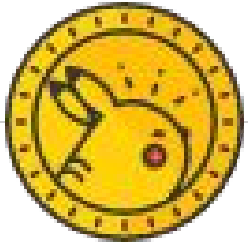
Web Scraper



Hey !

Eu sou Walter

Cursando de Sistemas recém apresentado as maravilhas do Python. Você pode me encontrar em overfront.netne.net

0

Web Scraper Vs Crawler

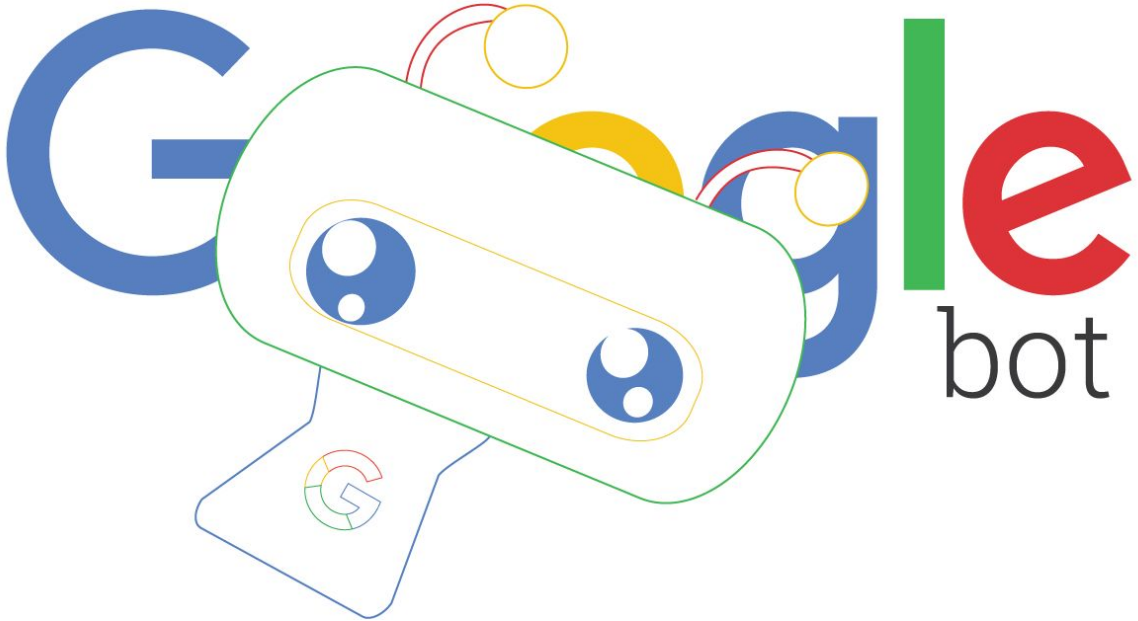
Qual é a diferença ?

Web Crawler

Programa que rastreia automaticamente de forma sistemática e recursiva a web, indexando os links com base nos seeds fornecidos em sua inicialização.

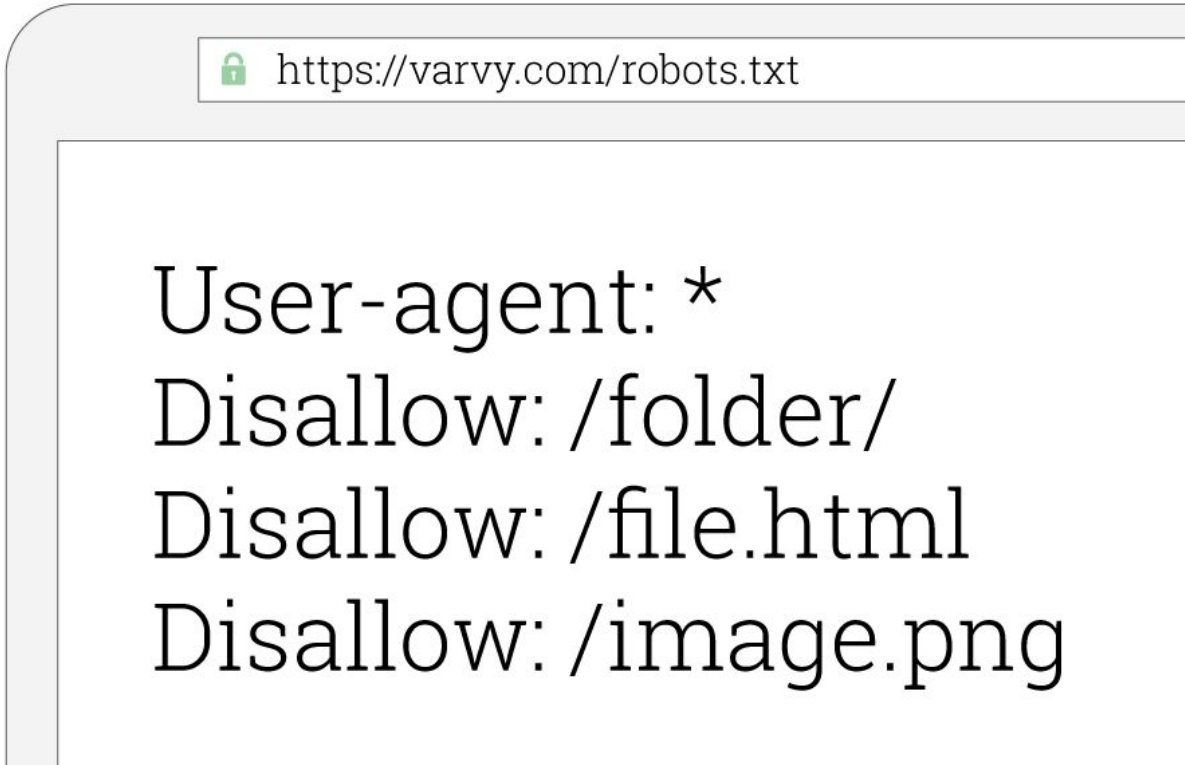
Google Translate

‘ Correia fotorreceptora da correia fotorreceptora ‘



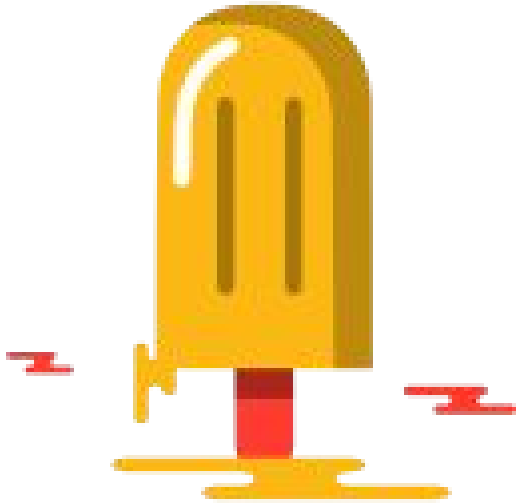
Melhor exemplo de Web Crawler : Google Bot

Como Bloquear



Web Scraper

Extração de dados da web realizada de forma automática com auxílio de Crawlers ou manual pelo browser e ou diretamente pelo protocolo Http



Um web scraper é uma Api para extrair dados de um web site.

Principais Aplicações

Mineração de Dados

Monitoramento

Diversão

Google & Amazon

As duas empresas fornecem Web Scrapers

Free (:

Google

Scraping.compunect.com/?scrape-google-search

Amazon

Github.com/adamlwgriffiths/amazon_scraper

Obs : Ainda que forneçam gratuitamente algumas tem um limite, como foi dito na talk do Marco Tulio

Youtu.be/PS8dm2NiDtg?t=187

A Rede Social - Filme

Qual é o poder que você tem ?

Cuidados

- Violação de direitos autorais
- Violação do computador, fraude e abuso

Let’s Code

Vamos obter a cotação da bitcoin e a quantidade de bitcoins de uma carteira na block chain

Requisitos

- Python
- Requests (Pip install ...)
- Adicional : Orientação a Objetos + Tk Inter

1 º Passo - Import

Tkinter é uma biblioteca da linguagem Python que acompanha a instalação padrão e permite desenvolver interfaces gráficas.

Requests também é uma biblioteca, porém não é padrão e tem como foco ‘ Http para humanos ‘.

```
from Tkinter import *

import requests as rq
```

Formas Recomendadas

Import Tkinter

From Tkinter import Tk

2 º Passo - Dados

Vamos obter os dados a partir de três parâmetros : site, tag html que indica onde o conteúdo começa & o número de caracteres que desejamos obter. Obtemos o código html da página e convertemos o mesmo para texto para que seja possível utilizar a função index e então encontrar o início do texto que juntamente com o número de caracteres e aplicação de slicing nós é retornado o conteúdo em questão.

```
def data ( site, param, houses ) :

    if ( not site.startswith ( 'http://' ) ):

        site = 'http://' + site

    html = rq.get ( site )

    html = html.text

    posi = html.index ( param ) + len ( param )

    return html [ posi : posi + houses ]
```

Alternativa Expressão Regular

```
# import re; bitcoin = re.findall ( r'\d\.\d\d', html.text ); bitcoin [2]
```

3 º Passo - Instância

Chamamos a função e então criamos duas variáveis que vão ser utilizadas posteriormente para mostrar os valores na tela a partir da interface gráfica.

```
bitcoin = data ( 'dolarhoje.com/bitcoin/', 'id="nacional" value="', 6 )

blockch = data ( 'blockchain.info/address/1QAc9S5EmycqjzzWDc1yiWzr9jJLC8sLiY', '<span data-c="', 9 )
```

4 º Passo - Output

Utilizando a interface gráfica Tkinter criamos uma classe juntamente com seus widgets e então utilizamos a mesma para mostrar a informação obtida nos passos anteriores.

```
class aplication ( object ) :

    def __init__ ( self, root, btc, blk ) :

        self.fr = Frame ( root ).pack ()
```

```
self.wd = Label ( self.fr, text = btc + ' Bitcoin', width = 25, height = 10, bg = 'grey16', fg = 'white' ).pack ( side = LEFT )
self.dw = Label ( self.fr, text = blk + ' Blockchain', width = 25, height = 10, bg = 'grey16', fg = 'white' ).pack ( side = RIGHT )
```

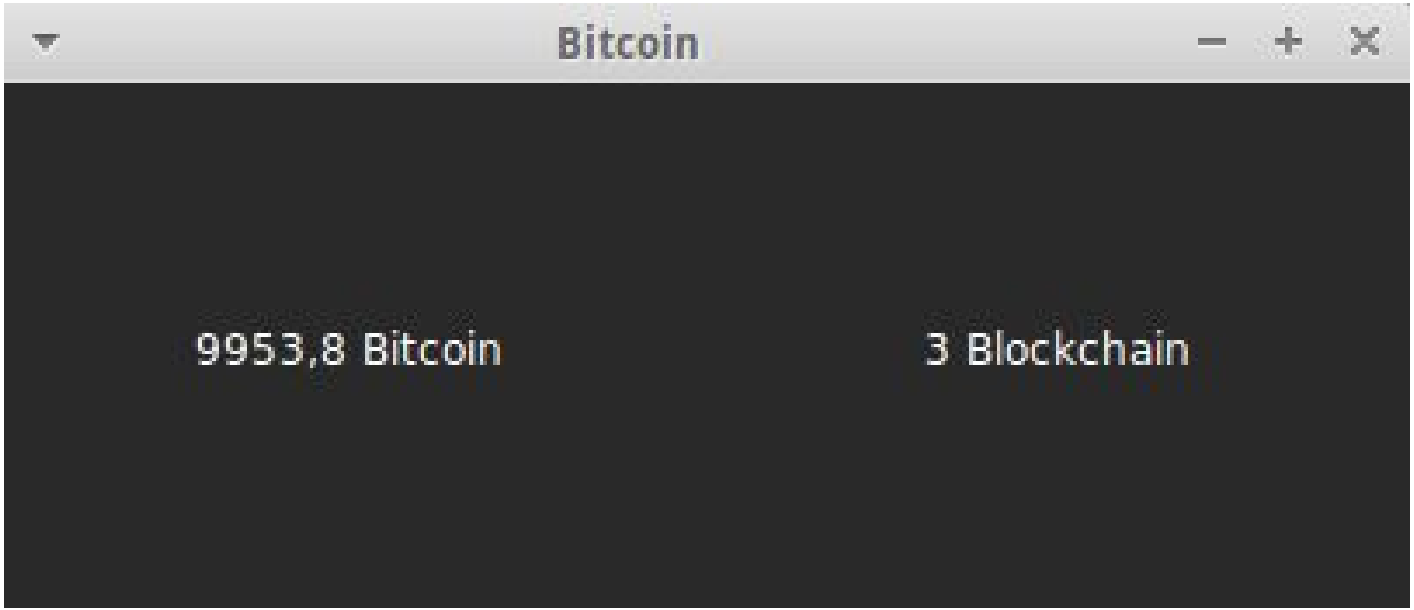
```
root = Tk ()


root.wm_title ( 'Bitcoin' )

aplication ( root, bitcoin, blockch )


root.mainloop ()
```

5 º Passo - Interpretamos



Código

Bit.ly/2tjqONT

Desafios

Buscar os valores de uma Api ou fazer a implementação atual ficar em real time

Utilização de regex na busca dos valores no ‘ arquivo ‘ html

Criar a Gui com Kivy para que seja multi plataforma

Libs

Beautiful Soup is a Python library for pulling data out of HTML and XML files.

Alternativas

Py Spider

lxml

Scrapy

Github.com/BruceDone/awesome-crawler

Thanks !

Alguma dúvida ?

- Overfront.netne.net.