Introdução

Primeiro Web Scraper

Introdução – Primeiro Web Scraper

Executar um web scraping é realizar solicitações GET para um servidor web em busca de uma página específica, ler a saída HTML desta página e extrair dados para isolar o conteúdo que precisamos.

A princípio parecerá meio complicado visualizar uma resposta do servidor sem a formatação feita pelo browser, mas é isso que fazemos quando realizamos um scraping.

Introdução – Primeiro Web Scraper

Temos duas tarefas básicas a realizar quando fazermos web scraping.

1. Carregar páginas web como strings.
2. Analisar o HTML para localizar as informações que interessam.

Introdução – Primeiro Web Scraper

O Python oferece duas excelentes ferramentas para realizar as duas tarefas.

Podemos usar *requests* através da biblioteca *urllib* para carregar as páginas web e *BeautifulSoup* para realizar a análise das informações.

Nesta aula veremos como carregar o conteúdo de páginas web e na próxima aula veremos como usar o *BeautifulSoup*.

Introdução – Primeiro Web Scraper

urllib é uma biblioteca padrão do Python e contém funções para solicitação de dados web, manipulação de cookies e até alteração de metadados como cabeçalhos e o agente do usuário.

A função ***urlopen*** é usada para abrir um objeto remoto por meio de uma rede e ler tal objeto. Este objeto pode ser arquivos HTML, arquivos de imagem ou qualquer outro fluxo de arquivos.

Introdução – Primeiro Web Scraper

Veja como fazer uma requisição a um site com Python.

```
from urllib.request import urlopen
html = urlopen("https://evaldowolkers.wordpress.com/")
print(html.read())
```

Foi importada a função urlopen do módulo request que pertence à biblioteca urllib.

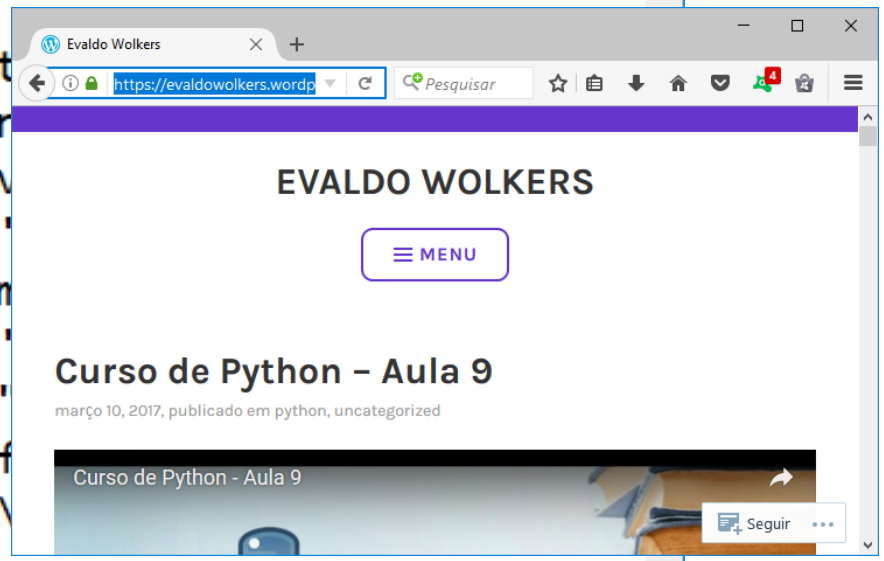
Será exibido o código HTML retornado pelo servidor web.

Introdução – Primeiro Web Scraper

Veja como fazer uma requisição a um site com Python.

Command Prompt

```
>>> from urllib.request import urlopen
>>> html = urlopen("https://evaldowolkers.wordpress.com/")
>>> print(html.read())
b'<!DOCTYPE html>\n<html lang="pt-BR">\n<head>\n<meta charset
iewport" content="width=device-width, initial-scale=1">\n<lir
tp://gmpg.org/xfn/11">\n<link rel="pingback" href="https://ev
m/xmlrpc.php">\n\n<title>Evaldo Wolkers</title>\n<link rel=\'
s2.wp.com\' />\n<link rel=\'dns-prefetch\' href=\'//s0.wp.com
refetch\' href=\'//s1.wp.com\' />\n<link rel=\'dns-prefetch\'
is.com\' />\n<link rel="alternate" type="application/rss+xml"
 Wolkers &raquo;" href="https://evaldowolkers.wordpress.com/f
ernate" type="application/rss+xml" title="Feed de coment\xc3\
```



FIM