Fatec
Rio Claro

Linguagem de Programação II

Prof. Orlando Saraiva Júnior
orlando.nascimento@fatec.sp.gov.br

Pilares do Paradigma

Fatec
Rio Claro

O paradigma é sustentado por quatro pilares principais:

Abstração: focar nas características essenciais do objeto, ignorando detalhes irrelevantes.

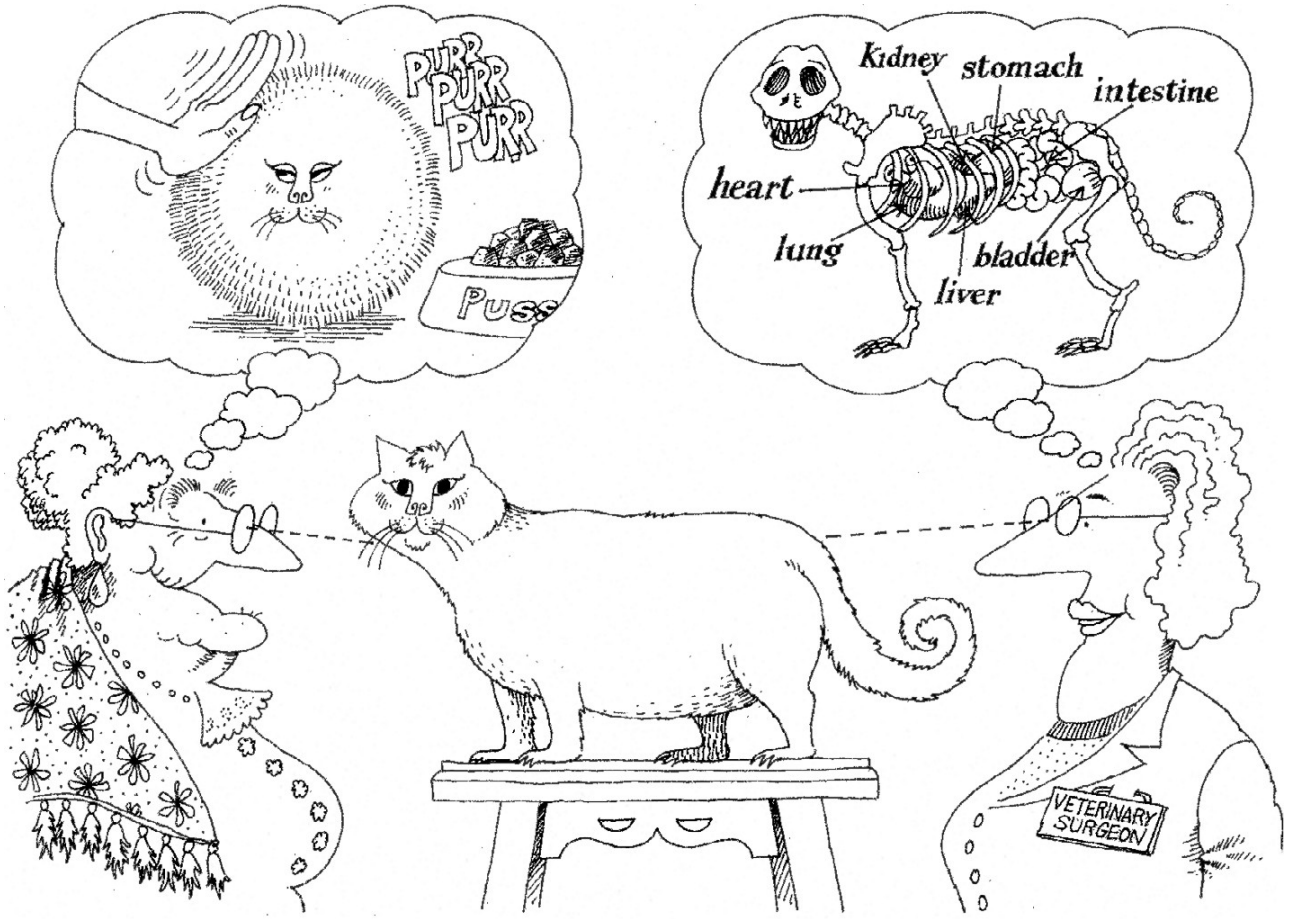
Encapsulamento: proteger os dados, controlando como são acessados ou modificados.

Herança: permitir que uma classe herde atributos e métodos de outra, facilitando o reuso de código.

Polimorfismo: possibilitar que diferentes objetos respondam de formas distintas a uma mesma mensagem/método.

Pilares do Paradigma

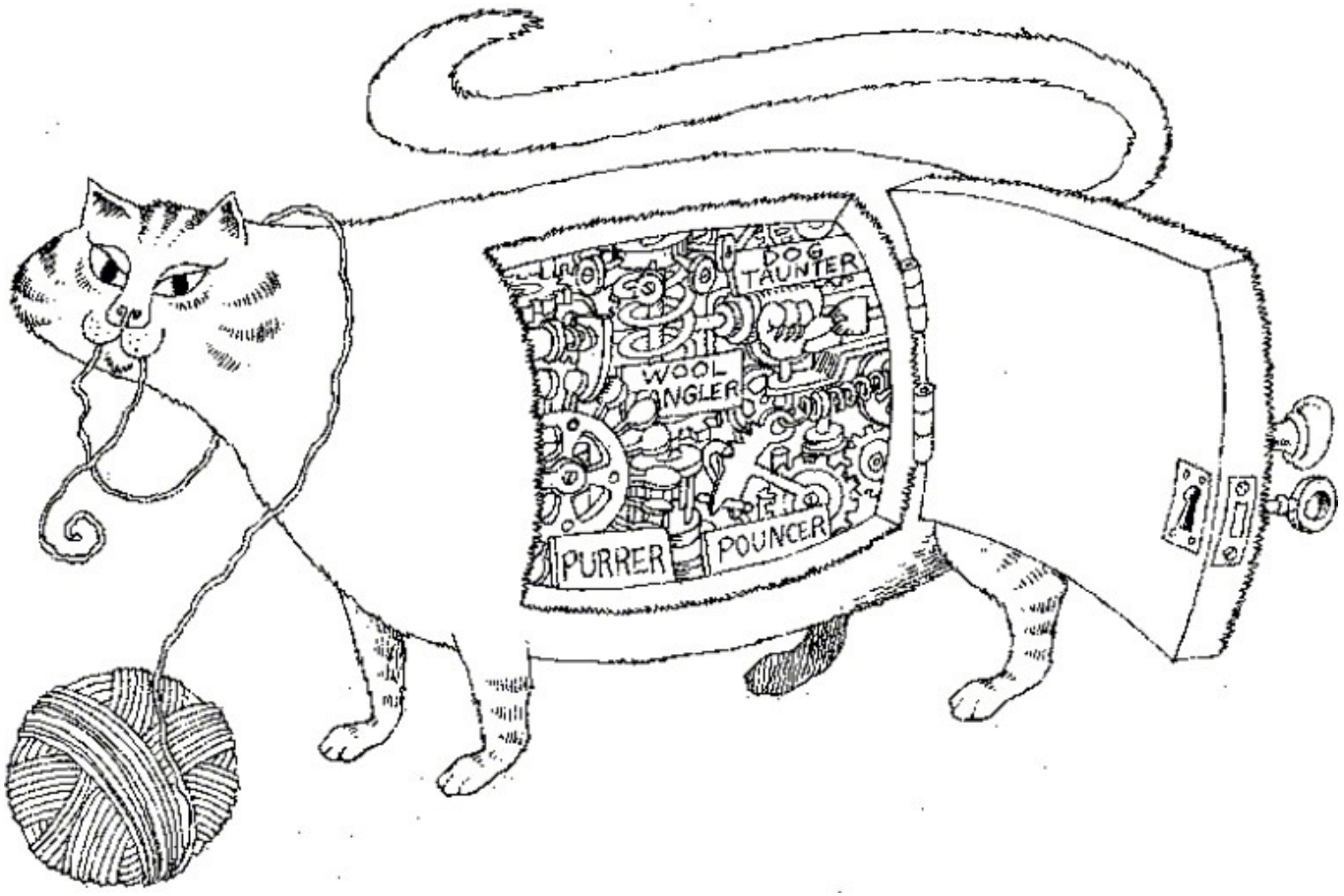
Fatec
Rio Claro



Abstração: focar nas características essenciais do objeto, ignorando detalhes irrelevantes.

Pilares do Paradigma

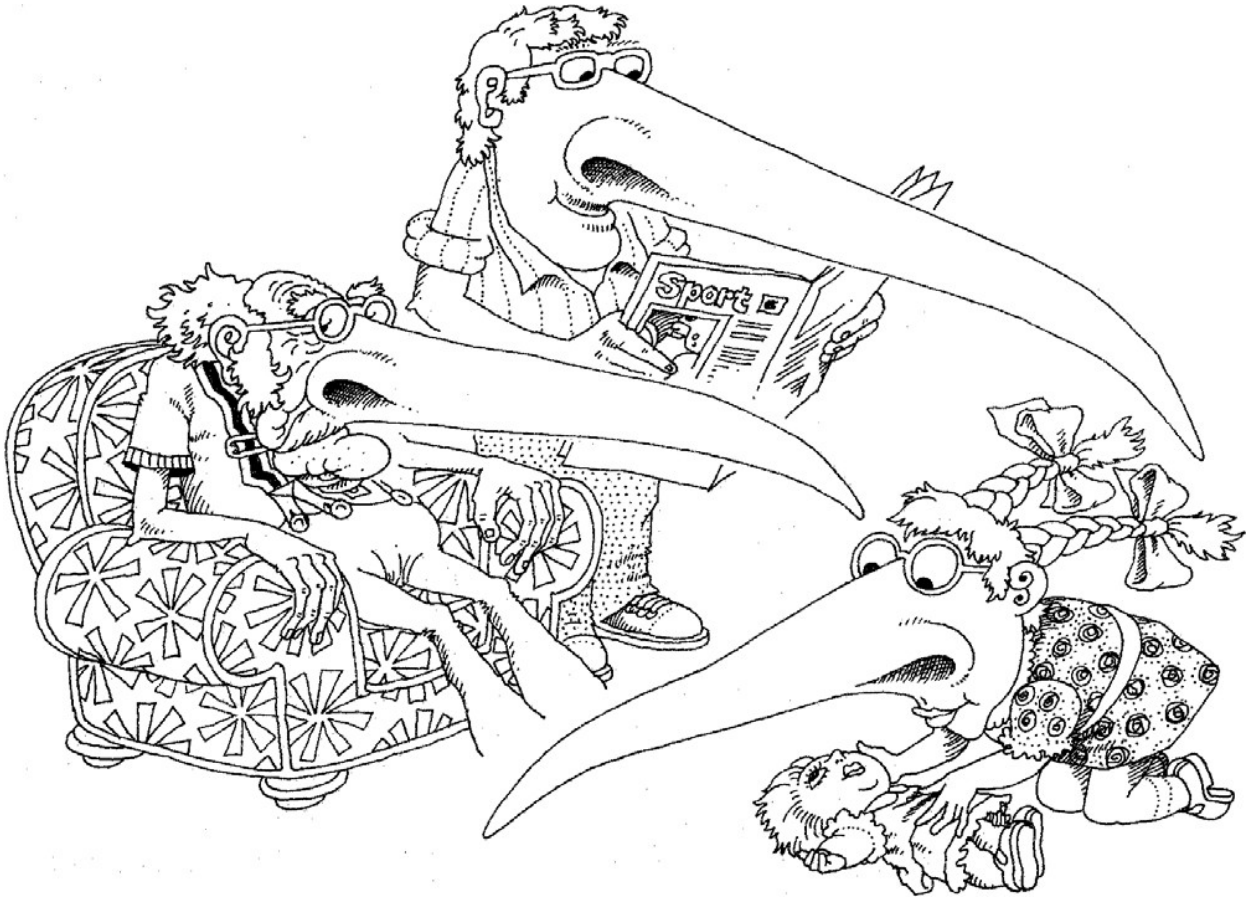
Fatec
Rio Claro



Encapsulamento: proteger os dados, controlando como são acessados ou modificados.

Pilares do Paradigma

Fatec
Rio Claro



Herança: permitir que uma classe herde atributos e métodos de outra, facilitando o reuso.

Web Scraping

Web Scraping

Fatec
Rio Claro

Web scraping é o termo usado para descrever o uso de um programa para baixar e processar conteúdo da web.

Por exemplo, o Google utiliza diversos programas de *web scraping* para indexar páginas da web para seu mecanismo de busca.

Existem diversos módulos que facilitam a raspagem de páginas da web em Python.

Web Scraping

Fatec
Rio Claro

Webbrowser: Vem com Python e abre um navegador em uma página específica.

Requests: Baixa arquivos e páginas da internet.

Beautiful Soup: Analisa HTML, o formato em que as páginas da web são escritas.

Prática

Fatec
Rio Claro

```
>>> import webbrowser
>>> webbrowser.open('http://inventwithpython.com/')
```

WebBrowser

Fatec
Rio Claro

Uma aba do navegador web será aberta no URL http://inventwithpython.com/. Esta é praticamente a única coisa que o módulo webbrowser pode fazer.

Mesmo assim, a função open() possibilita algumas coisas interessantes.

Por exemplo, é tedioso copiar um endereço de rua para a área de transferência e exibir um mapa dele no Google Maps. Você pode simplificar essa tarefa escrevendo um script simples para iniciar automaticamente o mapa no seu navegador usando o conteúdo da sua área de transferência. Dessa forma, você só precisa copiar o endereço para a área de transferência e executar o script, e o mapa será carregado para você.

Prática

Fatec
Rio Claro

```
#! python3
# mapIt.py - Launches a map in the browser using an address from the
# command line or clipboard.
import webbrowser, sys, pyperclip
if len(sys.argv) > 1:
    # Get address from command line.
    address = ' '.join(sys.argv[1:])
else:
    # Get address from clipboard.
    address = pyperclip.paste()
webbrowser.open('https://www.google.com/maps/place/' + address)
```

WebBrowser

Fatec
Rio Claro

Se não houver argumentos de linha de comando, o programa assumirá que o endereço está armazenado na área de transferência.

Você pode obter o conteúdo da área de transferência com pyperclip.paste() e armazená-lo em uma variável chamada address.

Por fim, para iniciar um navegador com o URL do Google Maps, chame webbrowser.open().

Request

Fatec
Rio Claro

O módulo requests foi escrito porque o módulo urllib2 do Python é muito complicado de usar. Se precisar baixar coisas da web, basta usar o módulo requests.

Prática

Fatec
Rio Claro

```
>>> import requests
>>> res = requests.get('https://automatetheboringstuff.com/files/rj.txt')
>>> type(res)
>>> res.status_code == requests.codes.ok
>>> len(res.text)
>>> print(res.text[:250])
```

Request

Fatec
Rio Claro

A URL leva a uma página web com a peça completa de Romeu e Julieta, disponível no site do livro.

Você pode verificar se a solicitação para esta página web foi bem-sucedida verificando o atributo status_code do objeto Response.

Se for igual ao valor de requests.codes.ok, então tudo correu bem.

Se a solicitação for bem-sucedida, a página da web baixada será armazenada como uma string na variável text do objeto Response.

Essa variável contém uma grande string com a peça inteira; a chamada para len(res.text) mostra que ela tem mais de 178.000 caracteres. Por fim, a chamada para print(res.text[:250]) exibe apenas os primeiros 250 caracteres.

Prática

Fatec
Rio Claro

```
>>> import requests
>>> r = requests.get('http://inventwithpython.com/page_that_does_not_exist')
>>> r.raise_for_status()
try:
    r.raise_for_status()
except Exception as exc:
    print('There was a problem: %s' % (exc))
```

Request

Fatec
Rio Claro

A partir daqui, você pode salvar a página da web em um arquivo no seu disco rígido com a função padrão open() e o método write().

No entanto, existem algumas pequenas diferenças. Primeiro, você deve abrir o arquivo em modo binário de escrita, passando a string 'wb' como segundo argumento para open().

Mesmo que a página esteja em texto simples (como o texto de Romeu e Julieta que você baixou anteriormente), você precisa escrever dados binários em vez de dados de texto para manter a codificação Unicode do texto.

Prática

Fatec
Rio Claro

```
>>> import requests
>>> res = requests.get('https://automatetheboringstuff.com/files/rj.txt')
>>> res.raise_for_status()
>>> playFile = open('RomeoAndJuliet.txt', 'wb')
>>> for chunk in res.iter_content(100000):
        playFile.write(chunk)
>>> playFile.close()
```

Request

Fatec
Rio Claro

O método iter_content() retorna "pedaços" do conteúdo em cada iteração do loop. Cada pedaço é do tipo de dado bytes, e você pode especificar quantos bytes cada pedaço conterá. Cem mil bytes geralmente é um bom tamanho, então passe 100000 como argumento para iter_content().

O arquivo RomeoAndJuliet.txt agora existirá no diretório de trabalho atual. Observe que, embora o nome do arquivo no site fosse rj.txt, o arquivo no seu disco rígido tem um nome diferente. O módulo requests simplesmente lida com o download do conteúdo das páginas da web. Uma vez que a página é baixada, ela é simplesmente um dado no seu programa. Mesmo se você perder a conexão com a Internet após baixar a página da web, todos os dados da página ainda estarão no seu computador.

Request

Fatec
Rio Claro

Para revisar, aqui está o processo completo para baixar e salvar um arquivo:

1) Chame requests.get() para baixar o arquivo.
2) Chame open() com 'wb' para criar um novo arquivo no modo de gravação binária.
3) Realize loop com o método iter_content() do objeto Response.
4) Chame write() em cada iteração para gravar o conteúdo no arquivo.
5) Chame close() para fechar o arquivo.

Request

Fatec
Rio Claro

O método raise_for_status() é uma boa maneira de garantir que um programa pare caso ocorra um download incorreto.

Isso é bom: você quer que seu programa pare assim que ocorrer um erro inesperado. Se uma falha no download não for um problema para o seu programa, você pode encapsular a linha raise_for_status() com instruções try e except para lidar com esse caso de erro sem travar.

HTML

HTML

Fatec
Rio Claro

Linguagem de Marcação de Hipertexto (HTML) é o formato em que as páginas da web são escritas.

Um arquivo HTML é um arquivo de texto simples com a extensão .html. O texto nesses arquivos é cercado por tags, que são palavras entre colchetes angulares.

As tags informam ao navegador como formatar a página da web. Uma tag inicial e uma tag de fechamento podem envolver algum texto para formar um elemento.

O texto (ou HTML interno) é o conteúdo entre as tags inicial e final. Por exemplo, o seguinte HTML exibirá "Olá, mundo!" no navegador, com "Olá" em negrito:

HTML

Fatec
Rio Claro

Linguagem de Marcação de Hipertexto (HTML) é o formato em que as páginas da web são escritas. Um arquivo HTML é um arquivo de texto simples com a extensão .html. O texto nesses arquivos é cercado por tags, que são palavras entre colchetes angulares.

As tags informam ao navegador como formatar a página da web. Uma tag inicial e uma tag de fechamento podem envolver algum texto para formar um elemento.

O texto (ou HTML interno) é o conteúdo entre as tags inicial e final. Por exemplo, o seguinte HTML exibirá "Olá, mundo!" no navegador, com "Olá" em negrito:

```
<strong>Hello</strong> world!
```

Fatec Rio Claro

Existem muitas tags diferentes em HTML. Algumas dessas tags têm propriedades extras na forma de atributos dentro dos colchetes angulares. Por exemplo, a tag <a> envolve um texto que deveria ser um link. A URL para a qual o texto direciona é determinada pelo atributo href.

Veja um exemplo:

```
free <a href="http://inventwithpython.com">Python books</a>.
```

HTML

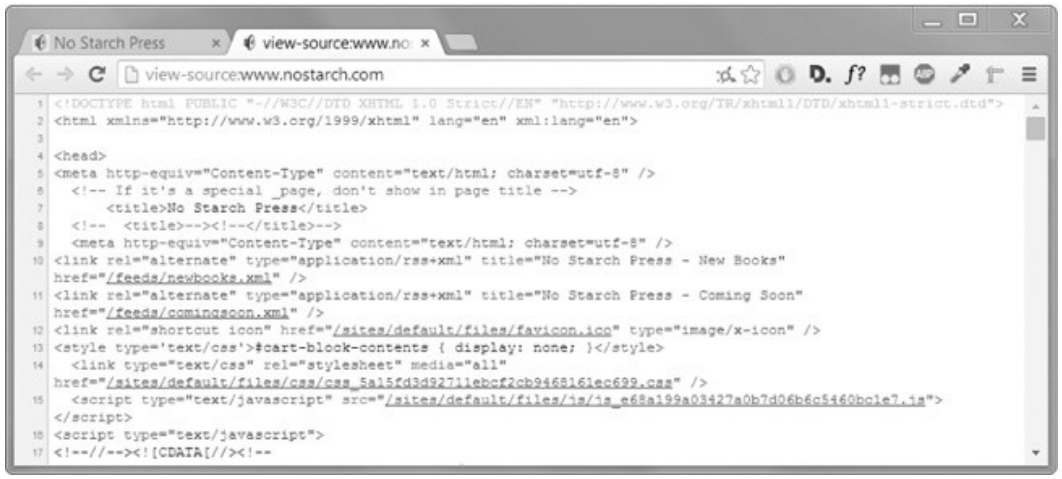
Fatec
Rio Claro

Você precisará verificar o código-fonte HTML das páginas da web com as quais seus programas trabalharão. Para isso, clique com o botão direito do mouse (ou pressione CTRL + clique no OS X) em qualquer página da web no seu navegador e selecione

Exibir Código-fonte ou Exibir código-fonte da página para ver o texto HTML da página. Este é o texto que o seu navegador realmente recebe. O navegador sabe como exibir, ou renderizar, a página da web a partir desse HTML.

HTML

Fatec
Rio Claro



Beautiful Soup

Fatec
Rio Claro

Beautiful Soup é um módulo para extrair informações de uma página HTML. O nome do módulo BeautifulSoup é bs4 (de Beautiful Soup, versão 4).

A função bs4.BeautifulSoup() precisa ser chamada com uma string contendo o HTML que será analisado. A função bs4.BeautifulSoup() retorna um objeto BeautifulSoup.

Prática

Fatec
Rio Claro

```
>>> import requests, bs4
>>> res = requests.get('http://nostarch.com')
>>> res.raise_for_status()
>>> noStarchSoup = bs4.BeautifulSoup(res.text)
>>> type(noStarchSoup)
>>> exampleFile = open('example.html')
>>> exampleSoup = bs4.BeautifulSoup(exampleFile)
>>> type(exampleSoup)
```

Beautiful Soup

Fatec
Rio Claro

O código usa requests.get() para baixar a página principal do site, em seguida, passa o atributo text da resposta para bs4.BeautifulSoup(). O objeto BeautifulSoup retornado é armazenado em uma variável chamada noStarchSoup.

Você também pode carregar um arquivo HTML do seu disco rígido passando um objeto File para bs4.BeautifulSoup().

Digite o seguinte no shell interativo (certifique-se de que o arquivo example.html esteja no diretório de trabalho):

Beautiful Soup

Fatec
Rio Claro

Você pode recuperar um elemento de página web de um objeto BeautifulSoup chamando o método select() e passando uma string de um seletor CSS para o elemento que você está procurando. ***Seletores*** são como expressões regulares: eles especificam um padrão a ser procurado, neste caso, em páginas HTML em vez de strings de texto comuns.

O método select() retornará uma lista de objetos ***Tag***, que é como o Beautiful Soup representa um elemento HTML. A lista conterá um objeto Tag para cada correspondência no HTML do objeto BeautifulSoup.

Os valores das tags podem ser passados para a função str() para exibir as tags HTML que eles representam. Os valores das tags também têm um atributo attrs que exibe todos os atributos HTML da tag como um dicionário.

Prática

Fatec
Rio Claro

```
>>> import requests, bs4
>>> res = requests.get('http://nostarch.com')
>>> res.raise_for_status()
>>> noStarchSoup = bs4.BeautifulSoup(res.text)
>>> for link in noStarchSoup.select('a'):
        print(str(link))
>>> tag =  noStarchSoup.select('a')[0]
>>> tag.attrs
```

Beautiful Soup

Fatec
Rio Claro

O objeto mais comumente usado na biblioteca BeautifulSoup é, apropriadamente, o objeto BeautifulSoup.

O código a seguir retorna apenas a primeira instância da tag h1 encontrada na página. Por convenção, apenas uma tag h1 deve ser usada em uma única página, mas as convenções costumam ser quebradas na web, então você deve estar ciente de que isso retornará apenas a primeira instância da tag, e não necessariamente aquela que você está procurando.

Prática

Fatec
Rio Claro

```
from urllib.request import urlopen

from bs4 import BeautifulSoup

html = urlopen('http://www.pythonscraping.com/pages/page1.html')

bs = BeautifulSoup(html.read(), 'html.parser')

print(bs.h1)
```

Beautiful Soup

Fatec
Rio Claro

Este conteúdo HTML é então transformado em um objeto BeautifulSoup com a seguinte estrutura:

html → <html><head>...</head><body>...</body></html>

head → <head><title>Uma Página Útil<title></head>

title → <title>Uma Página Útil</title>

body → <body><h1>Um Int...</h1> Lorem ip... </body>

h1 → <h1>Um Título Interessante</h1>

div → Lorem Ipsum dolor...

Beautiful Soup

Fatec
Rio Claro

Observe que a tag h1 que você extrai da página está aninhada em duas camadas na estrutura do objeto BeautifulSoup (html → body → h1). No entanto, ao buscá-la no objeto, você chama a tag h1 diretamente:

```
bs.h1
```

Na verdade, qualquer uma das seguintes chamadas de função produziria a mesma saída:

```
bs.html.body.h1
```

```
bs.body.h1
```

```
bs.html.h1
```

Mini-Projeto

Mini Projeto

Fatec
Rio Claro

Blogs e outros sites atualizados regularmente geralmente têm uma página inicial com a postagem mais recente, bem como um botão "Anterior" na página que leva à postagem anterior. Essa postagem também terá um botão "Anterior", e assim por diante, criando um caminho da página mais recente até a primeira postagem do site.

Se você quisesse uma cópia do conteúdo do site para ler quando não estivesse online, poderia navegar manualmente por todas as páginas e salvar cada uma.

Mas isso é um trabalho bem chato, então vamos escrever um programa para fazer isso.

Mini Projeto

Fatec
Rio Claro

XKCD é uma webcomic geek popular com um site que se encaixa nessa estrutura. A página inicial em http://xkcd.com/ tem um botão "Anterior" que guia o usuário de volta para as histórias em quadrinhos anteriores. Baixar cada história em quadrinhos manualmente levaria uma eternidade, mas você pode escrever um script para fazer isso em alguns minutos.

Veja o que o programa faz:

- Carrega a página inicial do XKCD.

- Salva a imagem da história em quadrinhos nessa página.

- Segue o link da história em quadrinhos anterior.

- Repete até chegar à primeira história em quadrinhos.

Mini Projeto

Fatec Rio Claro

Isso significa que o código precisará fazer o seguinte:

- Baixar páginas com o módulo requests.
- Encontrar a URL da imagem da história em quadrinhos de uma página usando o Beautiful Soup.
- Baixar e salvar a imagem da história em quadrinhos no disco rígido com iter_content().
- Encontrar a URL do link da história em quadrinhos anterior e repetir.

Prática

Fatec
Rio Claro

Analisar o código **xkcd.py**

Outras ideias para Mini Projeto

Fatec Rio Claro

- Baixar páginas e seguir links são a base de muitos programas de rastreamento da web. Programas semelhantes também podem fazer o seguinte:
- Fazer backup de um site inteiro seguindo todos os seus links.
- Copiar todas as mensagens de um fórum da web.
- Duplicar o catálogo de itens à venda em uma loja online.

Os módulos ***requests*** e ***BeautifulSoup*** são ótimos, desde que você consiga descobrir a URL que precisa passar para requests.get().

No entanto, às vezes isso não é tão fácil de encontrar. Ou talvez o site que você deseja que seu programa navegue exija que você faça login primeiro. O módulo ***Selenium*** dará aos seus programas o poder de executar essas tarefas sofisticadas.

Fatec
Rio Claro

Dúvidas

Prof. Orlando Saraiva Júnior
orlando.nascimento@fatec.sp.gov.br

Prática

Fatec
Rio Claro

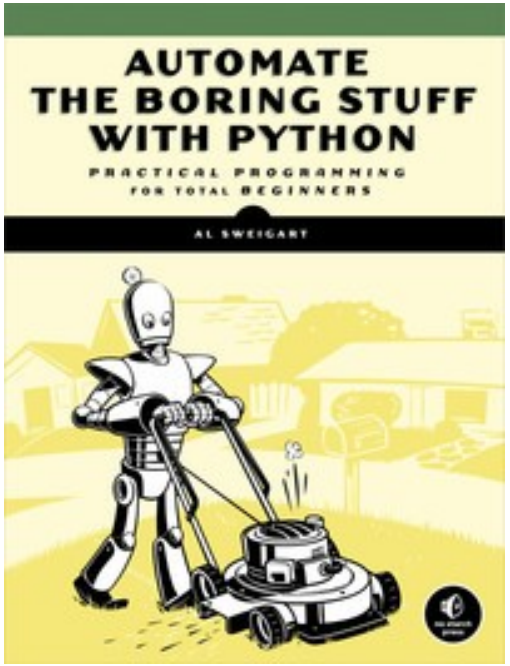
Desenvolva um mini projeto pessoal, demonstrando uma forma de realizar *Web Scraping*.

Faça uso das bibliotecas apresentadas, e outras bibliotecas que julgar importante adicionar.

Não esqueça de gerar o arquivo *requirements.txt* com as bibliotecas utilizadas por você.

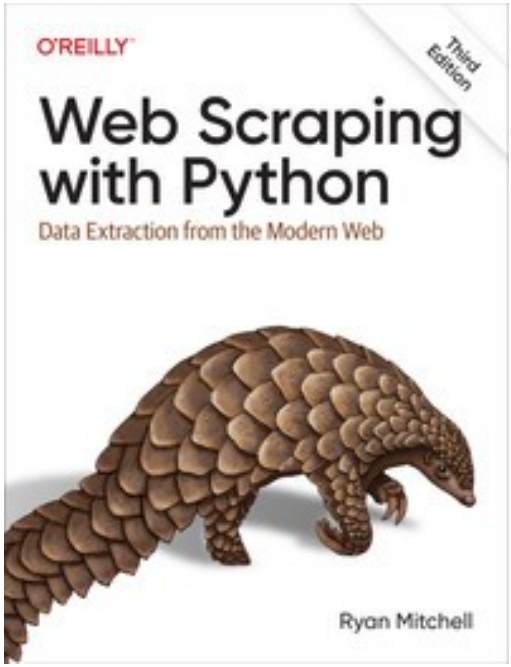
Bibliografia

SWEIGART, Al. Automate the boring stuff with Python. San Francisco: No Starch Press, 2015. 504 p.



Bibliografia

MITCHELL, Ryan. Web scraping with Python. 3. ed. Sebastopol: O'Reilly Media, 2024. 352 p.



Bibliografia

NAIR, Vineeth G. N. Getting Started with Beautiful Soup. Birmingham: Packt Publishing, 2014. 130 p.