

The logo for 'globo play' is displayed in a bold, red, lowercase sans-serif font. The word 'globo' is followed by a space and then 'play'. The entire logo is centered within a dark gray rectangular box.

globo play

Web Scraping

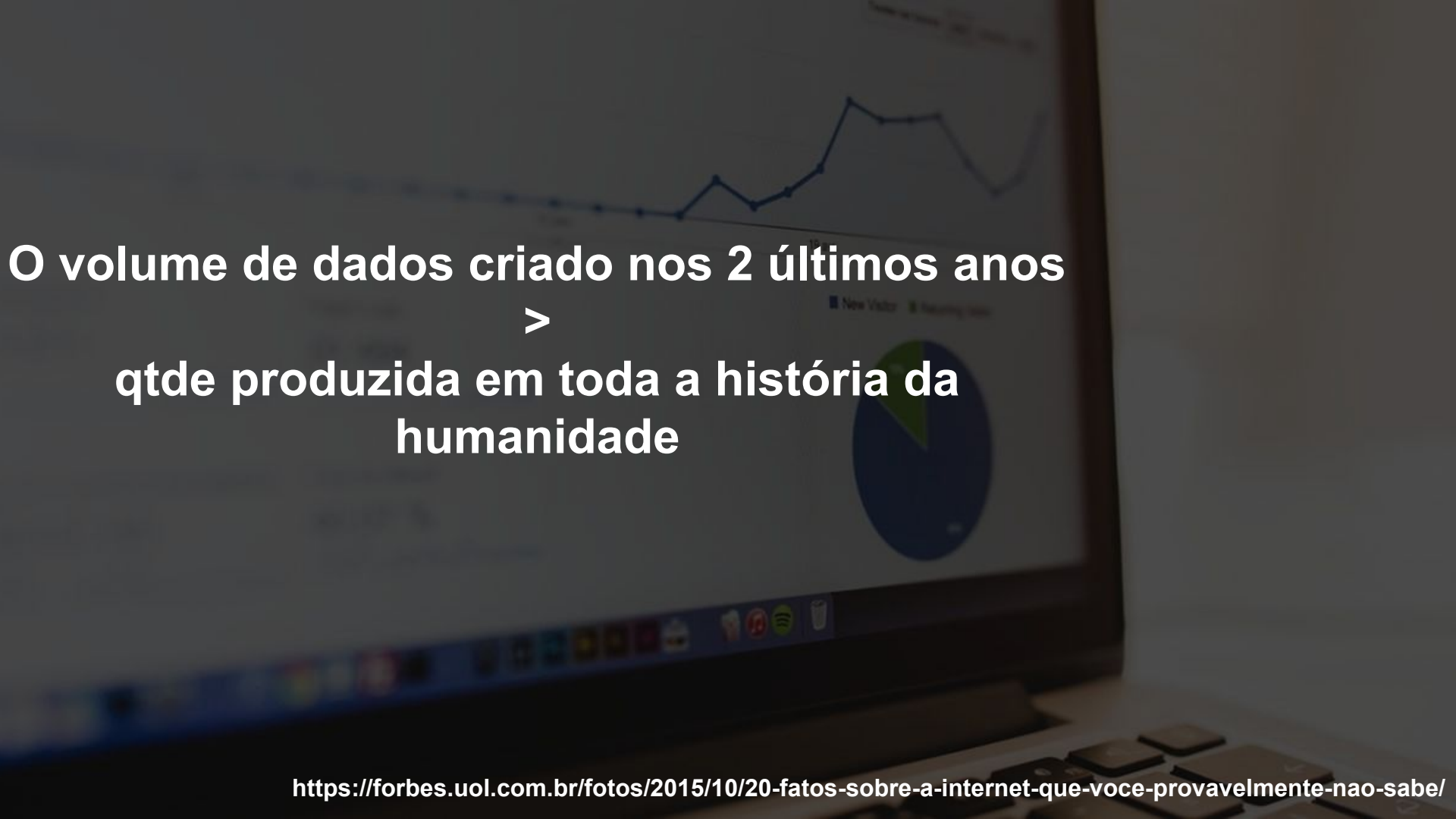


Priscilla Lusie

Growth Hacking Developer

www.linkedin.com/in/priscilla-lusie



The background image shows a laptop screen with a dark overlay. On the screen, there is a line graph with a blue line showing an upward trend and a pie chart with a green slice. The text is overlaid in white.

**O volume de dados criado nos 2 últimos anos
>
qtde produzida em toda a história da
humanidade**

Leis

Lei da transparência

(lei complementar 131/2009)

Toda entidade pública (da União, dos Estados, do Distrito Federal e dos Municípios) deve divulgar na internet em tempo real (prazo máximo de 24h) sua receita e despesas.

Lei do acesso à informação

(LAI 12.527/2011)

Direito constitucional de acesso às informações públicas para qualquer pessoa física ou jurídica sem necessidade de exposição de motivo.

.....

Redes Sociais no Brasil

79%
dos
internautas
brasileiros
possuem
redes sociais

Impacto nas
eleições de
forma
mundial

2017 - 57,8%
dos brasileiros
na web,
segundo IBGE

5:25 h/dia
conectado à
internet
em média

3:47 h/dia
em
Redes
Sociais

A word cloud of various file formats on a dark background. The words are in different colors and sizes. 'PDF' is the largest word in the center, in red. 'XLS' is at the top left in orange. 'XML' and 'PNG' are to its right in red. 'API' is below 'XLS' in orange. 'HTML' is to the left of 'API' in pink. 'ZIP' is in the middle, rotated, in purple. 'CSV' is to the right of 'ZIP' in purple. 'JSON' is at the bottom right in orange.

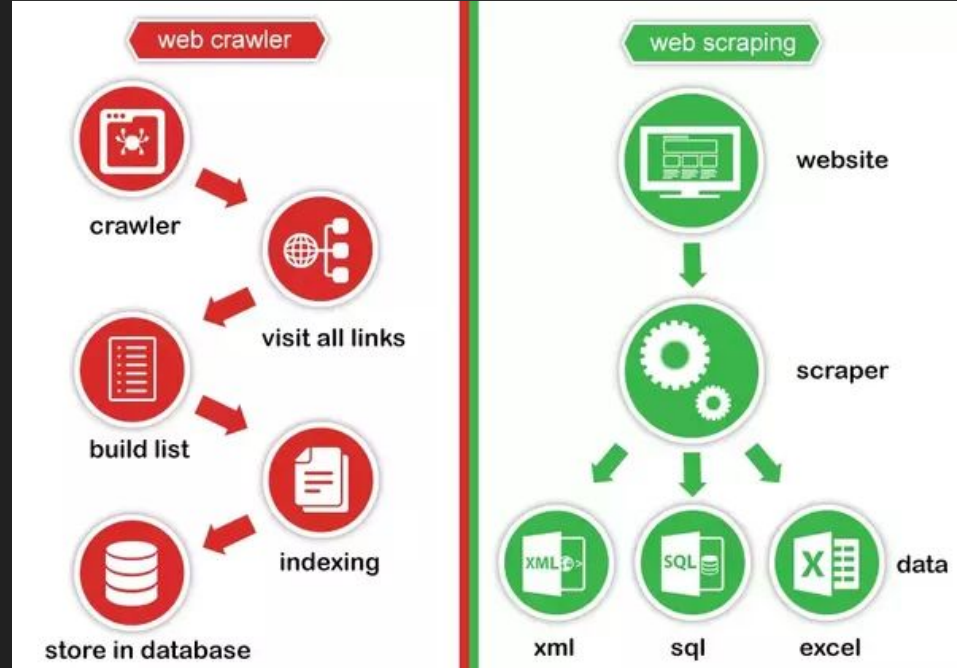
XLS
XML PNG
PDF
API ZIP CSV
HTML JSON

Não há definição de formato

Web Crawler

X

Web Scraping



Visão Geral



Exercício 1

*Buscar as propostas de campanhas
dos candidatos*

Exercício 1

***Buscar as propostas de campanhas
dos candidatos***

R.

<http://www.tse.jus.br/eleicoes/eleicoes-2018/propostas-de-candidatos>

Buscas Manuais

- Monitora mudanças em páginas:
 - https://visualping.io/cdlp.html?utm_source=cd&utm_medium=redirtocdlp&utm_campaign=mig1

START FREE MONITORING!

How it works



We take a
screenshot



We wait 1 day
or 5 min



We take another
screenshot



We compare
them



We notify you if
different

- Monitora mudanças em páginas:
 - <https://versionista.com/>

Monitor Website Changes.

Track Web Page Edits.
Change Detection & Alerts For Entire Sites.



Web change monitor for teams

- Hundreds of brands trust us to watch website changes
- Monitor changes for entire sites, not just a few pages
- Auto-crawl and monitor changes to key web pages
- Track changes and get notified by email & Slack
- Competitive insights, SEO, compliance, and more



Cloud-based web change detection

- Monitor changes to HTML, PDFs, dynamic content
- Simple, powerful SaaS solution
- Detailed email summaries and instant change alerts
- Discover, crawl, and monitor changes to websites
- Powerful filters to avoid irrelevant content

Get Started – See Our Change Tracking in Action



- Monitora mudanças em páginas:
 - <http://archive.org/web/> (possui histórico de várias URLs)



- Redes Sociais:
 - <https://tweetdeck.twitter.com/> (buscas no twitter)
 - <https://findmyfbid.com/> (buscar facebook ID)
 - <https://stalkscan.com/> (buscar facebook ID)
 - <https://inteltechniques.com/menu.html> (cruzar perfis para ver se possuem conexão ... encontrar laranjas)

- Buscas por imagens:
 - <https://www.google.com.br/imghp?hl=pt-PT>
 - <https://yandex.com/images/>
 - <https://www.tineye.com/> (mostra a primeira vez que uma imagem apareceu na rede)
 - <http://suncalc.net/#/51.508,-0.125,2/2018.03.08/16:51> (mostra a sombra em determinado local e horário)
- Outros:
 - <http://www.lemonde.fr/verification/> (alertas sobre informações falsas e artigos enganosos a partir da análise de URLs)
 - <https://firstdraftnews.org/> (coalisão de empresas que lutam contra notícias falsas)

Formatos

XML

Extensible Markup Language

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<receita nome="pão" tempo_de_preparo="5 minutos" tempo_de_cozimento="1 hora">
  <titulo>Pão simples</titulo>
  <ingredientes>
    <ingrediente qtde="3" unidade="xícaras">Farinha de Trigo</ingrediente>
    <ingrediente qtde="7" unidade="gramas">Fermento</ingrediente>
    <ingrediente qtde="1.5" unidade="xícaras" estado="morna">Água</ingrediente>
    <ingrediente qtde="1" unidade="colheres de chá">Sal</ingrediente>
  </ingredientes>
  <instrucoes>
    <passo>Misture todos os ingredientes, e dissolva bem.</passo>
    <passo>Cubra com um pano e deixe por uma hora em um local morno.</passo>
    <passo>Misture novamente, coloque numa bandeja e asse num forno.</passo>
  </instrucoes>
</receita>
```

779 car.

Fonte: <https://pt.wikipedia.org/wiki/XML>

Validador: https://www.w3schools.com/xml/xml_validator.asp

JSON

JavaScript

Object

Notation

```
{  
  "receita": {  
    "nome": "pão",  
    "tempo_de_preparo": "5 minutos",  
    "tempo_de_cozimento": "1 hora",  
    "titulo": "Pão simples",  
    "ingredientes": [  
      {"qtde": "3", "unidade": "xícaras", "nome": "Farinha de Trigo"},  
      {"qtde": "7", "unidade": "gramas", "nome": "Fermento"},  
      {"qtde": "1.5", "unidade": "xícaras", "estado": "morna", "nome": "Água"},  
      {"qtde": "1", "unidade": "colheres de chá", "nome": "Sal"}  
    ],  
    "instrucoes": [  
      {"passo": "Misture todos os ingredientes, e dissolva bem."},  
      {"passo": "Cubra com um pano e deixe por uma hora em um local morno."},  
      {"passo": "Misture novamente, coloque numa bandeja e asse num forno."}  
    ]  
  }  
}
```

Validador: <https://jsonlint.com/>

653 car.

HTML Hypertext Markup Language

```
<h1 style="color: #5e9ca0;">Receita</h1><h2 style="color: #2e6c80;">&nbsp;</h2><table  
class="editorDemoTable"><thead><tr><td>nome</td><td>tempo de preparo</td><td>tempo de  
cozimento</td><td>titulo</td><td>ingredientes</td><td>instrucoes</td></tr></thead><tbody><tr>  
<td>P&atilde;o</td><td>5 minutos</td><td>&nbsp;<td>1 hora</td><td>P&atilde;o  
simples</td><td>&nbsp;</td></tr></tbody></table><table><thead><tr><td>qtde</td><td>unidade</td><td>nome</td><td>est  
ado</td></tr></thead><tbody><tr><td>3</td><td>xcaras</td><td>Farinha de  
Trigo</td><td>&nbsp;</td></tr><tr><td>7</td><td>gramas</td><td>Fermento</td><td>&nbsp;</td>  
></tr><tr><td>1.5</td><td>xcaras</td><td>&Aacute;gua</td><td>morna&nbsp;</td></tr><tr><td>  
1</td><td>colheres de  
ch&aacute;ca</td><td>Sal</td><td>&nbsp;</td></tr></tbody></table></td><td>&nbsp;<td><table><thead>  
><tr><td>passo</td></tr></thead><tbody><tr><td>Misture todos os ingredientes e dissolva  
bem</td></tr><tr><td>Cubra com um pano e deixe por uma hora em um local  
morno.</td></tr><tr><td>Misture novamente, coloque numa bandeja e asse num  
forno.</td></tr></tbody></table></td></tr></tbody></table>
```

Editor: <https://html-online.com/editor/>

1088 car.

API

Application Programming Interface

```
curl -XGET www.google.com
```

```
curl -H "Content-Type: application/json" -XPOST www.api.com -d '{ "nome": "Priscilla", "empresa": "Globo.com" }'
```

```
curl -H "Content-Type: application/json" -XPUT www.api.com/Priscilla -d '{ "nome": "Priscilla Lusie" }'
```

```
curl -XDELETE www.api.com/Priscilla
```

GET

POST

PUT

DELETE

Códigos de retorno:

- **200** = Success.
- **301** = Moved Permanently.
- **400** = Bad Request. Algo errado na página ou nas configurações do servidor.
- **401** = Unauthorized
- **403** = Forbidden
- **404** = Not Found. Página não encontrada.
- **500** = Internal Server Error.

Exercício 2

*Procurar mensagens do Globoplay
sobre "Assédio" no Twitter*

Exercício 2

*Procurar mensagens do Globoplay
sobre "Assédio" no Twitter*

R. <https://twitter.com/search?q=globoplay%20assedio>

Exercício 3

*Procurar mensagens do Globoplay
sobre "Assédio" no Twitter via API*

Exercício 3

*Procurar mensagens do Globoplay
sobre "Assédio" no Twitter via API*

R. <https://developer.twitter.com/>



<https://www.getpostman.com/>

Autenticação

Twitter dev

Consumer API keys

XIIgfWgJX8ubCE7R5XxayaGN9 (API key)

Blh5wYEC8EWcBfdrP9qtKJOF4zg04hVOfxAt4uZ7AVINe
JJnt (API secret key)

Access token & access token secret

2678203501-jZFsJjk0LO7dywVZRLcEZgC3jzZCn3XogPSP
SPa (Access token)

MiAwMgECsMBNKhgecFVAxaQcK7wDxJdbroMgxEFzwim
pY (Access token secret)



GET <https://api.twitter.com/1.1/search/tweets.json?q=globoplay%20assedio>



Python

- Linguagem de Alto Nível
- Orientada a objetos
- Web
- Tipagem dinâmica
- Facilita a leitura do código
- Está entre as 5 linguagens mais populares, de acordo com uma pesquisa conduzida pela RedMonk

<https://pt.wikipedia.org/wiki/Python>

<https://www.anaconda.com/download/>

<https://anaconda.org/account/register>

Exercício 4

*Procurar mensagens do Globoplay
sobre "Assédio" no Twitter via API
usando Python*

Estudos Futuros

Selenium: <https://www.seleniumhq.org/>

Twitter: <https://developer.twitter.com/en/docs.html>