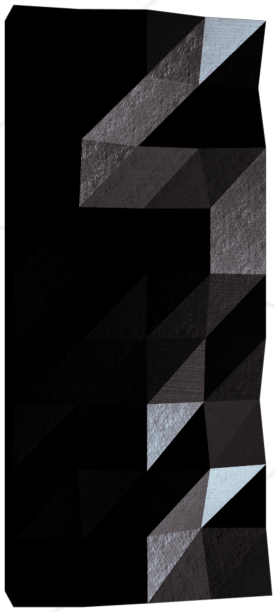

Bootcamp Ciberseguridad | 42 Madrid

arachnida

Resumen: Web scraping y manejo de metadatos

Versión: 1

Índice general

Capítulo I

Introducción

Los `metadatos` son información que se utiliza para describir otros datos. Son esencialmente `datos sobre datos`. Frecuentemente se utililizan en imágenes y documentos, pudiendo llegar a revelar información sensible de quienes lo han creado o manipulado. En este proyecto, crearás dos instrumentos que te permitirán extraer información automáticamente de la `web` y después analizarla para conocer o eliminar datos sensibles.

Capítulo II

Prólogo

Los arácnidos son una clase de artrópodos quelicerados entre los que hay más de 100.000 especies diferentes poblando el planeta. Entre ellos se encuentran las arañas, pero también las garrapatas, los escorpiones o los ácaros. El rasgo común más característico de los arácnidos son sus cuatro pares de patas, así como sus `quelíceros`, unos apéndices puntiagudos que utilizan para agarrar el alimento.

Theridion Grallator · CC Wikimedia Commons*

Capítulo III

Parte Obligatoria

Los dos programas pueden ser scripts o binarios. En caso de lenguajes compilados, debes el código fuente y compilarlo durante la evaluación. Puedes utilizar funciones o librerías que te permitan crear peticiones HTTP y manejar archivos, pero la lógica de cada programa debe estar desarrollada por ti. Es decir, utilizar `wget` o `scrapy` será considerado `cheat` y supondrá suspender el proyecto.

Capítulo IV

Ejercicio 1 - Spider

Nombre de función	`spider`
Archivos a entregar	`spider.c`
Funciones autorizadas	`Nada`
Descripción	`Extraer todas las imágenes de un sitio web`

El programa `spider` permitirá extraer todas las imágenes de un sitio web, de manera recursiva, proporcionando una url como parámetro. Gestionarás las siguientes opciones del programa:

```
./spider [-rlpS] URL
```

- Opción `-r` : descarga de forma recursiva las imágenes en una URL recibida como parámetro.
- Opción `-r -l [N]` : indica el nivel profundidad máximo de la descarga recursiva. En caso de no indicarse, será 5.
- Opción `-p [PATH]` : indica la ruta donde se guardarán los archivos descargados. En caso de no indicarse, se utilizará `./data/`.

  El programa descargará por defecto las siguientes extensiones:

  - .jpg/jpeg
  - .png
  - .gif
  - .bmp

Capítulo V

Ejercicio 2 - Scorpion

Nombre de función	`scorpion`
Archivos a entregar	`scorpion.c`
Funciones autorizadas	`Nada`
Descripción	`Buscar datos EXIF y otros metadatos`

El segundo programa `scorpion` recibirá archivos de imagen como parámetros y será capaz de analizarlos en busca datos EXIF y otros metadatos, mostrándolos en pantalla. El programa será compatible, al menos, con las mismas extensiones que gestiona `spider`. Deberá mostrar atributos básicos como la fecha de creación, así como otros datos EXIF. El formato en el que se muestren los metadatos queda a tu elección.

```
./scorpion FILE1 [FILE2 ...]
```

Capítulo VI

Parte Bonus

La evaluación de los bonus se hará `SI Y SOLO SI` la parte obligatoria es `PERFECTA`. De lo contrario, los bonus serán totalmente `IGNORADOS`.

Puedes mejorar tu proyecto con las siguientes características:

- Compatibilidad de ambos programas con `.docx` y `.pdf`.
- Interfaz gráfica para la visualización y el manejo de los metadatos.
- Eliminación de los metadatos
- Modificación de los metadatos

Capítulo VII

Evaluación por pares

Este proyecto será corregido por tus compañeros. Entrega los archivos en el repositorio Git y asegúrate de que todo funciona como se espera.