Web Scraping

1. Visão Bem Simplificada das Camadas de Um Sistema

Banco de Dados

- Oracle
- MySQL
- PostgreSQL
- ...

Linguagem de Programação

- Python
- PHP
- Java
- ...

Codificação da interface

- HTML
- CSS
- JavaScript
- Angular
- ...

Interface do Usuário:

- Chrome
- Opera
- Firefox
- ...

2. O que recebemos de um servidor quando acessamos um site?

Sempre que acessamos um site, o navegador faz uma solicitação chamada de solicitação GET, que recebe arquivos (HTML, CSS, JS, imagens, ...) do servidor, e renderizar a referida página. Lembrando que em HTML as tags são aninhadas e podem aparecer dentro de outras tags.

De acordo com a posição em relação à outra tag, as mesmas podem ser classificadas como:

- child (filha): uma tag dentro de outra tag.
- parent (pai): uma tag que tem outras tags dentro.
- sibling (irmão): uma tag que está aninhada dentro do mesmo pai que outra tag.

Exemplo:

```
<html><!--tag pai-->
  <head><!--tag filha-->
  </head>
  <body>
    <p><!--tag irmã-->
      texto
    </p>
    <p><!--tag irmã-->
      texto
    </p>
  </body>
</html>
```

As tags mais comuns são as seguintes:

- a: links que redirecionam para uma outra página.
- div: divide uma área na página.
- table: cria uma tabela.
- form: cria um formulário.

2.1. Saiba mais em:

@Curso em Vídeo de HTML5 e CSS3
https://youtu.be/Ejkb_YpuHWs

3. E se eu quiser obter os dados de um determinado site?

Os dados de um site não podem ser baixados como um formato facilmente legível por máquina, como JSON, HTML ou XML. Vamos analisar um site de ações como exemplo:

https://twitter.com

3.1. Posso usar o Python?

O Python é uma linguagem de programação que permite que você trabalhe rapidamente e integre sistemas de forma mais eficaz, e tem muitas bibliotecas para leitura e escrita de dados no HTML.

3.2. Onde escrever o código?

- **Anaconda Navigator**: é uma interface gráfica de usuário (GUI) de desktop incluída na distribuição Anaconda® que permite iniciar aplicativos e gerenciar facilmente pacotes, ambientes e canais conda sem usar comandos de linha de comando. O Navigator pode pesquisar pacotes no Anaconda.org ou em um repositório local do Anaconda. Está disponível para Windows, macOS e Linux. Se quiser mais detalhes sobre o Anaconda, pode consultar:

  https://docs.anaconda.com/anaconda/navigator/

- **Jupyter Notebook**: a interface clássica do notebook. Um aplicativo web original para criar e compartilhar documentos computacionais. Ele oferece uma experiência simples, simplificada e centrada em documentos. Se quiser mais detalhes sobre o Jupyter, pode consultar:

  https://jupyter.org/

- **Outras possibilidades:**
  - **Try Jupyter**: https://jupyter.org/try
  - **Colab**: https://colab.research.google.com/
  - **PyCharm**: https://www.jetbrains.com/pt-br/pycharm/

3.3. Posso usar códigos prontos?

Os programas do mundo real são complexos. Então, visando simplificar o processo de desenvolvimento e torná-lo mais eficaz, os desenvolvedores aproveitam a programação modular, reutilizando códigos pré-existentes. Podemos começar destacando estas:

- **Requests:** uma biblioteca HTTP para Python simples e elegante, para baixar a página. Assim, não teremos a necessidade para adicionar query string nas suas URLs manualmente, ou de codificar seus dados de formulário POST. Se quiser mais detalhes sobre essa biblioteca, pode consultar:

  https://requests.readthedocs.io/pt_BR/latest/index.html

- **Beautiful Soup:** uma biblioteca que trabalha com seu analisador favorito para fornecer maneiras idiomáticas de navegar, pesquisar e modificar a árvore de análise. Geralmente, economiza horas ou dias de trabalho para os programadores. Se quiser mais detalhes sobre essa biblioteca, pode consultar:

  https://www.crummy.com/software/BeautifulSoup/bs4/doc/

- **urllib3**: emitirá vários avisos diferentes com base no nível de suporte à verificação de certificado. Esses avisos indicam situações particulares e podem ser resolvidos de diferentes maneiras. Se quiser mais detalhes sobre essa biblioteca, pode consultar:

  https://urllib3.readthedocs.io/en/stable/

- **pandas**: é uma ferramenta de análise e manipulação de dados de código aberto rápida, poderosa, flexível e fácil de usar, construída sobre a linguagem de programação Python.
  https://pandas.pydata.org/

3.4.E se o módulo não estiver instalado?

Podemos utilizar o pip, que é um instalador de pacotes para Python. Você pode usar pip para instalar pacotes do Python Package Index e outros índices.

3.5.Já ouviu falar sobre API (Application Programming Interface)?

APIs são mecanismos que permitem que dois componentes de software se comuniquem usando um conjunto de definições e protocolos. Por exemplo, o sistema de software do instituto meteorológico contém dados meteorológicos diários. O aplicativo meteorológico em seu telefone “fala” com este sistema por meio de APIs e mostra atualizações meteorológicas diárias no telefone.

Fonte: https://aws.amazon.com/pt/what-is/api/

3.6.Posso acessar qualquer dado de sites da internet?

O que é Web Scraping? É essencialmente extrair e reunir conjuntos de dados da web (o que pode ser considerado Big Data em alguns casos), dados esses que são a pedra angular do Big Data Analytics, Machine Learning e Inteligência Artificial.

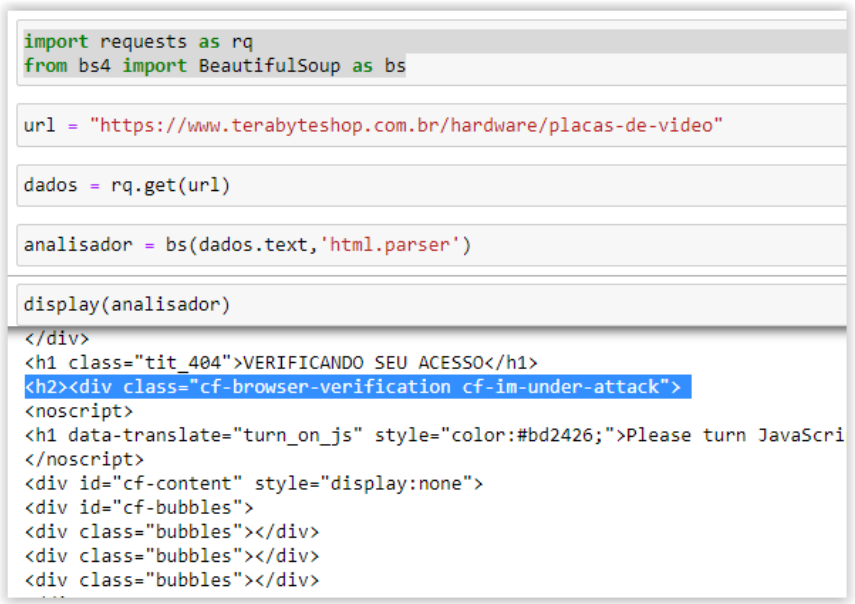
A reputação do web scraping ficou muito pior nos últimos anos e por boas razões.

É cada vez mais usado para fins comerciais para obter uma vantagem competitiva e normalmente há um motivo financeiro por trás disso. Muitas vezes, é feito com total desconsideração das leis de direitos autorais e dos Termos de Serviço.

Fonte: https://blog.dsacademy.com.br/web-scraping-e-web-crawling-sao-legais-ou-ilegais/

3.6.1. Vamos analisar um exemplo

Imagine que você queira obter os dados dos produtos à venda em uma loja de componentes de computador.



```
import requests as rq
from bs4 import BeautifulSoup as bs

url = "https://www.terabyteshop.com.br/hardware/placas-de-video"

dados = rq.get(url)

analisador = bs(dados.text,'html.parser')

display(analisador)
```

```
</div>
<h1 class="tit_404">VERIFICANDO SEU ACESSO</h1>
<h2><div class="cf-browser-verification cf-im-under-attack">
<noscript>
<h1 data-translate="turn_on_js" style="color:#bd2426;">Please turn JavaScri
</noscript>
<div id="cf-content" style="display:none">
<div id="cf-bubbles">
<div class="bubbles"></div>
<div class="bubbles"></div>
<div class="bubbles"></div>
```

O Cloudflare Under Attack Mode executa verificações de segurança adicionais para ajudar a mitigar os ataques DDoS de camada 7. Os usuários validados acessam seu site e o tráfego suspeito é bloqueado. Ele foi projetado para ser usado como um dos últimos recursos quando uma zona está sob ataque (e pausará temporariamente o acesso ao seu site e afetará a análise do site).

Fonte: https://support.cloudflare.com/hc/en-us/articles/200170076-Understanding-Cloudflare-Under-Attack-mode-advanced-DDOS-protection-

3.7. Um pouco sobre segurança antes de começar

OAuth 2.0: é o protocolo padrão do setor para autorização. OAuth 2.0 se concentra na simplicidade do desenvolvedor do cliente, ao mesmo tempo em que fornece fluxos de autorização específicos para aplicativos da Web, aplicativos de desktop, telefones celulares e dispositivos de sala de estar.

Fonte: https://oauth.net/2/

Bearer Token (autenticação do portador): alguns servidores emitirão tokens, que são uma sequência curta de caracteres hexadecimais, enquanto outros podem usar tokens estruturados, como JSON Web Tokens. Em outras palavras, o Bearer Token é um esquema de autenticação HTTP que envolve tokens de segurança chamados tokens de portador, que pode ser entendido como "dar acesso ao portador deste token". O cliente deve enviar este token no cabeçalho Authorization ao fazer solicitações para recursos protegidos.

4. Vamos programar...

IMPORTANTE: vamos analisar os códigos do ponto de vista técnico e não político.

4.1. Usando a API do Google News

Podemos os módulos BeautifulSoup e/ou Selenium para realizar a coleta de dados na Web (web scraping), mas esse processo pode ser muito problemático. Então, podemos adaptar o código para um site específico ou de correr o risco de ser bloqueado por algum sistema de segurança.

Então, podemos usar métodos mais simples de capturar notícias com Python, usando API's (Application Programming Interface) dos próprios fornecedores de dados, como no caso do GoogleNews. Saiba mais em:

https://pypi.org/project/GoogleNews/

4.1.1. O código em python

```
!pip install GoogleNews
```

```
from GoogleNews import GoogleNews
import pandas as pd
```

```
googlenews = GoogleNews()
googlenews.set_lang('pt')
googlenews.set_period('5d')
```

```
googlenews.clear()
```

```
googlenews.search('Bolsonaro')
```

```
googlenews.total_count()
```

```
pesquisa_bolsonaro = googlenews.results()
```

```
df_bolsonaro = pd.DataFrame(pesquisa_bolsonaro)
```

```
display(df_bolsonaro)
```

```
googlenews.clear()
```

```
pesquisa_lula = googlenews.search('Lula')
```

```
googlenews.total_count()
```

```
pesquisa_lula = googlenews.results()
```

```
df_lula = pd.DataFrame(pesquisa_lula)
```

```
display(df_lula)
```

4.1.1. Exemplos de resultados do código

	title	media	date	datetime	desc	link	img
0	Deputado do PT representa contra Michelle Bols...		3 mins ago	NaT	Michelle fez um pronunciamento de Dia das Mães...	https://www.brasil247.com/brasil/deputado-repr...	...
1	Bolsonaro se une à União Brasil e ao PSDB nos ...		8 mins ago	NaT	Em busca da reeleição, Jair Bolsonaro (PL) ini...	https://oantagonista.uol.com.br/brasil/bolsona...	...
2	Bolsonaro cita Jessi ao falar de Fies e ex-BBB...		1 hour ago	2022-05-09 14:09:40.075574	Em uma postagem, Jessilane Alves comemorou após...	https://www.brasil247.com/tanostrends/bolsonar...	...
3	Deputado aciona Justiça por pronunciamento de ...		1 hour ago	2022-05-09 14:09:40.083581	O deputado federal Rui Falcão (PT-SP) protocol...	https://oantagonista.uol.com.br/brasil/pt-acio...	...

	title	media	date	datetime	desc	link	img
0	"Lula respeita a candidatura de Ciro"		3 mins ago	NaT	Em meio a rumores de que Ciro Gomes poderia se...	https://oantagonista.uol.com.br/brasil/lula-re...	...
1	Em BH, Lula conversa com representantes de par...		6 mins ago	NaT	Pré-candidato do PT à presidência da República...	https://www.em.com.br/app/noticia/politica/202...	...
2	Até o jingle de Lula é velho		9 mins ago	NaT	Lula reeditou o jingle Lula lá, de 1989. Apesa...	https://oantagonista.uol.com.br/opiniao/ate-o-...	...
3	Papo reto com André Constantine - Lançamento d...		1 hour ago	2022-05-09 14:11:02.905184	Dafne Ashton e André Constantine debatem o lan...	https://www.youtube.com/watch?v=1fRaGUNg4kc	...

5. Vamos programar...

IMPORTANTE: vamos analisar os códigos do ponto de vista técnico e não político.

5.1. Usando snscrape com Twitter

O snscrape é um raspador para serviços de redes sociais (SNS). Ele raspa coisas como perfis de usuários, hashtags ou pesquisas e retorna os itens descobertos, por exemplo, as postagens relevantes. Saiba mais em:

https://github.com/JustAnotherArchivist/snscrape

5.1.1. O código em python

```
!pip install python-twitter-v2
!pip install snscrape
```

```
import snscrape.modules.twitter as dados
import pandas as pd
import datetime
```

```
lista_twittes_bolsonaro = []
```

```
data_final = datetime.date.today()
data_inicial = '2022-1-1'
```

```
for i,tweet in enumerate(dados.TwitterSearchScraper(f'{"Bolsonaro"} + since:{data_inicial} until:{data_final}').get_items()):
    if i>500:
        break
    lista_twittes_bolsonaro.append([tweet.date, tweet.id, tweet.content, tweet.username])
```

```
df_bolsonaro = pd.DataFrame(lista_twittes_bolsonaro, columns=['Datetime', 'Tweet Id', 'Text', 'Username'])
```

```
display(df_bolsonaro)
```

```
lista_twittes_lula = []
```

```
for i,tweet in enumerate(dados.TwitterSearchScraper(f'{"Lula"} + since:{data_inicial} until:{data_final}').get_items()):
    if i>500:
        break
    lista_twittes_lula.append([tweet.date, tweet.id, tweet.content, tweet.username])
```

```
df_lula = pd.DataFrame(lista_twittes_lula, columns=['Datetime', 'Tweet Id', 'Text', 'Username'])
```

```
display(df_lula)
```

5.1.2. Exemplos de resultados do código

```
display(df_bolsonaro)
```

	Datetime	Tweet Id	Text	Username
0	2022-05-04 23:59:57+00:00	1522003128522059777	@qualvim @jairbolsonaro Então, eu pedi desculp...	bheatrizeh
1	2022-05-04 23:59:57+00:00	1522003126521323521	@monark Eu tenho certeza #bolsonaro	TeresinhaARocha
2	2022-05-04 23:59:57+00:00	1522003125275697154	@Jac_Bolsonaro Ridículo	Reinald52174741
3	2022-05-04 23:59:56+00:00	1522003124348755968	@ClaudiaRamosDu @RitaGue27205379 Só espero que...	Robhynson1
4	2022-05-04 23:59:56+00:00	1522003124088885249	@randolfeap Tenho três filhos, todos os três v...	SilvaCa50080377
...	...	...	...	...
496	2022-05-04 23:54:04+00:00	1522001645084811264	EDUARDO BOLSONARO VOLTOU NÃO PRA PL FILHO DA PUTA	thevivica
497	2022-05-04 23:54:03+00:00	1522001642916298754	@AnndersonnRR @jairbolsonaro 👍👍👍	vinirmv22
498	2022-05-04 23:54:03+00:00	1522001641104363520	@CarlaZambelli38 @jairbolsonaro Agora que vcs ...	RationalRightBR
499	2022-05-04 23:54:01+00:00	1522001633231704067	@jairbolsonaro Presidente,eu fui na Paulista n...	Mavergueiro1947
500	2022-05-04 23:54:00+00:00	1522001631038091264	@Jac_Bolsonaro Kkkkk justiça kkkkk	dribighi

501 rows × 4 columns

```
display(df_lula)
```

	Datetime	Tweet Id	Text	Username
0	2022-05-04 23:59:59+00:00	1522003135430017031	@ciaraCnugent @LulaOficial Não temos interesse...	Marcelo_PNeves
1	2022-05-04 23:59:58+00:00	1522003132661837825	@HugoGloss @LulaOficial @photoluisadorr @TIME ...	KyrillosRoselle
2	2022-05-04 23:59:58+00:00	1522003129503563777	@MarkHamill @LulaOficial @cynaramenezes o dia ...	fernanda_OSS
3	2022-05-04 23:59:57+00:00	1522003125074382851	@Paola01242703 @senadorhumberto @LulaOficial O...	sandro_vxs
4	2022-05-04 23:59:56+00:00	1522003122759127046	@USParmera @donairene13 Lula é santista kkkkk	Ale84853364
...	...	...	...	...
496	2022-05-04 23:56:15+00:00	1522002196581265409	@LulaOficial @ricardostuckert Corre que a Estr...	GibaTricolor
497	2022-05-04 23:56:14+00:00	1522002192139444225	@Zuletalleras Si lo dice Lula que es de izquie...	ALVAROLON13
498	2022-05-04 23:56:14+00:00	1522002189803270147	@Flaviamaynarte Em outubro será curado quando ...	Kleberrfb
499	2022-05-04 23:56:14+00:00	1522002189492928512	@senadorhumberto É Lula neles kkkkk	ValdileiCorrei1
500	2022-05-04 23:56:13+00:00	1522002186930114562	@MarkHamill @LulaOficial the real cover my fri...	CS70alter

501 rows × 4 columns

6. Exercício

1) Encontre as publicações do whindersson nunes no twitter
2) Pesquise sobre a API do próprio twitter:
   a) Qual o nome da API?
   b) Quais são as restrições de acesso?
   c) Como utilizar?
3) Pesquise sobre API para realizar web scraping do twitter utilizando a linguagem R