O PROBLEMA

Foi proposto a coleta de dados sobre relatos de pessoas que avistaram ovinis em todas as partes do mundo através do site NATIONAL UFO REPORTING CENTER *(http://www.nwlink.com/~ufocntr/)*, na qual, tem uma vasta base de dados sobre relatos de aparições de ovinis. O objetivo é coletar todos os dados dos vinte anos, entre setembro 1997 e agosto de 2017 e exportá-los para um arquivo csv.

A SOLUÇÃO

Para solucionar o problema foi criado um script em python para capturar os dados utilizando a técnica de web scraping, que é uma forma de mineração que permite a extração de dados de sites da web convertendo-os em informação estruturada para posterior análise.

Para desenvolver o script, foi necessário a utilização de algumas bibliotecas:

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

- A biblioteca requests foi utilizada para fazer as requisições HTTP ao serviço.
- Já a biblioteca BeautifulSoup foi utilizada para coletar informações das páginas web. Com essa biblioteca é possível extrair tabelas, listas, parágrafos e você também pode colocar filtros para extrair informações de páginas da web.
- Pandas também foi uma biblioteca muito importante no processo de desenvolvimento do script, através dela, foi possível agrupar e combinar dados.
- E por último, mas não menos importante, foi utilizado a biblioteca time, através dela foi possível controlar a velocida das requisições http, a fim de limitar o tempo das requisições para que não houvessem problemas relacionados a banimento no servidor do site da base de dados.

Vamos ao código!

```
1  link = 'http://www.nuforc.org/webreports/ndxe'
2  final ='.html'
3
4  df_1 = []
5
6  for i in range(1997,2018):
7    mes = ['01','02','03','04','05','06','07','08','09','10','11','12']
8    if(i==1998):
9      mes = ['01','02','03','04','05','06','07','08','09']
10   if(i==1997):
11     mes = ['10','11','12']
12   for j in mes:
13     url_it = requests.get('{0}{1}{2}{3}'.format(link,i,j,final))
14     soup = BeautifulSoup(url_it.content, 'html.parser').find('table')
15     df = pd.read_html(str(soup))[0]
16     df_1.append(df)
17     time.sleep(1)
18
19
20 df_2 = pd.concat(df_1, ignore_index=True)
21 df_2.to_csv('OvNi.csv',index=False)
22
```

Na linha 1 está o link do site e na linha 2 o final da url do site, ambos armazenados em variáveis, na qual serão concatenadas logo mais.

A variável *df_1* é um vetor onde será armazenado os dados das diversas páginas da base de dados.

Foi criado um loop for para definir o intervalo de tempo em anos.

E outro for para percorrer os meses dos anos.

Foi necessário utilizar dois *if's* porque o intervalo de tempo solicitado foi entre *setembro 1997 e agosto de 2017*. E isso foi ajustado pela lista de meses de nome *mes.*

A variável *url_it* armazena o link criado a partir da url do site mais a variação de meses e anos definidas pelo intervalo solicitado. Para criar a url, foi necessário concatenar as variáveis link, i, j e final, e a cada iteração teremos um link referente ao mês do relato óvni, a partir do link que criamos.

Na variável *soup* utilizamos a biblioteca *BeautifulSoup* para capturar os dados das tabelas contidas nas página html, acessadas pelas urls que montamos.

Após isso, usamos a biblioteca *pandas* para ler o *html* da página e logo em seguida para anexar os dados coletados das tabelas para nosso *dataframe*.

Depois disso, a biblioteca *time* foi utilizada para moderar o tempo das requisições.

Na linha seguinte, utilizamos o *pandas* novamente, dessa vez, com a finalidade de concatenar todas as nossas tabelas capturadas, transformando em apenas uma grande tabela.

E por fim, exportamos a tabela para um arquivo CSV.

https://github.com/HugoCalisto/datascience