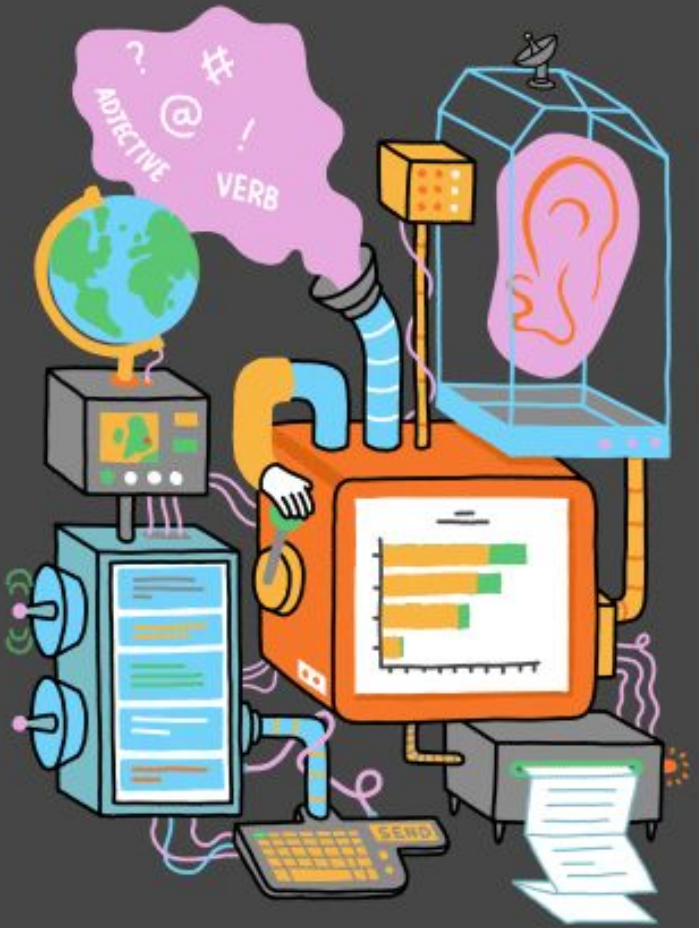
Conseguindo dados de portais públicos com Scraping!

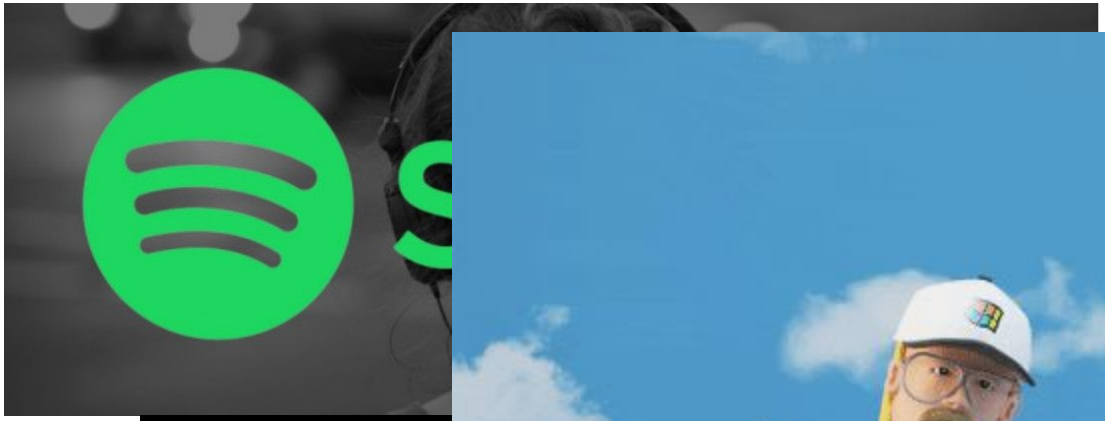
I Hackathon do CERES

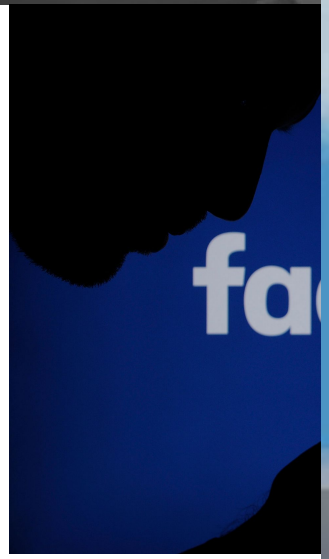
~Who am I?

JERIMUM HACKERSPACE

PARKWAYES


ADJECTIVE
VERB
SEND





Windows 95


NETFLIX
CROWN
ORANGE BLACK
how i met your

DADOS É O NOVO PETRÓLEO DO SEC. XXI?

- Comportamento de Pessoas
- Comportamento de Sociedade
- Comportamento para Tendência Moda

Você pode dominar o MUNDO!!!!!

IMPORTANCIA!

A transparência e o acesso à informação são direitos do cidadão, tornando-se dever da Administração Pública assegurar que estes tenham condições de monitorar as decisões de interesse público e fiscalizar órgãos do governo.

LEI Nº 12.527 - Lei de Acesso à Informação (LAI)

Em maio de 2012, entrou em vigor no Brasil, a Lei nº 12.527/2011, sancionada pela então presidente do país, Dilma Rousseff. A partir desta nova norma toda e qualquer pessoa pode ter acesso a qualquer informação pública de orgãos e entidades.

PROJETOS!

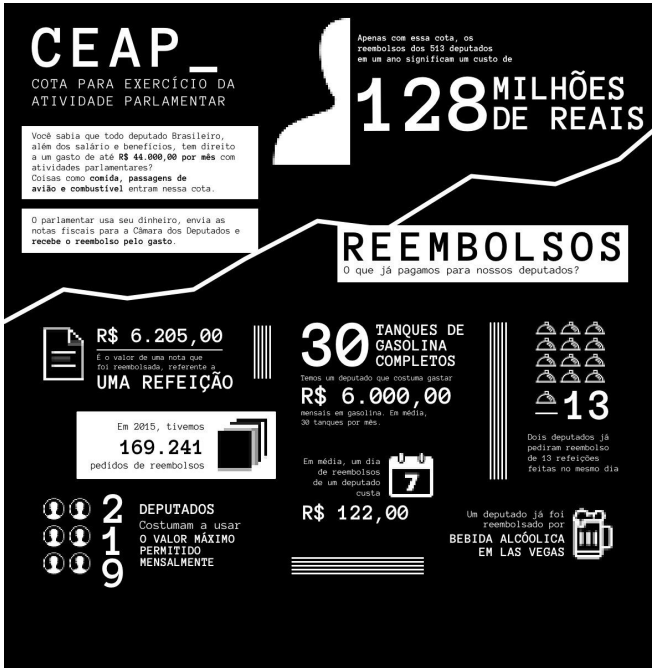
Colaboradados!

OPERAÇÃO
SERENATA
DE AMOR

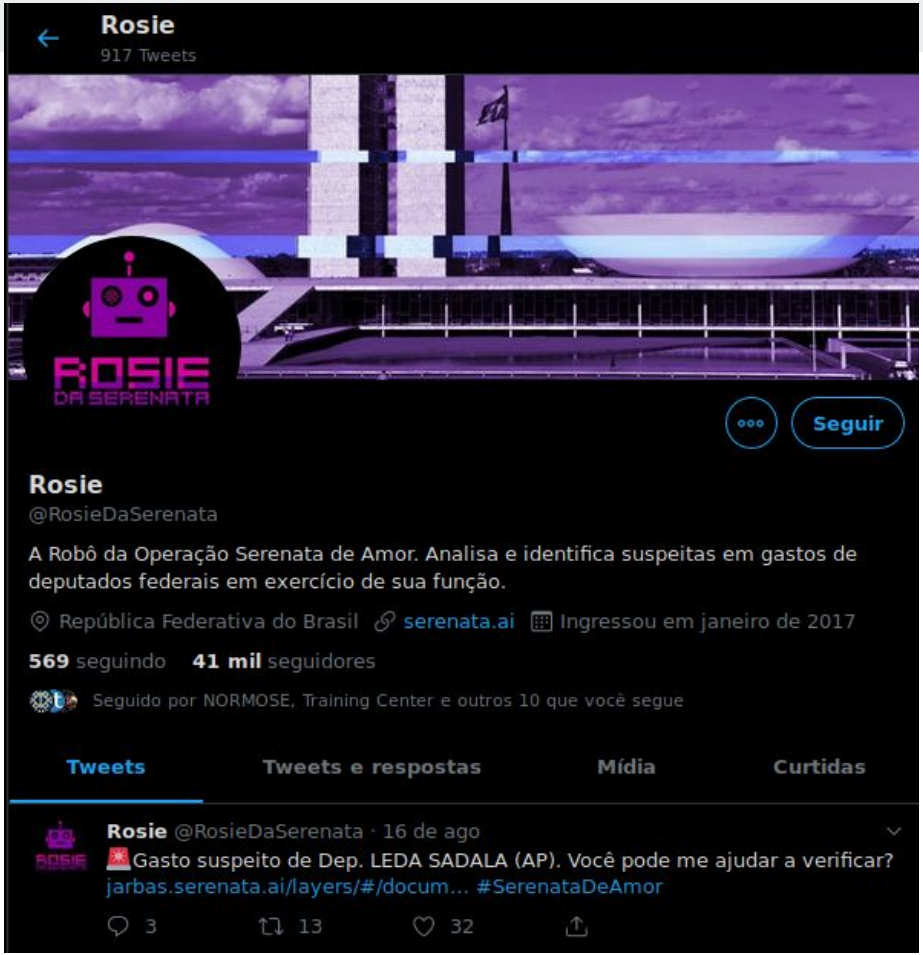
OPERAÇÃO

SERENATA DE AMOR

Inteligência artificial para controle social da administração pública



https://serenata.ai/


Rosie
917 Tweets
ROSIE
DA SERENATA
Seguir
Rosie
@RosieDaSerenata
A Robô da Operação Serenata de Amor. Analisa e identifica suspeitas em gastos de deputados federais em exercício de sua função.
República Federativa do Brasil
serenata.ai
Ingressou em janeiro de 2017
569 seguindo
41 mil seguidores
Seguido por NORMOSE, Training Center e outros 10 que você segue
Tweets
Tweets e respostas
Mídia
Curtidas
Rosie @RosieDaSerenata · 16 de ago
🚨Gasto suspeito de Dep. LEDA SADALA (AP). Você pode me ajudar a verificar?
jarbas.serenata.ai/layers/#/docum… #SerenataDeAmor
3
13
32

Jarbas

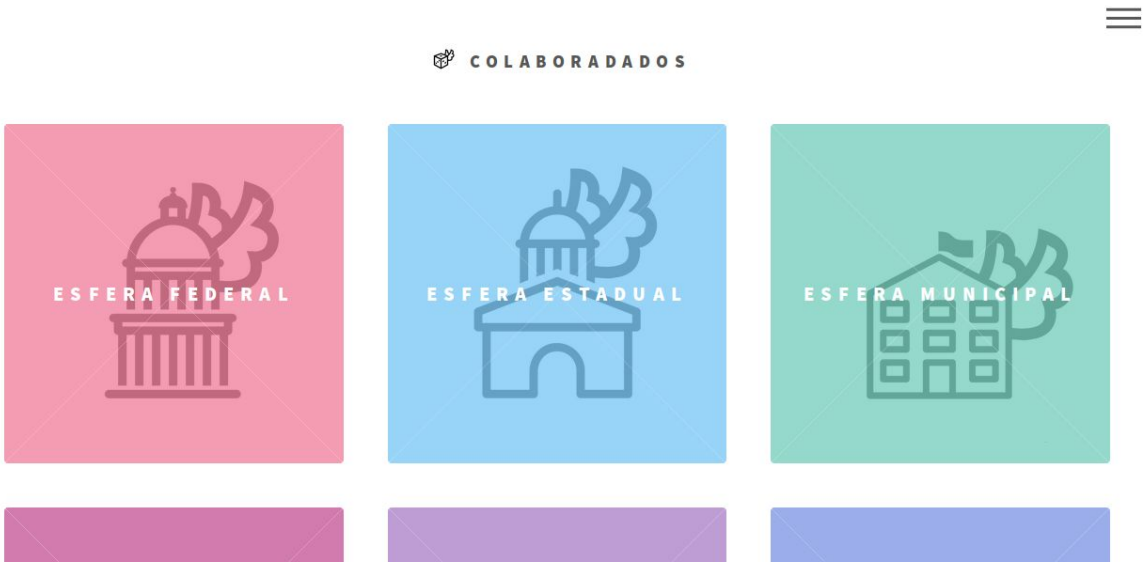
Document #6875907

Summary

CNPJ invalid or not found.

Congressperson	LEDA SADALA (AVANTE/AP)
Expense date	Apr 25th, 2019
Claim date	4/2019
Subquota	Congressperson meal (13)
Company	CASARAO DA VILA RESTAURANTE LTDA - ME (27.522.923/0001-60)
Expense value	162.45 BRL
Remark value	0.00 BRL
Total net value	162.45 BRL
Total reimbursement value	0.00 BRL
Suspicions	Meal price is an outlier

Colaboradados!



Um grupo de amigos que resolveu, por meio da programação, investigar, monitorar e, acima de tudo, informar a população sobre transparência governamental e assuntos que envolvam dados.

http://colaboradados.com.br

Extração de Dados!

PORTAIS:
Dados de Mobilidade Urbana Natal/RN (2018)
Dados de Relação de Servidores da Instituição - IFRN

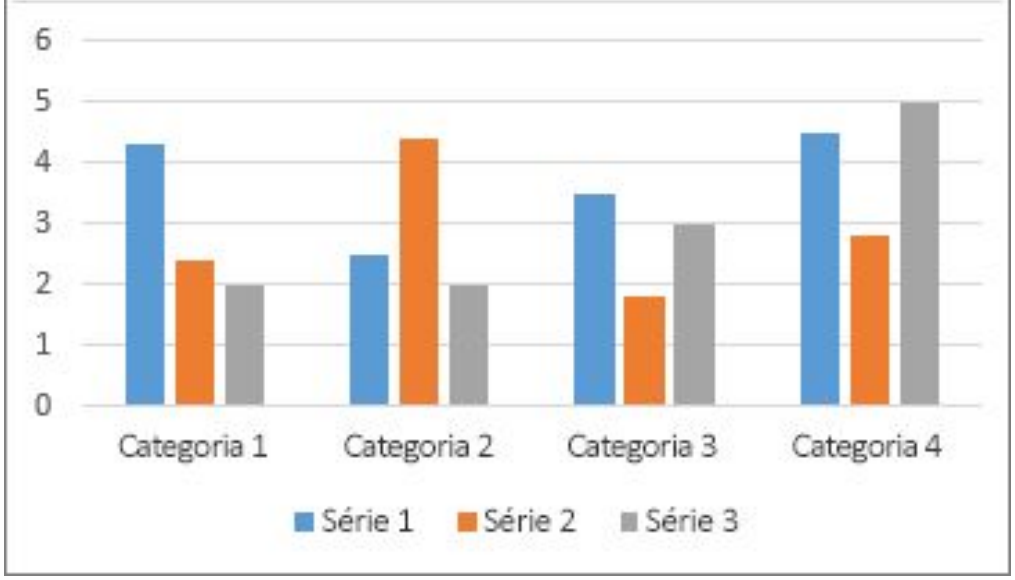
Dados:

Bilhetagem Eletrônica - Natal

Dados disponibilizados pelo IFRN

Gráficos > Tabelas!

	Série 1	Série 2	Série 3
Categoria 1	4,3	2,4	2
Categoria 2	2,5	4,4	2
Categoria 3	3,5	1,8	3
Categoria 4	4,5	2,8	5



Web Scraping!

Mas o que é Web Scraping?

Pode ser uma coleta de dados web, ou raspagem web, é uma forma de mineração que permite a extração de dados de sites da web convertendo-os em informação estruturada para uma análise posterior.

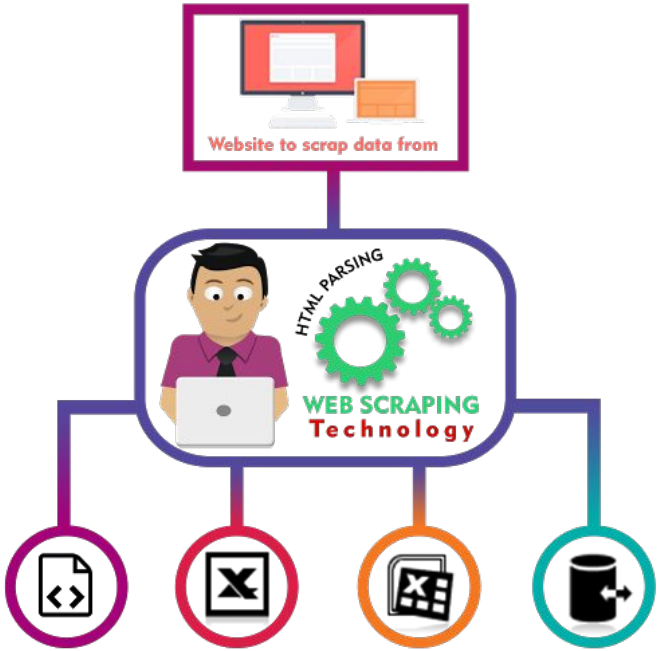
Resumindo!

Scraping e pegar muitos dados de sites na internet, centralizar em um local, estruturar e usar com algum propósito.

Como funciona?

De forma beeeem resumida:

- Código acessa o site
- Realiza a raspagem desejada
- Converte os dados:
  - CVS
  - Json
  - XML
  - SQL

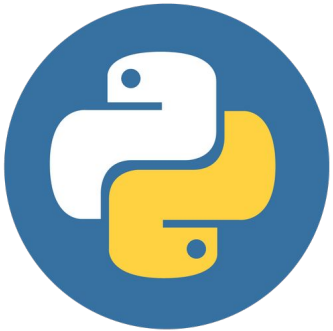


~Ferramentas:

Python

BeautifulSoup

Selenium



BeaultifulSoup?

É uma biblioteca do Python serve para extrairmos dados de HTML e XML, de forma fácil e descomplicada podemos acessar os ‘nós’ da estrutura do HTML da página ou até mesmo classes e pegar suas informações.

Selenium?

Selenium é uma ferramenta para testar aplicações web pelo browser de forma automatizada. Selenium se refere ao *Acceptance Testing* (ou functional testing) que envolve rodar testes num sistema finalizado. Os testes rodam diretamente em um browser, exatamente como o usuário faria.

<code/>

Loteria

Jupyter Notebook:

LoteriasCaixa.ipybn

Configuração de ambiente!

 # Instação do PIP:

```
$ sudo apt install python3-pip
```

 # Instalação do Virtualenv:

```
$ sudo pip3 install virtualenv
```

 # Inicializando o Virtualenv:

```
$ virtualenv <nome_do_caminho>
```

 # Instalação Selenium:

```
$ sudo pip install selenium
```

Configuração de ambiente!

 # Download do WebDriver - Geckdrive (Monzilla)

https://github.com/mozilla/geckodriver/releases

```
<!-- Colocar Geckodrive na pasta bin - virtualenv
     aqui funcionou assim! -->                    .
```

 # Instalação do BeautifulSoup:

```
sudo pip install bs4
```

<code/>

intro

Google Básico

basico.py

<code/>

scraping


Downloading Geocities...
ARCHIVE
team