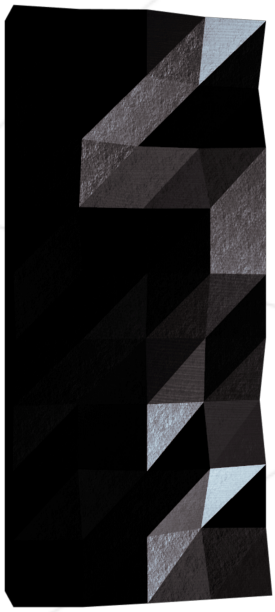

Cybersecurity Piscine

Arachnida

Summary: Introductory project to web scraping and metadata.

Version: 1.00

Contents

Chapter I

Introduction

This project will allow you to process data from the web.

- First you will create a small program to automatically extract information from the web.
- Then you will create a second program to analyze these files and manipulate the metadata.

Metadata is information which purpose is to describe other data. It's essentially **data about data**. It is frequently used to describe information contained on images and documents, and can reveal sensitive information about those who have created or manipulated them.

Chapter II

Prologue

Arachnids are a class of chelicerate arthropods among which there are more than 100,000 different species populating the planet. Among them are spiders, but also ticks, scorpions or mites. The most characteristic common feature of arachnids is their four pairs of legs, as well as their `chelicerae`, pointed appendages that they use to grab food.

Chapter III

Mandatory Part

You must create two programs. The two programs can be scripts or binaries.

In the case of compiled languages, you must include all the source code and compile it during evaluation.

You can use functions or libraries that allow you to create HTTP requests and handle files, but the logic of each program must be developed by yourself.

So, using wget or scrapy will be considered cheating and this project will be graded 0.

Chapter IV

Exercice 1 - Spider

The `spider` program allow you to extract all the images from a website, recursively, by providing a url as a parameter.

You have to manage the following program options:

`./spider [-rlpS] URL`

- Option `-r` : recursively downloads the images in a URL received as a parameter.
- Option `-r -l [N]` : indicates the maximum depth level of the recursive download. If not indicated, it will be 5.
- Option `-p [PATH]` : indicates the path where the downloaded files will be saved. If not specified, `./data/` will be used.

The program will download the following extensions by default:

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

Chapter V

Exercice 2 - Scorpion

The second `scorpion` program receive image files as parameters and must be able to parse them for EXIF and other metadata, displaying them on the screen.

The program must at least be compatible with the same extensions that `spider` handles.

It display basic attributes such as the creation date, as well as EXIF data. The output format is up to you.

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

Chapter VI

Bonus Part

You can improve your project with the following features:

- Add an option to your scorpion program. This option must allow you to modify/delete the metadata of a given file.
- A nice graphical interface for viewing and managing metadata.
- Add an option to your scorpion program that allows searching by dorks on a search engine.

The bonus part will only be assessed if the mandatory part is PERFECT. Perfect means the mandatory part has been integrally done and works without malfunctioning. If you have not passed ALL the mandatory requirements, your bonus part will not be evaluated at all.

Chapter VII

Submission and peer-evaluation

Turn in your assignment in your `Git` repository as usual. Only the work inside your repository will be evaluated during the defense. Don't hesitate to double check the names of your folders and files to ensure they are correct.