Creepy Crawlies

Web crawling is vast and intricate, but you don't have to embark on this journey alone. A plethora of web crawling tools are available to assist you, each with its own strengths and specialties. These tools automate the crawling process, making it faster and more efficient, allowing you to focus on analyzing the extracted data.

Popular Web Crawlers

1. `Burp Suite Spider`: Burp Suite, a widely used web application testing platform, includes a powerful active crawler called Spider. Spider excels at mapping out web applications, identifying hidden content, and uncovering potential vulnerabilities.
2. `OWASP ZAP (Zed Attack Proxy)`: ZAP is a free, open-source web application security scanner. It can be used in automated and manual modes and includes a spider component to crawl web applications and identify potential vulnerabilities.
3. `Scrapy (Python Framework)`: Scrapy is a versatile and scalable Python framework for building custom web crawlers. It provides rich features for extracting structured data from websites, handling complex crawling scenarios, and automating data processing. Its flexibility makes it ideal for tailored reconnaissance tasks.
4. `Apache Nutch (Scalable Crawler)`: Nutch is a highly extensible and scalable open-source web crawler written in Java. It's designed to handle massive crawls across the entire web or focus on specific domains. While it requires more technical expertise to set up and configure, its power and flexibility make it a valuable asset for large-scale reconnaissance projects.

Adhering to ethical and responsible crawling practices is crucial no matter which tool you choose. Always obtain permission before crawling a website, especially if you plan to perform extensive or intrusive scans. Be mindful of the website's server resources and avoid overloading them with excessive requests.

Scrapy

We will leverage Scrapy and a custom spider tailored for reconnaissance on `inlanefreight.com`. If you are interested in more information on crawling/spidering techniques, refer to the "Using Web Proxies" module, as it forms part of CBBH as well.

Installing Scrapy

Before we begin, ensure you have Scrapy installed on your system. If you don't, you can easily install it using pip, the Python package installer:

Creepy Crawlies

```
MisaelMacias@htb[/htb]$ pip3 install scrapy
```

This command will download and install Scrapy along with its dependencies, preparing your environment for building our spider.

ReconSpider

First, run this command in your terminal to download the custom scrapy spider, `ReconSpider`, and extract it to the current working directory.

Creepy Crawlies

```
MisaelMacias@htb[/htb]$ wget -O ReconSpider.zip https://academy.hackthebox.com/storage/modules/144/ReconSpider.v1.2.zi
MisaelMacias@htb[/htb]$ unzip ReconSpider.zip
```

With the files extracted, you can run `ReconSpider.py` using the following command:

Creepy Crawlies

```
MisaelMacias@htb[/htb]$ python3 ReconSpider.py http://inlanefreight.com
```

Replace `inlanefreight.com` with the domain you want to spider. The spider will crawl the target and collect valuable information.

results.json

After running `ReconSpider.py`, the data will be saved in a JSON file, `results.json`. This file can be explored using any text editor. Below is the structure of the JSON file produced:

Code: json

```
{
    "emails": [
        "lily.floid@inlanefreight.com",
        "cvs@inlanefreight.com",
        ...
    ],
    "links": [
        "https://www.themeansar.com",
        "https://www.inlanefreight.com/index.php/offices/",
        ...
    ],
    "external_files": [
        "https://www.inlanefreight.com/wp-content/uploads/2020/09/goals.pdf",
        ...
    ],
    "js_files": [
        "https://www.inlanefreight.com/wp-includes/js/jquery/jquery-migrate.min.js?ver=3.3.2",
        ...
```

Cheat Sheet

Go to Questions

Table of Contents

Introduction
Introduction

WHOIS
WHOIS
Utilizing WHOIS

DNS & Subdomains
DNS
Digging DNS
Subdomains
Subdomain Bruteforcing
DNS Zone Transfers
Virtual Hosts
Certificate Transparency Logs

Fingerprinting
Fingerprinting

Crawling
Crawling
robots.txt
.Well-Known URIs
Creepy Crawlies

Search Engine Discovery
Search Engine Discovery

Web Archives
Web Archives

Automating Recon
Automating Recon

Skills Assessment
Skills Assessment

My Workstation

OFFLINE

Start Instance

∞ / 1 spawns left

```
    ],
    "form_fields": [],
    "images": [
        "https://www.inlanefreight.com/wp-content/uploads/2021/03/AboutUs_01-1024x810.png",
        ...
    ],
    "videos": [],
    "audio": [],
    "comments": [
        "<!-- #masthead -->",
        ...
    ]
}
```


Each key in the JSON file represents a different type of data extracted from the target website:

JSON Key	Description
`emails`	Lists email addresses found on the domain.
`links`	Lists URLs of links found within the domain.
`external_files`	Lists URLs of external files such as PDFs.
`js_files`	Lists URLs of JavaScript files used by the website.
`form_fields`	Lists form fields found on the domain (empty in this example).
`images`	Lists URLs of images found on the domain.
`videos`	Lists URLs of videos found on the domain (empty in this example).
`audio`	Lists URLs of audio files found on the domain (empty in this example).
`comments`	Lists HTML comments found in the source code.

By exploring this JSON structure, you can gain valuable insights into the web application's architecture, content, and potential points of interest for further investigation.

Connect to Pwnbox
Your own web-based Parrot Linux instance to play our labs.

Pwnbox Location

UK 164ms

Terminate Pwnbox to switch location

Start Instance

∞ / 1 spawns left

Waiting to start...

Enable step-by-step solutions for all questions

Questions

Answer the question(s) below to complete this Section and earn cubes!

Cheat Sheet

+1 After spidering inlanefreight.com, identify the location where future reports will be stored. Respond with the full domain, e.g., files.inlanefreight.com.

inlanefreight-comp133.s3.amazonaws.htb

Submit Hint

Previous Next

Mark Complete & Next

Powered by HACKTHEBOX