

TECNOLOGIA EM SISTEMAS PARA INTERNET

**André Luiz Gomes dos Santos
Brenda Lopes Miranda Teixeira
Luiz Henrique de Paiva Ventura**

**RELATÓRIO DE PRÁTICA INTEGRADA
DE
CIÊNCIA DE DADOS E APRENDIZADO DE MÁQUINA**

Brasília - DF

05/03/2020

Sumário

1. Objetivos	3
2. Descrição do problema	3
3. Desenvolvimento	3
3.1 Código implementado	3
3.1.1 Base de dados	4
4. Considerações Finais	8
Referências	10

1. Objetivos

Este relatório tem por objetivo extrair dados sobre avistamento de objetos voadores não identificados do website: The National UFO Reporting Center (nuforc.org).

2. Descrição do problema

A The National UFO Reporting Center é uma organização dos EUA (Estados Unidos da América) que investiga avistamento de objetos voadores não identificados que surgem no céu.

O site nuforc.org como mencionado no tópico anterior, possui diversos dados. Na qual são disponibilizados relatórios de Ovnis. Tais dados são disponibilizados desde da década 1974, quando foi criada.

Segundo a Wikipedia, a The National UFO Reporting Center já identificou 90.000 OVNIS avistados durante esse período. A mesma disponibiliza os dados gráficos e estatísticas para quaisquer pessoas que se interessem possam ver. Além disso, a organização fornece um número telefônico para que pessoas que avistarem quaisquer objeto voador, possam ligar diretamente ao centro e avisar.

A partir do conhecimento da Nuforc e do que ela disponibiliza, neste trabalho será extraído esses dados a devorante de uma linguagem de programação.

3. Desenvolvimento

Este trabalho está sendo desenvolvido usando um Script Python por ser uma linguagem orientada a objetos é bastante maleável, o grupo G2 está utilizando a plataforma Google Colaboratory, assim todos podem modificar e acrescentar o código quando necessário. Todos os códigos estão sendo disponibilizados no github.

3.1 Código implementado

Abaixo será mostrada imagens do código já implementado com descrições.

3.1.1 Base de dados

Para iniciar a atividade de coleta de dados, primeiro será realizado um teste de requisições no site www.nuforc.org e a será realizado às importações necessárias para a execução da coleta.

A imagem a seguir mostra a importação de bibliotecas, extração de dados de forma tabular dos vinte anos e também é realizado testes de requisição.

Figura 1 - Base de Dados

```
5.1 - Coleta de dados

1° Base de dados

[2] 1 #Base de dados: http://www.nuforc.org/
    2 #Extração de dados de forma tabular de vinte anos (Set/1997 - Agosto/2017 = 240 pág)
    3
    4 import pandas as pd
    5 from bs4 import BeautifulSoup
    6 import requests
    7 import time
    8
    9 #Teste de requisição HTTP
   10 requisicao = requests.get('http://www.nuforc.org/webreports/ndxevent.html')
   11 requisicao

<Response [200]>
```

Fonte: Própria

3.1.2 Coleta de dados

Aqui criamos uma lista vazia e depois foi feito um for para o mês de Setembro a Dezembro e após foi feito um if para verificar se é menor que o mês 10.

Figura 2 - Lista do ano de 1997

```
] #lista com os dados
lista10 = []

#for para percorrer as páginas
#range (início, fim, intervalo)
mes=0
#Ano 1997
for mes in range (9,13,1):
    if mes<10:

        data=str(1997)+str(0)+str(mes)
        if requisicao.status_code == 200:
            lista10.append(f'http://www.nuforc.org/webreports/ndxe{data}.html')
```

```

else:
    data=str(1997)+str(mes)
    if requisicao.status_code == 200:

        lista10.append(f'http://www.nuforc.org/webreports/ndxe{data}.html')
time.sleep(1)

#Quantidade de páginas na lista
len(lista10)

```

Fonte: Própria

Nesta parte o código faz o teste de requisição informando se a resposta foi bem sucedida.

Figura 3 - Requisição

```

1 #Anos 1998 até 2016
2 lista11 = []
3 for ano in range (1998,2017,1):
4     #print(f'Página: {ano}')
5     mes=0
6     data=0
7     for mes in range (1,13,1):
8         if mes<10:
9
10            data=str(ano)+str(0)+str(mes)
11            if requisicao.status_code == 200:
12                lista11.append(f'http://www.nuforc.org/webreports/ndxe{data}.html')
13
14
15
16     else:
17         data=str(ano)+str(mes)
18         if requisicao.status_code == 200:
19
20            lista11.append(f'http://www.nuforc.org/webreports/ndxe{data}.html')
21
22     time.sleep(1)
23
24 #Quantidade de páginas na lista
25 len(lista11)

```

Fonte: Própria

Janeiro a Agosto.

Figura 4 - Ano 2017

```
[10] 1 #Ano 2017
2 lista12 = []
3
4 #for para percorrer as páginas
5 #range (início, fim, intervalo)
6 mes=0
7 #Ano 1997
8 for mes in range (1,9,1):
9     if mes<10:
10
11         data=str(2017)+str(0)+str(mes)
12         if requisicao.status_code == 200:
13             lista12.append(f'http://www.nuforc.org/webreports/ndxe{data}.html')
14
15
16
17     else:
18         data=str(ano)+str(mes)
19         if requisicao.status_code == 200:
20
21             lista12.append(f'http://www.nuforc.org/webreports/ndxe{data}.html')
22 time.sleep(1)
23
24 #Quantidade de páginas na lista
25 len(lista12)
```

8

Fonte: Própria

Nesta imagem criamos uma lista a partir das listas anteriores conforme mostrado abaixo. Ela faz a união de todas as listas.

Figura 5 - Lista 2

```
1 #Criar uma lista das listas anteriores
2 #lista nova para criação de tabelas - lista2
3 #Variável conte - controle de listas
4 # reunir todas as lista em uma (lista2)
5 # lista 2 possui todos os relatos por mês
6
7 lista2 = []
8 for conte in lista10:
9     pag = pd.read_html(conte, header = 0)
10     lista2.append(pag[0])
11
12 lista3 = []
13 for conte in lista11:
14     pag = pd.read_html(conte, header = 0)
15     lista3.append(pag[0])
16
17 lista4 = []
18 for conte in lista12:
19     pag = pd.read_html(conte, header = 0)
20     lista4.append(pag[0])
21
22 lista2.append(lista4)
23 lista2.append(lista3)
24 lista2
```

	Date / Time	...	Posted
0	9/30/97 20:00	...	10/30/06
1	9/22/97 20:00	...	8/5/09
2	9/19/97	...	12/1/19
3	9/18/97 20:15	...	3/13/12
4	9/15/97 00:00	...	7/19/10
5	9/15/97 20:00	...	11/9/17
6	9/15/97 20:00	...	3/17/17
7	9/15/97 20:00	...	7/28/10
8	9/15/97 20:00	...	2/24/07
9	9/7/97 12:00	...	1/10/09
10	9/1/97 23:00	...	2/1/07
11	9/1/97 22:30	...	12/12/09

[12 rows x 7 columns] Date / Time ... Posted

Fonte: Própria

Figura 6- Criação do DataFrame

```
[13] 1 #Criação do dataframe
2
3 df_relatos = pd.DataFrame()
4 for i in lista2:
5     df_relatos = df_relatos.append(i).reset_index(drop = True)
6
7
8 df_relatos
```

	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
...
71991	12/1/16 18:00	Amherst	NY	Oval	5 minutes	Odd, mostly stationary lights, in night sky, w...	12/5/16
71992	12/1/16 17:30	Montgomery	AL	Circle	1 minute	2 extremely bright lights moving slowly and di...	12/5/16
71993	12/1/16 16:30	Asbury Park	NJ	Oval	2 minutes	Jersey shore sighting.	12/5/16
71994	12/1/16 12:46	Escondido	CA	Other	120 seconds	Reddish orange trail.	1/6/17
71995	12/1/16 07:00	Raton	NM	Cylinder	5-10 minutes	UFO sighting just south of Raton, New Mexico, ...	4/7/17

71996 rows × 7 columns

Criar arquivo CSV

Por fim, o arquivo em formato csv é criado usando um comando. Formato que pode ser visualizado no excel.

Figura 7 - CSV

```
3° Salvar em um arquivo .csv (OVNIS.csv)
```

```
[14] 1 #Código para salvar o dataframe
2 df_relatos.to_csv('OVNIS.csv')
```

Fonte: Própria

4. Considerações Finais

A princípio as dificuldades enfrentadas eram de como extrair os dados de forma eficaz e completa a partir dos anos de 1997 a 2017, contabilizando vinte anos, utilizando os parâmetros da Url. Como foi mostrado no tópico 3.1 esse dados foi possível extrair usando dataframes. Em tempos de pandemia, com os alunos trabalhando e estudando em horários diferentes, também foi identificado um impasse para que todos pudessem se organizar e se reunir em horário específico. Usando a plataforma meet conseguimos finalizar esta parte do trabalho.

Referências

Wikipedia. **National UFO Reporting Center**. 2021. Disponível em:

<https://en.wikipedia.org/wiki/National_UFO_Reporting_Center>. Acesso em: 06 de mar 2021.

Towardsdatascience. **All Pandas read_html() you should know for scraping data from HTML tables**. Disponível em:

<<https://towardsdatascience.com/all-pandas-read-html-you-should-know-for-scraping-data-from-html-tables-a3cbb5ce8274>>. Acesso em: 8 mar. 2021.