



## Python II: Web Scraping

Abordando web scraping em Python, como utilizar e as ferramentas que podem ser aplicadas para extrair dados.

---

Augusto Goulart

*14 de outubro de 2022*

Realização:



Pró-Reitoria  
de Graduação  
(PROGRAD)

# Apresentação

---

# O que é *web scraping*?



1. Processo de extração de dados de páginas da web para serem usados para análise, comparação e muitos outros propósitos.
2. O tipo de dados que pode ser coletado varia dentre textos, imagens, classificações, URLs e muito mais.

# Legalidade



Contanto que os dados que você está coletando não exijam uma conta para acesso, não sejam bloqueados por um arquivo **robots.txt** e estejam disponíveis publicamente, **provavelmente** não há problema.

Para que o *web scraping* pode ser usado?

## Apresentação

---

O que vamos implementar hoje?

*Web scraping* envolve três etapas:

1. **Coleta de dados:** nesta etapa, os dados são coletados de páginas da web (geralmente com um web crawler)
2. **Análise e transformação de dados:** envolve a transformação do conjunto de dados coletados em um formato que pode ser usado para análise posterior (como uma planilha ou arquivo JSON)
3. **Armazenamento de dados:** o último estágio envolve o armazenamento dos dados transformados

Vamos começar!



# Etapas para criar um *web scraper* com Python



1. Selecione os URLs que você deseja extrair e como será definido no código
2. Encontre o conteúdo HTML que você deseja extrair
3. Escolha suas ferramentas e bibliotecas

# Implementação



Utilizar os códigos disponíveis no repositório abaixo para elaborar um *web scraper*:

Repo: <https://github.com/oAGoulart/workshops>

**PEÇA AJUDA!**

# Desafio

# Extrair textos e artigos em *Markdown*



Tentar implementar um *web scraper* que possa extrair textos e artigos no formato **Markdown** mantendo o estilo dos títulos e parágrafos.

# Referências



Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2009).

***Introduction to algorithms.***

Mit Press.



Localhost (2022).

**Basic training: Intro to python skills for ai, part ii.**

[Online; acessado em 23-Julho-2022].



Moroney, L. (2021).

***AI and Machine Learning for Coders a programmer's Guide to Artificial Intelligence.***

O'Reilly.