Hoja informativa: Recuperación de datos de recursos en línea

Práctica

```
# Recuperar información de una página web usando una URL
import requests

req = requests.get(URL)
print(req.text) # mostrar el contenido de la página
print(req.status_code) # mostrar código de estado
```

```
# Busca una cadena para la primera subcadena que coincida con una expresión regular

import re
print(re.search(pattern, string).group())
```

```
# Divide una cadena en subcadenas por ocurrencias de patrones
# maxsplit - número máximo de divisiones, maxsplit=0 por defecto

import re
print(re.split(pattern, string, maxsplit=num_split))
```

```
# Encuentra una subcadena y reemplazarla con subcadena repl
import re
print(re.sub(pattern, repl, string))
```

```
# Encuentra todas las subcadenas coincidentes

import re
print(re.findall(pattern, string))
```

```
# Genera una estructura de árbol para una página web
```

```
from bs4 import BeautifulSoup
soup = BeautifulSoup(req.text, parser)
```

```
# Encuentra la primera etiqueta 'tag'
# Devuelve una cadena con etiqueta, atributos y contenido
# attrs - diccionario de atributos tag

tag_content = soup.find(tag, attrs={"attr_name": "attr_value"})
print(tag_content.text) # contenido sin etiqueta
```

```
# Operaciones con todas las etiquetas tag
# attrs - diccionario de atributos tag

for tag_content in soup.find_all(tag, attrs={"attr_name": "attr_value"}):
  # hacer algo
```

Teoría

Minería web: el proceso de buscar recursos en línea y recuperar datos de ellos

HTML: Lenguaje de marcado de hipertexto, un lenguaje utilizado para crear páginas web

HTTP: protocolo de transferencia utilizado para transmitir información en línea

HTTPS: una versión segura de HTTP

Etiqueta HTML: un elemento de lenguaje de marcado de hipertexto

Atributo tag: una función que permite realizar ajustes cuando se trabaja con etiquetas