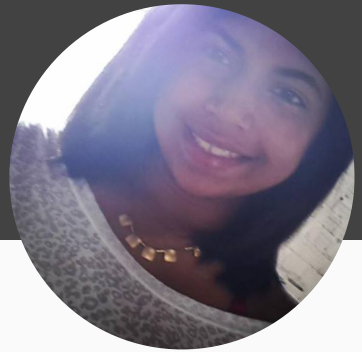
Raspagem de Dados

Workshop

About

PyLadies São Luís: mulheres líderes na comunidade de código aberto Python



Ana Clara Cavalcante

Técnica em Informática
Graduanda em Sistemas de Informação

Meiryanne Martins

Graduada em Ciência da Computação

Salete Farias

Mestra em Ciência da Computação
Doutoranda em Educação

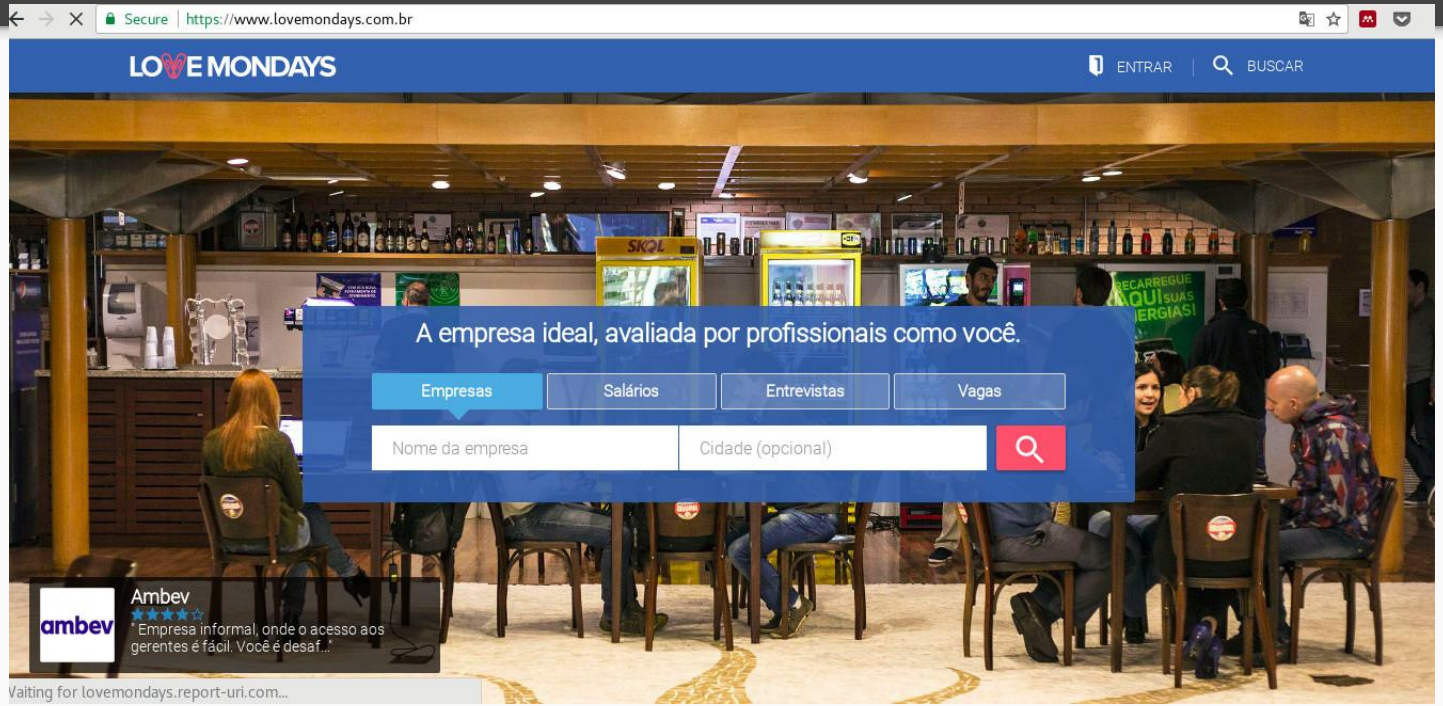
pyladies
São Luís

Raspagem de Dados:

Conceito: também conhecido como Data Scraping, é a atividade de extrair dados de sites e transportá-los para um formato mais simples e maleável para que possam ser analisados e cruzados com facilidade.

Exemplos - OPERAÇÃO SERENATA DE AMOR https://serenata.ai/

Exemplos - LOVE MONDAYS https://serenata.ai/



Etapas da Raspagem de Dados

- *Entender a fonte de dados*
- *Planejar como será a obtenção dos dados*
- *Programar a captura (montar os scripts)*
- *Verificar os dados capturados*
- *Armazenar em bancos de dados*
- *Pensar na visualização para o usuário final**

Onde obtenho dados?

- Dados Públicos
  - Dados Abertos Governamentais
    - metodologia para a publicação de dados do governo em formatos reutilizáveis, visando o aumento da transparência e maior participação política por parte do cidadão, além de gerar diversas aplicações desenvolvidas colaborativamente pela sociedade.
  - http://www.w3c.br/divulgacao/pdf/dados-abertos-governamentais.pdf

Onde obtenho dados?

- http://www.governoeletronico.gov.br/acoes-e-projetos/Dados-Abertos/como-abrir-os-dados
- Kit de Dados Abertos
  - http://kit.dados.gov.br/
  - Conversão de Dados (ex.: json para csv)
    - http://konklone.io/json/
- Manual dos Dados Abertos: Governo
  - **http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf**

Onde obtenho dados?

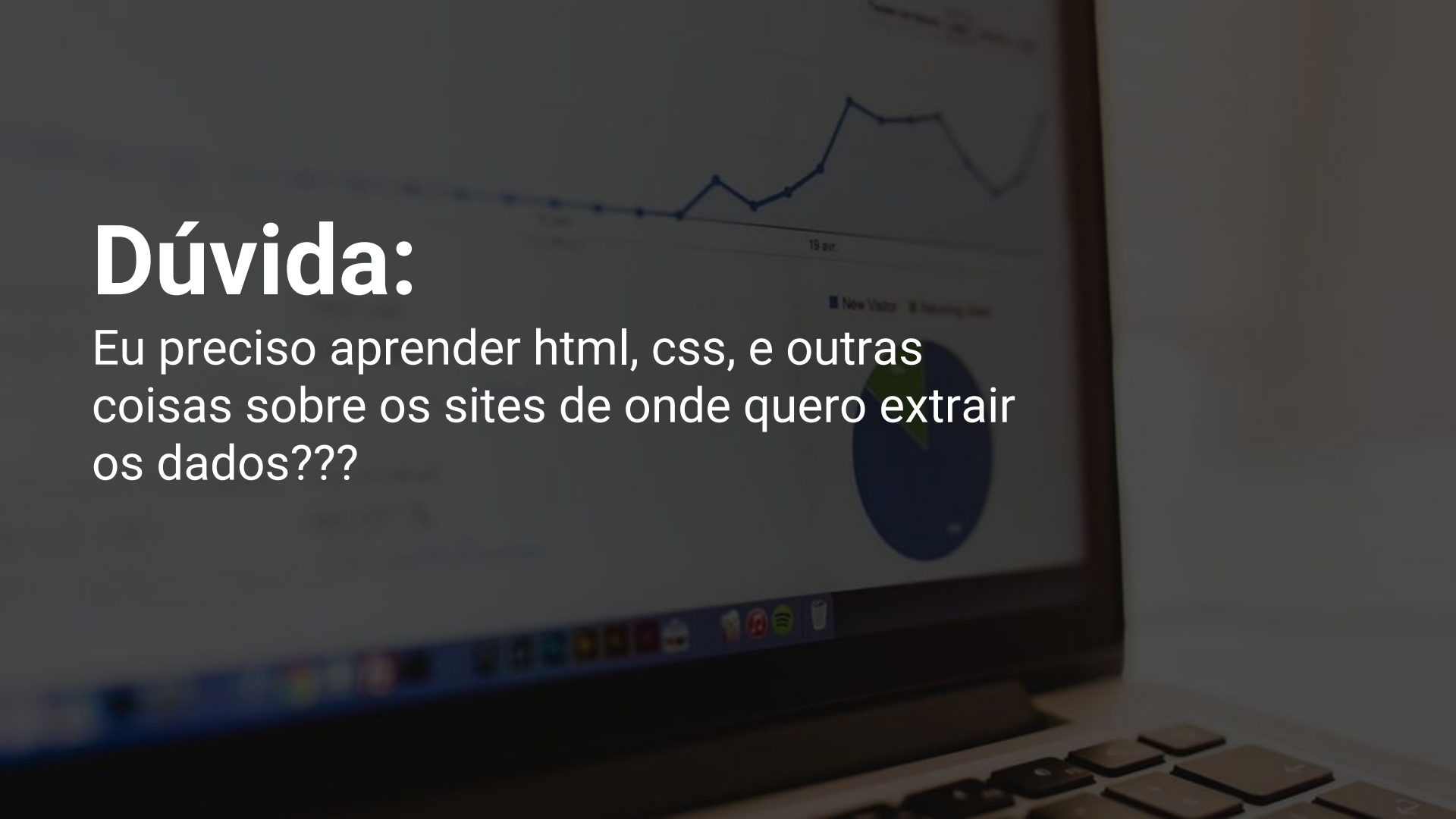
- Base de Dados - Formatos
  - XML
  - JSON
  - RDF
  - CSV

Onde obtenho dados?

Onde obtenho dados?

Por que Python?

- *Curva rápida de aprendizado*
- *Foco no problema, sem perder tempo na sintaxe*
- *Interativa*
- *Alta produtividade*
- *Baterias inclusas*
- *Comunidade livre, forte, diversificada, alegre e acolhedora*



Dúvida:

Eu preciso aprender html, css, e outras coisas sobre os sites de onde quero extrair os dados???

Biblioteca

Beautiful Soup

- *Nome do Poema da Alice no País das Maravilhas*
- *Tenta dar sentido àquilo que parece não ter sentido nenhum*
- *Quando coletamos dados na web eles vem como uma sopa caótica de dados, bagunçados*
- *A biblioteca transforma esses dados caóticos em objetos Python para manipular de acordo com as necessidades*

Bibliotecas

urllib

- Um módulo do Python que define funções e classes(Objetos) que manipulam URL(s), seja ela uma URL simples baseado no protocolo HTTP, envio de dados GET e POST download de arquivos, cookies, sessão etc.
- Para a manipulação de dados pela internet ou recursos para um servidor em intranet, a utilização deste módulo é essencial para o funcionamento de sua aplicação
-


Show me the
code

Bibliotecas

urllib

escolas.py - /home/saletefarias/Documents/IFMA/Aulas/2016-1/Python/Aulas/urllib/escolas....

File Edit Format Run Options Window Help

```
import urllib.request
import json
url = 'http://educacao.dadosabertosbr.com/api/escolas?nome='
escola = 'MARANHAO'
resp = urllib.request.urlopen(url+escola).read()
resp = json.loads(resp.decode('utf-8'))
for x in resp[1]:
  print (x['nome'])
  print ('Código:', x['cod'])
  print (x['cidade'], x['estado'])
  print (x['enemMediaGeral'], x['situacaoFuncionamentoTxt'])
  print ()
```

Bibliotecas

BeautifulSoup

14 lines (9 sloc) | 398 Bytes

Raw Blame History

```
1  import sys
2  import requests
3  from bs4 import BeautifulSoup
4
5  url = 'http://www2.correios.com.br/sistemas/rastreamento/resultado.cfm'
6
7  params = { 'objetos': sys.argv[1] }
8  headers = { 'Referer': 'http://www2.correios.com.br/sistemas/rastreamento/default.cfm' }
9
10 html = requests.post(url, data=params, headers=headers)
11
12 soup = BeautifulSoup(html.text, 'html.parser')
13 print(soup.strong.text)
```

Web Scrapping

Let’s code!!!!!

1-basicExample.py
2-beautifulSoup.py
script_basico1.py
script_basico2.py

red_or_green.py
http://www.pythonscraping.com/pages/warandpeace.html

Bibliotecas

urllib

- Códigos (prática)
  - copadomundo.py
  - escolas.py
  - escolas2.py
  - escolas3.py

Bibliotecas

Beautiful Soup

- Códigos (prática)
  - Empresas Internet
  - Educação Mundial

Obrigada!

saletefarias@gmail.com
anaclaracavalcante14@gmail.com@gmail.com
meirymartinsp.rodrigues@gmail.com@gmail.com