Instituto Federal, Ciência e Tecnologia de Brasília – campus Brasília

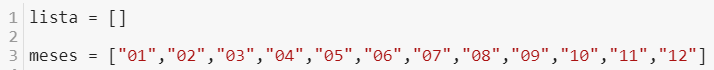
Estudante : Samuel Ferreira

COLETA DE DADOS

A atividade realizada de durante a ministração da matéria de Data Science, foi realizada a coleta de dados do website “ THE NATIONAL UFO REPORTING CENTER ”. O objetivo da coleta de dados é construir um Data Frame com os dados disponibilizado pelo website. A forma que os dados foram disponibilizados foi por ***strings*** relatados por pessoas que colocavam o depoimento no site. As informações a serem coletadas foram em um período de 20 anos os depoimentos das pessoas que relatavam ter visto um OVIN.

1º ETAPA

Esta 1ª etapa para resolução da coleta de dados durante os 20 anos de dados foi criar uma lista vazia para armazenar os dados do Data Frame e uma lista com a quantidade de meses dos anos.



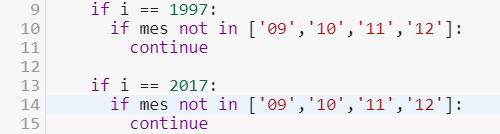
2º ETAPA

Foi criado um “for” referente aos (20 anos de dados) que inicia em 1997 e termina em 2017 (contabilizando os 20 anos de dados) e o segundo “for” percorrerá a variável ‘meses’ e dentro do “for” com a variável ‘link’ recebe dois colchetes referente ao anos e meses e o ( ***.format*** recebe o endereço da URL) convertendo os anos em ***str*** concatenando aos meses.



3º ETAPA

Aqui foi feito uma verificação onde se o ano for igual a 1997 continua o fluxo. E irá passar para a segunda verificação onde se o mês não conter a sequência de ***str*** o fluxo irá continuar. Pois para completar os 20 anos certinho é necessário pular os meses da seguinte sequência para completar os 20 anos como nós queremos.



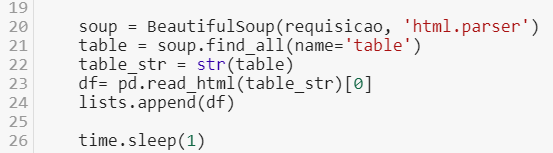
4º ETAPA

Nesta etapa faremos a requisição da página na linha 14 passando o ***.text*** para obtermos os elementos em TXT.



5º ETAPA

Depois de fizemos a requisição da página, foi usado a biblioteca ***BeautifulSoup*** para extrair a tabela. A biblioteca ***BeautifulSoup*** é usada para Extração de dados de arquivos HTML e XML. Primeiro, devemos criar um objeto que irá salvar o documento de maneira estruturada de acordo com as tags na linha 20, e depois podemos acessar o elemento chamando o método ***find*** passando como argumento o nome da tag, no caso “ ***table*** ”. Depois transformamos a tabela em formato ***String***, passamos para uma variável ***“df”*** lendo o arquivo em sua posição 0 e incrementando a lista com o ***append***. E por último colocamos um ***“ time.sleep*** ” com 1 segundo para o site não receber muitas requisições e travar o site. Por ventura causando um ataque DDos.



Passamos para variável “ ***file\_df*** ”uma sequência de objetos em serio com a função ***pd.concat*** ignorando o index. E por fim nós criamos o nosso arquivo “.csv ”



Link para o GitHub: <https://github.com/samuelflopes/Coleta>