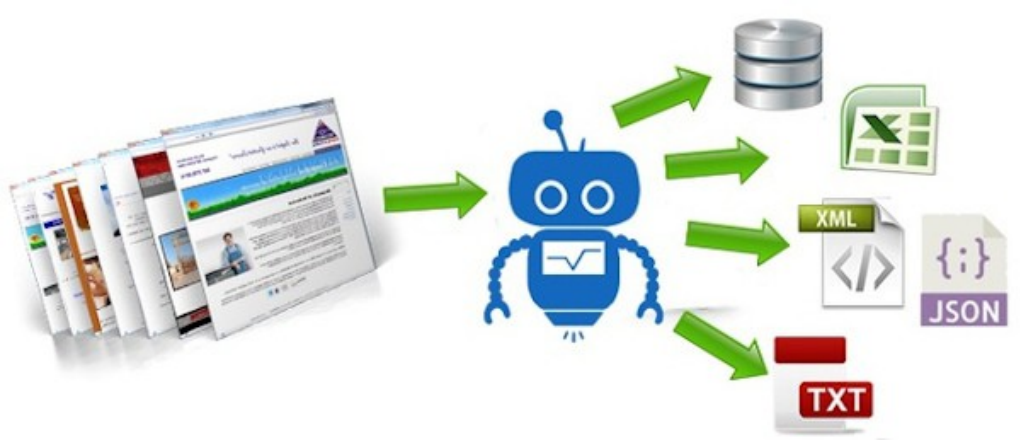
Python

Web Scraping



Rogelio Ferreira Escutia

Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

Leer páginas Web

Página Web

- **Página de prueba:**

Página Web

- **Código de la página de prueba:**

```html
<!DOCTYPE html>

<!-- Rogelio Ferreir Escutia - 10/marzo/2020 - Pruebas de Web Scraping -->

<html lang="es">
    <head>
        <meta charset="utf-8" />
        <title>Web Scraping</title>
        <meta name="keywords" content="web, scraping, búsquedas" />
        <meta name="description" content="Página para hacer pruebas de Web Scraping" />
        <meta name="author" content="Rogelio Ferreira Escutia" />
        <link rel="icon" type="image/png" href="favicon.ico" />
        <link rel="stylesheet" type="text/css" href="scraping.css" />
    </head>
    <body>
        <h1>Bienvenido al mundo del Web Scraping!</h1>
        <h2>Esta es una página de pruebas sobre Web Scraping</h2>
        <p>Web Scraping es un conjunto de técnicas para extraer información de páginas Web.</p>
        <h2>Enlaces de Prueba</h2>
        <a href="http://www.xumarhu.net/">Xumarhu</a>
        <br /><br /><a href="http://www.itmorelia.edu.mx/">Tec de Morelia</a>
        <br /><br /><a href="http://dsc.itmorelia.edu.mx/">Departamento de Sistemas y Computación</a>
    </body>
</html>
```

Página Web

- **Código Python para descargar e imprimir la página Web:**

```
from urllib.request import urlopen

print("Imprimir el código de una página")
html = urlopen('http://sagitario.itmorelia.edu.mx/~rogelio/hola.htm')
print(html.read())
```

- **Ejecución en consola:**

```
MacBook-Pro-de-Rogelio-2:web_scraping rogelioferreiraescutia$ python3 python_web_scr
aping_leer_pagina.py
Imprimir el código de una página
b'<!DOCTYPE html>\n\n<!-- Rogelio Ferreira - 9/octubre/2019 - Pruebas de Web Scrapin
g -->\n\n<html lang="es">\n    <head>\n        <meta charset="utf-8" />\n        <ti
tle>Web Scraping</title>\n        <meta name="keywords" content="web, scraping, b\xc
3\xbasquedas" /> \n\t\t<meta name="description" content="P\xc3\xa1gina para hacer pr
uebas de Web Scraping" />\n        <meta name="author" content="Rogelio Ferreira Esc
utia" />\n        <link rel="icon" type="image/png" href="favicon.ico" />\n        <
link rel="stylesheet" type="text/css" href="scraping.css" />\n    </head>\n\n    <bo
dy>\n        <h1>Bienvenido al mundo del Web Scraping!</h1>\n        <h2>Esta es una
 p\xc3\xa1gina de pruebas sobre Web Scraping</h2>\n        <p>Web Scraping es un con
junto de t\xc3\xa9cnicas para extraer informaci\xc3\xb3n de p\xc3\xa1ginas Web.</p>\
n    </body>\n</html>\n'
```

Extraer título

Página Web - Título

- **Código Python para extraer el título de una página Web:**

```
#   Se requiere instalar "Beautifulsoup"
#   Mac: pip3 install beautifulsoup4
#   Linux: sudo apt-get install python-bs4
#

from urllib.request import urlopen
from bs4 import BeautifulSoup

print("\nExtraer el título de una página Web")
pagina = "http://sagitario.itmorelia.edu.mx/~rogelio/hola.htm"
html = urlopen(pagina)
bs = BeautifulSoup(html.read(), 'html.parser')
titulo = str(bs.title)
print("\nPágina: "+pagina)
print("Título: "+titulo+"\n")
```

- **Ejecución en consola:**

```
MacBook-Pro-de-Rogelio-2:web_scraping rogelioferreiraescutia$ python3 python_web_scr
aping_extraer_titulo.py

Extraer el título de una página Web

Página: http://sagitario.itmorelia.edu.mx/~rogelio/hola.htm
Título: <title>Web Scraping</title>
```

Extraer etiquetas

Página Web - Etiquetas

- **Código Python para extraer el título de una página Web:**

```
#   Se requiere instalar "Beautifulsoup"
#   Mac: pip3 install beautifulsoup4
#   Linux: sudo apt-get install python-bs4
#

from urllib.request import urlopen
from bs4 import BeautifulSoup

print("Extraer el contenido de algunas etiquetas de una página Web")
html = urlopen('http://sagitario.itmorelia.edu.mx/~rogelio/hola.htm')
bs = BeautifulSoup(html.read(), 'html.parser')
print(bs.title)
print(bs.h1)
print(bs.h2)
print(bs.h3)
```

- **Ejecución en consola:**

```
[MacBook-Pro-de-Rogelio-2:web_scraping rogelioferreiraescutia$ python3 python_web_scr]
aping_extraer_etiquetas.py
Extraer el contenido de algunas etiquetas de una página Web
<title>Web Scraping</title>
<h1>Bienvenido al mundo del Web Scraping!</h1>
<h2>Esta es una página de pruebas sobre Web Scraping</h2>
None
```

Extraer enlaces

Página Web - Enlaces

- **Código Python para extraer los enlaces de una página:**

```
from urllib.request import urlopen
from bs4 import BeautifulSoup

pagina_inicial = "http://sagitario.itmorelia.edu.mx/~rogelio/hola.htm"

url = urlopen(pagina_inicial)
print("\nExtraer los enlaces de la página Web: " + pagina_inicial + "\n")

bs = BeautifulSoup(url.read(), 'html.parser')
for enlaces in bs.find_all("a"):
    print("href: {}".format(enlaces.get("href")))
print("\nFin de enlaces encontrados\n")
```

- **Ejecución en consola:**

```
Extraer los enlaces de la página Web: http://sagitario.itmorelia.edu.mx/~rogelio/hola.htm

href: http://www.xumarhu.net/
href: http://www.itmorelia.edu.mx/
href: http://dsc.itmorelia.edu.mx/

Fin de enlaces encontrados
```

Rogelio Ferreira Escutia

Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

rogelio.fe@morelia.tecnm.mx

rogeplus@gmail.com

xumarhu.net

@rogeplus

https://www.youtube.com/channel/UC0on88n3LwTKxJb8T09sGjg

rogelioferreiraescutia