**Atividade de Ciências de Dados**

**Aluno:** Heberth Vinícius Amarante Severo

**Descrição do problema:**

Foi proposto aos alunos do curso de sistemas uma atividade de Web Scraping, onde os alunos devem fazer a coleta de alguns dados do site da The National UFO Reporting Center, esse site é para que as pessoas possam compartilhar seus relatos sobre suas experiencias sobre algum objeto voador não identificado (OVNI), os dados do site foram separados em várias páginas, cada página representa um mês referente a um determinado ano.

O objetivo principal dessa atividade é acessar cada página do período de setembro de 1997 a agosto de 2017 e armazenar todos os dados em um arquivo csv, utilizando a linguagem de programação python.

Para a realização da atividade foi utilizado o Colab, uma ferramenta para desenvolvimento da Google.

A seguir há uma explicação sobre cada etapa do processo da coleta dos dados, armazenamento dos dados e criação do arquivo csv.

**1ª Etapa**

Para o início do arquivo no Colab é necessário fazer a importação de algumas funções do python, a Figura 1 mostra todas as funções utilizadas para a nossa atividade.

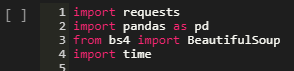


Figura 1 Importação das bibliotecas e das funções

Essas funções são necessárias para fazer o Web Scraping (coleta de dados na Web), com elas podemos fazer a análise da URL de desejamos coletar os dados, fazer a criação de Data Frames para colocar os dados e também para a criação do arquivo csv, a única função que foge a regra é a *time,* que só é necessário para grandes coletas de dados como é o nosso caso. Sem a utilização da função *time* o site onde está ocorrendo a coleta dados poderá achar que está sofrendo um ataque DDoS e irá bloquear o seu IP e você não terá mais acesso ao site.

**2ª Etapa**

Nessa etapa da atividade foi criado um loop para percorrer todos os meses de todos os anos que foram informados na descrição da atividade, como mostra a Figura 2.

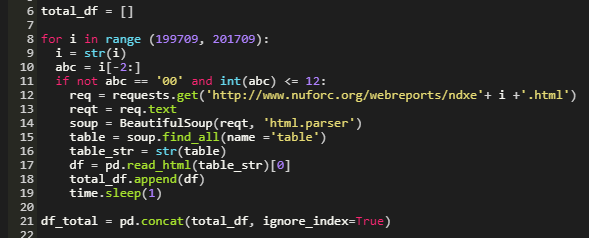
****

Figura Criando um loop para verificar e armazenar os dados

No *for* foi passado o ano e o mês de início e do fim que desejamos coletar os dados. Fez-se necessário a criação de um *if* para verificar se o loop estava coletando os dados apenas dos anos desejados, dentro do *if* é onde ocorre toda a coleta dos dados, desde a URL desejada até a criação de um data frame para armazenar todos os dados coletados que foram concatenados na variável *df\_total*.

**3ª Etapa**

A última etapa dessa a atividade é a criação do arquivo csv, como mostra a Figura 3.



Figura Criação do arquivo csv

Para a criação do arquivo foi necessário pegar a variável onde foram armazenados os dados, escolher o tipo de arquivo desejado, no nosso caso csv, e dentro dos parênteses foi colocado o nome desejado para o nosso arquivo.

**Observação**

O arquivo csv assim como este relatório estarão no GitHub, cujo link está a seguir: <https://github.com/heberth-hvas/icd_5>

**Referências**

**Google Colab:** <https://colab.research.google.com/notebooks/welcome.ipynb>

**Documentação Pandas:** <https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.append.html>

**Time Sleep:** <https://www.tutorialspoint.com/python3/time_sleep.htm>

**Exemplo de Web Scraping:** <https://medium.com/data-hackers/como-fazer-web-scraping-em-python-23c9d465a37f>