

BeautifulSoup

Expressões Regulares com *BeautifulSoup*

Expressões Regulares com BeautifulSoup

Podemos utilizar expressões regulares com o *BeautifulSoup* para realização de *web scraping*.

A maioria das funções que recebe uma *string* como argumento também pode receber uma expressão regular.

Expressões Regulares com BeautifulSoup

Vamos utilizar como exemplo uma página do site disponibilizado pelo autor do livro ***Web Scraping with Python***.

O endereço da página de exemplo é:

<http://www.pythonscraping.com/pages/page3.html>

Expressões Regulares com BeautifulSoup

← → ↻ ⓘ www.pythonscraping.com/pages/page3.html



Totally Normal Gifts






Here is a collection of totally normal, totally reasonable gifts that your friends are sure to love! Our collection is ha

We haven't figured out how to make online shopping carts yet, but you can send us a check to:

123 Main St.

Abuja, Nigeria

We will then send your totally amazing gift, pronto! Please include an extra \$5.00 for gift wrapping.

Item Title	Description	Cost	Image
Vegetable Basket	This vegetable basket is the perfect gift for your health conscious (or overweight) friends! <i>Now with super-colorful bell peppers!</i>	\$15.00	
Russian Nesting Dolls	Hand-painted by trained monkeys, these exquisite dolls are priceless! And by "priceless," we mean "extremely expensive"! <i>8 entire dolls per set! Octuple the presents!</i>	\$10,000.52	
Fish Painting	If something seems fishy about this painting, it's because it's a fish! <i>Also hand-painted by trained monkeys!</i>	\$10,005.00	
Dead Parrot	This is an ex-parrot! <i>Or maybe he's only resting?</i>	\$0.50	
Mystery Box	If you love surprises, this mystery box is for you! Do not place on light-colored surfaces. May cause oil staining. <i>Keep your friends guessing!</i>	\$1.50	

Nesta página de exemplo existem imagens de produtos que estão definidas com a seguinte tag:

```

```

Porém, existe uma imagem que não é imagem de produto, que é a:

```

```

Se nossa intenção é retornar todas imagens de produtos e buscarmos por tags `img` (`.findall("img")`), teremos uma imagem que não é de um produto.

Quando estivermos realizando scraping podemos nos deparar com uma situação parecida, porque em uma página podem existir diversas imagens além das que precisamos. Sejam imagens ocultas, imagens em branco utilizadas para espaçamento e alinhamento de elementos, etc.

Expressões Regulares com BeautifulSoup

Neste exemplo identificamos uma forma que diferencia imagens de produtos de outras imagens. As imagens dos produtos possuem no nome as letras “*img*” seguidas de um número, o que as diferencia da imagem “logo.jpg”.

Expressões Regulares com BeautifulSoup

Este é o caminho da imagem de um produto:

```

```

Podemos utilizar a seguinte expressão regular para encontrar estas imagens:

```
"\.{2}/img/gifts/img\d*\.jpg"
```

Onde:

\. = Representa o caractere "." literalmente .

{2} = Duas ocorrências do caractere anterior (ponto)

/img/gifts/img = String literal.

\d = Representa dígito de zero a nove.

* = Zero ou mais ocorrências do caractere anterior (no caso, zero ou mais ocorrências de um número de 0 a 9).

\.jpg = Ponto literal e a string jpg.

Expressões Regulares com BeautifulSoup

Executando o exemplo

```
from urllib.request import urlopen
from bs4 import BeautifulSoup
import re

html = urlopen("http://www.pythonscraping.com/pages/page3.html")
soup = BeautifulSoup(html, "html.parser")

imagens = soup.findAll("img", {"src":re.compile("\.{2}/img/gifts/img\d*\.{2}.jpg")})

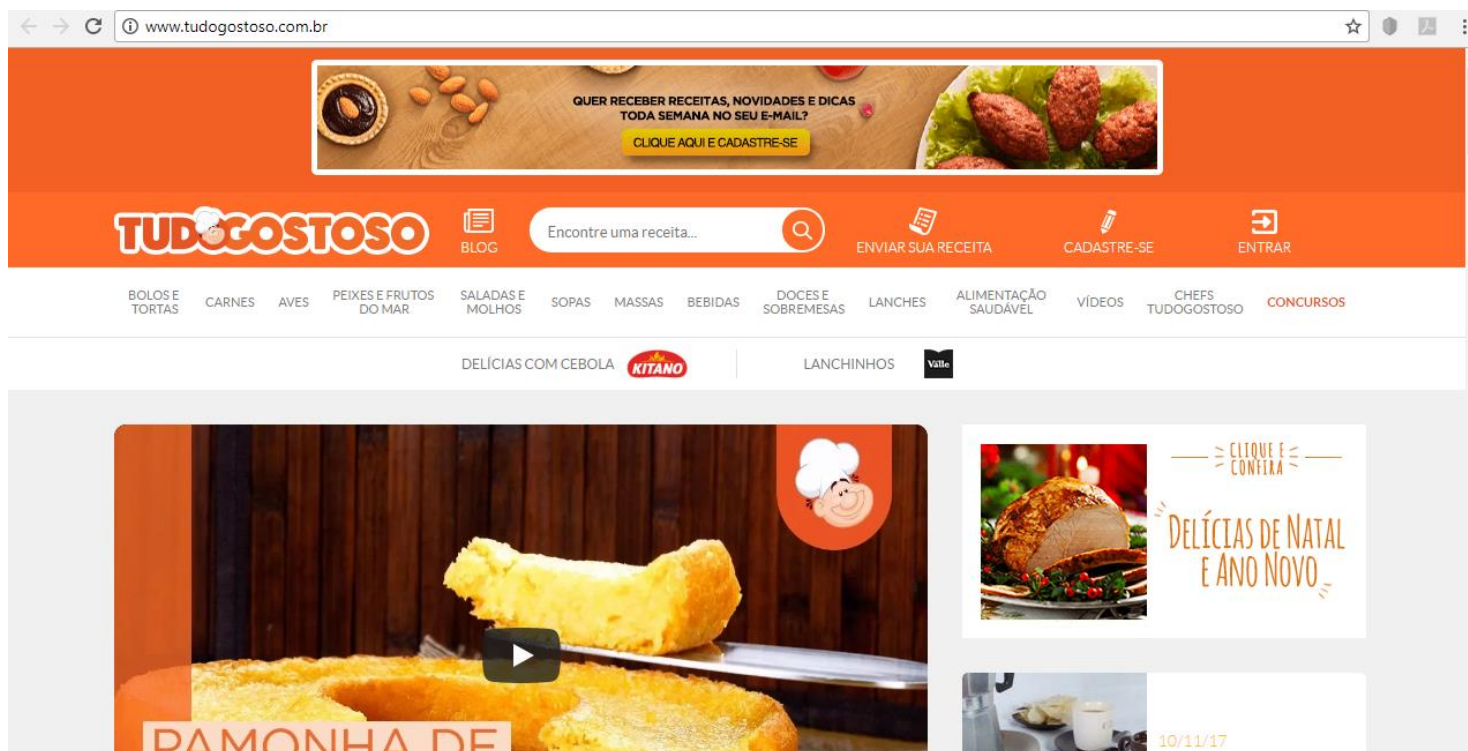
for img in imagens:
    print(img["src"])
```

Resultado:

```
../img/gifts/img1.jpg
../img/gifts/img2.jpg
../img/gifts/img3.jpg
../img/gifts/img4.jpg
../img/gifts/img6.jpg
```

Expressões Regulares com BeautifulSoup

Segundo exemplo:



Retornando links de categorias de um site de receitas.

Expressões Regulares com BeautifulSoup

Executando o exemplo:

```
from urllib.request import urlopen, Request
from bs4 import BeautifulSoup
import re

# Erro 403 porque alguns sites tratam scraping
#html = urlopen("http://www.tudogostoso.com.br")

req = Request("http://www.tudogostoso.com.br", headers={'User-Agent': 'Mozilla/5.0'})
html = urlopen(req).read()

soup = BeautifulSoup(html, "html.parser")

links = soup.findAll("a", {"href":re.compile("/categorias/.*\\.php")})

for link in links:
    print(link["href"])
```

Resultado

```
/categorias/bolos-e-tortas-doces.php
/categorias/carnes.php
/categorias/aves.php
/categorias/peixes-e-frutos-do-mar.php
/categorias/saladas-molhos-e-acompanhamentos.php
/categorias/sopas.php
/categorias/massas.php
/categorias/bebidas.php
/categorias/doces-e-sobremesas.php
/categorias/lanches.php
/categorias/alimentacao-saudavel.php
```

Expressões Regulares com BeautifulSoup

Terceiro exemplo:

The screenshot shows the homepage of the Folha de São Paulo website. The main content area features several news articles with headlines and images. On the right side, there is a sidebar with a list of top stories, each with a rank number and a brief description. Below the list is a 'VER ÍNDICE' link. At the bottom right, there is a 'folhashop' section with a search bar and a product listing for a Samsung LED 65 inch TV.

URL: <https://www1.folha.uol.com.br/poder/2017/11/1935507-globo-pede-a-funcionarios-que-se-definam-sobre-a-eleicao.shtml>

Buscar links do site Folha de São Paulo que apontem para o próprio `folha.uol.com.br`, da categoria “Mercado”, do ano de 2017, mês 11.

Expressões Regulares com BeautifulSoup

Executando o exemplo:

```
from urllib.request import urlopen, Request
from bs4 import BeautifulSoup
import re

html = urlopen("https://www.folha.uol.com.br/")
soup = BeautifulSoup(html, "html.parser")
links = soup.findAll("a", {"href": re.compile(".*\.folha\.uol\.com\.br/mercado/2017/11/.*\.shtml")})

for link in links:
    print(link["href"])
```

Resultado:

```
//www1.folha.uol.com.br/mercado/2017/11/1935512-trabalhador-teria-de-contribuir-44-anos-para-ter-teto-da-aposentadoria.shtml
//www1.folha.uol.com.br/mercado/2017/11/1935511-instrucao-maior-eleva-fosso-salarial-entre-branco-e-negro.shtml
//www1.folha.uol.com.br/mercado/2017/11/1935513-judiciario-do-rio-recebe-auxilio-peru-de-r-2000.shtml
//www1.folha.uol.com.br/mercado/2017/11/1935604-nao-me-pagaram-pelo-meu-trabalho-a-inusitada-queixa-trabalhista-deixada-em-roupas-da-zara-na-turquia.shtml
```

FIM