Web Scraping com Python e Selenium

Sedir Morais

Entusiasta de Python 🐍

Analista e Desenvolvedor de Sistemas (IFRN)

Pesquisador do LAIS/UFRN

Mestre em Eng. de Computação (UFRN)

Doutorando em Eng. de Computação (UFRN)

Web Scraping

"(...) O *web scraping* é a prática de coletar dados por qualquer meio que não seja um programa que interage com uma API."

–Ryan Mitchell - Web Scraping with Python

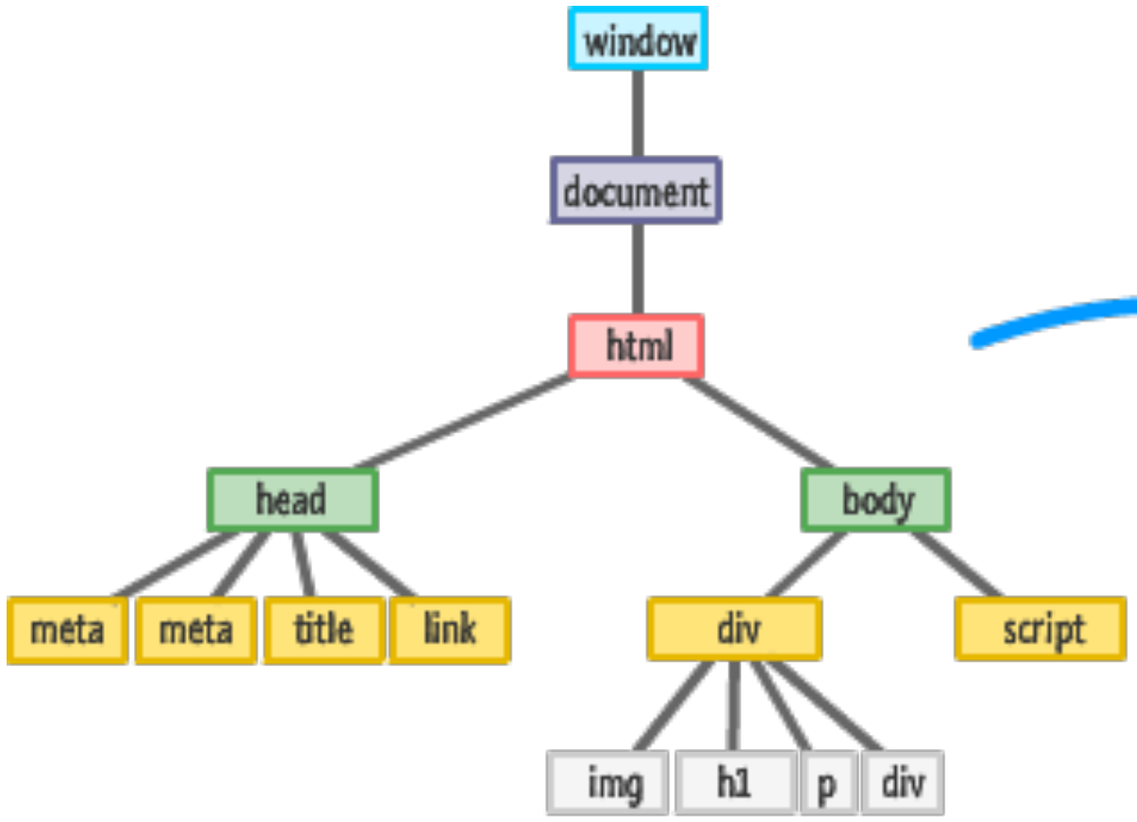
Cenários de utilização

- Busca de informações pessoais
- Comparação de preços de produtos e serviços
- Detecção de mudanças em sites
- Integração de dados
- Pesquisa

Técnicas

- Identificação de padrões com expressões regulares
- DOM *parsing*
- DOM *parsing* + JavaScript
- Entre outras...

DOM



The DOM

The Browser

(aka what you see)

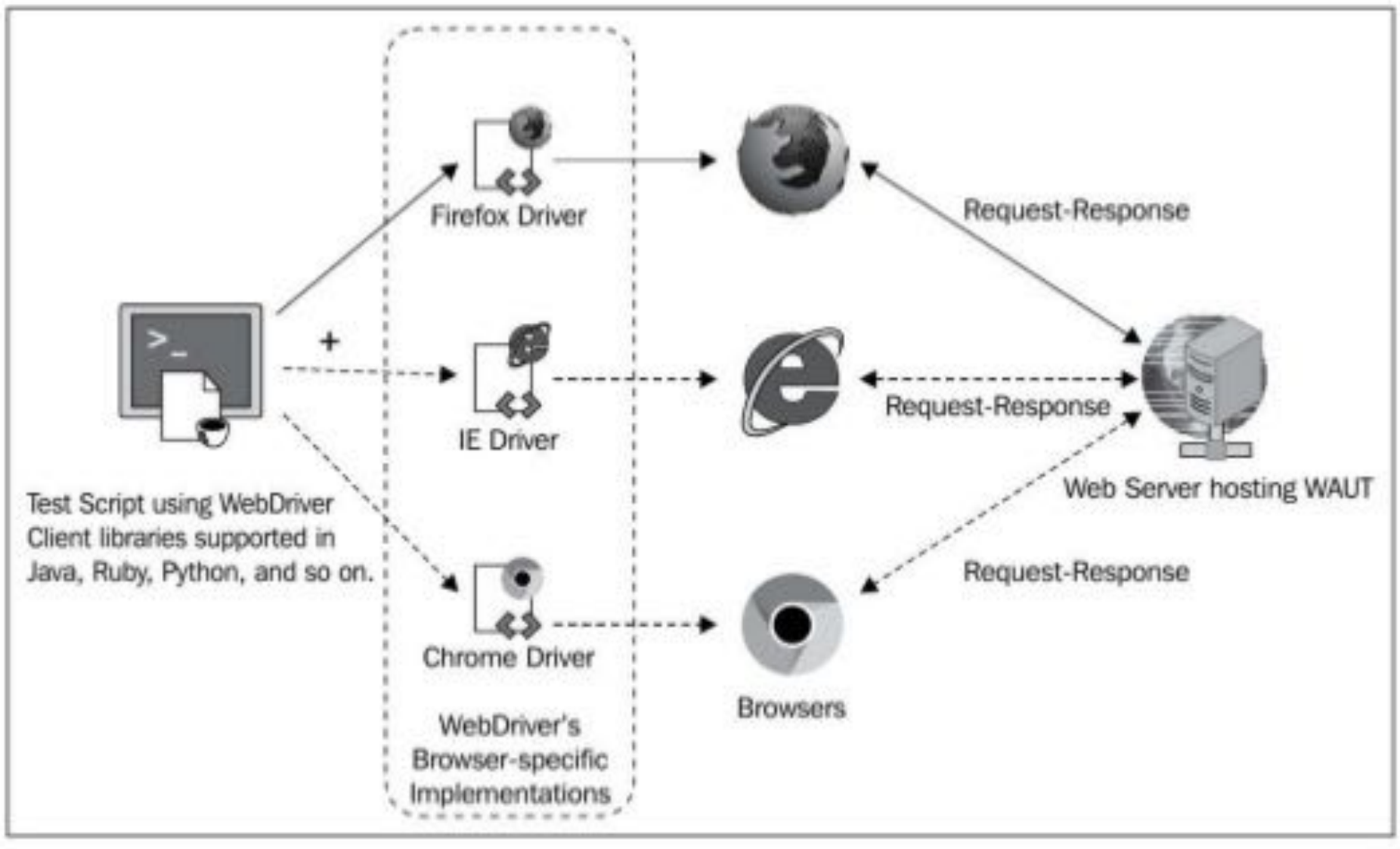
Fonte: https://www.kirupa.com/html5/traversing_the_dom.htm

Selenium

Selenium

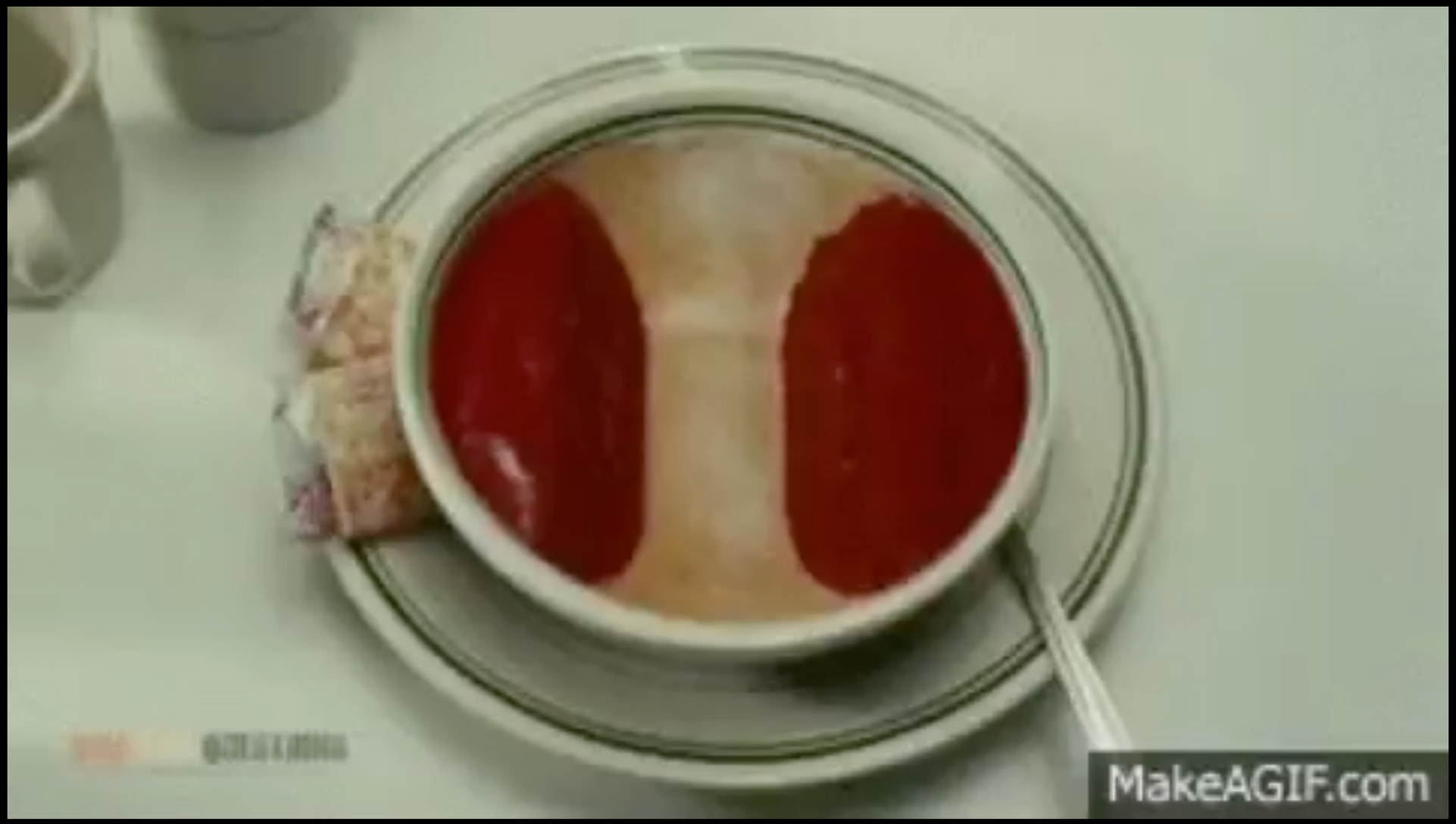
- É uma ferramenta que:
    - Permite controlar um browser/navegador web
    - Pode acessar endereços na web, ler e interagir com todo o conteúdo de um site, incluindo JavaScript
    - É bastante usada em testes automatizados
    - Suporta várias linguagens de programação e também pode ser usado de forma independente

Selenium



Fonte: https://www.kirupa.com/html5/traversing_the_dom.htm

Selenium



Selenium

- Vantagens:
    - Suporta JavaScript
    - Escuta mudanças no DOM

- Desvantagem:
    - Bastante lento e pesado em relação as outras opções

Requisitos de instalação

Python 2.7 ou 3.4+

```
pip install selenium
```

Driver do browser no PATH do sistema:

Chrome:	https://sites.google.com/a/chromium.org/chromedriver/downloads
Edge:	https://developer.microsoft.com/en-us/microsoft-edge/tools/webdriver/
Firefox:	https://github.com/mozilla/geckodriver/releases
Safari:	https://webkit.org/blog/6900/webdriver-support-in-safari-10/

```
export PATH=$PATH:/caminho/da/pasta/do/driver
```

Código básico

```
from selenium import webdriver

url = "https://www.fifa.com/worldcup/statistics/players/
goal-scored"

# Inicializa webdriver
driver = webdriver.Chrome(chrome_options=None)
# Aguarda o browser
driver.implicitly_wait(30)
# Entra na URL
driver.get(url)
```

Buscando elementos

```
# Entra na URL
driver.get("https://www.uol.com.br")

# Aguarda carregamento da página
driver.implicitly_wait(30)

# Seleciona campo de busca cujo form name é ‘q'
input = driver.find_element_by_name('q')

# Envia string para o campo
input.send_keys('tecnologia')

# Submete o formulário
input.submit()

# Aguarda
driver.implicitly_wait(30)

# Obtém os resultados através do nome da classe 'result-item'
results = driver.find_elements_by_class_name('result-item')

for result in results:
    title = result.find_element_by_tag_name('a')
    print(title.text, title.get_attribute('href'))
```

Buscando elementos

```
# Busca um único elemento pelo atributo id
goals_table = driver.find_element_by_id(‘goal-scored')

pagination_base = driver.find_element_by_id('goal-scored_paginate')
```

```
# Busca vários elementos pelo nome de tag a partir do
# elemento anterior
rows = goals_table.find_elements_by_tag_name(‘tr')



for row in rows:
    cols = row.find_elements_by_tag_name(‘.//td’)
```

Live Code!

Live Code!

```
# Saldo de gols dos jogadores na copa do mundo 2018

url = "https://www.fifa.com/worldcup/statistics/players/goal-scored"
```


Web Scraping com Python e Selenium

Obrigado

Contato

Sedir Morais

Telegram: @sedirmorais
GitHub: @sedir
E-mail: philippi.sedir@gmail.com