Python

Archivos HTML

Rogelio Ferreira Escutia

Profesor / Investigador
Tecnológico Nacional de México
Campus Morelia

Tipos de Archivos

Tipos de Archivos

- **Básicamente hay 2 tipos de archivos:**

Text File

- **Plain Text**: .txt, .csv
- **Source Code**: .py, .html, .css. .js
- **Data**: .json, .xml

Binary File

- **Executable**: .exe, .dmg, .bin
- **Images**: .jpg, .png, .gif, .tiff, .ico
- **Video**: .mp4, .m4v, .mp4, .mov
- **Audio**: .aif, .mp3, .mpa, wav
- **Compressed**: .zip, .deb, .tar.gz
- **Font**: .woff, .otf, .ttf
- **Document**: .pdf, .docx, .xlsx

HTML

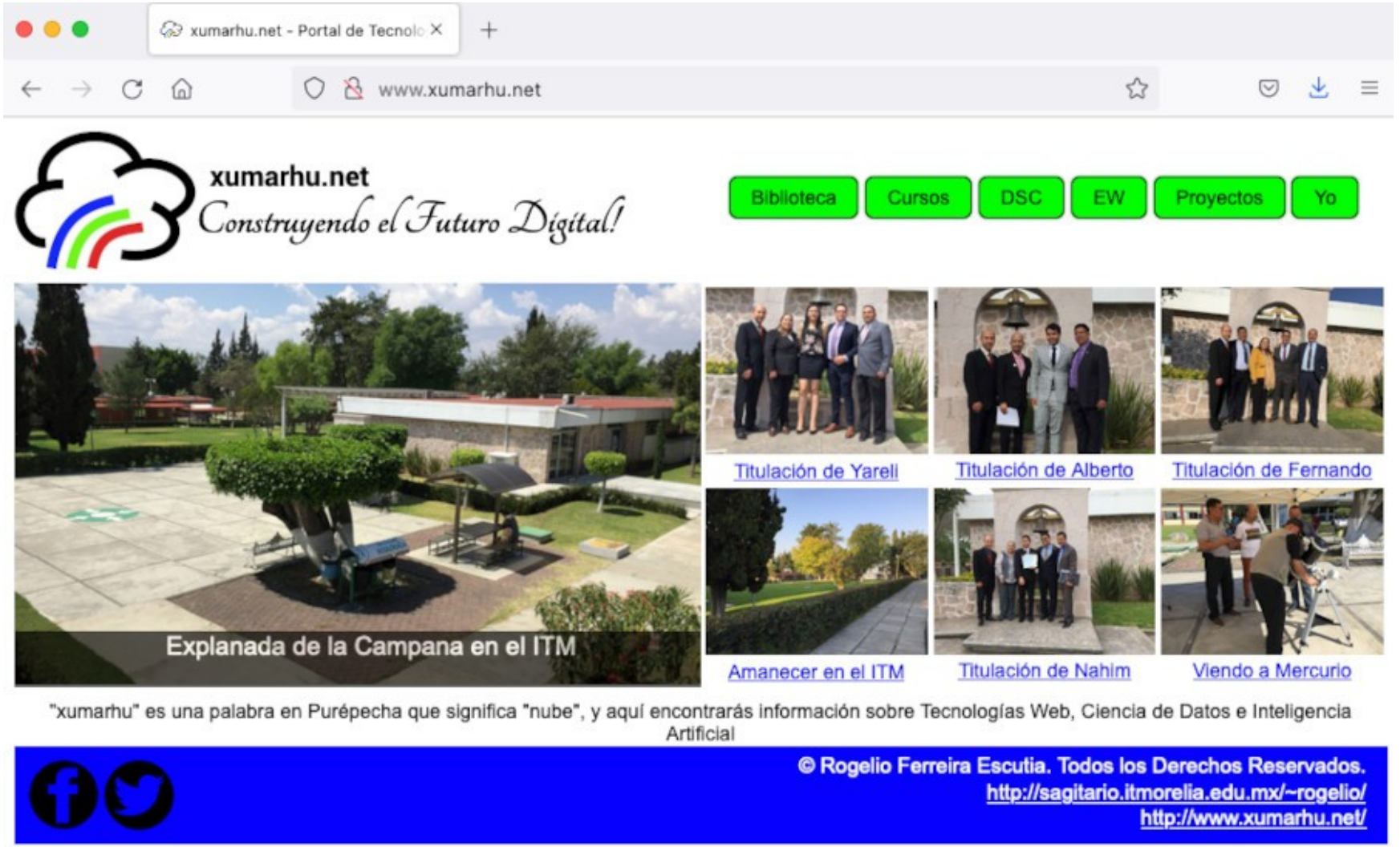
HTML

- **"HTML", siglas en inglés de “HyperText Markup Language” (“Lenguaje de Marcado de Hipertexto”), hace referencia al lenguaje de marcado para la elaboración de páginas Web.**

- **Es un estándar que sirve de referencia del software que conecta con la elaboración de páginas web en sus diferentes versiones, define una estructura básica y un código (denominado código HTML) para la definición de contenido de una página web, como texto, imágenes, videos, juegos, entre otros.**

HTML

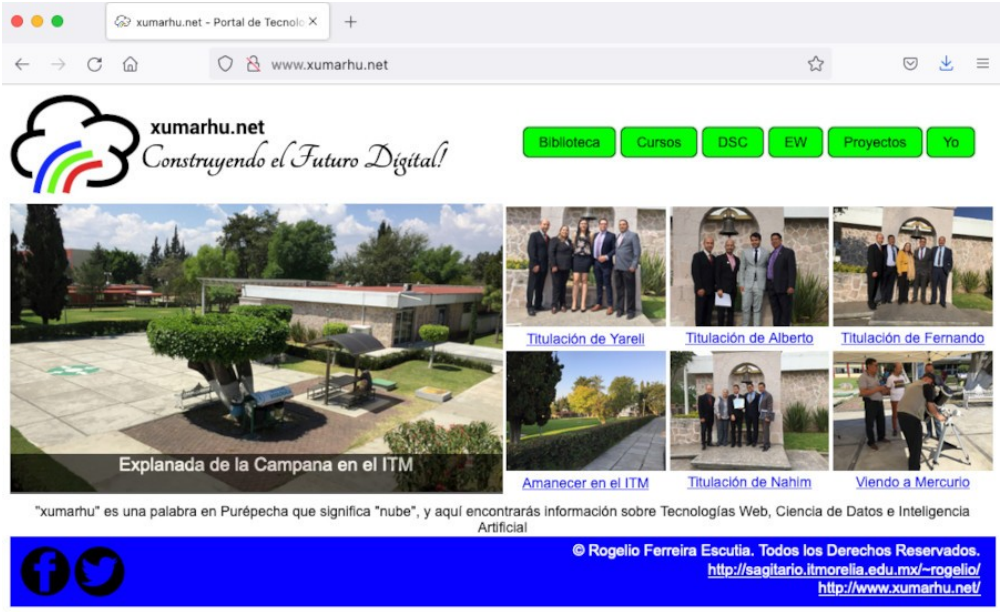
- **http://www.xumarhu.net/**



Ejemplo de uso de archivos HTML

Ejemplo HTML (1)

- **Descargar una página y extraer su información.**
- **Para este ejemplo utilizaremos la página anterior:**
  - **http://www.xumarhu.net/**



Ejemplo HTML (2)

- **Bibliotecas a utilizar:**

```
# Bibliotecas a utilizar
from urllib.request import urlopen  # Importamos "urlopen" para abrir una página Web
from bs4 import BeautifulSoup       # Importamos "BeautifulSoup" para hacer WebScraping
```

Ejemplo HTML (3)

- **Definimos la página a utilizar, la cargamos a memoria y la parseamos:**

```
# Definir e imprimir página a leer
pagina_inicial = "http://www.xumarhu.net/"
print("\nPAGINA WEB: " + pagina_inicial)

# Cargar página en memoria
url = urlopen(pagina_inicial)

# Parsear la página usando HTML
bs = BeautifulSoup(url.read(), 'html.parser')
```

Ejemplo HTML (4)

- **Extraemos el título de la página:**

```
# Extraer e imprimir el título de la página
print("\nTITULO: " + bs.title.text + "\n")
```

```
PAGINA WEB: http://www.xumarhu.net/

TITULO: xumarhu.net - Portal de Tecnología
```

Ejemplo HTML (5)

- **Extraemos todos los enlaces y sus textos, y los almacenamos en listas:**

```
# Extraer todos los enlaces de la página
print("\nExtraer todo los enlaces de la página Web: ", pagina_inicial)
enlaces = []
texto = []
for ciclo in bs.find_all("a"):
    print("href: {}".format(ciclo.get("href")))
    enlaces.append(str(format(ciclo.get("href"))))
    texto.append(str(format(ciclo.get_text())))
print("\nFin de enlaces encontrados\n")
```

Ejemplo HTML (6)

- **Salida:**

```
Extraer todo los enlaces de la página Web:  http://www.xumarhu.net/
href: index.html
href: bib_inic.htm
href: cur_inic.htm
href: dsc_inic.htm
href: ewe_inic.htm
href: pro_inic.htm
href: rfe_inic.htm
href: 2020-03-17_itm_1024x768.jpg
href: 2020-03-05_titulacion_yareli_yazmin_duran_garcia_1024x768.jpg
href: 2020-03-05_titulacion_yareli_yazmin_duran_garcia_1024x768.jpg
href: 2020-03-06_titulacion_alberto_perez_villegas_1024x768.jpg
href: 2020-03-06_titulacion_alberto_perez_villegas_1024x768.jpg
href: 2020-03-02_titulacion_fernando_diaz_neri_1024x768.jpg
href: 2020-03-02_titulacion_fernando_diaz_neri_1024x768.jpg
href: 2019-11-11_itm_jardin_1024x768.jpg
href: 2019-11-11_itm_jardin_1024x768.jpg
href: 2019-10-29_titulacion_nahim_angelo_gomez_ceja_1024x768.jpg
href: 2019-10-29_titulacion_nahim_angelo_gomez_ceja_1024x768.jpg
href: 2019-11-11_itm_viendo_mercurio_1024x768.jpg
href: 2019-11-11_itm_viendo_mercurio_1024x768.jpg
href: https://www.facebook.com/xumarhu.net
href: https://twitter.com/rogeplus
href: http://sagitario.itmorelia.edu.mx/~rogelio/
href: http://www.xumarhu.net/

Fin de enlaces encontrados
```

Ejemplo HTML (7)

- **Imprimir el primer enlace:**

```
# Imprimir el primer enlace y su texto
print("Primer enlace: ", enlaces[0], " - ", texto[0])
```

```
Primer enlace:  index.html  -
```

Ejemplo HTML (8)

- **Extraer la primer noticia que aparece en pantalla:**

```
# Extraer la PRIMER noticia importante
print("NOTICIA 1 - Texto: " + texto[9] + " Enlace: " + enlaces[9])
```

```
NOTICIA 1 - Texto: Titulación de Yareli Enlace: 2020-03-05_titulacion_yareli_yazmin_duran_garci
a_1024x768.jpg
```

Ejemplo HTML (9)

- **Descargar la foto de la primer noticia:**

```
# Descargar imagen a la computadora
import requests
url_imagen = pagina_inicial + enlaces[9]
nombre_local_imagen = enlaces[9]
imagen = requests.get(url_imagen).content
with open(nombre_local_imagen, 'wb') as handler:
    handler.write(imagen)
```

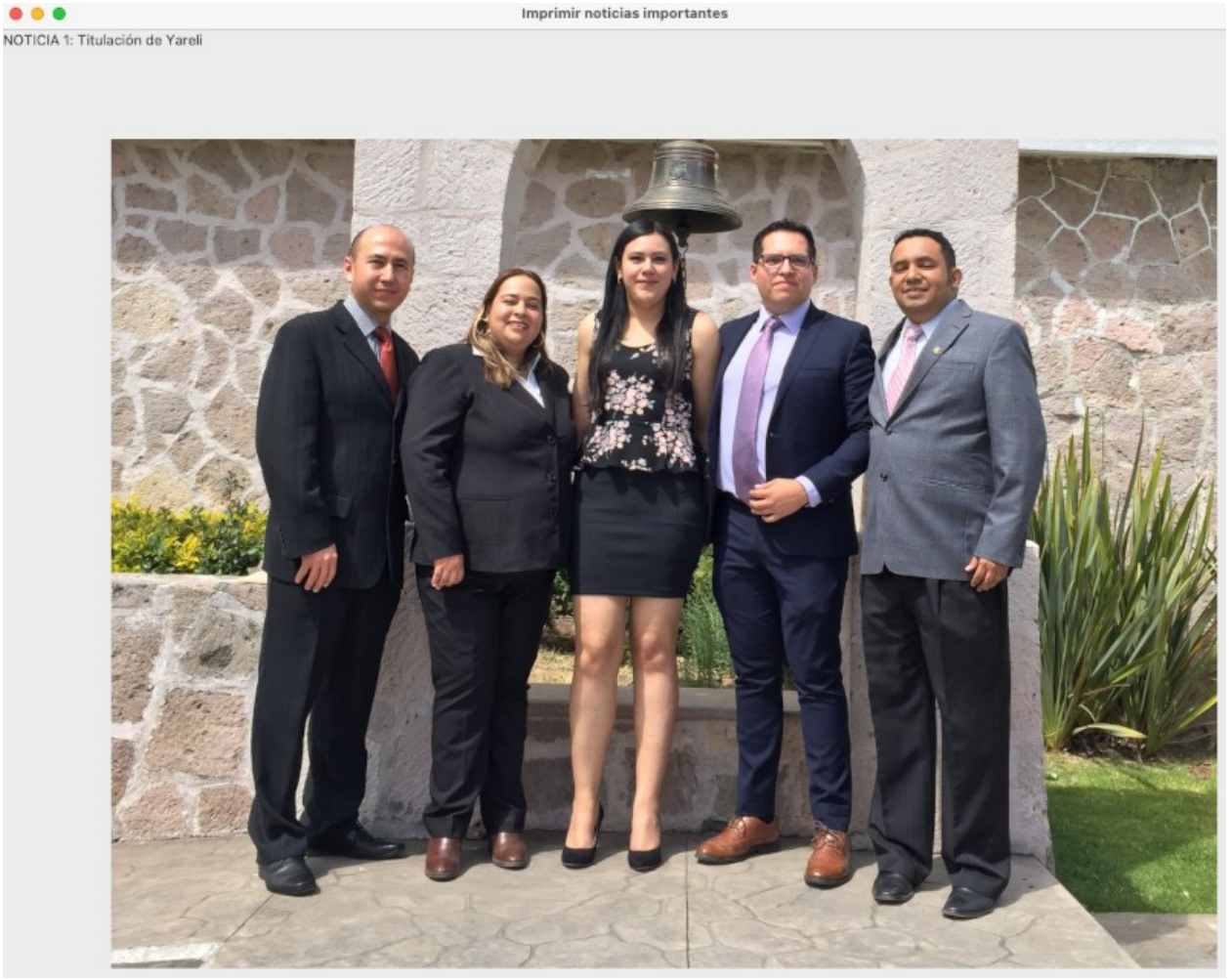
Ejemplo HTML (10)

- **Abrir una ventana e imprimir la foto que se acaba de descargar de la primer noticia usando Tkinter (modo gráfico):**

```
# Abrir Modo Gráfico usando la biblioteca "Tkinter"
import tkinter as tk
from tkinter import ttk, PhotoImage, Label
from PIL import ImageTk, Image
ventana = tk.Tk()
ventana.title("Imprimir noticias importantes")
ttk.Label(ventana, text = "NOTICIA 1: " + texto[9]).grid(column = 0, row = 0)
path = enlaces[9]
img = ImageTk.PhotoImage(Image.open(path))
fondo=Label(ventana,image=img).place(x=100,y=100)
ventana.mainloop()
```

Ejemplo HTML (11)

- **En la pantalla aparecerá la siguiente ventana:**



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