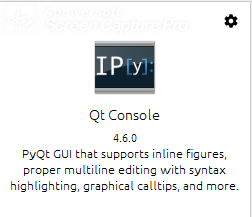
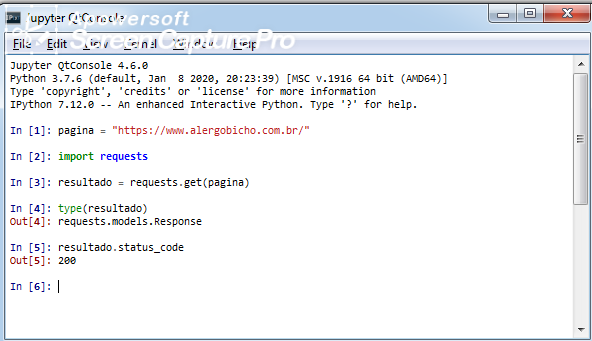
WEB SCRAPING

Passo a Passo de como é realizado o WEBSCRAPING. Utilizei o Anaconda esse aplicativo para a realização dessa atividade





# Verifico se está ok para a utilização

Jupyter QtConsole 4.6.0

Python 3.7.6 (default, Jan 8 2020, 20:23:39) [MSC v.1916 64 bit (AMD64)]

Type 'copyright', 'credits' or 'license' for more information

IPython 7.12.0 -- An enhanced Interactive Python. Type '?' for help.

# Defino a página que eu desejo analisar. Nesse caso eu optei pela minha página. Lembrando que tudo o que eu faço fica armazenado no servidor do cliente. Utilizei para esse exercício o site da alergobicho.

pagina = “ https://www.alergobicho.com.br/ ”

#Pacotes Python que permitem manipular paginas web

import requests

# Defino a página que eu desejo analisar. Get ( método). Nesse momento eu conecto a página. Estando ok eu recebo uma mensagem

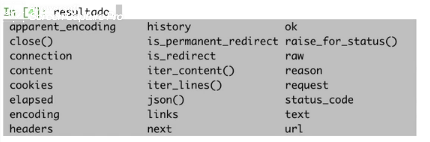
resultado = requests.get(pagina)  
type(resultado)  
Out[4]: requests.models.Response

# código 200 indica que a conexão foi feita com sucesso.  
# código 404 indica que a conexão não foi feita com sucesso.Vericamos se a pagina é válida

resultado.status\_code  
Out[5]: 200

# retorna os métodos e atributos do resultado, que podemos manipular o objeto.

resultado. (apertar a tecla TABLE)



# o tipo de resultado que ele retorna é bytes. Para poder manipular todo o conteúdo como HTML devemos transformar em texto, logo uma string ..“str”.

type(resultado.content)

Out[7]: bytes

fonte = resultado.text

type(fonte)

Out[9]: str

**# o pacote Request do Python serve apenas para obter os dados.**

2- Temos que saber o que é código HTML dentro da nossa página. Para isso utilizamos o   
“Parse HTML” pacote BeautifulSoup.

from bs4 import BeautifulSoup

soup = BeautifulSoup(fonte, 'html.parser')

#Para ver se o resultado se deu certo. Ele mostra todo o código.

print(soup.prettify)

Mostra toda a pagina em HTML a partir desse comando acima.



Exemplo do código do Professor

****

**Para poder executar o webscrapping se faz necessário o entendimento mínimo de HTM e CSS (código de marcação). Um bom site para conhecimento é W3schools.com  
Como localizar o código fonte de um site. Com o mouse direito dentro do site (Inspecionar o código fonte ou ver o código fonte da página)**

**Contruindo um webscrapper**

Para esse exercício, iremos utilizar o modelo do professor. Extrair a (pasta 3 –Parte2)

Como transformar o meu computador em local host para a realização do exercício.

Na minha máquina executar o cmd (como administrador da máquina).

**Comandos do cmd : cd ../ ( volta um nível pasta)**

No cmd da minha máquina eu digito onde está o documento que é esse caminho que está abaixo: C:\Users\Taraneh\Documents\Webscraping\Parte2

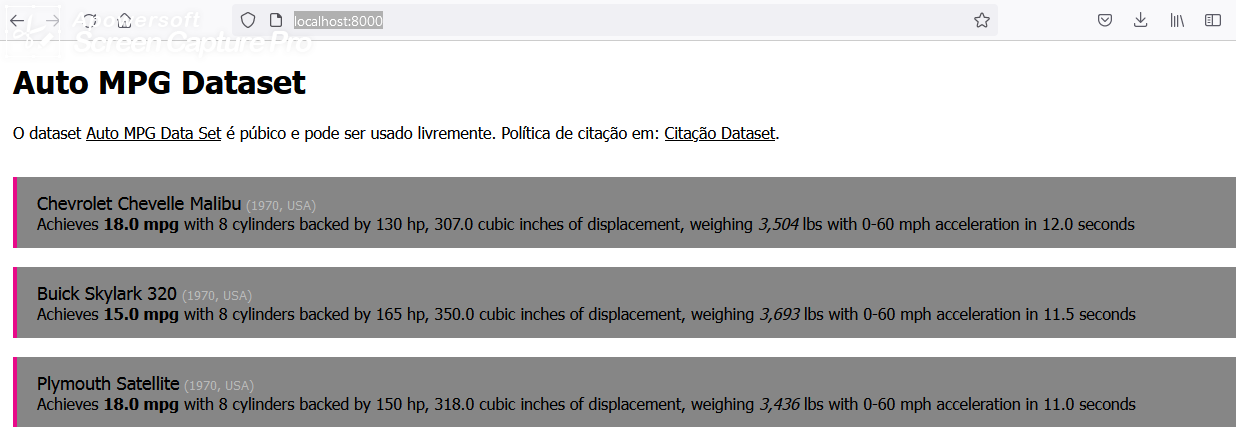


Quando eu chegar nesse nível de pasta eu terei que inicializar o servidor web pelo computador dentro dessa pasta (cmd) utilizando o seguinte caminho: **python -m http.server**



Isso mostra que o servidor foi inicializado. Agora é só abrir a pasta e digitar:

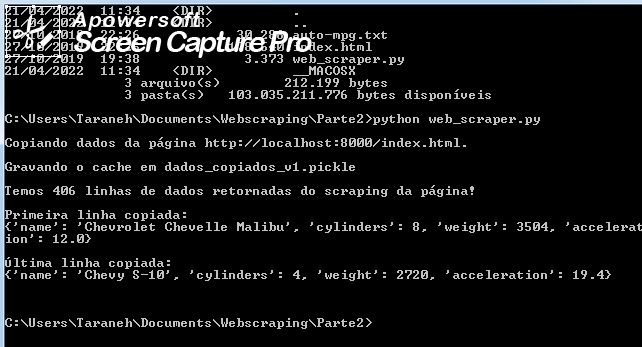
<http://localhost:8000/>



Isso mostra que o passo foi executado com sucesso.

Agora vamos abrir uma nova janela CMD (abrir como adm) e chegar até onde está a pasta index.

Digitar : **python web\_scraper.py**



Isso mostra o resultado da etapa executada com sucesso.

1-) A cópia dessa página da internet gerou um arquivo.csv.

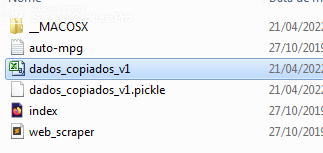
2-) A quantidade de linhas que foram copiadas.(406)

3-) A primeira e última linha também copiadas.

Gera um dicionário {} com chave e valor.

Resultando em um arquivo chamado **dados\_copiados\_v1** na mesma pasta (PARTE2)

O arquivo



O arquivo **web\_scraper.py** mostra como foi feito esse script para a cópia desse documento.

Vamos abrir agora um Jupyter Notebook e abrir o arquivo web\_scraper.py

<http://localhost:8888/edit/Documents/Webscraping/Parte2/web_scraper.py>

