

全部课程 (/courses/) / Python打造漏洞扫描器 (/courses/761) / CMS识别(web指纹识别)扫描器开发

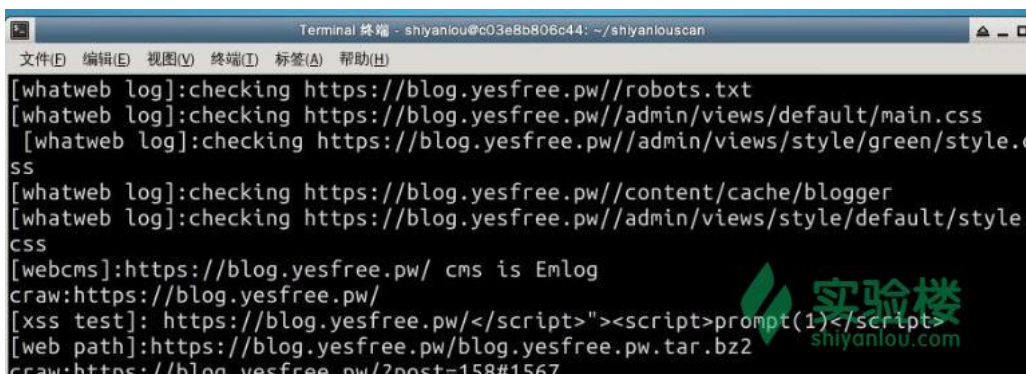
在线实验，请到PC端体验

CMS识别(web指纹识别)扫描器开发

一、实验介绍

1.1 实验内容

在网站渗透的过程中，我们可以对网站进行扫描识别出使用的程序。今天我们就来做这个WEB指纹识别工具。



```
Terminal 终端 - shiyanlou@c03e8b806c44: ~/shiyanlouscan
[whatweb log]:checking https://blog.yesfree.pw//robots.txt
[whatweb log]:checking https://blog.yesfree.pw//admin/views/default/main.css
[whatweb log]:checking https://blog.yesfree.pw//admin/views/style/green/style.css
[whatweb log]:checking https://blog.yesfree.pw//content/cache/blogger
[whatweb log]:checking https://blog.yesfree.pw//admin/views/style/default/style.css
[webcms]:https://blog.yesfree.pw/ cms is Emlog
craw:https://blog.yesfree.pw/
[xss test]: https://blog.yesfree.pw/</script>"><script>prompt(1)</script>
[web path]:https://blog.yesfree.pw/blog.yesfree.pw.tar.bz2
craw:https://blog.yesfree.pw/?post=158#1567
```

1.2 实验知识点

- web指纹收集
- 指纹分析
- md5校验
- 关键词校验

1.3 实验环境

- Python2.7
- Xfce终端
- Sublime

1.4 适合人群

本课程难度为一般，属于初级级别课程，适合具有Python基础的用户，熟悉python基础知识加深巩固。

1.5 代码获取

你可以通过下面命令将代码下载到实验楼环境中，作为参照对比进行学习。

```
$ wget http://labfile.oss.aliyuncs.com/courses/761/shiyanloucan5.zip
$ unzip shiyanloucan5.zip
```

二、实验原理

CMS识别原理

CMS英文全称是：Content Management System 中文名称是：网站内容管理系统

动手实践是学习 IT 技术最有效的方式！

开始实验

CMS识别原理就是得到一些CMS的一些固有特征，通过得到这个特征来判断CMS的类别。

这里我们使用MD5识别和正则表达式识别的方式，就是用特定的文件路径访问网站，获得这个文件的MD5或者用正则表达式匹配某个关键词，如果匹配成功就说明这个是这个CMS。

所以，这个识别的成功率是根据我们的字典来的，这里，作者给大家提供了作者精心收集的1400+国内外网络常见指纹，拥有这些指纹，相信识别主流网站程序已经没有问题。怎么了？激动了吗，来看看我们怎么编写的。

三、实验步骤

3.1 指纹格式

这里截取一些web指纹作为参考：

```
{
  "url": "/install/",
  "re": "aspcms",
  "name": "AspCMS",
  "md5": ""
},
{
  "url": "/about/_notes/dwsync.xml",
  "re": "aspcms",
  "name": "AspCMS",
  "md5": ""
},
{
  "url": "/admin/_Style/_notes/dwsync.xml",
  "re": "aspcms",
  "name": "AspCMS",
  "md5": ""
},
{
  "url": "/apply/_notes/dwsync.xml",
  "re": "aspcms",
  "name": "AspCMS",
  "md5": ""
},
{
  "url": "/tpl/green/common/images/notebg.jpg",
  "re": "",
  "name": "自动发卡平台",
  "md5": "690f337298c331f217c0407cc11620e9"
},
{
  "url": "/images/download.png",
  "re": "",
  "name": "全程oa",
  "md5": "9921660baaf9e0b3b747266eb5af880f"
},
{
  "url": "/kindeditor/license.txt",
  "re": "",
  "name": "T-Site建站系统",
  "md5": "b0d181292c99cf9bb2ae9166dd3a0239"
},
{
  "url": "/public/ico/favicon.png",
  "re": "",
  "name": "悟空CRM",
  "md5": "834089ffa1cd3a27b920a335d7c067d7"
},
{
  "url": "/public/js/php/file_manager_json.php",
  "re": "",
  "name": "悟空CRM",
  "md5": "c64fd0278d72826eb9041773efa1f587"
},
{
  "url": "/plugins/weathermap/images/exclamation.png",
  "re": "",
  "name": "CactiEZ插件",
  "md5": "2e25cb083312b0eabfa378a89b07cd03"
}
```

动手实践是学习 IT 技术最有效的方式！

开始实验

可以看到，我们提供的是json格式，好处是可以方便以后再其他语言上的复用。

3.2 指纹文件

我们在 data 目录下存放了 data.json 文件格式的web指纹，总共有1400+的国内常见指纹，大家可以在终端上输入。

```
wget http://labfile.oss.aliyuncs.com/courses/761/shiyanlouscan5.zip
unzip shiyanlouscan5
```

获取源码后进入源码目录，然后再 data/data.json 文件就是我们的web指纹识别文件。

3.3 记录

思路虽然简单，但实现起来还是有很多问题的，比如效率问题，1000+指纹说明需要访问1000+的网页，单步的话速度太慢了，所以我们会使用线程，等用多了也会发现线程也太慢了，所以我们可以用协程，不过这个得等到后面扫描器升级的时候再说到，我们现在只是做出雏形，不需要太过于专注于效率，所以我就使用多线程来完成这个过程了。

3.4 代码编写

新建文件 lib/core/webcms.py，代码如下：

```
#!/usr/bin/env python
# __author__ = 'w8ay'
import json,os,sys,hashlib,threading,Queue
from lib.core import Download

class webcms(object):
    workQueue = Queue.Queue()
    URL = ""
    threadNum = 0
    NotFound = True
    Downloader = Download.Downloader()
    result = ""

    def __init__(self,url,threadNum = 10):
        self.URL = url
        self.threadNum = threadNum
        filename = os.path.join(sys.path[0], "data", "data.json")
        fp = open(filename)
        webdata = json.load(fp,encoding="utf-8")
        for i in webdata:
            self.workQueue.put(i)
        fp.close()

    def getmd5(self, body):
        m2 = hashlib.md5()
        m2.update(body)
        return m2.hexdigest()

    def th_whatweb(self):
        if(self.workQueue.empty()):
            self.NotFound = False
            return False

        if(self.NotFound is False):
            return False
        cms = self.workQueue.get()
        _url = self.URL + cms["url"]
        html = self.Downloader.get(_url)
        print "[whatweb log]:checking %s"%_url
        if(html is None):
            return False
        if cms["re"]:
            if(html.find(cms["re"])!=-1):
                self.result = cms["name"]
                self.NotFound = False
                return True
        else:
            md5 = self.getmd5(html)
            if(md5==cms["md5"]):
                self.result = cms["name"]
                self.NotFound = False
                return True

    def run(self):
        while(self.NotFound):
            th = []
            for i in range(self.threadNum):
                t = threading.Thread(target=self.th_whatweb)
                t.start()
                th.append(t)
            for t in th:
                t.join()
            if(self.result):
                print "[webcms]:%s cms is %s"%(self.URL,self.result)
            else:
                print "[webcms]:%s cms NOTFound!"%self.URL
```

首先我们要读取cms指纹列表到队列中:

```
filename = os.path.join(sys.path[0], "data", "data.json")
fp = open(filename)
webdata = json.load(fp,encoding="utf-8")
for i in webdata:
    self.workQueue.put(i)
fp.close()
```

动手实践是学习 IT 技术最有效的方式!

开始实验

然后 run 方法就是创建线程用的：

```
def run(self):
    while(self.NotFound):
        th = []
        for i in range(self.threadNum):
            t = threading.Thread(target=self.th_whatweb)
            t.start()
            th.append(t)
        for t in th:
            t.join()
    if(self.result):
        print "[webcms]:%s cms is %s"%(self.URL,self.result)
    else:
        print "[webcms]:%s cms NOTFound!"%self.URL
```

线程调用的是 th_whatweb 方法：

```
def th_whatweb(self):
    if(self.workQueue.empty()):
        self.NotFound = False
        return False

    if(self.NotFound is False):
        return False
    cms = self.workQueue.get()
    _url = self.URL + cms["url"]
    html = self.Downloader.get(_url)
    print "[whatweb log]:checking %s"%_url
    if(html is None):
        return False
    if cms["re"]:
        if(html.find(cms["re"])!=-1):
            self.result = cms["name"]
            self.NotFound = False
            return True
    else:
        md5 = self.getmd5(html)
        if(md5==cms["md5"]):
            self.result = cms["name"]
            self.NotFound = False
            return True
```

当然了，线程我们需要考虑下情况进行退出。

```
if(self.workQueue.empty()):
    self.NotFound = False
    return False

if(self.NotFound is False):
    return False
```

然后后面的代码就是下载网页源码，然后进行分析了。

如何调用呢？

```
from lib.core import webcms

if __name__ == "__main__":
    webcms = webcms.webcms("http://blog.yesfree.pw/")
    webcms.run()
```

得到了CMS名称会自动打印出来。

3.5 调用

重写下主文件 w8ay.py 即可：

动手实践是学习 IT 技术最有效的方式！

开始实验

```
#!/usr/bin/env python
#-*- coding:utf-8 -*-
'''
Name:w8ayScan
Author:w8ay
Copyright (c) 2017
'''

import sys
from lib.core.Spider import SpiderMain
from lib.core import webcms

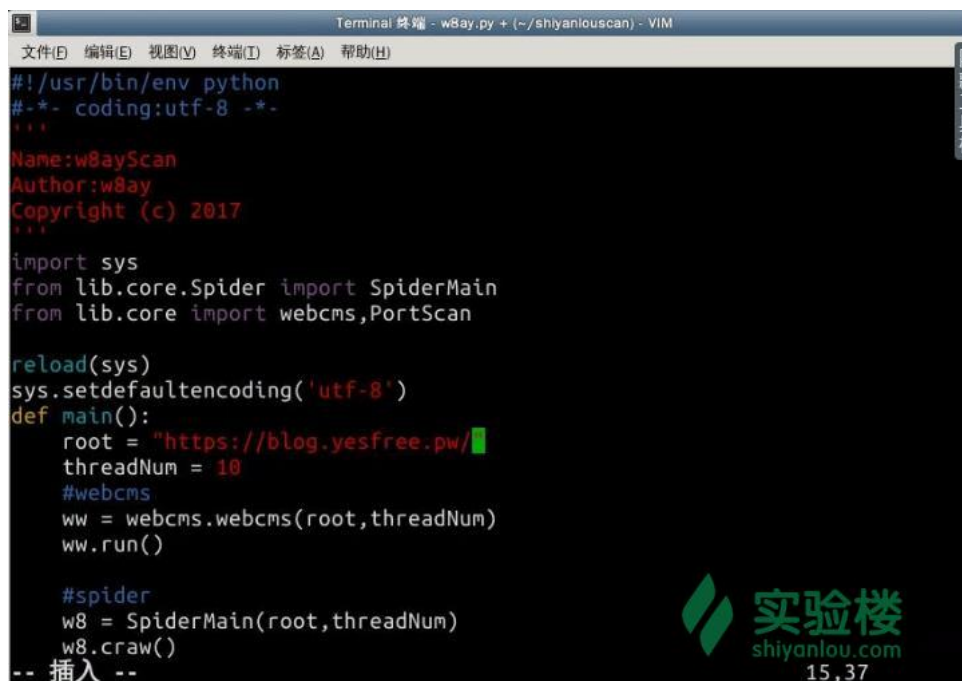
reload(sys)
sys.setdefaultencoding('utf-8')
def main():
    root = "https://www.shiyanlou.com/"
    threadNum = 10
    #webcms
    ww = webcms.webcms(root,threadNum)
    ww.run()

    #spider
    w8 = SpiderMain(root,threadNum)
    w8.craw()

if __name__ == '__main__':
    main()
```

当然，如果我们要测试cms识别的话可以把 root = "https://www.shiyanlou.com/" 改成 root = "https://blog.yesfree.pw/"。

因为实验楼是探测不到使用了什么程序的，可以探测下我的博客是什么程序？



因为我们是在开发扫描器，所以我们尽可能多的把信息输出方便以后的调试，下面是正在测试的webcms的时候进行访问的地址：

```
Terminal 终端 - shiyanlou@c03e8b806c44: ~/shiyanlouscan
文件(F) 编辑(E) 视图(V) 终端(T) 标签(A) 帮助(H)
[whatweb log]:checking https://blog.yesfree.pw//dede/templates/article_coonepage_
rule.htm
[whatweb log]:checking https://blog.yesfree.pw//include/alert.htm
[whatweb log]:checking https://blog.yesfree.pw//plus/sitemap.html
[whatweb log]:checking https://blog.yesfree.pw//setup/license.html
[whatweb log]:checking https://blog.yesfree.pw//member/js/box.js[whatweb log]:ch
ecking https://blog.yesfree.pw//php/modpage/readme.txt

[whatweb log]:checking https://blog.yesfree.pw//templates/default/style/dedecms.c
ss
[whatweb log]:checking https://blog.yesfree.pw//special/index.html
[whatweb log]:checking https://blog.yesfree.pw//company/template/default/search_
list.htm
[whatweb log]:checking https://blog.yesfree.pw//robots.txt
[whatweb log]:checking https://blog.yesfree.pw//
[whatweb log]:checking https://blog.yesfree.pw//
[whatweb log]:checking https://blog.yesfree.pw//bbcode.js
[whatweb log]:checking https://blog.yesfree.pw//u2upopup.js
[whatweb log]:checking https://blog.yesfree.pw//admin/discuzfiles.md5
[whatweb log]:checking https://blog.yesfree.pw//templates.cdb
[whatweb log]:checking https://blog.yesfree.pw//images/admindcp/admincp.js
[whatweb log]:checking https://blog.yesfree.pw//newsfader.js
[whatweb log]:checking https://blog.yesfree.pw//api/manyou/cloud_channel.htm
[whatweb log]:checking https://blog.yesfree.pw//include/javascript/ajax.js
```

最后我们看到我的博客程序被成功识别了出来：

```
Terminal 终端 - shiyanlou@c03e8b806c44: ~/shiyanlouscan
文件(F) 编辑(E) 视图(V) 终端(T) 标签(A) 帮助(H)
[whatweb log]:checking https://blog.yesfree.pw//robots.txt
[whatweb log]:checking https://blog.yesfree.pw//admin/views/default/main.css
[whatweb log]:checking https://blog.yesfree.pw//admin/views/style/green/style.c
ss
[whatweb log]:checking https://blog.yesfree.pw//content/cache/blogger
[whatweb log]:checking https://blog.yesfree.pw//admin/views/style/default/style.
css
[webcms]:https://blog.yesfree.pw/ cms is Emlog
craw:https://blog.yesfree.pw/
[xss test]: https://blog.yesfree.pw/</script>"><script>prompt(1)</script>
[web path]:https://blog.yesfree.pw/blog.yesfree.pw.tar.bz2
craw:https://blog.yesfree.pw/?post=158#1567
```

◀ 上一节 (/courses/761/labs/2649/document)

下一节 ▶ (/courses/761/labs/2670/document)

课程教师



new4
共发布过1门课程

查看老师的所有课程 > (/teacher/102428)



动手做实验，轻松学IT



公司 (http://weibo.com/shiyanlou2013)

合作

- 关于我们 (/aboutus)
- 联系我们 (/contact)
- 加入我们 (http://www.simplecloud.cn/jobs.html)
- 技术博客 (https://blog.shiyanlou.com)

- 我要投稿 (/contribute)
- 教师合作 (/labs)
- 高校合作 (/edu/)
- 友情链接 (/friends)
- 开发者 (/developer)
- 学习路径
 - Python学习路径 (/paths/python)
 - Linux学习路径 (/paths/linuxdev)
 - 大数据学习路径 (/paths/bigdata)
 - Java学习路径 (/paths/java)
 - PHP学习路径 (/paths/php)

- 服务
- 企业版 (/saas)
 - 实战训练营 (/bootcamp/)
 - 会员服务 (/vip)
 - 实验报告 (/courses/reports)
 - 常见问题 (/questions/?)

动手实践是学习 IT 技术最有效的方式，开始实验