CIÊNCIA DE DADOS (BIG DATA PROCESSING AND ANALYTICS)

RECUPERAÇÃO DA INFORMAÇÃO NA WEB E EM REDES SOCIAIS

Professor curador
Luciano Moreira Camilo e Silva

TRILHA 2
RECUPERAÇÃO DE INFORMAÇÃO POR RASPAGEM – INTRODUÇÃO

Qual é a diferença entre internet e web?

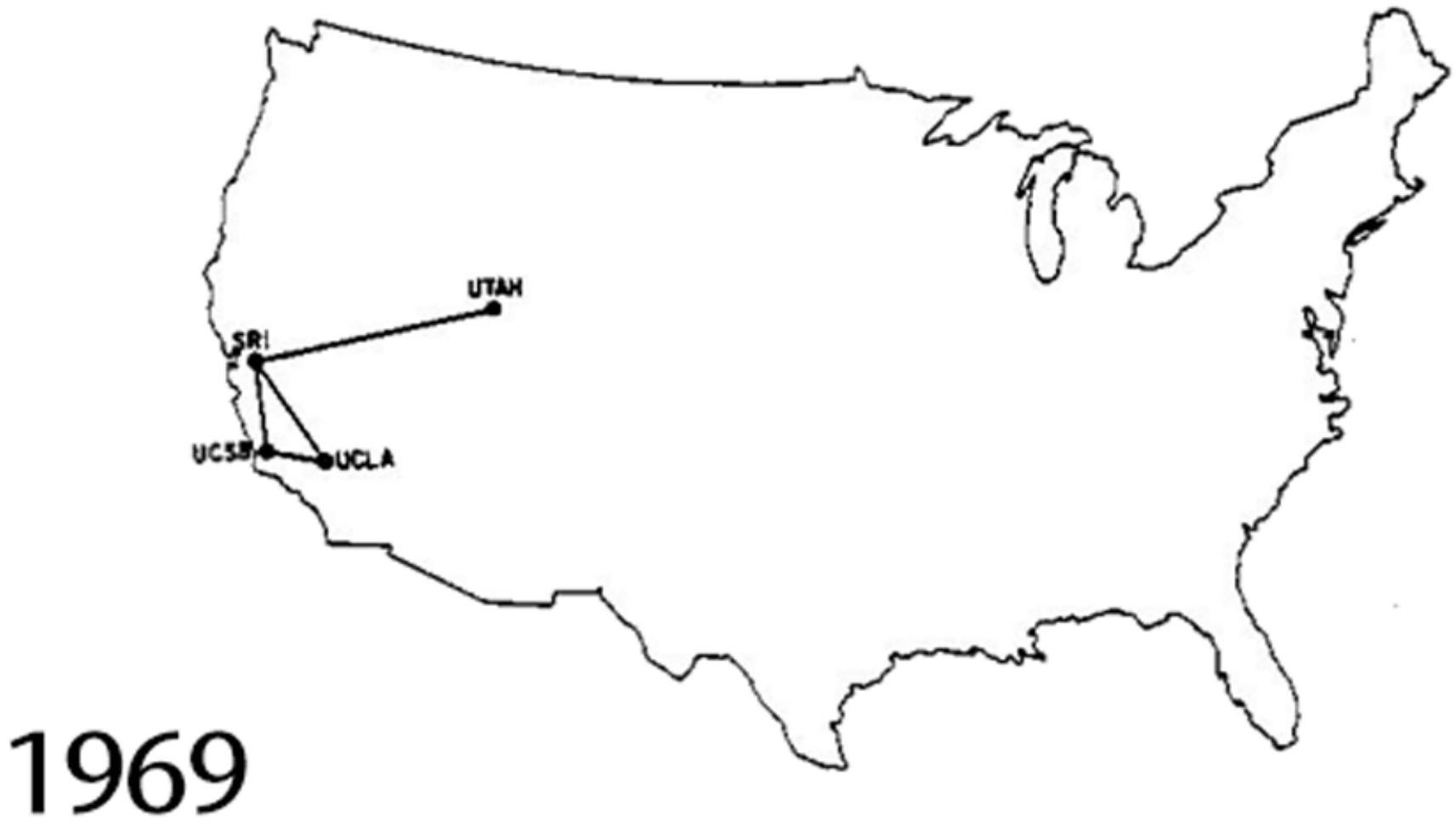
Internet

conexão • rede de rede • tcp/ip

World Wide Web

aplicação • distribuição de conteúdo • http

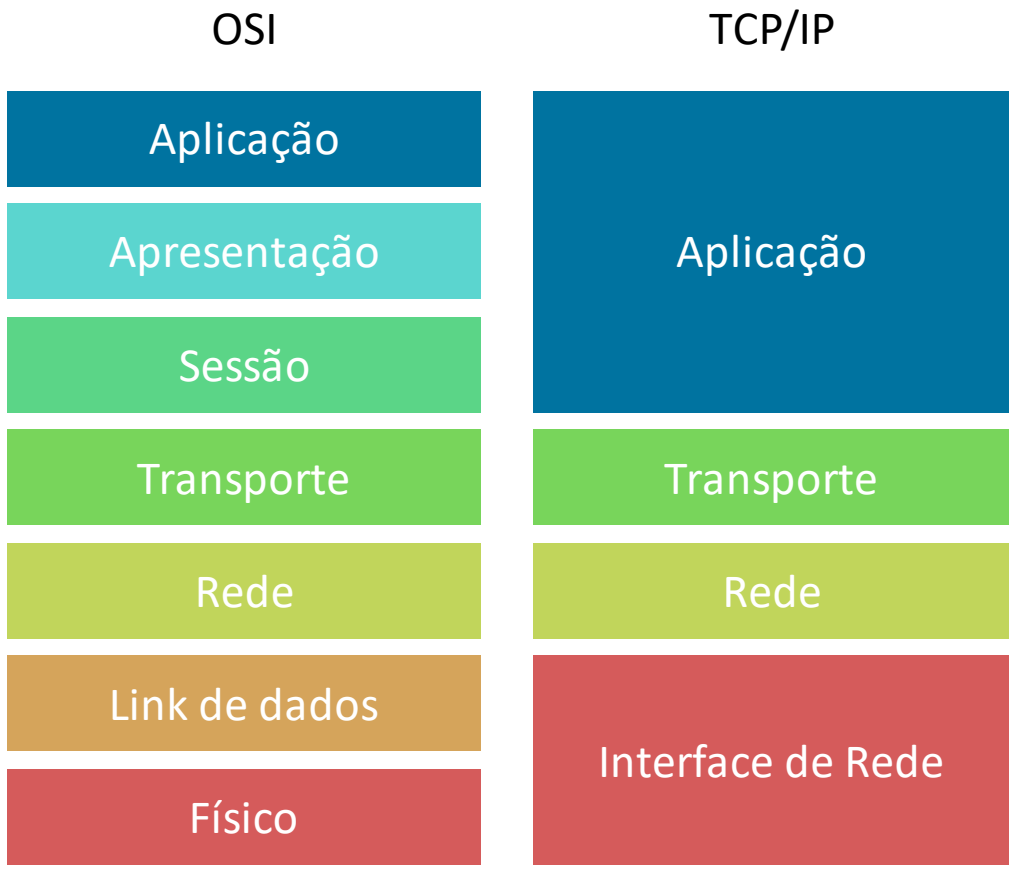
ARPANET



ROSER, Max; RITCHIE, Hannah; ORTIZ-OSPINA, Esteban. Internet. *OurWorldInData.org*, 2015. Disponível em: <https://ourworldindata.org/internet>.

Algumas aplicações famosas da internet

- ***File Transfer Protocol (FTP) e SSH File Transfer Protocol (SFTP)***
- ***Simple Mail Transfer Protocol (SMTP), POP e IMAP***
- ***Network News Transfer Protocol (NNTP)***
- ***Internet Relay Chat (IRC)***
- ***Gopher***
- ***HTTP***



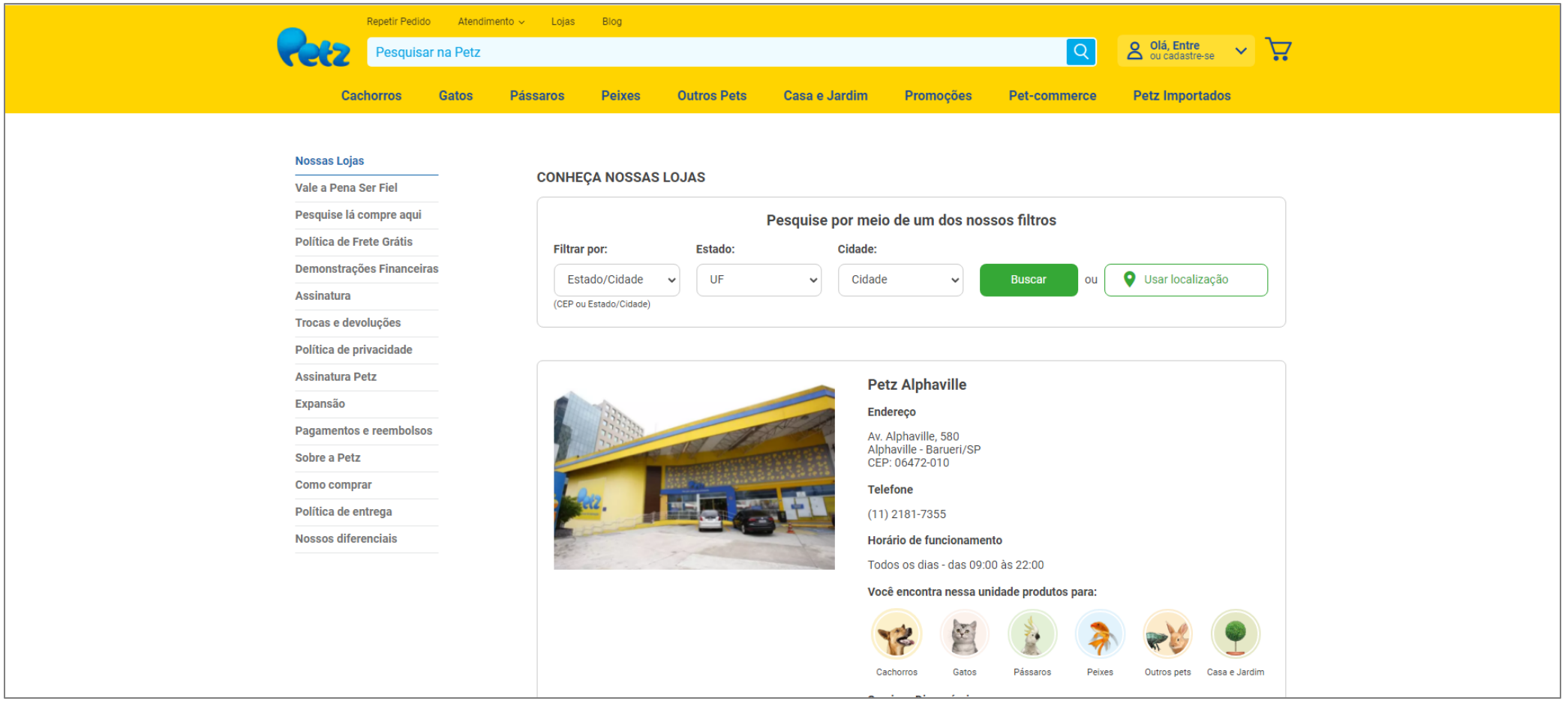
Fonte: Elaborada pelo autor.

A web como nós a conhecemos

HTTP

HTML • CSS • Javascript • Imagens • Vídeos • etc.

Uma página da internet



Fonte: <https://www.petz.com.br/nossas-lojas>.

Uma página da internet – Olá, mundo!

hello.html > html

```
1   <!DOCTYPE html>
2   <html>
3       <head>
4           <title>Exemplo de html</title>
5       </head>
6       <body>
7           <h1>Exemplo de HTML</h1>
8           <h2>O classico Olá mundo</h2>
9           <p>Este paragrafo vai aparecer abaixo do <i>heading</i> nivel 2</p>
10          <h2>Exemplo com lista</h2>
11          <ul>
12              <li> Item 1 </li>
13              <li> Item 2 </li>
14              <li> Item 3 </li>
15          </ul>
16      </body>
17  </html>
```

title:Exemplo de HTML

Exemplo de HTML

O classico Olá mundo

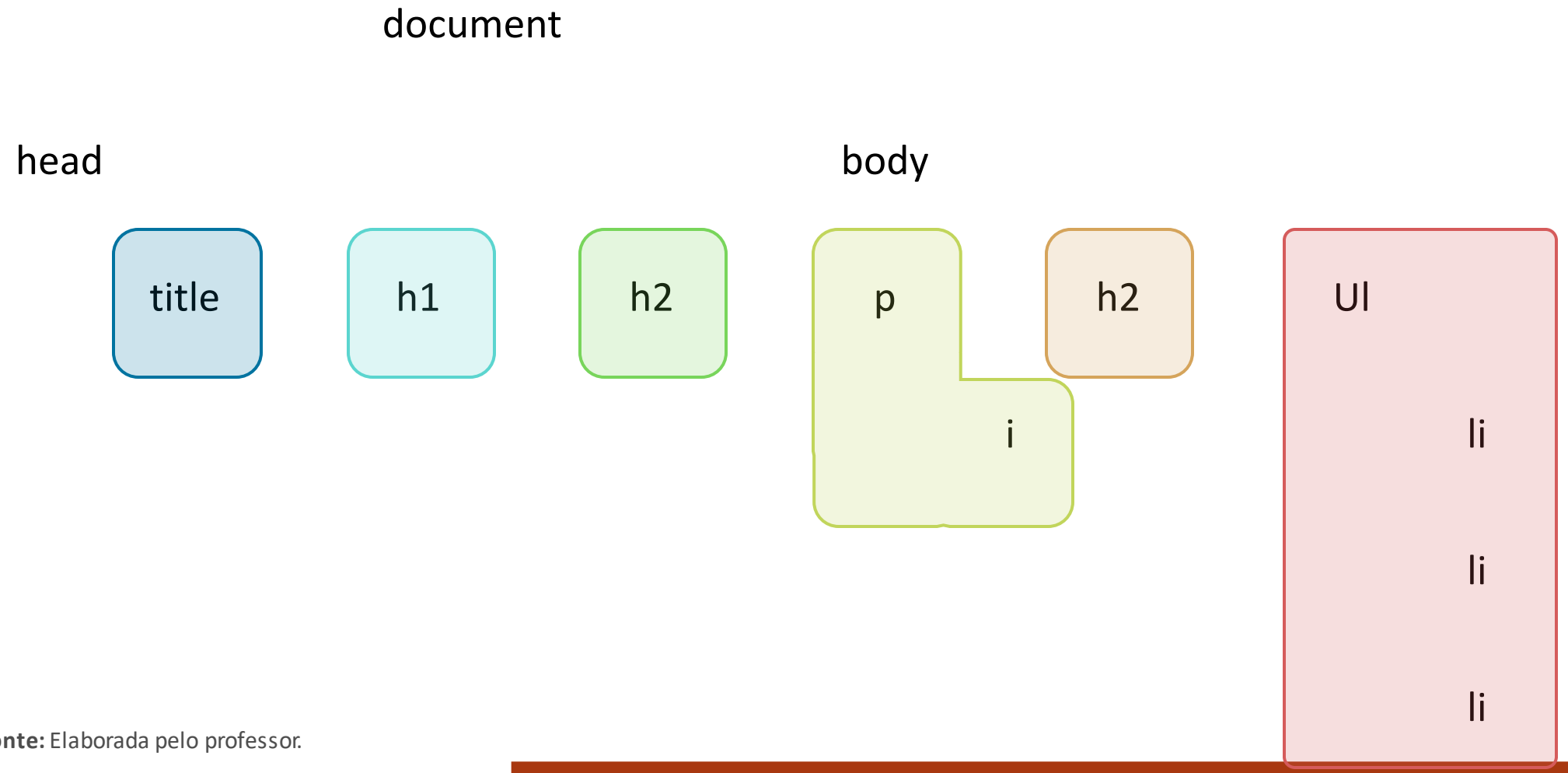
Este paragrafo vai aparecer abaixo do *heading* nivel 2 de Olá Mundo

Exemplo com lista

- Item 1
- Item 2
- Item 3

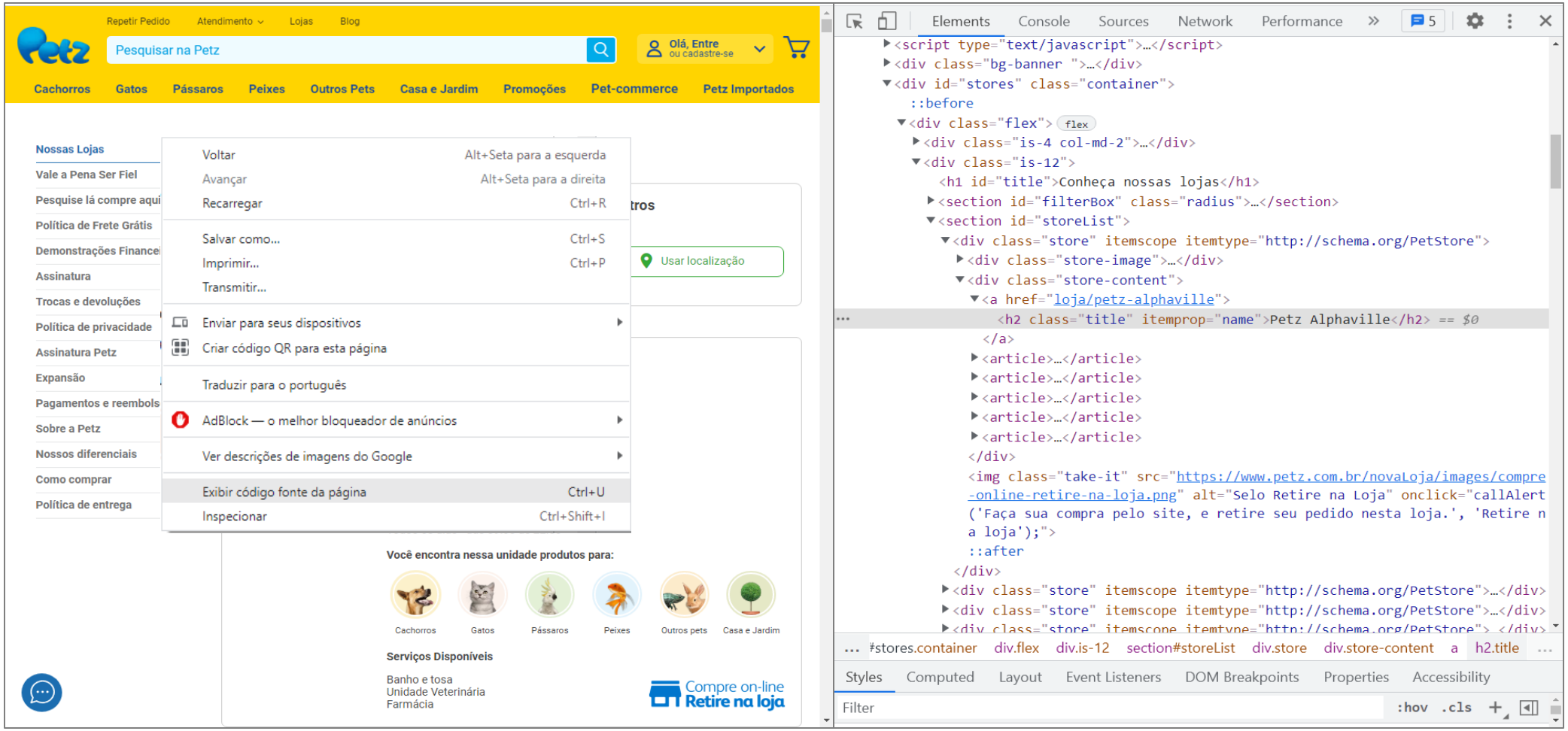
Fonte: Elaborada pelo professor.

Uma página da internet – Olá, mundo! – Árvore DOM



Fonte: Elaborada pelo professor.

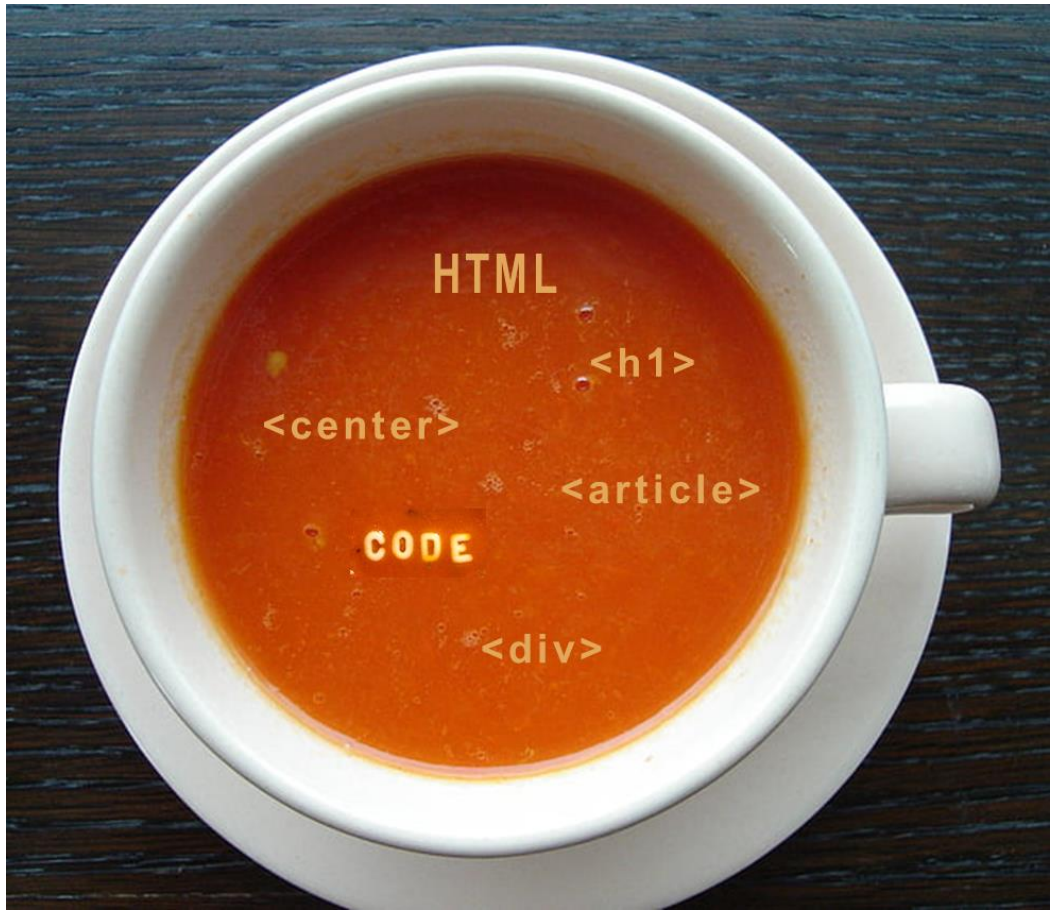
Uma página da internet e seu HTML



Fonte: <https://www.petz.com.br/nossas-lojas>; Chrome DevTools.

A biblioteca *Beautiful Soup 4.0*

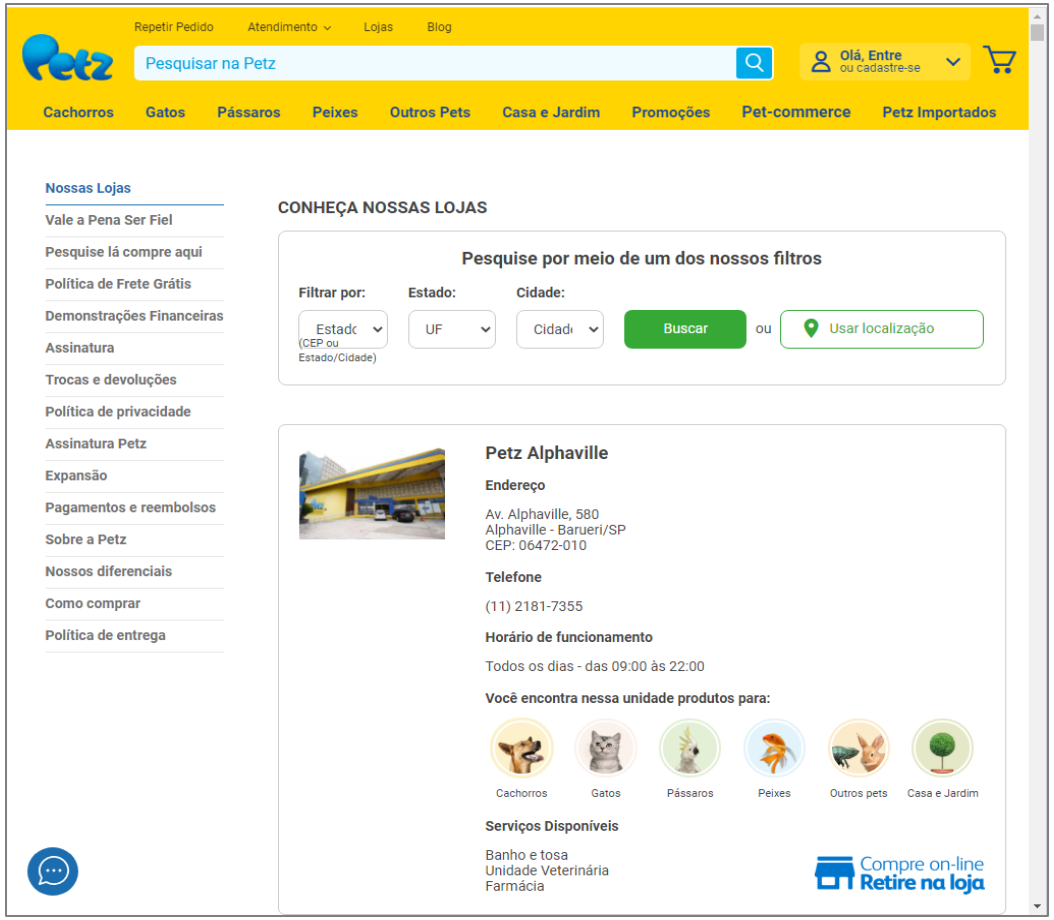
- Muitos documentos possuem erro de sintaxe (sopa de etiquetas).
- Leonard Richardson criou um interpretador que trata alguns dos erros mais comuns.
- Ideal para pequenos projetos de raspagem de documentos.



Adaptação em foto de domínio público.
Fonte: <https://www.hippopx.com/>.

Exercício prático

A equipe de marketing de uma empresa de distribuição de produtos veterinários precisava do endereço de todas as lojas da rede Petz para uma campanha. Sabe-se que todos os endereços estão disponíveis no site da empresa, porém, como utilizar a raspagem de tela e o *Beautiful Soup 4.0* para obter esses dados de forma estruturada?



Fonte: <https://www.petz.com.br/nossas-lojas>.

Passo 1 – Baixando o documento para uma *string python*

```
1.  import requests
2.
3.  url = 'https://www.petz.com.br/nossas-lojas'
4.  res = requests.get(url)
5.  html_page = res.text
```

Passo 2 – Invocando o *Beautiful Soup* 4.0

```
1.  import requests
2.
3.  url = 'https://www.petz.com.br/nossas-lojas'
4.  res = requests.get(url)
5.  html_page = res.text
```

```
1.  from bs4 import BeautifulSoup
2.  soup = BeautifulSoup(html_page, 'html.parser')
```

Passo 3 – Fazendo a raspagem

```
1.  import requests
2.
3.  url = 'https://www.petz.com.br/nossas-lojas'
4.  res = requests.get(url)
5.  html_page = res.text
```

```
1.  from bs4 import BeautifulSoup
2.  soup = BeautifulSoup(html_page, 'html.parser')
```

```
1.  import re
2.  stores = []
3.  for index, store in enumerate(soup.find_all("div", {"class": "store"}), start=1):
4.      name = store.find("h2", {"itemprop": "name"}).get_text().strip()
5.      address = store.find("p", {"itemprop": "address"})
6.      address = address.get_text()
7.      address = re.sub(' +', ' ', address)
8.      address = " ".join([s for s in address.strip().splitlines(False) if s.strip()])
9.      addressLoc = store.find("span", {"itemprop": "addressLocality"}).get_text().strip()
10.     region = store.find("span", {"itemprop": "addressRegion"}).get_text().strip()
11.     telephone = store.find("p", {"itemprop": "telephone"}).get_text().strip()
12.     openingHours = store.find("p", {"itemprop": "openingHours"}).get_text().strip()
13.     row = { 'name': name,                          'address':address,
14.             'telephone': telephone,                'openingHours':openingHours,
15.             'addressLocality': addressLoc,         'region': region}
16.     stores.append(row)
```

Passo 4 – Exportando para o Excel

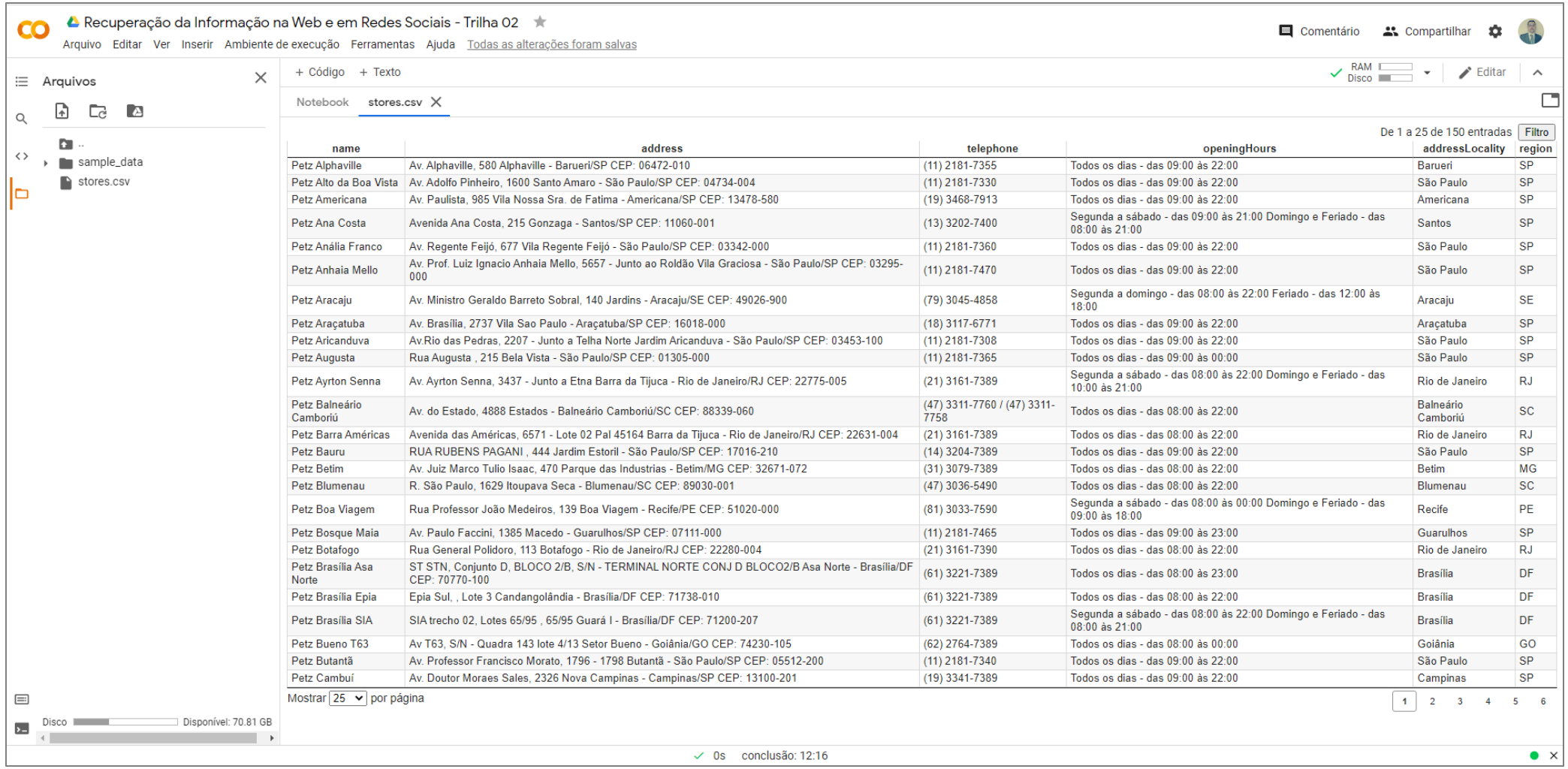
```
1.  import requests
2.
3.  url = 'https://www.petz.com.br/nossas-lojas'
4.  res = requests.get(url)
5.  html_page = res.text
```

```
1.  from bs4 import BeautifulSoup
2.  soup = BeautifulSoup(html_page, 'html.parser')
```

```
1.  import re
2.  stores = []
3.  for index, store in enumerate(soup.find_all("div", {"class": "store"}), start=1):
4.      name = store.find("h2", {"itemprop": "name"}).get_text().strip()
5.      address = store.find("p", {"itemprop": "address"})
6.      address = address.get_text()
7.      address = re.sub(' +', ' ', address)
8.      address = " ".join([s for s in address.strip().splitlines(False) if s.strip()])
9.      addressLoc = store.find("span", {"itemprop": "addressLocality"}).get_text().strip()
10.     region = store.find("span", {"itemprop": "addressRegion"}).get_text().strip()
11.     telephone = store.find("p", {"itemprop": "telephone"}).get_text().strip()
12.     openingHours = store.find("p", {"itemprop": "openingHours"}).get_text().strip()
13.     row = { 'name': name,                          'address':address,
14.             'telephone': telephone,                'openingHours':openingHours,
15.             'addressLocality': addressLoc,         'region': region}
16.     stores.append(row)
```

```
1.  import csv
2.  keys = stores[0].keys()
3.  with open('stores.csv', 'w', newline='')  as output_file:
4.    dict_writer = csv.DictWriter(output_file, keys)
5.    dict_writer.writeheader()
6.    dict_writer.writerows(stores)
```

Arquivo final tabulado



De 1 a 25 de 150 entradas | Filtro

name	address	telephone	openingHours	addressLocality	region
Petz Alphaville	Av. Alphaville, 580 Alphaville - Barueri/SP CEP: 06472-010	(11) 2181-7355	Todos os dias - das 09:00 às 22:00	Barueri	SP
Petz Alto da Boa Vista	Av. Adolfo Pinheiro, 1600 Santo Amaro - São Paulo/SP CEP: 04734-004	(11) 2181-7330	Todos os dias - das 09:00 às 22:00	São Paulo	SP
Petz Americana	Av. Paulista, 985 Vila Nossa Sra. de Fatima - Americana/SP CEP: 13478-580	(19) 3468-7913	Todos os dias - das 09:00 às 22:00	Americana	SP
Petz Ana Costa	Avenida Ana Costa, 215 Gonzaga - Santos/SP CEP: 11060-001	(13) 3202-7400	Segunda a sábado - das 09:00 às 21:00 Domingo e Feriado - das 08:00 às 21:00	Santos	SP
Petz Anália Franco	Av. Regente Feijó, 677 Vila Regente Feijó - São Paulo/SP CEP: 03342-000	(11) 2181-7360	Todos os dias - das 09:00 às 22:00	São Paulo	SP
Petz Anhaia Mello	Av. Prof. Luiz Ignacio Anhaia Mello, 5657 - Junto ao Roldão Vila Graciosa - São Paulo/SP CEP: 03295-000	(11) 2181-7470	Todos os dias - das 09:00 às 22:00	São Paulo	SP
Petz Aracaju	Av. Ministro Geraldo Barreto Sobral, 140 Jardins - Aracaju/SE CEP: 49026-900	(79) 3045-4858	Segunda a domingo - das 08:00 às 22:00 Feriado - das 12:00 às 18:00	Aracaju	SE
Petz Araçatuba	Av. Brasília, 2737 Vila Sao Paulo - Araçatuba/SP CEP: 16018-000	(18) 3117-6771	Todos os dias - das 09:00 às 22:00	Araçatuba	SP
Petz Aricanduva	Av.Rio das Pedras, 2207 - Junto a Telha Norte Jardim Aricanduva - São Paulo/SP CEP: 03453-100	(11) 2181-7308	Todos os dias - das 09:00 às 22:00	São Paulo	SP
Petz Augusta	Rua Augusta , 215 Bela Vista - São Paulo/SP CEP: 01305-000	(11) 2181-7365	Todos os dias - das 09:00 às 00:00	São Paulo	SP
Petz Ayrton Senna	Av. Ayrton Senna, 3437 - Junto a Etna Barra da Tijuca - Rio de Janeiro/RJ CEP: 22775-005	(21) 3161-7389	Segunda a sábado - das 08:00 às 22:00 Domingo e Feriado - das 10:00 às 21:00	Rio de Janeiro	RJ
Petz Balneário Camboriú	Av. do Estado, 4888 Estados - Balneário Camboriú/SC CEP: 88339-060	(47) 3311-7760 / (47) 3311-7758	Todos os dias - das 08:00 às 22:00	Balneário Camboriú	SC
Petz Barra Américas	Avenida das Américas, 6571 - Lote 02 Pal 45164 Barra da Tijuca - Rio de Janeiro/RJ CEP: 22631-004	(21) 3161-7389	Todos os dias - das 08:00 às 22:00	Rio de Janeiro	RJ
Petz Bauru	RUA RUBENS PAGANI , 444 Jardim Estoril - São Paulo/SP CEP: 17016-210	(14) 3204-7389	Todos os dias - das 09:00 às 22:00	São Paulo	SP
Petz Betim	Av. Juiz Marco Tulio Isaac, 470 Parque das Industrias - Betim/MG CEP: 32671-072	(31) 3079-7389	Todos os dias - das 08:00 às 22:00	Betim	MG
Petz Blumenau	R. São Paulo, 1629 Itoupava Seca - Blumenau/SC CEP: 89030-001	(47) 3036-5490	Todos os dias - das 08:00 às 22:00	Blumenau	SC
Petz Boa Viagem	Rua Professor João Medeiros, 139 Boa Viagem - Recife/PE CEP: 51020-000	(81) 3033-7590	Segunda a sábado - das 08:00 às 00:00 Domingo e Feriado - das 09:00 às 18:00	Recife	PE
Petz Bosque Maia	Av. Paulo Faccini, 1385 Macedo - Guarulhos/SP CEP: 07111-000	(11) 2181-7465	Todos os dias - das 09:00 às 23:00	Guarulhos	SP
Petz Botafogo	Rua General Polidoro, 113 Botafogo - Rio de Janeiro/RJ CEP: 22280-004	(21) 3161-7390	Todos os dias - das 08:00 às 22:00	Rio de Janeiro	RJ
Petz Brasília Asa Norte	ST STN, Conjunto D, BLOCO 2/B, S/N - TERMINAL NORTE CONJ D BLOCO2/B Asa Norte - Brasília/DF CEP: 70770-100	(61) 3221-7389	Todos os dias - das 08:00 às 23:00	Brasília	DF
Petz Brasília Epia	Epia Sul, , Lote 3 Candangolândia - Brasília/DF CEP: 71738-010	(61) 3221-7389	Todos os dias - das 08:00 às 22:00	Brasília	DF
Petz Brasília SIA	SIA trecho 02, Lotes 65/95 , 65/95 Guará I - Brasília/DF CEP: 71200-207	(61) 3221-7389	Segunda a sábado - das 08:00 às 22:00 Domingo e Feriado - das 08:00 às 21:00	Brasília	DF
Petz Bueno T63	Av T63, S/N - Quadra 143 lote 4/13 Setor Bueno - Goiânia/GO CEP: 74230-105	(62) 2764-7389	Todos os dias - das 08:00 às 00:00	Goiânia	GO
Petz Butantã	Av. Professor Francisco Morato, 1796 - 1798 Butantã - São Paulo/SP CEP: 05512-200	(11) 2181-7340	Todos os dias - das 09:00 às 22:00	São Paulo	SP
Petz Cambuí	Av. Doutor Moraes Sales, 2326 Nova Campinas - Campinas/SP CEP: 13100-201	(19) 3341-7389	Todos os dias - das 09:00 às 22:00	Campinas	SP

Mostrar 25 por página

1 2 3 4 5 6

Disponível no Google colab

https://bit.ly/3yOaPJM

Universidade Presbiteriana
Mackenzie