CeWL - Custom Word List generator

Based on a discussion on PaulDotCom episode 129 about creating custom word lists by spidering a targets website and collecting unique words I decided to write CeWL, the Custom Word List generator. CeWL is a ruby app which spiders a given url to a specified depth, optionally following external links, and returns a list of words which can then be used for password crackers such as John the Ripper.

CeWL also has an associated command line app, FAB (Files Already Bagged) which uses the same meta data extraction

Table of Contents

techniques to create author/creator lists from already downloaded.

Change Log

Version 5.1

Added the GPL-3+ licence to allow inclusion in Debian.

Added a Gemfile to make installing gems easier.

Version 5.0

Adds proxy support from the command line and the ability to pass in credentials for both basic and digest authentication. Usage is simple, check the help (`--help`) for full information.

A few other smaller bug fixes as well.

Version 4.3

CeWL now sorts the words found by count and optionally (new --count argument) includes the word count in the output. I've left the words in the case they are in the pages so "Product" is different to "product" I figure that if it is being used for password generation then the case may be significant so let the user strip it if they want to. There are also more improvments to the stability of the spider in this release.

By default, CeWL sticks to just the site you have specified and will go to a depth of 2 links, this behaviour can be changed by passing arguments. Be careful if setting a large depth and allowing it to go offsite, you could end up drifting on to a lot of other domains. All words of three characters and over are output to stdout. This length can be increased and the words can be written to a file rather than screen so the app can be automated.

Version 4.2

Version 4.2 fixes a pretty major bug that I found while fixing a smaller bug for @yorikv. The bug was related to a hack I had to put in place because of a problem I was having with the spider, while I was looking in to it I spotted this line which is the one that the

Categories

Support The Site

I don't get paid for any of the projects on this site so if you'd like to support my work please use the link below when buying from Amazon and I will receive a small commission on all purchases.

spider uses to find new links in downloaded pages:

```
web_page.scan(/href="(.*?)"/i).flatten.map do |link|
```

This is fine if all the links look like this:

```
<a href="test.php">link</a>
```

But if the link looks like either of these:

```
<a href='test.php'>link</a>
<a href=test.php>link</a>
```

the regex will fail so the links will be ignored.

To fix this up I've had to override the function that parses the page to find all the links, rather than use a regex I've changed it to use Nokogiri which is designed to parse a page looking for links rather than just running through it with a custom regex. This brings in a new dependency but I think it is worth it for the fix to the functionality. I also found another bug where a link like this:

```
<a href='#name'>local</a>
```

which should be ignored as it just links to an internal name was actually being translated to '/#name' which may unintentionally mean referencing the index page. I've fixed this one as well after a lot of debugging to find how best to do it.

A final addition is to allow a user to specify a depth of 0 which allows CeWL to spider a single page.

I'm only putting this out as a point release as I'd like to rewrite the spidering to use a better spider, that will come out as the next major release.

Version 4.1

Version 4.1 is mainly bug fixes but one important feature change is the addition of two new parameters, meta_file and email_file. Previously you specified the filename for email and metadata output as optional fields to the email and meta parameters but I found that if you used the parameters in a specific order you could end up with this:

```
./cewl.rb --email http://www.digininja.org
```

This would take the URL as the output filename for the email parameter which isn't what is meant, hence removing the optional filename from the email parameter and adding the email_file parameter instead.

The main change in version 4 is the upgrade to run with Ruby 1.9.x, this has been tested on various machines and on BT5 as that is a popular platform for running it and it appears to run fine. Another minor change is that Up to version 4 all HTML tags were stripped out before the page was parsed for words, this meant that text in alt and title tags were missed. I now grab the text from those tags before stripping the HTML to give those extra few works.

Version 3.0

Version 3 of CeWL addresses a problem spotted by Josh Wright. The Spider gem doesn't handle JavaScript redirection URLs, for exmaple an index page containing just the following:

```
<script language="JavaScript">
self.location.href =
'http://www.FOO.com/FOO/connect/FOONet/Top+Navigator
</script>
```

wasn't spidered because the redirect wasn't picked up. I now scan

through a page looking for any lines containing "location.href=" and then add the given URL to the list of pages to spider.

Version 2.0

Version 2 of CeWL can also create two new lists, a list of email addresses found in mailto links and a list of author/creator names collected from meta data found in documents on the site. It can currently process documents in Office pre 2007, Office 2007 and PDF formats. This user data can then be used to create the list of usernames to be used in association with the password list.

Pronunciation

Seeing as I was asked, CeWL is pronounced "cool".

Download

download cewl version 5.1
download cewl version 5.0
download cewl version 4.3
download cewl version 4.2
download cewl version 4.1
download cewl version 3.0

Installation

CeWL needs the rubygems package to be installed along with the following gems:

nokogiri

mime-types

mini_exiftool

rubyzip

spider

These can be installed by running

```
bundle install
```

from the cewl directory. The mini_exiftool gem also requires the exiftool application to be installed.

On BT5 there is a problem with the version of Ruby installed by default. To get around this I've found the following works well on a brand new BT5 install:

```
gem source -c
gem install --user-install spider http_configuration
```

To use the gems you may also need to set the following environment variable:

```
RUBYOPT="rubygems"
```

Then just save CeWL to a directory and make it executable.

Usage

cewl [OPTION] ... URL

--help, -h
    Show help
--depth x, -d x
    The depth to spider to, default 2
--min_word_length, -m
    The minimum word length, this strips out all words under the specified length, default 3
--offsite, -o
    By default, the spider will only visit the site specified. With this option it will also visit external sites
--write, -w file
    Write the ouput to the file rather than to stdout

--ua, -u user-agent
    Change the user agent
-v
    Verbose, show debug and extra output
--no-words, -n
    Don't output the wordlist
--meta, -a file
    Include meta data, optional output file
--email, -e file
    Include email addresses, optional output file
--meta_file file
    Filename for metadata output
--email_file file
    Filename for email output
--meta-temp-dir directory
    The directory used used by exiftool when parsing files, the
    default is /tmp
--count, -c:
    Show the count for each of the words found
--auth_type
    Digest or basic
--auth_user
    Authentication username
--auth_pass
    Authentication password
--proxy_host
    Proxy host
--proxy_port
    Proxy port, default 8080
--proxy_username
    Username for proxy, if required
--proxy_password
    Password for proxy, if required
--verbose, -v
    Verbose
URL
    The site to spider.

Common Problems

Here are a couple of the common problems people have seen while trying to use CeWL and FAB.

Missing exiftool

If you see this error while trying to run either CeWL or FAB

```
/usr/lib/ruby/gems/1.8/gems/mini_exiftool-1.0.1/lib/
from /usr/lib/ruby/gems/1.8/gems/mini_exiftool-1.0.1
from /usr/local/lib/site_ruby/1.8/rubygems/custom_re
from /usr/local/lib/site_ruby/1.8/rubygems/custom_re
from ./cewl_lib.rb:1
from /usr/local/lib/site_ruby/1.8/rubygems/custom_re
from /usr/local/lib/site_ruby/1.8/rubygems/custom_re
from ./cewl.rb:58
```

then the application can't access exiftool. Either install it or make sure it is in your path.

HTTPS Problem

It has been reported that if you see this problem

```
/usr/local/lib/site_ruby/1.8/rubygems/custom_require
from /usr/local/lib/site_ruby/1.8/rubygems/custom_re
from /usr/lib/ruby/gems/1.8/gems/spider-0.4.4/lib/sp
from /usr/local/lib/site_ruby/1.8/rubygems/custom_re
from /usr/local/lib/site_ruby/1.8/rubygems/custom_re
from /usr/lib/ruby/gems/1.8/gems/spider-0.4.4/lib/sp
from /usr/local/lib/site_ruby/1.8/rubygems/custom_re
from /usr/local/lib/site_ruby/1.8/rubygems/custom_re
from ./cewl.rb:56
```

Then you need the Ruby libopenssl package. In Debian the package is called libopenssl-ruby.

Spider Missing Pages

Someone has reported that the spider misses some pages which are have querystrings on them. I haven't been able to reproduce this in my tests. If anyone has this problem and can reproduce it please let me know and I'll investigate it further.

Change Log

Keeping track of history.

Version 4.3 - Various spider bug fixes and the introduction of the sorting the results by count

Version 4.2 - Fixed the Spider gem by overriding the function, also handling #name links correctly

Version 4.1 - Small bug fixes and added new parameter to set filenames for email and metadata output

Version 4 - Runs with Ruby 1.9.x and grabs text out of alt and title tags

Version 3 - Now spiders pages referenced in JavaScript location commands

Version 2.2 - Data from email addresses and meta data can be written to their own files

Version 2.1 - Fixed a bug some people were having while using the email option

Version 2 - Added meta data support

Version 1 - released

Ruby Doc

CeWL is commented up in Ruby Doc format.