



escola
britânica de
artes criativas
& tecnologia

Profissão: Analista de dados



COLETA DE DADOS II



GUIA DA AULA 1



Capte dados da web (crawling)



Acompanhe aqui
os temas que
serão tratados
na videoaula

• Estruturas de dados

• HTTP

• Pacote requests



Estruturas de dados

- **Não estruturado:** texto, imagem, áudio, etc.
- **Semi estruturado:** html, json, etc.
- **Estruturado:** tabelas, planilhas, etc.

Web Crawling

HTTP

O HTTP (*hypertext transfer protocol*) é um protocolo de transferência de hipertexto (texto, imagens, vídeos, etc.). É o protocolo padrão de transferência de informação pela internet: <http://www.google.com/>

Cliente / Servidor é a arquitetura da internet. Nela, um **cliente** (navegador web, código Python, etc.) utiliza um **método** HTTP para interagir com um servidor (requisitar dados, enviar dados, etc.). O **servidor**, por sua vez, envia uma resposta para o cliente com um código de retorno indicando se a interação ocorreu com sucesso. **Métodos** são as operações que podemos realizar com o protocolo para interagir com um servidor.



Encontre uma lista completa de métodos neste [link](#)



Os métodos mais importantes são:

- **GET**: Requisitar dados (acessar uma página web, carregar o feed do Instagram, etc.);
- **POST**: Enviar dados (login, cadastro, mensagem WhatsApp, tweet do Twitter).

Códigos de retorno são os números de 0 a 1000 que recebemos como resposta do servidor ao realizar uma operação qualquer.

Os códigos de retorno mais importantes são:

- Entre **200** e **299**: Sucesso; Entre **400** e **499**: Erro do cliente;
- Entre **500** e **599**: Erro do servidor.

Código **200** (sucesso) é o mais comum e o **404** (não encontrado) o mais famoso!



Encontre uma lista completa de códigos de retorno neste [link](#)



Pacote requests

Pacote Python para interagir com a web através do protocolo HTTP.

```
In [ ]: import requests

        print(requests.__version_)
```

Método:

```
In [ ]: resposta = requests.get('http://www.google.com')
```

Código de retorno:

```
In [ ]: print(resposta.status_code)
```



Encontre a
documentação
neste ([link](#)).

