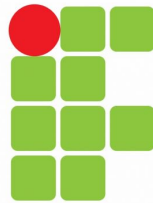


INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
SUL-RIO-GRANDENSE
CÂMPUS BAGÉ
SACI - 2019

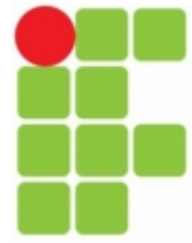
Introdução prática ao *web scraping* com *Python*

Prof. Alex Dias Camargo

alexcamargo@ifsul.edu.br



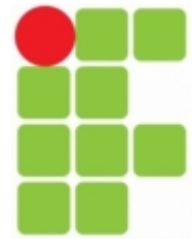
INSTITUTO FEDERAL DE
EDUCAÇÃO, CIÊNCIA E TECNOLOGIA
SUL-RIO-GRANDENSE



I. Agenda

Esta palestra está organizada da seguinte maneira:

- ☐ **Apresentação**
- ☐ **Objetivos**
- ☐ **Ferramentas**
- ☐ **Os componentes de uma página *web***
- ☐ ***Web scraping***
- ☐ ***Scraping* de uma página da *Wikipédia***
- ☐ ***Scraping* para montar *datasets* de esportes**
- ☐ **Exercícios**
- ☐ **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*

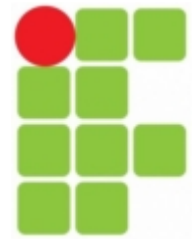
Orientador: Prof. Abner Guedes

□ **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***

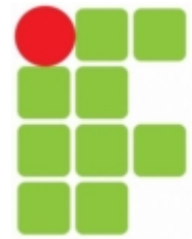
Orientadora: Profa. Karina Machado

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 do professor

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/UEPB**
Área: Bioinformática
- **Unihacker.Club: Programa Universidade Hacker**
Universidade: **UNIPAMPA**
Área: Segurança da informação



II. Apresentação do professor

Periódicos acadêmicos:

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

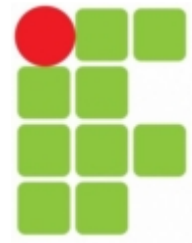
Universidade: **FURG**

Área: Multidisciplinar

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

Universidade: **URCAMP**

Área: Multidisciplinar



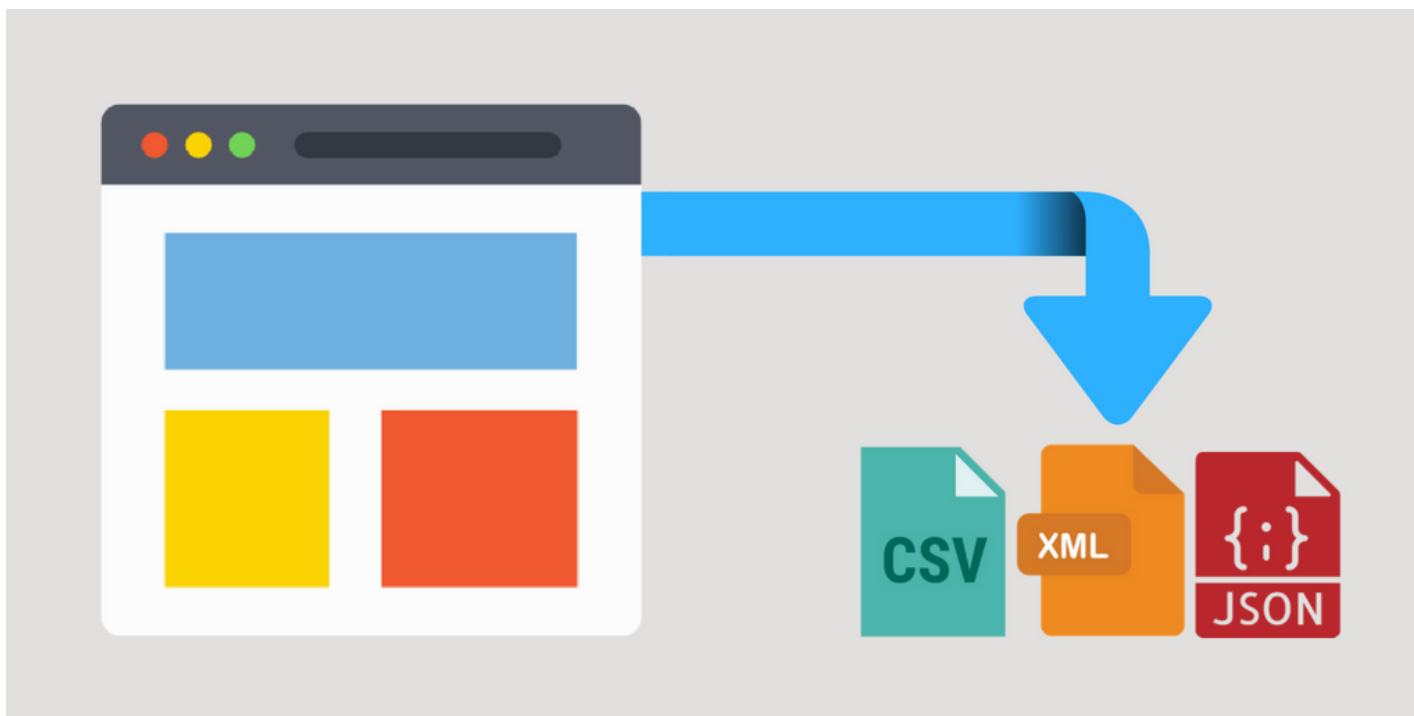
III. Objetivos

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.

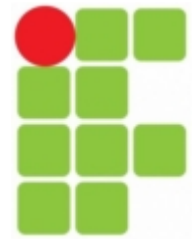
- **HTML**: contém o conteúdo principal da página.
- **CSS**: adiciona estilos para que a página fique customizada.
- **JS**: arquivos JavaScript adicionam interatividade à página.
- **Imagens**: formatos de imagem, tais como JPG e PNG.



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        [ <body>
12          |
13          [ Corpo do site.
14            |
15            [ </body>
16              |
17              [ </html>
```

Figura. Estrutura básica do HTML5.



1. Os componentes de uma página web

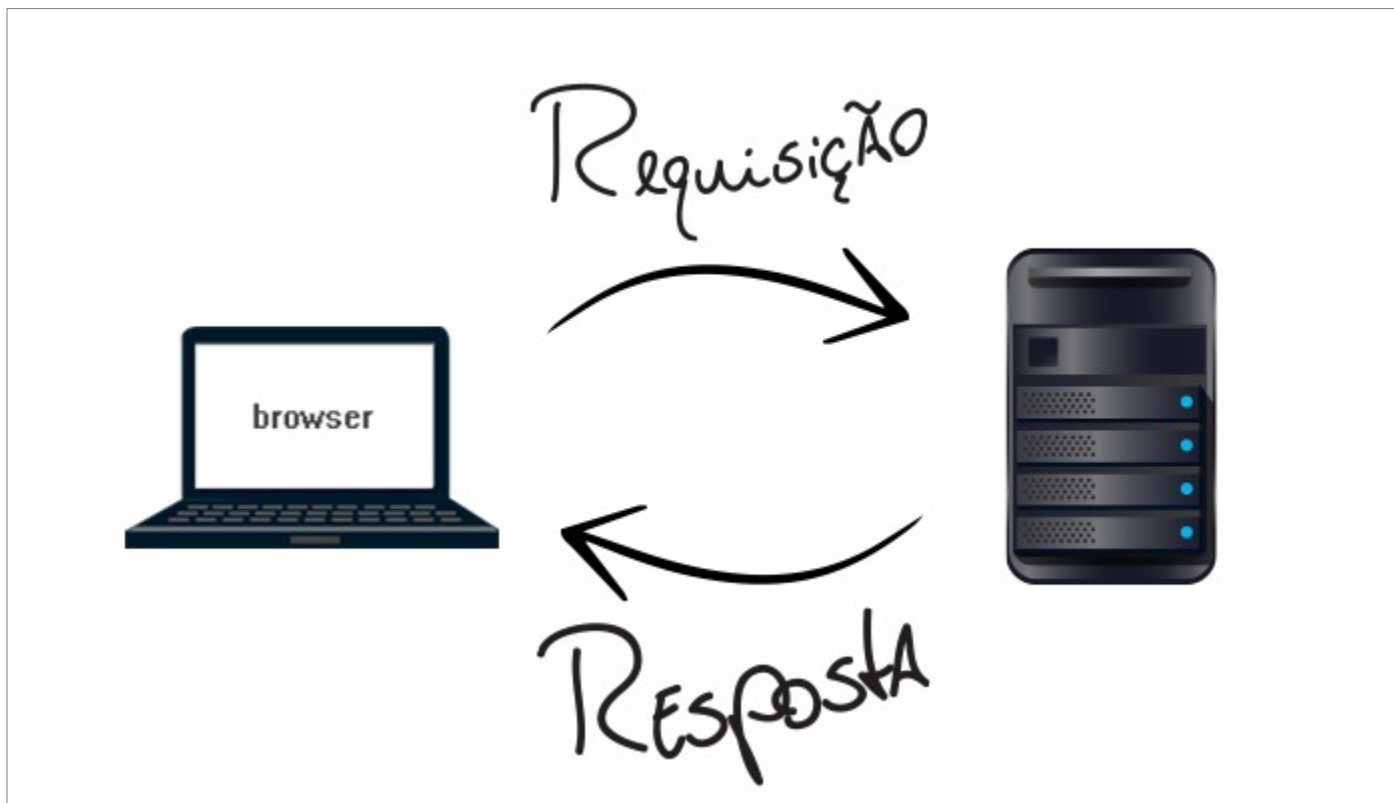
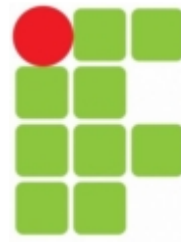
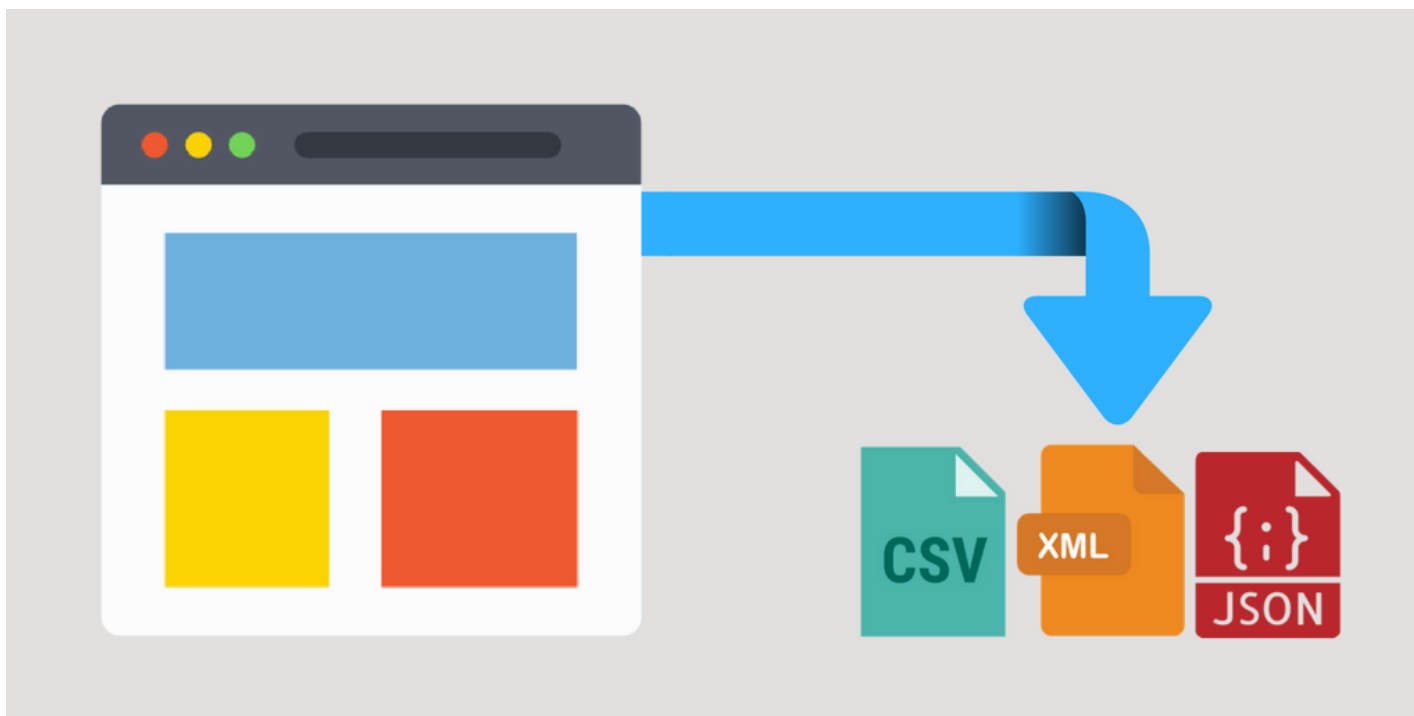
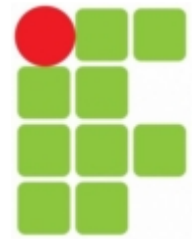


Figura. Arquitetura web padrão.



2. *Web Scraping*





2. 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"**.

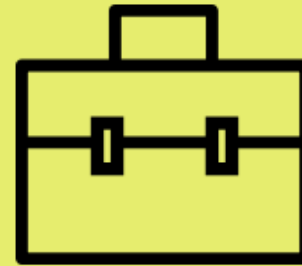
- Um uso popular do *scraping* na *web* é **procurar ofertas online**, como passagens aéreas, *shows*, etc.



E-commerce



Data Science



Job Boards

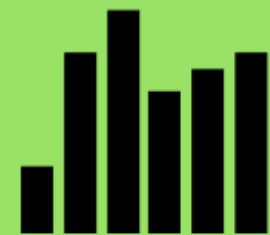


Marketing & Sales



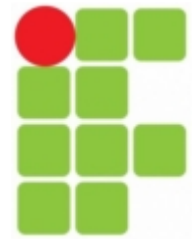
Data Journalism

Web Scraping Applications



Finance

Figura. Aplicações do *web scraping*.




2. 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 Web Scraping



Scraping Solutions
When the solutions you seek seems impossible

HOMEHOW IT WORKS?PRICINGFAQPRODUCTSGet A Quote

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

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

Call us for free!

FeaturesFeaturesFeaturesFeatures

Figura. Empresas de *web scraping*.



2 Web Scraping

Pricing

Subscription

Basic

Starting at

\$50

per month per website
maximum 1000 pages per site

Monthly subscription required

Enterprise

Starting at

\$1000

per month

Monthly subscription required

On Demand

On Demand

Starting at

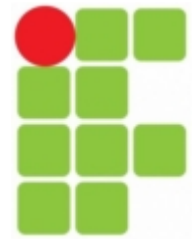
\$300

per website

No subscription required

Contact Us

Figura. Empresas de *web scraping*.



2. 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 Web Scraping

 [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 Web Scraping



 Sandbox Invites

 Manage Clients

 Entrar

Overview

Authentication >

Login Permissions >

Permissions Review >

Sandbox Mode >

Secure Requests >

Endpoints >

Embedding >

Mobile Sharing >

Libraries >

Support >

Changelog >

Platform Policy >

To continuously improve Instagram users' privacy and security, we are accelerating the deprecation of Instagram API Platform, making the following changes effective immediately. We understand that this may affect your business or services, and we appreciate your support in keeping our platform secure.

These **capabilities** will be disabled immediately (previously set for July 31, 2018 or December 11, 2018 deprecation). The following will be deprecated according to the timeline we **shared previously**:

- Public Content - all remaining capabilities to read public media on a user's behalf on December 11, 2018
- Basic - to read a user's own profile info and media in early 2020

For your reference, information on the **new Instagram Graph API**.

Hello Developers.

The Instagram API Platform can be used to build non-automated, authentic, high-quality apps and services that:

- Help **individuals share their own**



Figura. APIs para extração de dados.



2 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

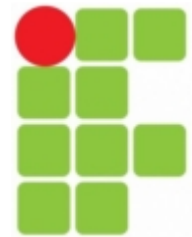
Depois de se familiarizar com os conceitos básicos, passe para tópicos mais avançados como estes:

[Saiba como nossos SDKs interagem com a Graph API lendo a documentação do SDK para iOS, Android](#)

Figura. APIs para extração de dados.

SACI - 2019

23



2. 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!**

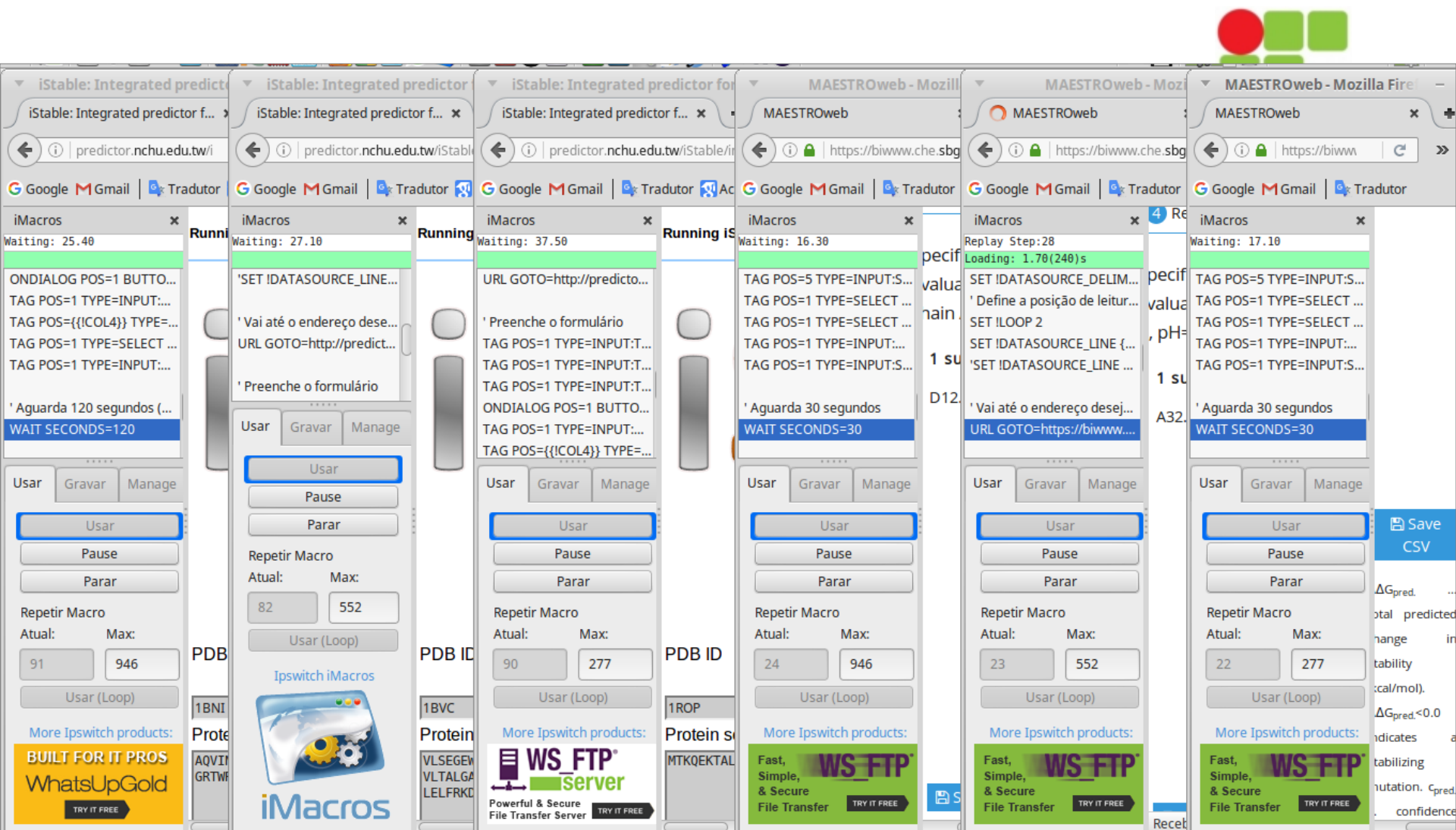


Figura. Muitas requisições "simultâneas".

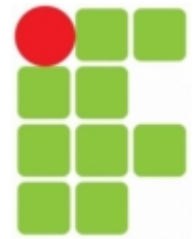


robots.txt

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

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

Figura. Exemplo de um robots.txt.



2. Web Scraping

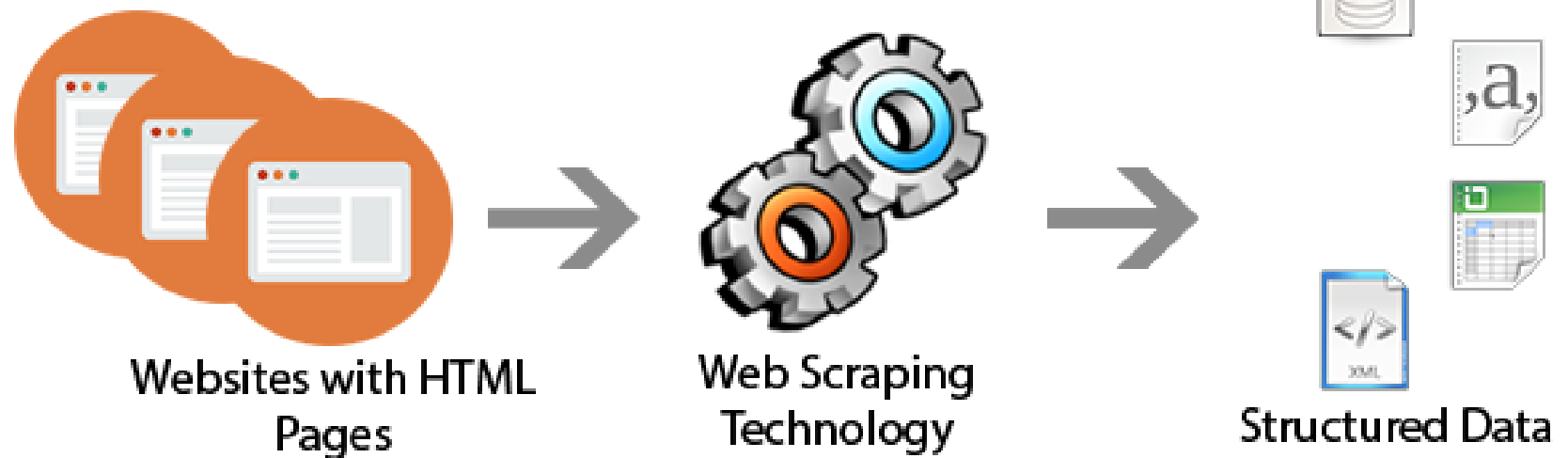


Figura. Visão geral de um *web scraping*.

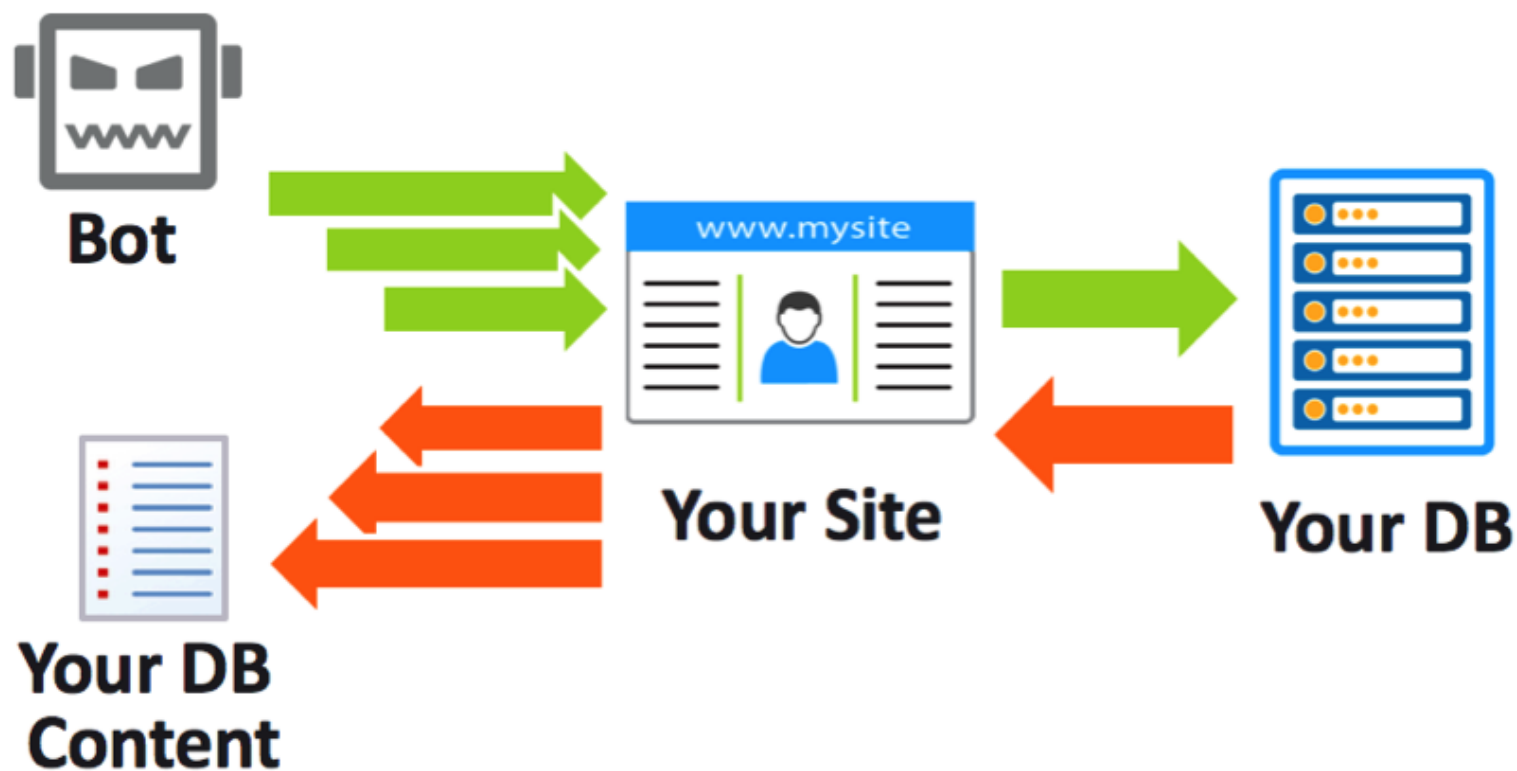
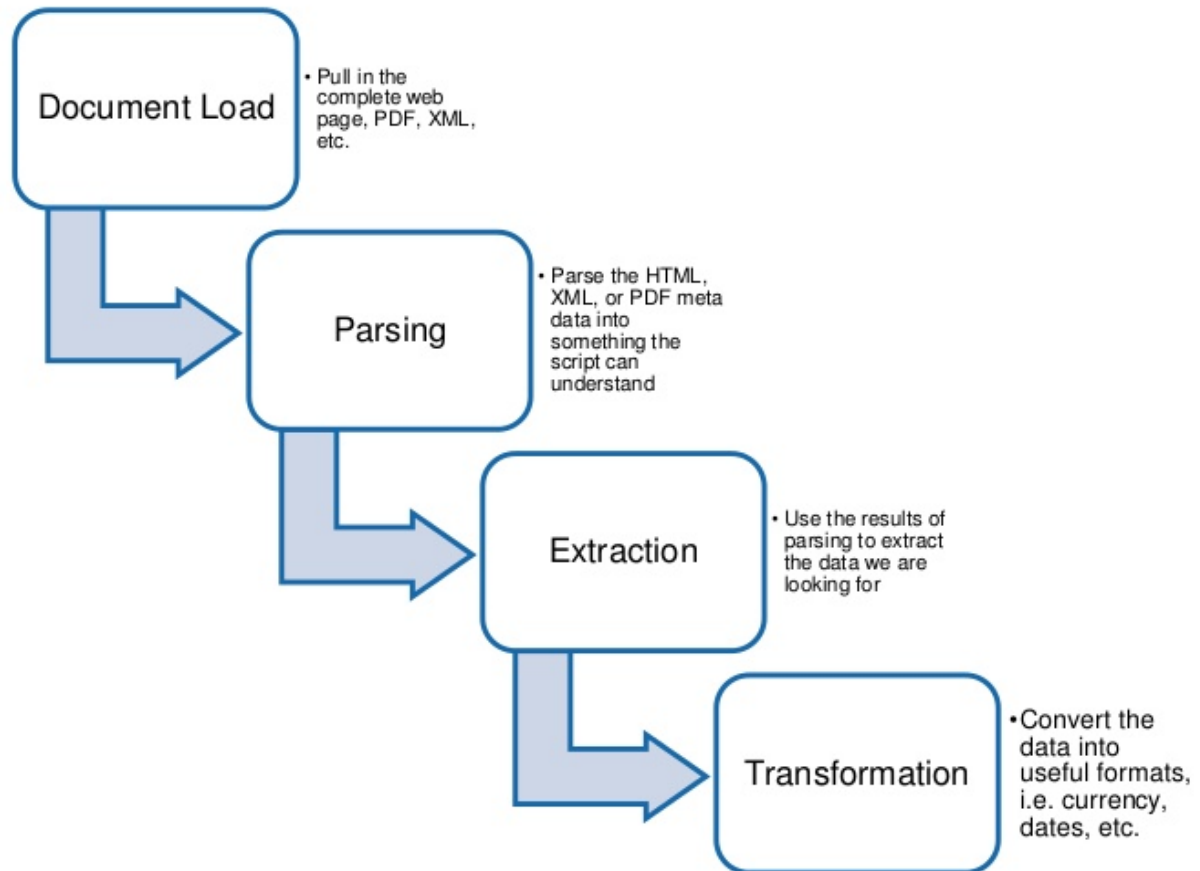


Figura. Visão geral de um *web scraping*.

Anatomy of a Scraper

2.



© 2014 Tommy Tavenner

Figura. Visão geral de um *web scraping*.

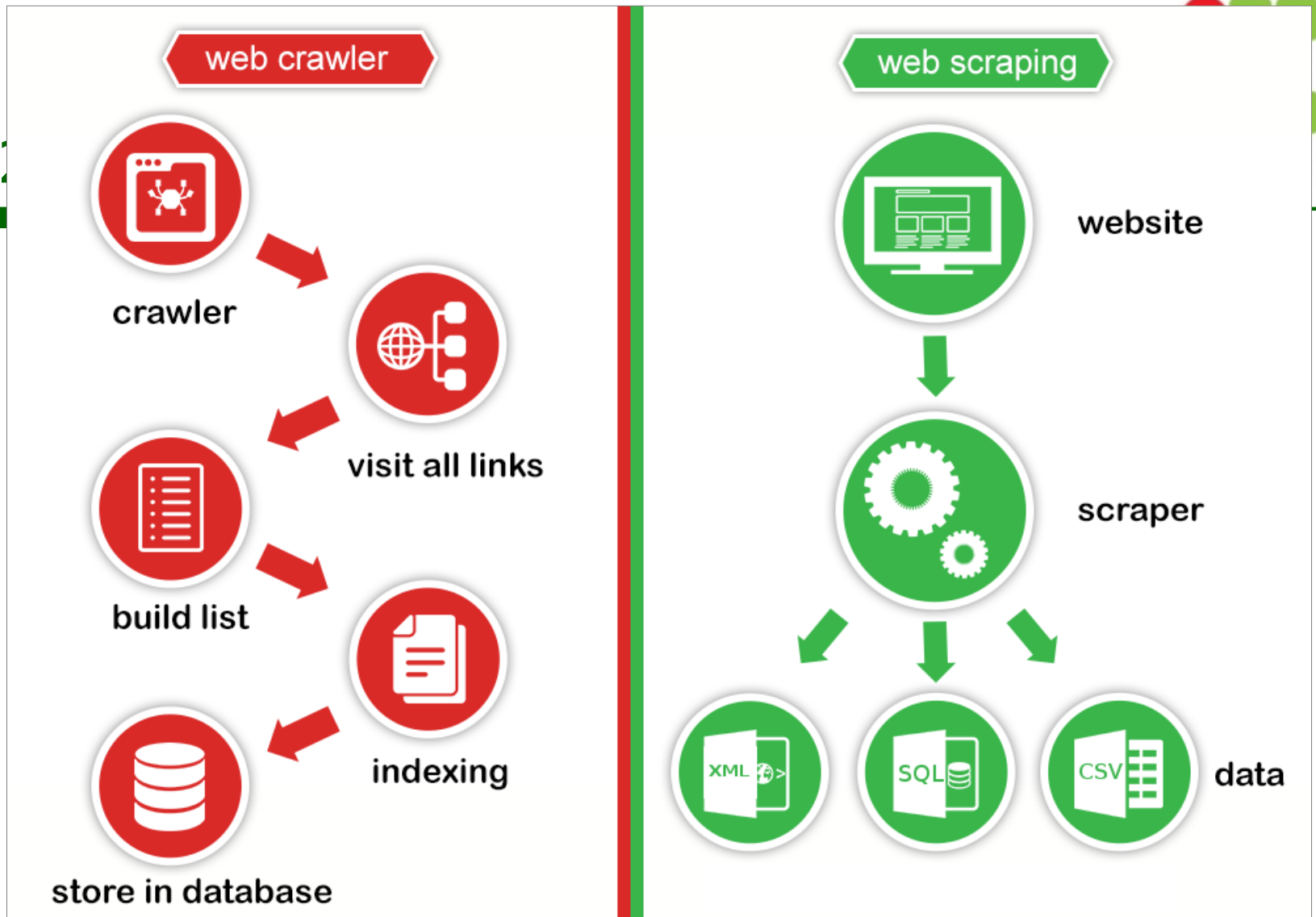
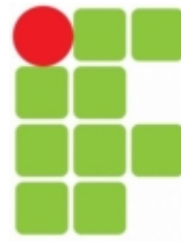
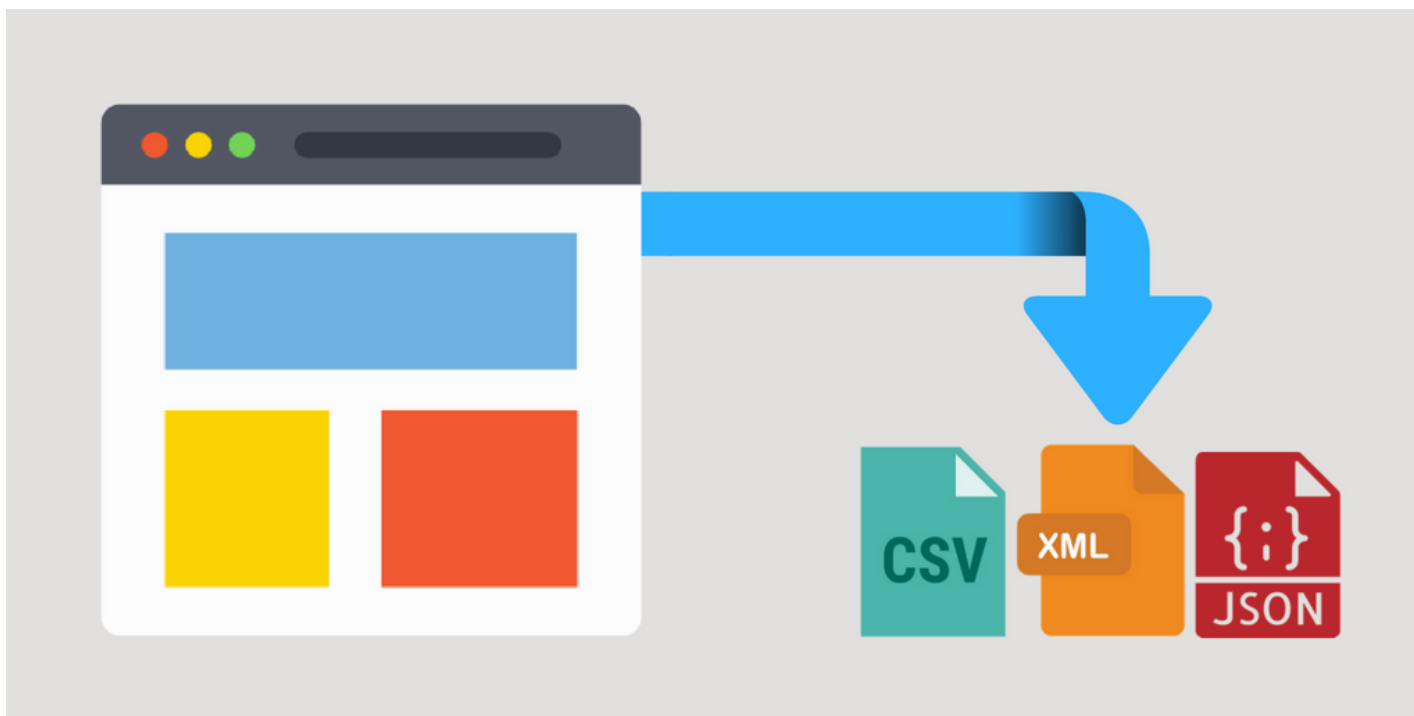
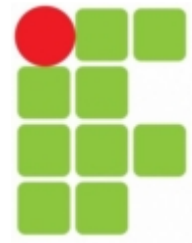


Figura. *Web Scraping versus Web Crawling.*



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/>



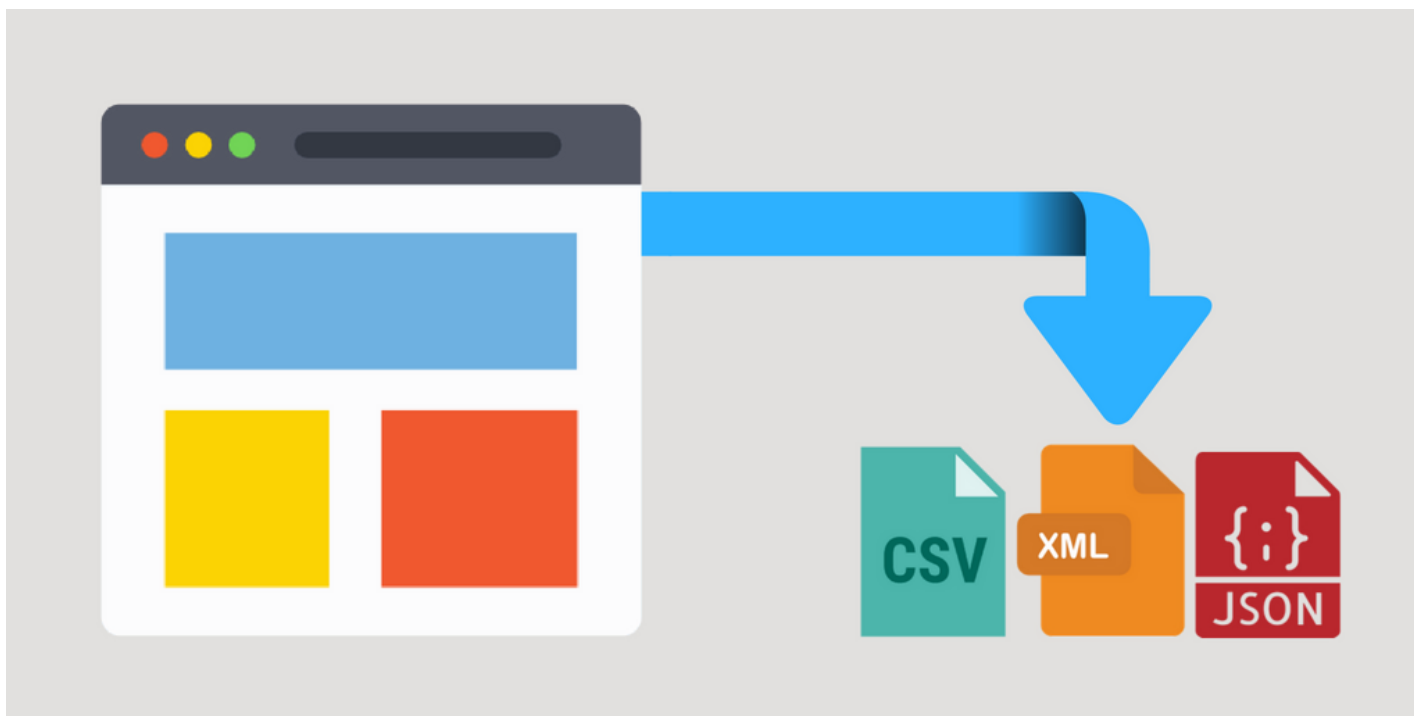
3. Ferramentas

KIT DO WEB SCRAPER:

- ***Selenium:***
<https://pypi.org/project/selenium/>
- ***Firefox geckodriver:***
<https://github.com/mozilla/geckodriver/releases/>



4. *Web scraping* na prática





4.1 *Scraping* de uma página da *Wikipédia*



A necessidade e importância de **extrair dados da Web** está se tornando cada vez mais importante. Cada vez mais me encontro em uma situação em que precisamos extrair dados de algum site.

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

4.2 *Web Scraping* para montar *datasets* de esportes



Upgrade



DATA HACKERS

ÚLTIMOS POSTS

ENGENHARIA

CIÊNCIA

ML

PODCAST

ESCREVA NO DH

CONHEÇA O DATA HACKERS!

Como fazer Web Scraping em Python

Um tutorial sobre Web Scraping para montar datasets de esportes para seus projetos pessoais



Luis Felipe Bueno

Follow

Mar 13 · 7 min read

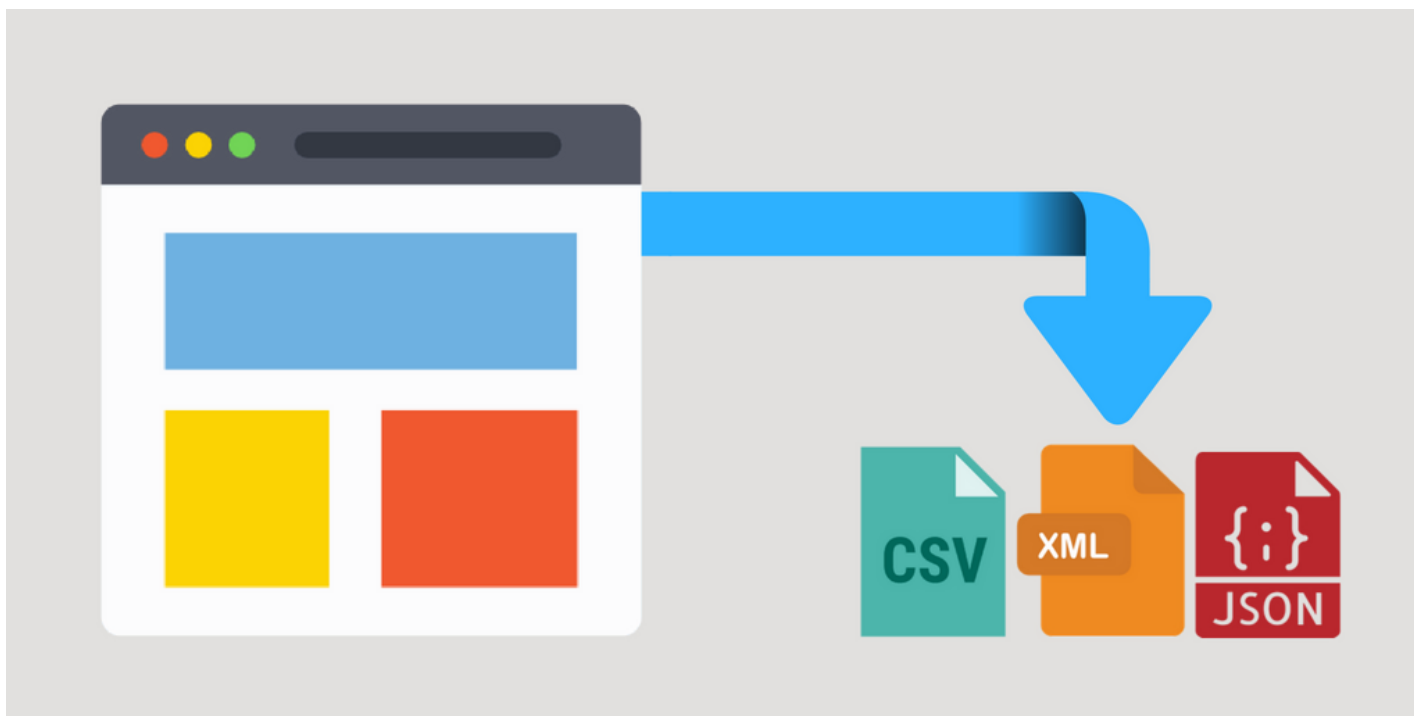


Link.

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



4. *Web scraping* na prática

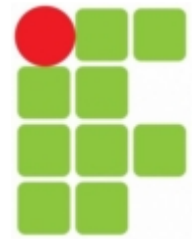




5. Exercícios

Tendo como ponto de partida o endereço: <http://www.ifsul.edu.br/>

1. Mostre o texto que exibe o CMS utilizado no desenvolvimento do portal.
2. Exiba as últimas notícias mostradas na página inicial (sem as tags HTML).
3. Faça uma extração da lista de cursos disponíveis no câmpus Bagé.
4. Armezene o conteúdo anterior no formato "txt" e "csv".
5. Salve uma captura da tela principal do portal com nome de "ifsul_website.png"



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>

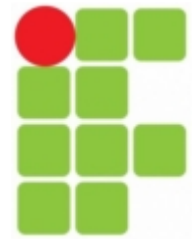


Agradecimentos

Obrigado pela sua participação! :)

- ❑ **Acadêmicos do Curso Técnico Integrado em Informática**
- ❑ **Comissão organizadora da Semana Acadêmica (SACI)**
- ❑ **Toda honra e glória ao Senhor Jesus!**

Abraços, Prof. Alex Dias Camargo
IFSul – Câmpus Bagé
25 de Setembro de 2019



Referências básicas

KOSCIANSKI, André; SOARES, Michel dos Santos. **Qualidade de software**. 2. ed. São Paulo: Novatec, 2007.

MOLINARI, Leonardo. **Testes de software - Produzindo sistemas melhores e mais confiáveis**. São Paulo: Érica, 2008.

RIOS, Emerson MOREIRA; MOREIRA, Trayahú. **Teste de software**. 3. ed. Rio de Janeiro: Alta Books, 2013.