

**TECNOLOGIA EM SISTEMAS PARA INTERNET**

**Daniel Evangelista Pereira  
Ribson Coelho Cardoch Valdés  
Douglas Seidi Shibata**

**RELATÓRIO DE PRÁTICA INTEGRADA  
DE  
CIÊNCIA DE DADOS E INTELIGÊNCIA ARTIFICIAL**

19/08/2020

**Brasília - DF**

**2020**

# Sumário

<b>1. Objetivos</b>	<b>3</b>
<b>2. Descrição do problema</b>	<b>4</b>
<b>3. Desenvolvimento</b>	<b>5</b>
3.1 Código implementado	5
<b>4. Considerações Finais</b>	<b>6</b>
<b>Referências</b>	<b>7</b>

# 1. Objetivos

Esta etapa do projeto tem como finalidade percorrer sobre a coleta de dados por meio de web scraping, com o script feito na linguagem python e gerando um arquivo Comma-separated values. Que será de suma importância para o restante do projeto. E armazenar os dados no repositório do Github com o versionamento de código do git.

## 2. Descrição do problema

Ao iniciar o desenvolvimento do projeto podemos destacar a forma para se coletar os dados e como estão distribuídos no site e a partir desta análise inicial, foi visto que havia uma grande quantidade de dados em diversas páginas que precisar ser extraídas do site para realizar a coleta de dados e armazenar em Data Frames para que se possa gerar um arquivo csv para começar o estudo dos dados na primeira etapa do projeto.

### 3. Desenvolvimento

As tecnologias utilizadas para a elaboração da primeira parte do projeto na fase de coleta de dados consiste na linguagem de programação python com o auxílio de suas bibliotecas, como o requests para fazer a requisição do site e retornar o status code, a biblioteca BeautifulSoup para a leitura das páginas em Hyper Text Markup Language(HTML) e para o armazenamento de dados foi utilizada a biblioteca pandas, por meio do notebook e o ambiente de desenvolvimento Google Colab.

#### 3.1 Código implementado

##### 5.2 - Script de coleta

###### 5.2 - Script da coleta

coleta de dados dos vinte anos, entre setembro 1997 e agosto de 2017. Coloque tudo em um DataFrame e depois salve em um arquivo .CSV com o nome OVNIS.csv.

```
1 import requests
2 from bs4 import BeautifulSoup
3 import pandas as pd
```

Importação das bibliotecas Requests, BeautifulSoup e Pandas.

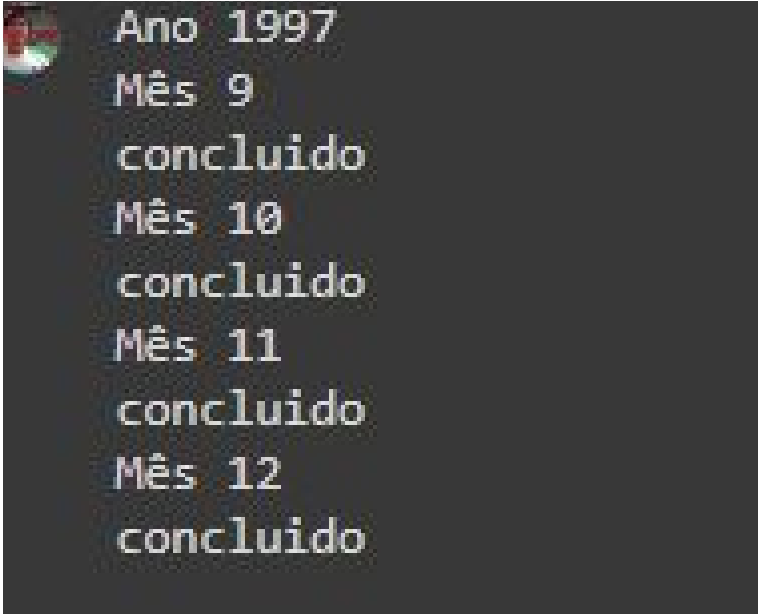
```
[ ] 1 td_data = list()
    2 th_data = list()
```

Declarando duas listas, uma para armazenar os dados da tabela que está no html da página. E segundo o cabeçalho da tabela.

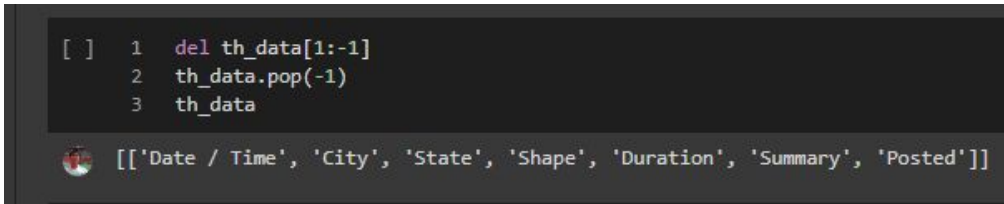
```
[ ] 1 #1997 mês de setembro
    2 print('Ano 1997')
    3 for p in range(9,13):
    4     print('Mês {}'.format(p))
    5     if p>9:
    6         URL = f'http://www.nuforc.org/webreports/ndxe1997{p}.html'
    7     else:
    8         URL = f'http://www.nuforc.org/webreports/ndxe19970{p}.html'
    9     headers={
   10         'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36'
   11     }
   12     page = requests.get(URL,headers=headers)
   13     soup = BeautifulSoup(page.content,'html.parser')
   14     table = soup.findAll('table')[0]
   15     tr = table.findAll(['tr'])
   16     try:
   17         for cell in tr:
   18             th = cell.find_all('th')
   19             th_data.append([col.text.strip('\n') for col in th])
   20             td = cell.find_all('td')
   21             row = [i.text.replace('\n','') for i in td]
   22             td_data.append(row)
   23     except:
   24         print('Erro')
   25     finally:
   26         print('concluido')
```

Neste pedaço de código foi feito a coleta de dados de setembro de 1997 até dezembro do mesmo ano. Na terceira linha foi utilizado o laço de repetição for para que possa se percorrer o site a

partir dos parâmetros passados na url do site. E para que se possa ter o controle dos parâmetros que são passados na url, foi printado os meses correspondentes. A variável Header pega os metadados para a coleta de dados do site, e a variável page recebe a url e o header dos metadados e faz a requisição da página html, e a variável soup irá utilizar a função BeautifulSoup para que se possa ler o conteúdo do site que foi requisitado. E a variável table receberá a variável soup que possui a função findall e irá buscar a primeira tag table que aparece no código html. E em seguida é declarada uma variável tr que receberá a variável table e será utilizado a função findall para que se possa encontrar a tag tr no código da página. E depois é feito um bloco de try para que se possa passar a requisição, se houver êxito, o código continuará. E foi implementado um laço de repetição for para percorrer as colunas da tabela para saber as informações, e a partir disso é adicionado ao array declarado inicialmente.



```
Ano 1997
Mês 9
concluido
Mês 10
concluido
Mês 11
concluido
Mês 12
concluido
```



```
[ ] 1 del th_data[1:-1]
    2 th_data.pop(-1)
    3 th_data

[['Date / Time', 'City', 'State', 'Shape', 'Duration', 'Summary', 'Posted']]
```

Nesta parte do código é feita a remoção de duplicadas no array do cabeçalho. Deixando apenas um lista com os cabeçalhos.

```
[ ] 1 df = pd.DataFrame(td_data)
    2 df1 = df.dropna()
    3 coluna = []
    4 for i in th_data:
    5     coluna = i
    6 df1.columns = coluna
    7 df1
```

	Date / Time	City	State	Shape	Duration	Summary	Posted
1	9/30/97 20:00	Nova Scotia (Canada)	NS	Light	8-10 seconds.	Single light resembling a star, but moving spu...	10/30/06
2	9/22/97 20:00	Solomons Island	MD	Disk	10 minutes	Close up at twilight, Stationary UFO.	8/5/09
3	9/19/97	Garden Grove	CA	Rectangle	4 mins.	Around 6:30 PM I was walking through a Vons Pa...	12/1/19
4	9/18/97 20:15	Panama City	FL	Unknown	30 seconds	Looked like stars in the sky so far up/moveing...	3/13/12
5	9/15/97 00:00	Houston	TX	Disk	5 minutes	Beautiful silver-colored flying saucer about t...	7/19/10
...	...	...	...	...	...	...	...
72136	8/1/17 02:45	Corcoran	MN	Light	Still going	Small light south west of Minneapolis maneuver...	8/4/17
72137	8/1/17 02:00	Moreno Valley	CA	Other	10 seconds	I was looking out the front windshield and loo...	8/4/17
72138	8/1/17 01:00	Bradenton	FL	Other	<20 seconds	I was walking my dog about 1am on August 1, 20...	5/9/19
72139	8/1/17 00:00	Springdale	AR		1 hour	Glowing flying people . seven of them flying L...	2/13/20
72140	8/1/17	Laurel	MD	Other		It was an alien project level 1 federal ran on...	6/25/20

71901 rows x 7 columns

Nesta parte do código é feita o armazenamento dos dados coletados no DataFrame. e depois renomeando o nome das colunas.

```
[ ] 1 df1.to_csv('OVNIS.csv',index=False)
    2 total = pd.read_csv('OVNIS.csv')
    3 total
```

	Date / Time	City	State	Shape	Duration	Summary	Posted
0	9/30/97 20:00	Nova Scotia (Canada)	NS	Light	8-10 seconds.	Single light resembling a star, but moving spu...	10/30/06
1	9/22/97 20:00	Solomons Island	MD	Disk	10 minutes	Close up at twilight, Stationary UFO.	8/5/09
2	9/19/97	Garden Grove	CA	Rectangle	4 mins.	Around 6:30 PM I was walking through a Vons Pa...	12/1/19
3	9/18/97 20:15	Panama City	FL	Unknown	30 seconds	Looked like stars in the sky so far up/moveing...	3/13/12
4	9/15/97 00:00	Houston	TX	Disk	5 minutes	Beautiful silver-colored flying saucer about t...	7/19/10
...	...	...	...	...	...	...	...
71896	8/1/17 02:45	Corcoran	MN	Light	Still going	Small light south west of Minneapolis maneuver...	8/4/17
71897	8/1/17 02:00	Moreno Valley	CA	Other	10 seconds	I was looking out the front windshield and loo...	8/4/17
71898	8/1/17 01:00	Bradenton	FL	Other	<20 seconds	I was walking my dog about 1am on August 1, 20...	5/9/19
71899	8/1/17 00:00	Springdale	AR	NaN	1 hour	Glowing flying people . seven of them flying L...	2/13/20
71900	8/1/17	Laurel	MD	Other	NaN	It was an alien project level 1 federal ran on...	6/25/20

Depois é a partir do Dataframe, foi utilizado a função to\_csv para gerar o arquivo csv.

## 4. Considerações Finais

Nesta primeira do projeto integrado de ciência de dados e inteligência artificial. Houve o estudo das bibliotecas sugeridas pelos professores para que se possa auxiliar na coleta de dados. E a partir do estudo dessas bibliotecas, a coleta de dados foi feita, e acabou gerando diversas versões de códigos de coletas, com melhorias em sua leitura e entendimento, tornando o código mais limpo.



## Referências

TAGLIAFERRI, Lisa. Como trabalhar com dados da web usando Requests e Beautiful soup com Python 3. **Community**. Estados Unidos, 09 de Jul de 2018. Disponível em: <<https://www.digitalocean.com/community/tutorials/como-trabalhar-com-dados-da-web-usando-requests-e-beautiful-soup-com-python-3-ptt>>. Acesso em: 08 de Set.2020

RICHARDSON, Leonard. Beautiful soup documentation. **Beautiful Soup 4.9.0 Documentation**. Estados Unidos, c2004-2020. Disponível em: <<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>>. Acesso em: 12 Set. de 2020

FIGUEIREDO, Vinicius. Seus Primeiros Passos com Data Scientist: Introdução ao Pandas. **Data Hackers**. São Paulo, 30 de maio de 2018. Disponível em: <<https://medium.com/data-hackers/uma-introdu%C3%A7%C3%A3o-simples-ao-pandas-1e15eea37fa1>>. Acesso em: 08 de Set. de 2020

GUIA Rápido. **Requests**. Estados Unidos, c2013. Disponível em: <[https://requests.readthedocs.io/pt\\_BR/latest/user/quickstart.html](https://requests.readthedocs.io/pt_BR/latest/user/quickstart.html)>. Acesso em: 11 Set. de 2020