Utilizing the search engine

The Kali Linux tools grouped in this category can be used to collect domain, e-mail address, and document metadata information from the target. These tools use a search engine to do their actions. The advantage of these tools is that they use search engine sites. So, you don't access the target website yourself, instead the search engine site will do that for you. As a result, the target website will not know about your action.

Let us explore several of these tools.

theharvester

The `theharvester` tool is an e-mail accounts, username, and hostname/subdomains gathering tool. It collects information from various public sources. As of Version 2.2, the public sources that are supported are as follows:

- Google
- Google profiles
- Bing
- PGP
- LinkedIn
- Yandex
- People123
- Jigsaw
- Shodan

To access `theharvester` in Kali Linux, you can use the console and type the following command:

```
# theharvester
```

This will display the usage information and example on your screen.

If we want to find the e-mail addresses and hostnames for a target domain using Google and limit the result to 100, the following is the appropriate command:

```
# theharvester -d example.com -l 100 -b google
```

The following e-mail addresses and hostnames are found:

```
[-] Searching in Google:
        Searching 0 results...

[+] Emails found:
------------------
info@ example .com
user1@ example .com
user2@ example .com
user3@ example .com

[+] Hosts found in search engines:
------------------------------------
192.168.118.14:sd1. example .com
192.168.118.14:sd2. example .com
192.168.118.14:event. example .com
192.168.118.14:test. example .com
```

```
203.34.118.7:nms. example .com
```

From the preceding result, we notice that we are able to get several e-mail addresses and hostnames from the Google search engine.

If we want to gather more information, let's say we want to collect the username from the target, we can use linkedin.com to do this. The following is the command for that:

```
# theharvester -d example.com -l 100 -b linkedin
```

The following is the result:

```
[-] Searching in Linkedin..
        Searching 100 results..
Users from Linkedin:

user1
user2
user3
user4
user5
user6

Total results:  6
```

The preceding list of usernames collected from LinkedIn will be useful in a penetration testing step later if we want to do an attack, such as a social engineering attack.

Metagoofil

Metagoofil is a tool that utilizes the Google search engine to get metadata from the documents available in the target domain. Currently, it supports the following document types:

- Word document (`.docx` , `.doc`)
- Spreadsheet document (`.xlsx` , `.xls` , `.ods`)
- Presentation file (`.pptx` , `.ppt` , `.odp`)
- PDF file (`.pdf`)

Metagoofil works by performing the following actions:

- Searching for all of the preceding file types in the target domain using the Google search engine
- Downloading all of the documents found and saving them to the local disk
- Extracting the metadata from the downloaded documents
- Saving the result in an HTML file

The metadata that can be found are as follows:

- Usernames
- Software versions
- Server or machine names

This information can be used later on to help in the penetration testing phase.

To access `Metagoofil` , go to the console and execute the following command:

```
# metagoofil
```

This will display a simple usage instruction and example on your screen.

As an example of `Metagoofil` usage, we will collect all the DOC and PDF documents (`-t` `.doc` , `.pdf`) from a target domain (`-d example.com`) and save them to a directory named `test` `(-o test`). We limit the search for each file type to 20 files (`-l 20`) and only download five files (`-n 5`). The report generated will be saved to `test.html` (`-f test.html`). We give the following command:

```
# metagoofil -d example.com -l 20 -t doc,pdf -n 5 -f test.html -o test
```

The redacted result of this command is as follows:

```
[-] Starting online search...
[-] Searching for doc files, with a limit of 200
        Searching 100 results...
        Searching 200 results...
Results: 191 files found
Starting to download 5 of them:
----------------------------------------

[1/5] /support/websearch/bin/answer.py?answer=186645&amp;%20form=bb&
amp;hl=en
Error downloading /support/websearch/bin/answer.py?answer=186645&
amp;%20form=bb&amp;hl=en
[2/5] http://www. example .com/documents/customerevidence
/27402_Cakewalk_final.doc
[3/5] http:// www. example .com/documents/customerevidence
/5588_marksspencer.doc
[4/5] http:// www. example .com/documents/uk/Ladbrokes.doc
[5/5] http:// www. example .com/~Gray/papers/PITAC_Interim_Report_8_98.doc

[-] Searching for pdf files, with a limit of 200
        Searching 100 results...
        Searching 200 results...
Results: 202 files found
Starting to download 5 of them:
----------------------------------------

[1/5] /support/websearch/bin/answer.py?answer=186645&amp;%20form=bb&
amp;hl=en
Error downloading /support/websearch/bin/answer.py?answer=186645&
amp;%20form=bb&amp;hl=en
[2/5] http:// www. example .com/pubs/77954/sl021801.pdf
[3/5] http:// www. example .com/pubs/152133/deepconvexnetwork-
interspeech2011-pub.pdf
[x] Error in the parsing process
[4/5] http:// www. example .com/en-us/collaboration/papers/uruguay.pdf
[5/5] http:// www. example .com/pubs/63611/2002-droppo-icslpb.pdf

[+] List of users found:
--------------------------
Benjamin Van Houten
Marketing
IT
```

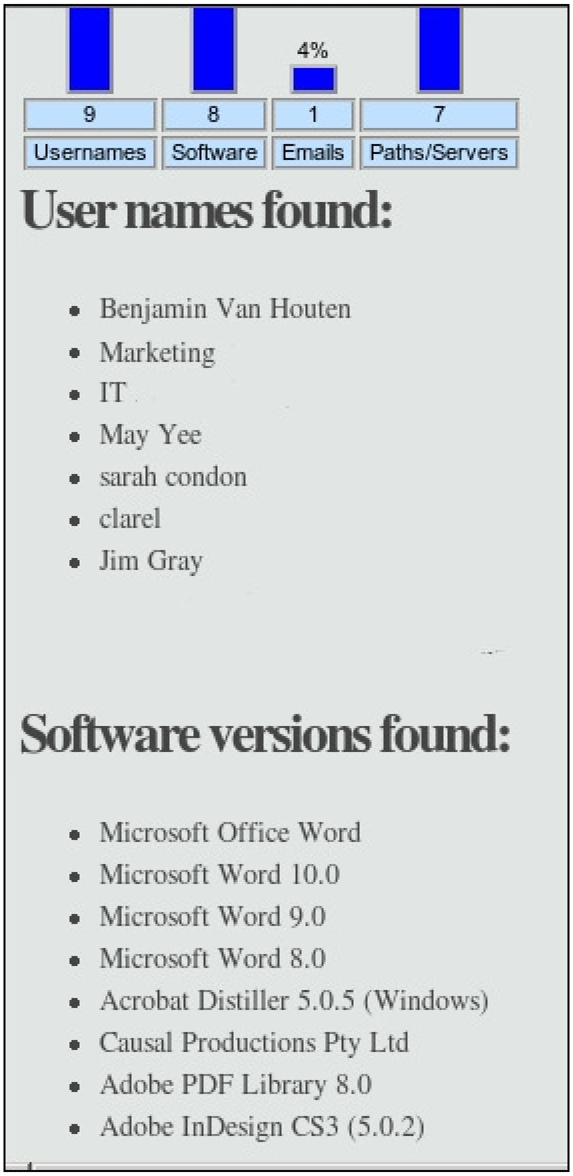
```
May Yee
sarah condon
clarel
Jim Gray

[+] List of software found:
-----------------------------
Microsoft Office Word
Microsoft Word 10.0
Microsoft Word 9.0
Microsoft Word 8.0
Acrobat Distiller 5.0.5 (Windows)
Adobe PDF Library 8.0
Adobe InDesign CS3 (5.0.2)
[+] List of paths and servers found:
---------------------------------------
'Macintosh HD:Temporary Items:AutoRecovery save of Congressio'
'NCO Server:Staff (NCO Staff):Yolanda Comedy:IR22July:IR10Aug'
'C:\jim\HPCC\PACIT_Report_8_98.doc'

[+] List of e-mails found:
----------------------------
gzweig@mail. example .com
```

You can see from the preceding result that we get a lot of information from the documents we have collected, such as the usernames and path information. We can use the obtained usernames to look for patterns in the username and for launching a brute force password attack on the usernames. But, be aware that doing a brute force password attack on an account may have the risk of locking the user accounts. The path information can be used to guess the operating system that is used by the target. We got all of this information without going to the domain website ourselves.

Metagoofil is also able to generate information in a report format. The following screenshot shows the generated report in HTML:



User names found:

- Benjamin Van Houten
- Marketing
- IT
- May Yee
- sarah condon
- clarel
- Jim Gray

Software versions found:

- Microsoft Office Word
- Microsoft Word 10.0
- Microsoft Word 9.0
- Microsoft Word 8.0
- Acrobat Distiller 5.0.5 (Windows)
- Causal Productions Pty Ltd
- Adobe PDF Library 8.0
- Adobe InDesign CS3 (5.0.2)

In the report generated, we get information about usernames, software version, e-mail address, and server information from the target domain.