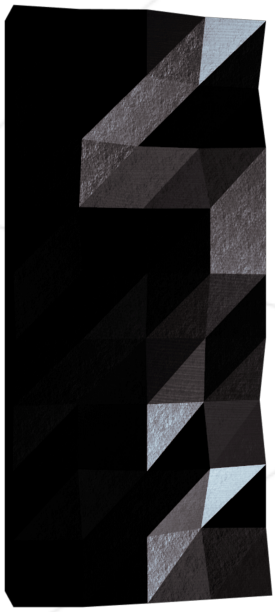

Cybersecurity Bootcamp | 42 Madrid

arachnida

Summary: Web scraping and metadata

Version: 1

Contents

I	**Introduction**	**2**
II	**Prologue**	**3**
III	**Mandatory Part**	**4**
IV	**Exercice 1 - Spider**	**5**
V	**Exercice 2 - Scorpion**	**6**
VI	**Bonus Part**	**7**
VII	**Peer evaluation**	**8**

Chapter I

Introduction

`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.

In this this project you will create two instruments that will allow you to automatically extract information from the `web` and then analyze it to discover or eliminate sensitive data.

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.

Theridion Grallator · CC Wikimedia Commons*

Chapter III

Mandatory Part

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

Program name	`spider`
Turn in files	`spider`
External functs.	`Nada`
Description	`Extract all images from a website`

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

Program name	`scorpion`
Turn in files	`scorpion`
External functs.	`Nada`
Description	`Search for EXIF data and other metadata`

The second `scorpion` program will receive image files as parameters and will be able to parse them for EXIF and other metadata, displaying them on the screen. The program will at least be compatible with the same extensions that `spider` handles. It will 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

The evaluation of the bonuses will be done `IF AND ONLY IF` the mandatory part is `PERFECT`. Otherwise, the bonuses will be totally `IGNORED`.

You can enhance your project with the following features:

- Compatibility with `.docx` and `.pdf`, both on Spider and Scorpion.
- A nice graphical interface for viewing and managing metadata.
- Metadata deletion.
- Metadata editing.
- Everything that comes to your mind... you will be able to justify everything during the defense.

Chapter VII

Peer evaluation

This project will be corrected by other students. Turn in the files inside the Git repository and make sure everything works as expected.