

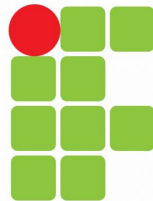
TCHELINUX 2019  
UNIVERSIDADE FEDERAL DO PAMPA  
CAMPUS BAGÉ  
CICLO DE PALESTRAS

---

## **Web Scraping: a arte de automatizar a recuperação de informação na Web**

**Prof. Alex Dias Camargo**

[alexcamargo@ifsul.edu.br](mailto:alexcamargo@ifsul.edu.br)



INSTITUTO FEDERAL DE  
EDUCAÇÃO, CIÊNCIA E TECNOLOGIA  
SUL-RIO-GRANDENSE



# I. Agenda

---

Esta palestra está organizada da seguinte maneira:

- ❑ **Apresentação**
- ❑ **Objetivos**
- ❑ **Ferramentas**
- ❑ **Componentes de uma página *web***
- ❑ **O que é *Web scraping*?**
- ❑ ***Scraping* na prática: [bage.rs.gov.br](http://bage.rs.gov.br)**
- ❑ ***Bônus: web scraping completo em 4 minutos (GUI mode)***
- ❑ **Onde estudar**
- ❑ **Agradecimentos**
- ❑ **Referências**



## II. Apresentação

---

Formação acadêmica:

□ **Bacharel em Sistemas de Informação (URCAMP, 2011)**

TCC: *Web sistema integrado a uma rede social para academias de ginástica*

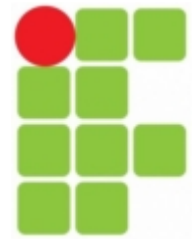
Orientador: Prof. Abner Guedes

□ **Especialista em Sistemas Distribuídos com Ênfase em Banco de Dados (UNIPAMPA, 2013)**

TCC: *Interligando bases de dados do sistema Controle de Marcas e Sinais utilizando o MySQL Cluster*

Orientador: Prof. Érico Amaral

Coorientador: Prof. Rafael Bastos (IDEAU)



## II. Apresentação

---

Formação acadêmica:

□ **Mestre em Engenharia de Computação (FURG, 2017)**

Dissertação: **EN-MUTATE: predição do impacto de mutações pontuais em proteínas utilizando *Ensemble Learning***

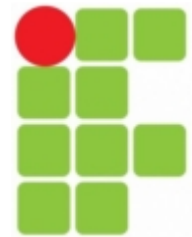
Orientadora: Profa. Karina Machado

Coorientador: Prof. Adriano Werhli

□ **Doutorado (iniciado) em Ciência da Computação (UFPEL)**

Tese: **Em definição**

Orientador: Aluno especial



## II. Apresentação

---

Experiência acadêmica:

□ **Professor (Ensino Técnico)**

Local: **Capacitar Escola Técnica**

Disciplinas: Banco de Dados e Análise de Sistemas

□ **Professor (Ensino Superior)**

Local: **Universidade Federal do Pampa - UNIPAMPA**

Disciplinas: Algoritmos e Programação, Laboratório  
Programação I e Laboratório de Programação II

□ **Professor (Ensino Básico, Técnico e Tecnológico)**

Local: **IFSUL Câmpus Bagé**

Disciplinas: Programação para Web II, Arquitetura de  
Computadores, Qualidade de Software, Desenvolvimento de  
Software, Informática (Eng. Agrônômica)



## II. Apresentação

---

Projetos acadêmicos:

- **Algo+: um portal para o apoio ao ensino de Algoritmos**  
Universidade: **UNIPAMPA**  
Área: Informática na educação
- **Bioinformática Estrutural de Proteínas: modelos, algoritmos e aplicações biotecnológicas**  
Universidade: **FURG/UFGM/UFPA**  
Área: Bioinformática
- **Unihacker.Club: Programa Universidade Hacker**  
Universidade: **UNIPAMPA**  
Área: Segurança da informação



## II. Apresentação

---

Periódicos acadêmicos:

❑ **Revisor do periódico ICCEEg (ISSN 2236-0093)**

Universidade: **FURG**

Área: Multidisciplinar

❑ **Revisor do periódico CCEI (ISSN 2356-6635)**

Universidade: **URCAMP**

Área: Multidisciplinar



## II. Apresentação

---

Histórico Tchelinix:

□ **Software Livre e Bioinformática: do DNA ao medicamento (2018)**

Local: **FURG**

Área: Bioinformática


TCHELINUX  
UNIVERSIDADE FEDERAL DO PAMPA  
CAMPUS BAGÉ  
CICLO DE PALESTRAS

---

**Software Livre e Bioinformática: do DNA ao medicamento**

**Alex Dias Camargo**

[alex@apus.digital](mailto:alex@apus.digital)



APUS DIGITAL

Setembro/2018



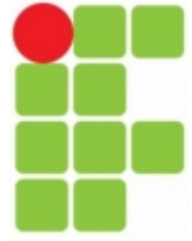


## III. Objetivos

---

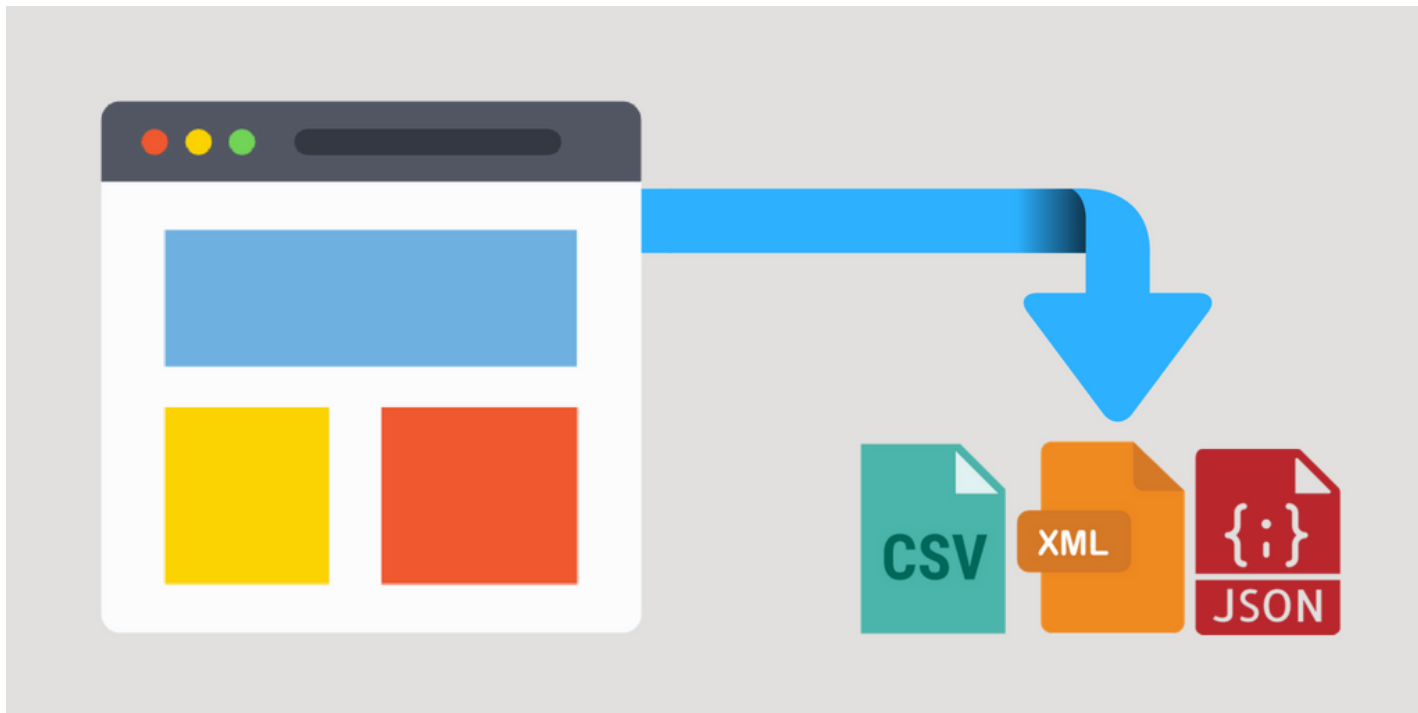
Dentre os objetivos, podem ser destacados:

- Compreender os conceitos e aplicações do **Web Scraping**.
- Explorar tecnologias no contexto **Python/Linux**.
- **Motivar novos cientistas de dados (e palestrantes).**



# 1. Os componentes de uma página *web*

---





# 1. Os componentes de uma página web

---

**Quando uma página na *Internet* é visitada**, o navegador faz uma solicitação à um servidor *web*. Essa solicitação é chamada de **GET**, pois são recebidos arquivos do servidor.

- **HTML**: contém o conteúdo principal da página.
- **CSS**: adiciona estilos para que a página fique customizada.
- **JS**: arquivos JavaScript adicionam interatividade à página.
- **Imagens**: formatos de imagem, tais como JPG e PNG.



# 1. Os componentes de uma página *web*

---



Figura. Camadas do desenvolvimento *web*.



# 1. Os componentes de uma página web

```
1 → <!DOCTYPE html>
2
3 → [ <html lang="pt-br">
4   |
5   [ <head>
6     |
7     [ <meta charset="utf-8">
8       |
9       [ <title>Título do site</title>
10        |
11        [ </head>
12         |
13         [ <body>
14           |
15           [ Corpo do site.
16             |
17             [ </body>
18              |
19              [ </html>
```

Figura. Estrutura básica do HTML5.



## O BACKEND WEB

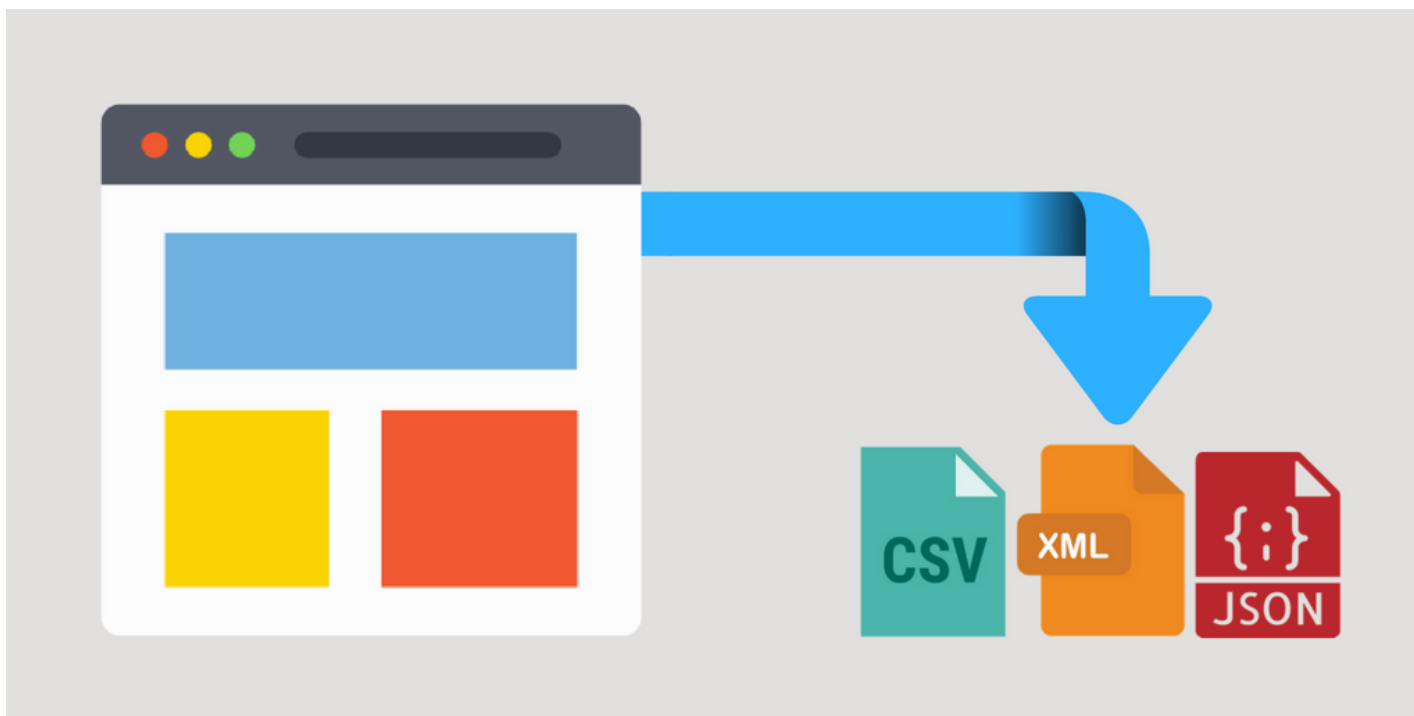


Figura. Arquitetura web padrão.



## 2. O que é *Web Scraping*?

---





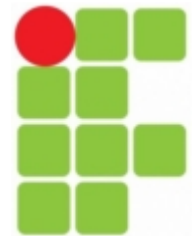
## 2. O que é *Web Scraping*?

---

*Web Scraping* é um **método de "raspagem" de dados** de *sites* que usa *scripts* para obter as informações necessárias, **simulando um comportamento "humano"**.

- Um uso popular do *scraping* na *web* é **procurar ofertas online**, como passagens aéreas, *shows*, etc.





## 2. O que é *Web Scraping*?

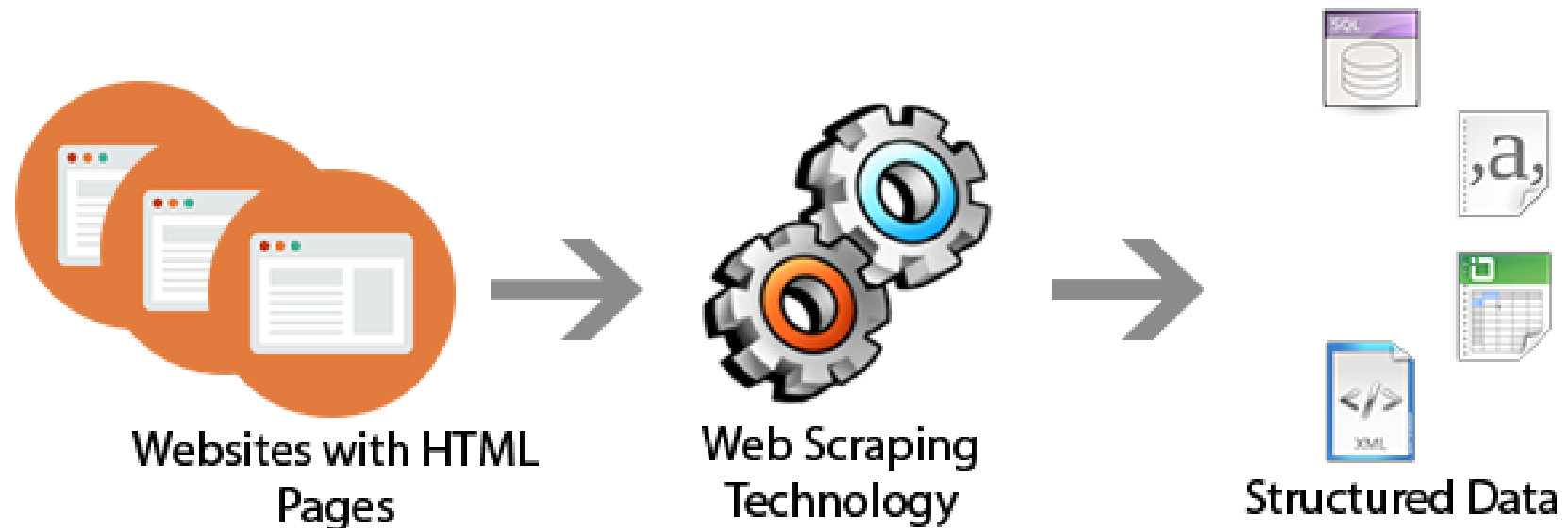


Figura. Visão geral de um *web scraping*.

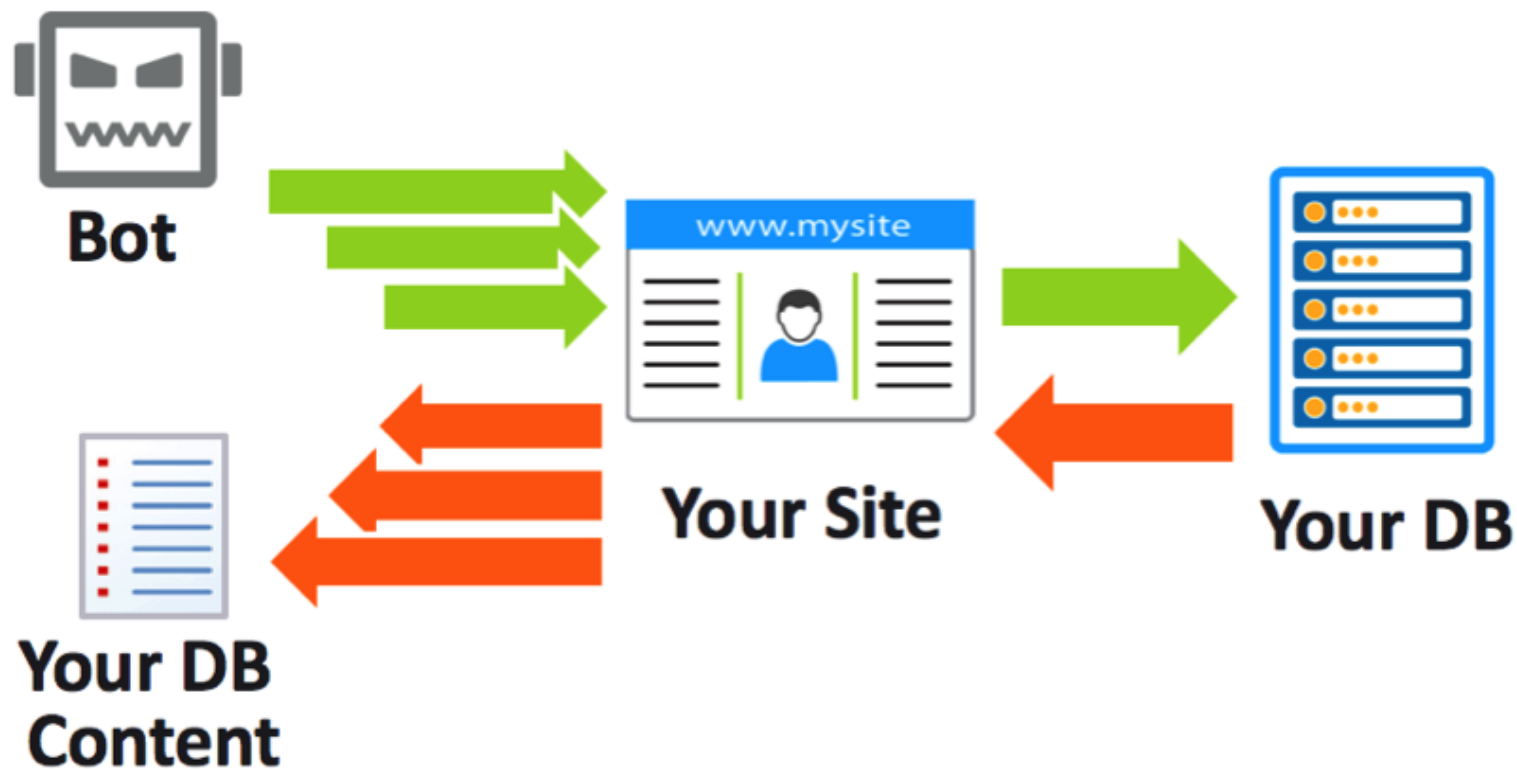
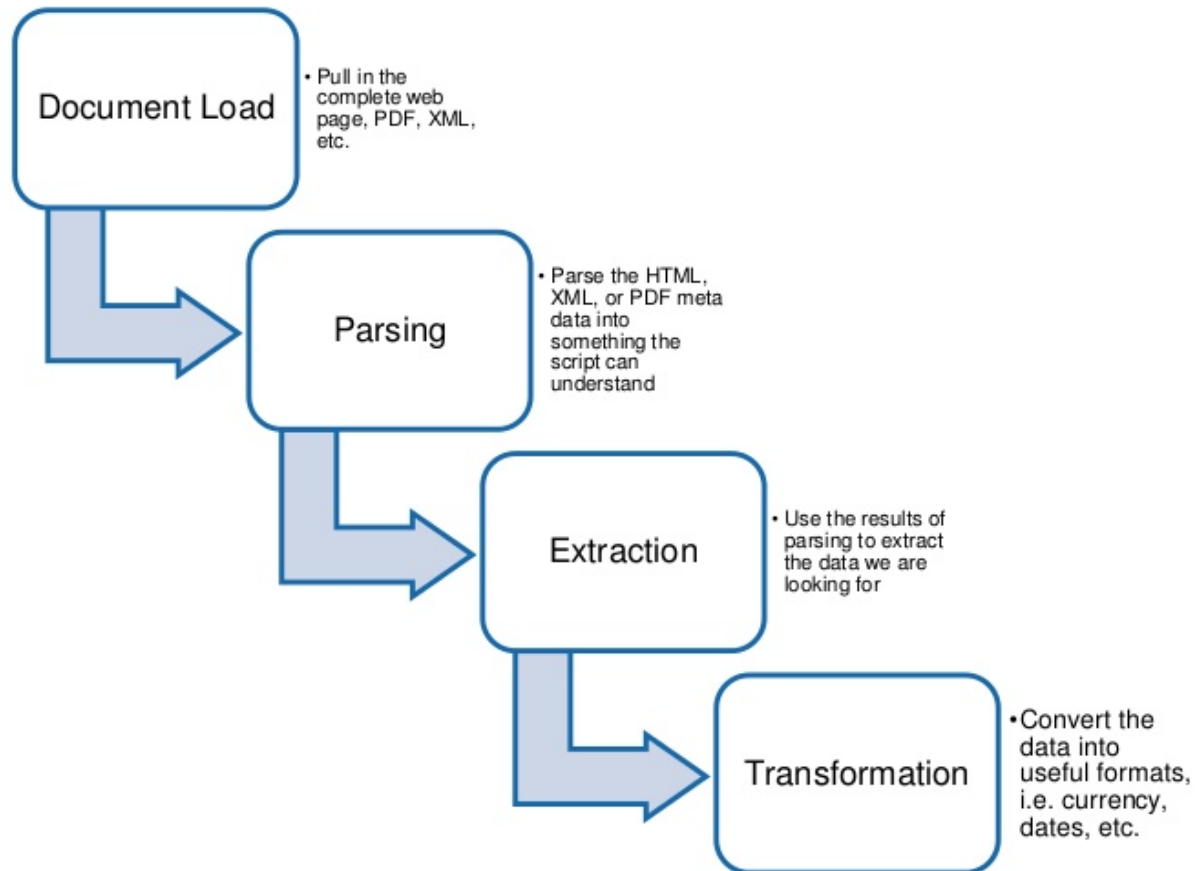


Figura. Visão geral de um *web scraping*.

# Anatomy of a Scraper

2.



© 2014 Tommy Tavenner

Figura. Visão geral de um *web scraping*.

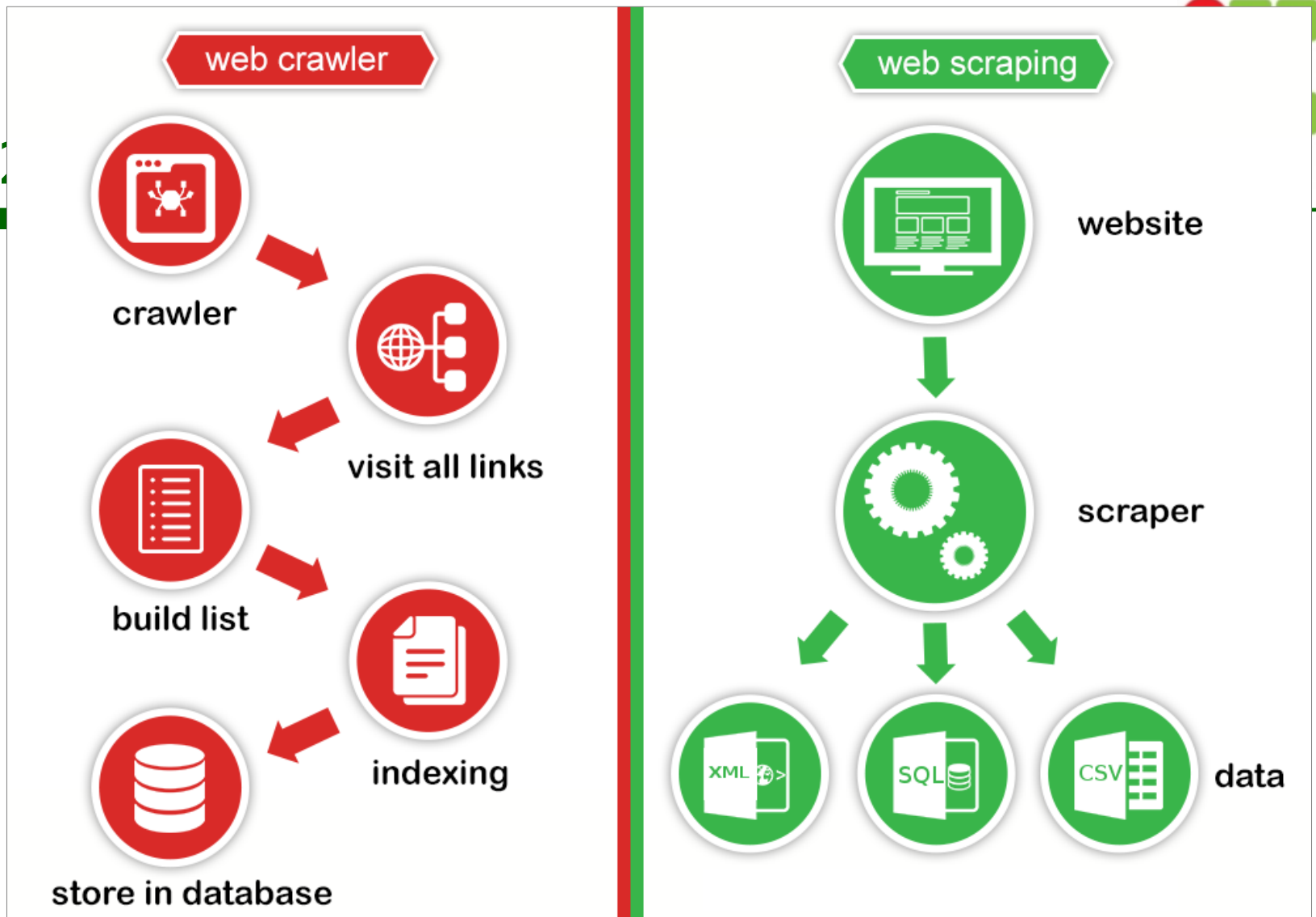


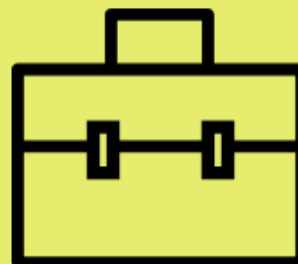
Figura. *Web Scraping versus Web Crawling.*



E-commerce



Data Science



Job Boards

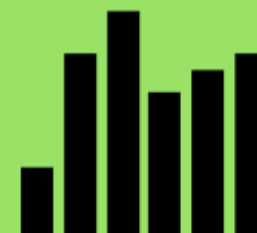


Marketing & Sales



Data Journalism

# **Web Scraping Applications**



Finance

Figura. Aplicações do *web scraping*.



## 2. O que é *Web Scraping*?

---

*Web Scraping* é um **método de "raspagem" de dados** de *sites* que usa *scripts* para obter as informações necessárias, **simulando um comportamento "humano"**.

- ❑ Um uso popular do *scraping* na *web* é **procurar ofertas online**, como passagens aéreas, *shows*, etc.
- ❑ Existem **empresas especializadas** no ramo?



## 2. ¿Qué es Web Scraping?

No Recurring Monthly Fees. Pay Only For What You Use.

STARTER	BEST VALUE	PROFESSIONAL	ENTERPRISE
<b>99\$</b> ONE TIME FEE	<b>\$179</b> ONE TIME FEE	<b>299\$</b> ONE TIME FEE	<b>CONTACT</b> FOR PRICING
<b>50,000</b> DATA RECORDS	<b>100,000</b> DATA RECORDS	<b>200,000</b> DATA RECORDS	<b>300,000+</b> DATA RECORDS
<b>NO EXCESS FEES</b> FIXED PRICE GUARANTEE	<b>NO EXCESS FEES</b> FIXED PRICE GUARANTEE	<b>NO EXCESS FEES</b> FIXED PRICE GUARANTEE	<b>NO EXCESS FEES</b> FIXED PRICE GUARANTEE
Free Trial	Free Trial	Free Trial	Free Trial
Features	Features	Features	Features

Figura. Empresas de *web scraping*.



## ¿Qué es Web Scraping?

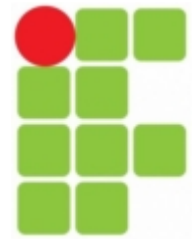
### Pricing

Subscription	On Demand	
<div><b>Basic</b></div> <div>Starting at</div> <div><b>\$50</b></div> <div>per month per website maximum 1000 pages per site</div> <div>Monthly subscription required</div>	<div><b>Enterprise</b></div> <div>Starting at</div> <div><b>\$1000</b></div> <div>per month</div> <div>Monthly subscription required</div>	<div><b>On Demand</b></div> <div>Starting at</div> <div><b>\$300</b></div> <div>per website</div> <div>No subscription required</div>

Contact Us

Figura. Empresas de *web scraping*.





## 2. O que é *Web Scraping*?

---

*Web Scraping* é um **método de "raspagem" de dados** de *sites* que usa *scripts* para obter as informações necessárias, **simulando um comportamento "humano"**.

- ❑ Um uso popular do *scraping* na *web* é **procurar ofertas online**, como passagens aéreas, *shows*, etc.
- ❑ Existem **empresas especializadas** no ramo?
- ❑ Uma alternativa para o *web scraping* é usar uma **API**, **se houver alguma disponível**. Ex.: *Twitter, Instagram, Facebook*, etc.



## 2 O que é Web Scraping?

The screenshot shows the Twitter Developer documentation page. The top navigation bar is purple with links for Developer, Use cases, Products, Docs, More, and Labs. On the right, there are links for Apply, Apps, a search icon, and a user profile icon. The main content area has a search bar on the left and a 'Docs' section on the right. The 'Docs' section is titled 'Stay Informed' and includes a 'Learn how' button. Below this, there are two columns: 'Search Tweets' and 'Account Activity API', each with a brief description of the API's capabilities.

Developer Use cases Products Docs More Labs Apply Apps

Search all documentation...

**Docs**

**Basics**

**Accounts and users**

**Tweets**

**Direct Messages**

**Media**

**Trends**

**Geo**

**Ads**

**Metrics**

**Stay Informed**

Staying informed about changes to our APIs is important for those developing on the platform and can be critical to maintaining your applications. We have a number of channels to help you stay in-the-loop.

[Learn how >](#)

**Search Tweets**

Use the Search API to find historical Tweets. Free to enterprise versions available.

**Account Activity API**

Have 15+ account activities delivered to you in realtime via a webhook connection.

Figura. APIs para extração de dados.



## 2 O que é Web Scraping?

**facebook** for developers

DocumentosFerramentasSuporteMeus aplicativos

Pesquisar documentação do desenvolvedor

**Graph API**  
Overview  
Using the Graph API  
FAQ  
Reference  
Webhooks  
Advanced  
Changelog  
Server-Sent Events

# Graph API

A versão mais recente é: **v4.0**

A Graph API é a principal forma de os aplicativos lerem e gravarem no gráfico social do Facebook. Todos os nossos SDKs e produtos interagem com a Graph API de algum modo, e nossas outras APIs são extensões da Graph API. Por isso, é crucial entender como ela funciona.

Se você não conhecer bem a Graph API, recomendamos que comece por estes documentos:

**Visão geral**  
Saiba como a Graph API está estruturada, o que são tokens de acesso e como funcionam as versões.

**Como usar a Graph API**  
Saiba como executar operações comuns.

**Explorador da Graph API**  
Saiba como fazer consultas e receber respostas da Graph API com nosso aplicativo Explorador da Graph API.

**Referência**  
Saiba como ler nossos documentos de referência para encontrar facilmente o que procura.

Nesta Página

Graph API

Este documento foi útil?  
☐ Sim ☐ Sim, mas... ☐ Não

Depois de se familiarizar com os conceitos básicos, passe para tópicos mais avançados como estes:  
[Saiba como nossos SDKs interagem com a Graph API lendo a documentação do SDK para iOS, Android](#)

Figura. APIs para extração de dados.

TCHELINUX - 2019

27



## 2. O que é *Web Scraping*?

---

*Web Scraping* é um **método de "raspagem" de dados** de *sites* que usa *scripts* para obter as informações necessárias, **simulando um comportamento "humano"**.

- ❑ Um uso popular do *scraping* na *web* é **procurar ofertas online**, como passagens aéreas, *shows*, etc.
- ❑ Existem **empresas especializadas** no ramo?
- ❑ Uma alternativa para o *web scraping* é usar uma **API**, **se houver alguma disponível**. Ex.: *Twitter, Instagram, Facebook*, etc.
- ❑ **"Be polite" (seja educado)**: um *scraping* pode sobrecarregar um servidor, principalmente, se o *script* estiver fazendo uma grande quantidade de solicitações. **Respeite o robots.txt!**

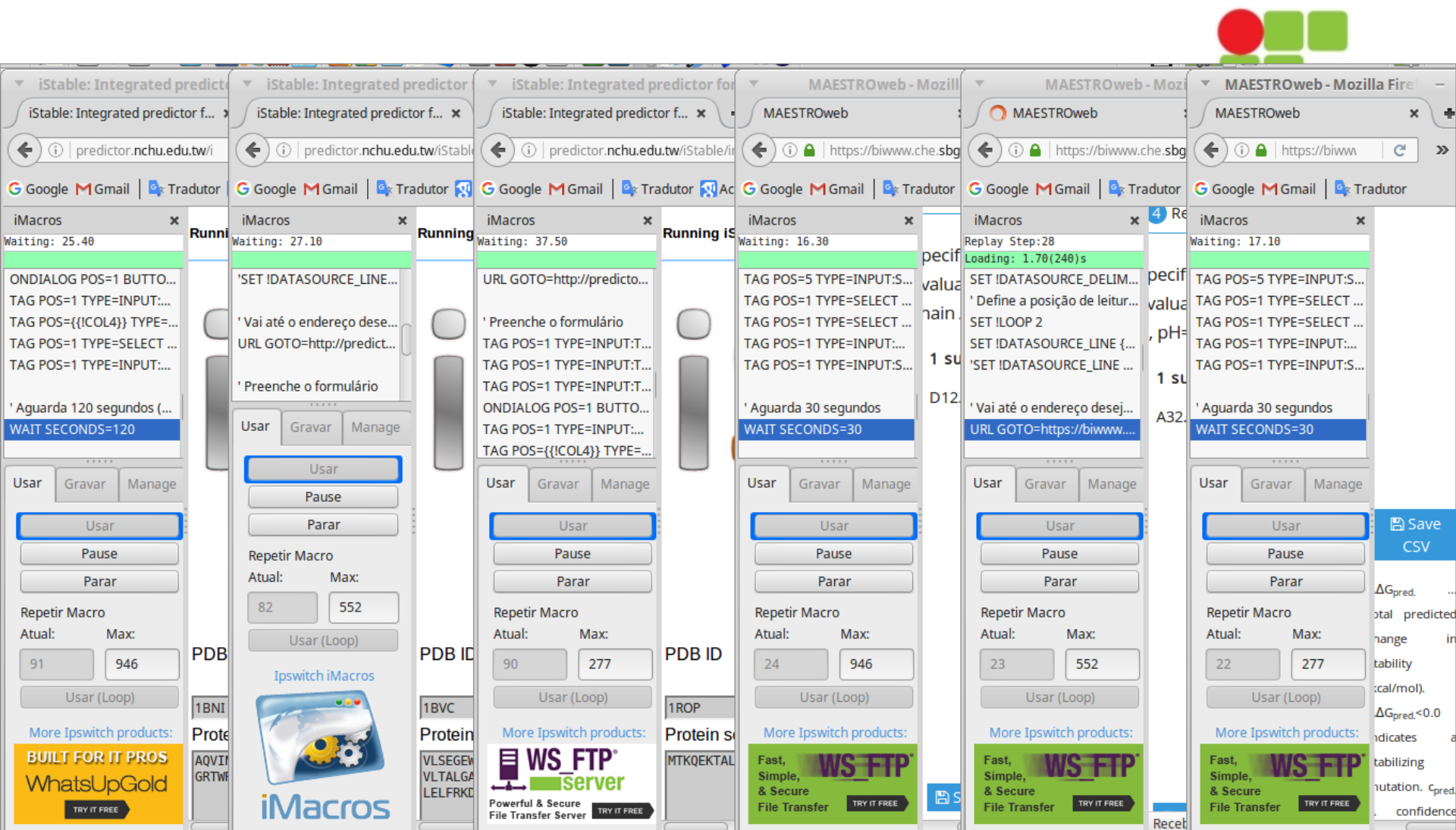


Figura. Muitas requisições "simultâneas".



robots.txt

 <https://varvy.com/robots.txt>

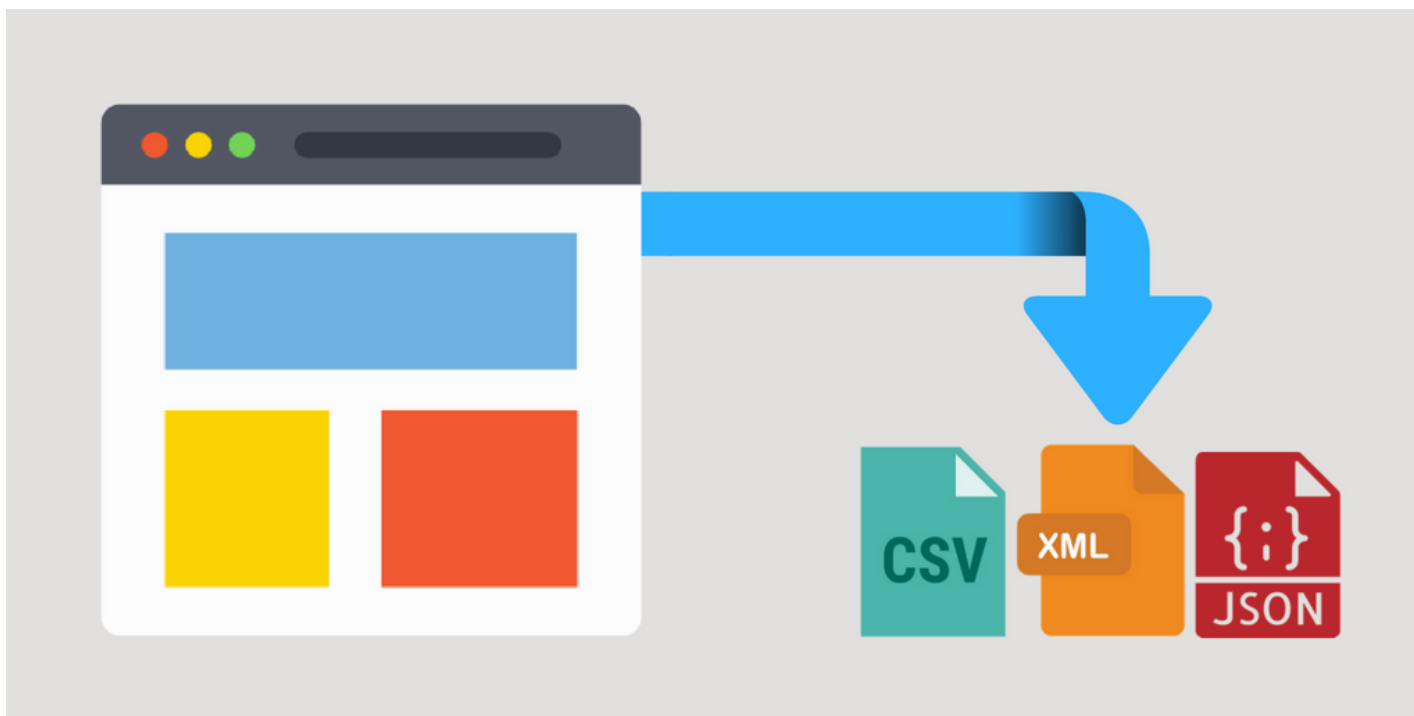
```
User-agent: *  
Disallow: /folder/  
Disallow: /file.html  
Disallow: /image.png
```

Figura. Exemplo de um robots.txt.



### 3. Ferramentas

---





## 3. Ferramentas

---

*KIT DO WEB SCRAPER:*

- ❑ **GitHub do projeto:**  
<https://github.com/alexcamargoweb/python-webscraping>
- ❑ **Linux Mint:**  
<https://linuxmint.com/download.php>
- ❑ **Python 3:**  
<https://www.python.org/downloads/>
- ❑ **Requests:**  
<https://pypi.org/project/requests/>
- ❑ **URLlib:**  
<https://pypi.org/project/urllib3/>
- ❑ **BeautifulSoup:**  
<https://pypi.org/project/beautifulsoup4/>





## 3. Ferramentas

---

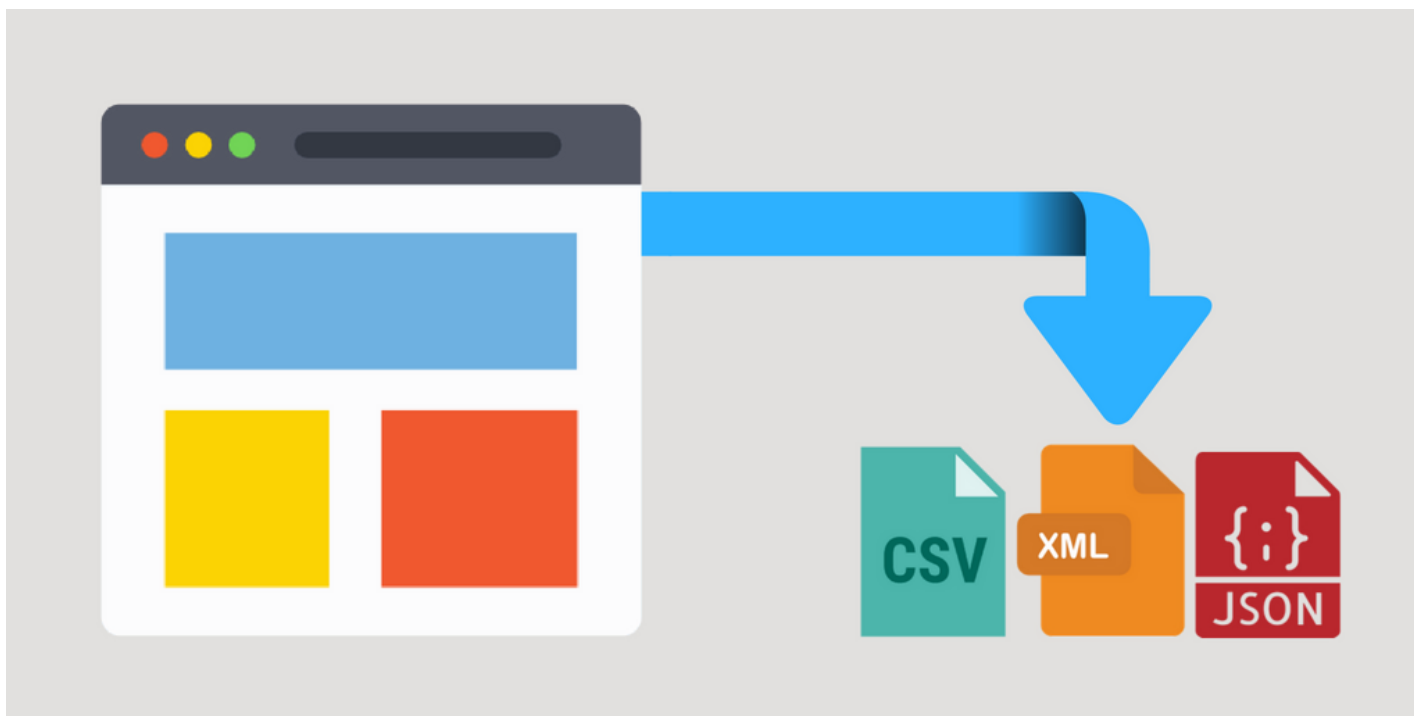
*KIT DO WEB SCRAPER:*

- ***Selenium:***  
<https://pypi.org/project/selenium/>
- ***Firefox geckodriver:***  
<https://github.com/mozilla/geckodriver/releases/>
- ***Selenium IDE (Firefox extension):***  
<https://addons.mozilla.org/pt-BR/firefox/addon/selenium-ide/>



## 4. *Scraping* na prática: bage.rs.gov.br

---





## 4. *Scraping* na prática: bage.rs.gov.br

PREFEITURA MUNICIPAL DE BAGÉ

INÍCIO O MUNICÍPIO ESTRUTURA ADMINISTRATIVA SERVIÇOS ON-LINE NOTÍCIAS PROGRAMAS MURAL OUVIDORIA

**PROCESSO SELETIVO**  
ASSISTENCIA SOCIAL

INSCRIÇÕES GRATUITAS  
**31.07**  
DAS 08 ÀS 23:59  
SITE PREFEITURA

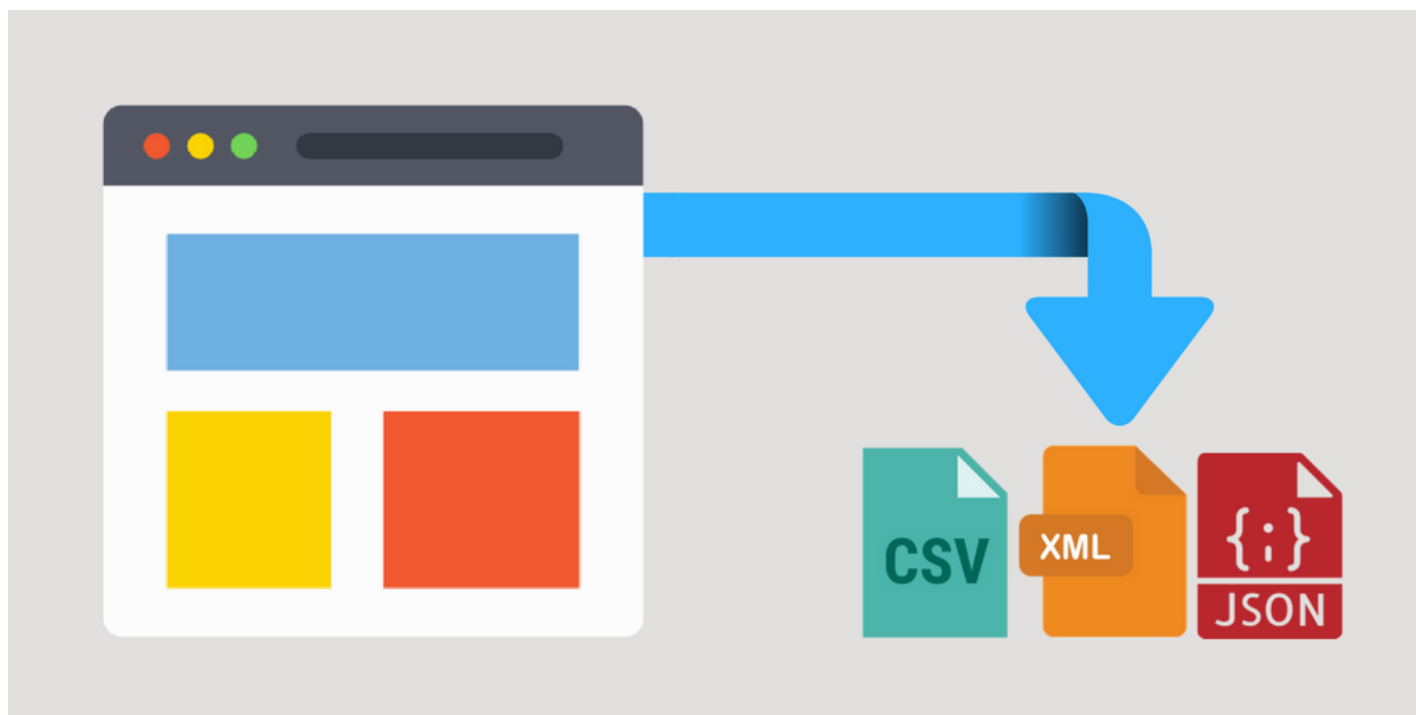
PROCESSO SELETIVO 2019 PORTAL DA TRANSPARÊNCIA 2ª VIA DO IPTU LICITAÇÕES

Link. <https://www.bage.rs.gov.br>



## 5. Bônus: *Web scraping* completo em 4 minutos (*GUI mode*)

---





## 5. Bônus: *Web scraping* completo em 4 minutos (*GUI mode*)

---

Demonstração...





## IV. Onde estudar

---

### Se interessou pelo assunto? :)

- ❑ *Learn Web Scraping with Python from Scratch* (15.851 alunos)  
<https://www.udemy.com/course/web-scraping-python-tutorial/>
- ❑ Anotações e *scripts* de *web scraping*, *screen scraping*, etc  
<https://github.com/ferreiraapfernanda/web-scraping>
- ❑ *Python Selenium WebDriver*  
<https://www.youtube.com/playlist?list=PLUY1IsOTtPeJNBuSweXS9pcSKbP4mr32S>
- ❑ MITCHELL, Ryan. **Web Scraping with Python: Collecting More Data from the Modern Web**. "O'Reilly, Inc.", 2018.  
<https://www.amazon.com/Web-Scraping-Python-Collecting-Modern/dp/1491910291>



# Agradecimentos

---

Obrigado pela sua participação! :)

- ❑ **Acadêmicos do curso de Engenharia de Computação**
- ❑ **Grupo de usuários de Software Livre Tchelinux**
- ❑ **Toda honra e glória ao Senhor Jesus!**



Abraços, Prof. Alex Dias Camargo  
Unipampa – Câmpus Bagé  
05 de Outubro de 2019



# Referências básicas

---

<https://likegeeks.com/python-web-scraping/>

<https://codeburst.io/web-scraping-101-with-python-beautiful-soup-bb617be1f486>

<https://imasters.com.br/back-end/aprendendo-sobre-web-scraping-em-python-utilizando-beautifulsoup>

<https://goomore.com/blog/web-scraping-python/>

<https://github.com/REMitchell/python-scraping>

<https://medium.com/data-hackers/como-fazer-web-scraping-em-python-23c9d465a37f>