

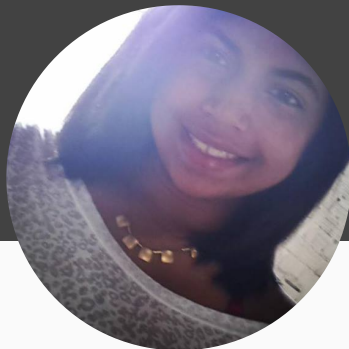
# Raspagem de Dados

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

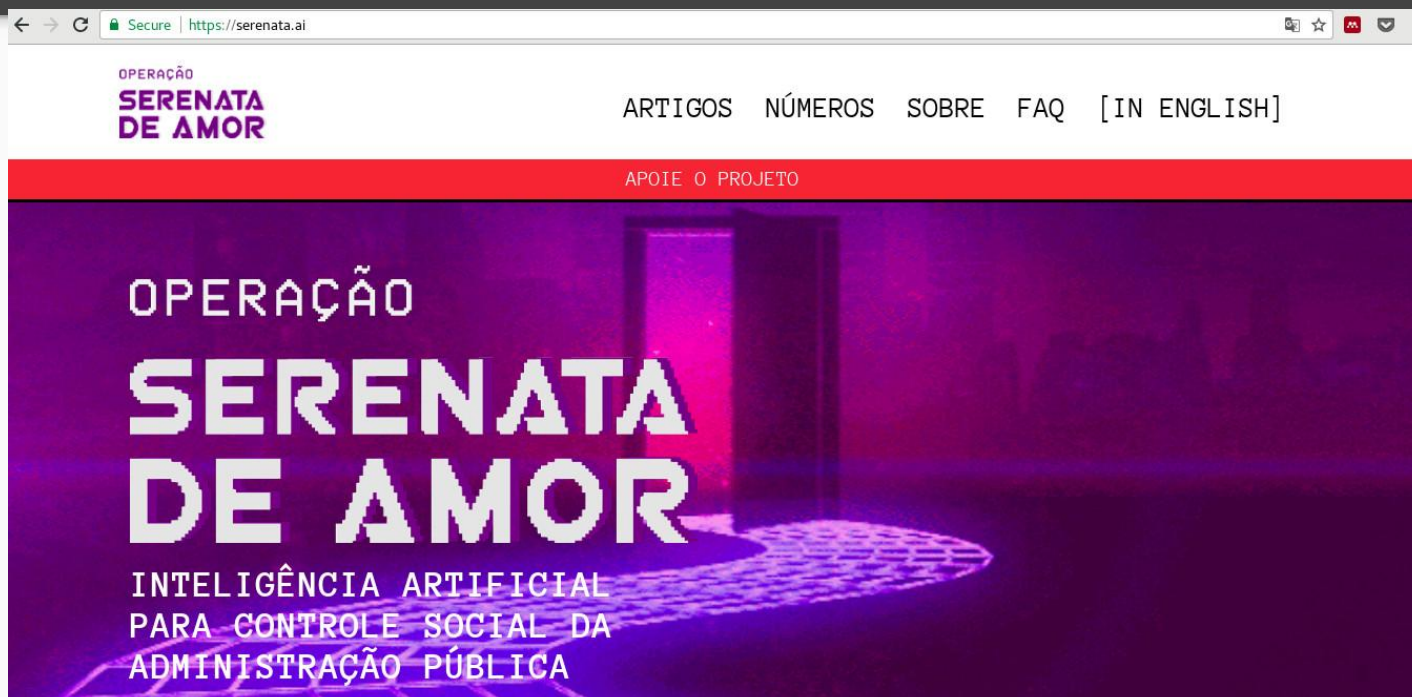


# Raspagem de Dados:

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.

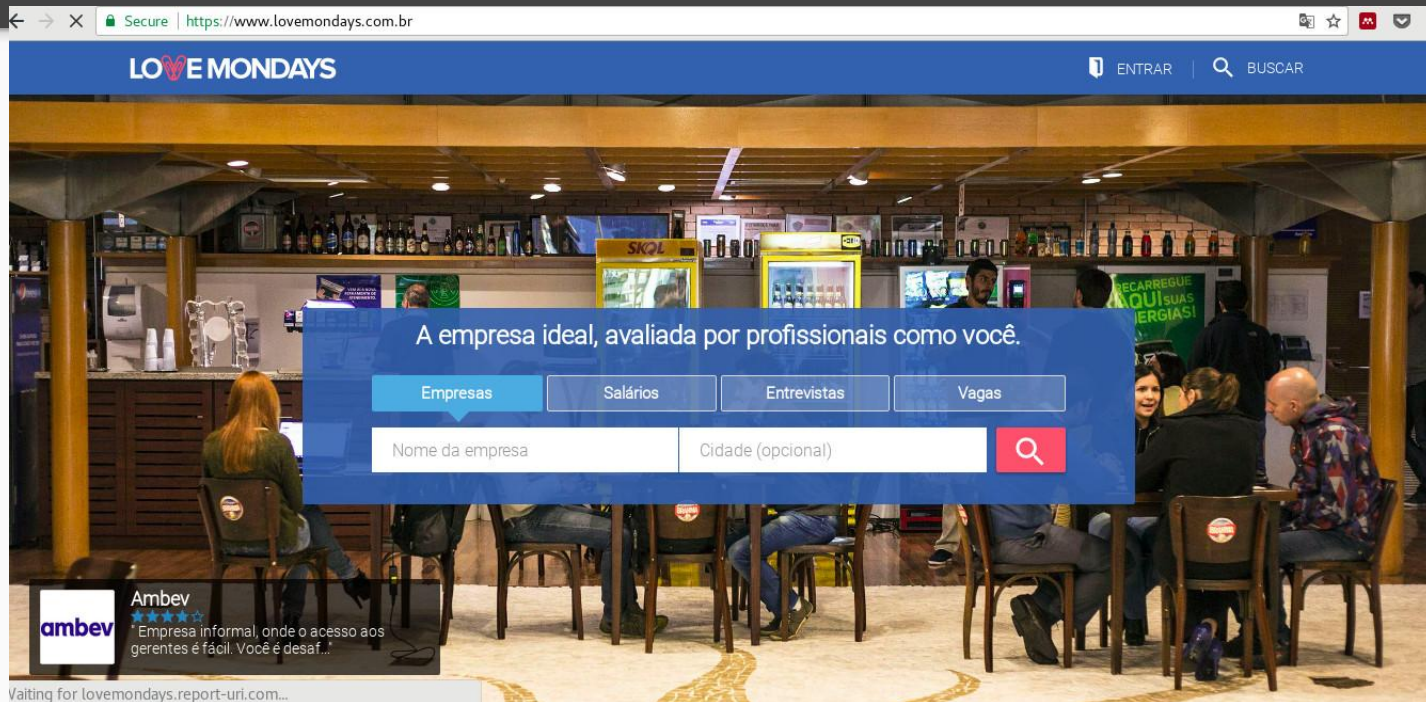
# Exemplos - OPERAÇÃO SERENATA DE AMOR

<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](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?

The screenshot displays the official website of the Portal da Transparência (Transparency Portal) of the Brazilian Government. The page features a green header with the portal's name and logo. Below the header, there is a navigation bar with links to 'Perguntas frequentes', 'Contato', 'Glossário', 'Links', and 'Manual de navegação'. The main content area is divided into several sections:

- CONSULTAS**: A sidebar on the left lists various data categories for consultation, including Despesas, Receitas, Convênios, Empresas Sancionadas, Empresas Punidas, Entidades Impedidas, Servidores, Imóveis Funcionais, and Beneficiários L10.559/02.
- GRÁFICOS E DOWNLOADS**: A section for downloading data and viewing graphics, with links for 'Download de Dados' and 'Portal em Gráficos'.
- INFORMAÇÕES**: A section for general information, including 'Sobre o Portal'.
- DESPESAS**: A central section for expenses, with sub-sections for 'Despesas - Empenho, liquidação e pagamento' and 'Despesas - Pagamento'. It includes a search bar and a 'consultar' button.
- RECEITAS**: A section for revenues.
- CONVÊNIOS**: A section for agreements.
- SAÇÕES**: A section for sanctions.
- SERVIDORES**: A section for servers.
- + CONSULTAS**: A link to additional consultation options.

On the right side of the page, there is a vertical menu with links to various services and information, including 'Painel de Municípios', 'Receba Informações de Liberação de Convênios', 'Banco de Preços', 'Jogos Rio 2016', 'Rede de Transparência', 'Portal de Acesso à Informação', 'Páginas de Transparência Pública', and 'Olho Vivo no Dinheiro Público'.

# Onde obtenho dados?

The screenshot shows the 'Portal da Transparência' website. The header includes the logo and name of the 'Ministério da Transparência e Controladoria-Geral da União'. Below the header, there is a navigation bar with links like 'Perguntas frequentes', 'Contato', 'Glossário', 'Links', and 'Manual de navegação'. The main content area is titled 'Detalhamento Diário das Despesas'. It contains a paragraph explaining the purpose of the consultation and a search form. The search form has fields for 'Período' (20/03/2018 to 20/03/2018), 'Fase da Despesa' (Empenho, Liquidação, Pagamento), 'Órgão Superior' (Todos), and 'Favorecido'. There are also buttons for 'Consultar' and 'Limpar campos'. At the bottom, there is a table titled 'Extrato das despesas do dia 20/03/2018:' and a pagination bar showing 'Página 1 de 696'.

Portal do Governo Brasileiro

Ministério da Transparência e Controladoria-Geral da União

## Portal da Transparência

GOVERNO FEDERAL

Perguntas frequentes | Contato | Glossário | Links | Manual de navegação

Acesso rápido Seleccione... OK

Você está em:  
Início » **Detalhamento Diário das Despesas**

### Detalhamento Diário das Despesas

A consulta "Detalhamento Diário das Despesas" do Portal da Transparência do Governo Federal apresenta dados detalhados e diariamente atualizados sobre os atos praticados pelas unidades gestoras do Poder Executivo Federal no decorrer da execução das suas despesas. Por meio da consulta, o cidadão poderá saber quanto e com o que está sendo comprometido o recurso do orçamento.

Por meio da pesquisa, é possível, inclusive, conhecer a fase em que a despesa se encontra: empenho, liquidação e pagamento. [Saiba mais](#)

**Consulta Rápida** [Consulta Avançada](#) | [Consulta por Documento](#)

Período: 20/03/2018 a 20/03/2018 Formato: dd/mm/aaaa

Fase da Despesa: ☒ Empenho ☐ Liquidação ☐ Pagamento

Órgão Superior: Todos (período de 1 dia ou favorecido específico)

Favorecido: Fornecer CNPJ, CPF, UG-Gestão ou outros (sem pontuações)

Consultar Limpar campos

Extrato das despesas do dia 20/03/2018:

Página 1 de 696 1 2 3 4 5 » Página: nº página Ir



# 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



An aerial photograph of the New York City skyline at dusk. The sky is a mix of dark purple, blue, and orange. The city is densely packed with skyscrapers, many of which are illuminated with their lights. The Empire State Building is prominent in the center, with its top lit in red and green. The Hudson River is visible on the right side of the image. The text "Show me the code" is overlaid in a large, white, sans-serif font on the left side of the image.

Show me the  
code



# Bibliotecas

*urllib*

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

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](mailto:saletefarias@gmail.com)

[anaclaracavalcante14@gmail.com](mailto:anaclaracavalcante14@gmail.com)

[meirimartinsp.rodriques@gmail.com](mailto:meirimartinsp.rodriques@gmail.com)

