando usado para o bem, pode ser aplicado em monitoramento de preços, monitoramento de notícias e pesquisa de marketing, entre outros.

O processo básico de raspagem, na verdade, é realizado diariamente por boa parte da população mundial, no ato de copiar e colar informações de um site para outro meio, a diferença é que a raspagem de rede faz isso em uma escala macroscópica e com automação inteligente, para extrair milhões de dados de páginas da internet.

No entanto, o processo de webscrapping não é uma tarefa simples. Os websites se apresentam em muitas formas e, como resultado, os webscrappers variam em funcionalidades e características. Também pode acontecer de encontrarmos captchas (aquele teste para verificar se você é humano) quando tentando realizar a raspagem de dados. Esses casos, infelizmente, não serão tratados nesse curso.

As principais etapas do processo de webscrapping são:

1. Identificar o website desejado;
2. Coletar as URLs das páginas das quais você deseja extrair os dados;
3. Fazer uma requisição para essas URLs para obter o HTML da página;
4. Usar os localizadores para encontrar os dados no HTML;
5. Salvar os dados em formato CSV ou XLSX, entre outros.

Alright, finally we are here. There are 2 different approaches for web scraping depending on how does website structure their contents.

Approach 1: If website stores all their information on the HTML front end, you can directly use code to download the HTML contents and extract out useful information.

There are roughly 5 steps as below:

Inspect the website HTML that you want to crawl

Access URL of the website using code and download all the HTML contents on the page

Format the downloaded content into a readable format

Extract out useful information and save it into a structured format

For information displayed on multiple pages of the website, you may need to repeat steps 2–4 to have the complete information.

Pros and Cons for this approach: It is simple and direct. However, if the website's front-end structure changes then you need to adjust your code accordingly.

Approach 2: If website stores data in API and the website queries the API each time when user visit the website, you can simulate the request and directly query data from the API

Steps:

Inspect the XHR network section of the URL that you want to crawl

Find out the request-response that gives you the data that you want

Depending on the type of request(post or get) and also the request header & payload, simulate the request in your code and retrieve the data from API. Usually, the data got from API is in a pretty neat format.

Extract out useful information that you need

For API with a limit on query size, you will need to use ‘for loop’ to repeatedly retrieve all the data

Pros and Cons for this approach: It is definitely a preferred approach if you can find the API request. The data you receive will be more structured and stable. This is because compared to the website front end, it is less likely for the company to change its backend API. However, it is a bit more complicated than the first approach especially if authentication or token is required.

## Quando é preciso realizar webscrapping?

O processo de webscrapping é necessário quando os dados desejados estão disponíveis dentro de um código HTML. Em outros casos, pode ser usado para obter arquivos que contém os dados desejados.

## Estrutura de uma página web

Páginas são criadas usando HTML (HyperText Markup Language)

<https://canaltech.com.br/seguranca/o-que-e-web-scraping/>

https://www.zyte.com/learn/what-is-web-scraping/