



分享

生活

关于

大家抢着看

嗨,我是mokeyjay,也可以叫我小紫

Pixiv每日排行榜Top50小部件 v4.2

Yande.re 图片爬虫

Pixiv原图半自动下载器 v1.1

推荐与分享

Windows下实用好用的软件推荐

JetBrains全家桶正版注册码分享

Linux下实用好用的软件推荐

Chrome下实用好用的扩展推荐

WordPress下实用好用的插件推荐

分享自己精选的动漫壁纸包

Pixiv每日榜Top50



最新评论

xiaobeii  
催更催更mokeyjay  
网盘.....我一会儿加密传一份吧 O...贰磊  
网盘无法访问, OD点不进去,....大...我爱吃米线  
这个线刷包直接用刷机软件刷可以...超能小粉红  
我13点搜的windows全局快捷键占...

笔记本

Linux Shell常用命令笔记



## wget 递归下载整个网站(网站扒皮必备)

2016-02-28 • Linux、技术 • 10 条评论

有时间看到别人网站的页面比较漂亮,就想给扒皮下来,学习学习。分享一个我常用网站扒皮命令 [wget](#)

这个命令可以以递归的方式下载整站,并可以将下载的页面中的链接转换为本地链接。

wget加上参数之后,即可成为相当强大的下载工具。

### wget -r -p -np -k

```
wget -r -p -np -k http://xxx.com/xxx
```

Bash

-r, --recursive (递归) specify recursive download. (指定递归下载)

-k, --convert-links (转换链接) make links in downloaded HTML point to local files. (将下载的HTML页面中的链接转换为相对链接即本地链接)

-p, --page-requisites (页面必需元素) get all images, etc. needed to display HTML page. (下载所有的图片等页面显示所需的内容)

-np, --no-parent (不追溯至父级) don't ascend to the parent directory.

另外断点续传用-n参数 日志 用-o参数

拿我自己的网站扒皮试一下吧

```
wget -r -p -np -k https://wujunze.com/
```

等网站递归下载完毕,你会发现你当前目录会有一个 wujunze.com的目录  
进入这个目录看一下

```
install.log  tmp-install.log  tmp1.2-root.tar.gz  wujunze.com
[root@VM_202_23_centos ~]# cd wujunze.com/
[root@VM_202_23_centos wujunze.com]# ll
total 808
-rw-r--r-- 1 root root 20866 Feb 27 22:34 201512_language_top.jsp
drwxr-xr-x 2 root root 4096 Feb 27 22:33 action
-rw-r--r-- 1 root root 20367 Feb 27 22:34 alter.jsp
drwxr-xr-x 19 root root 4096 Feb 27 22:34 category
-rw-r--r-- 1 root root 25993 Feb 27 22:34 chrome_skill.jsp
drwxr-xr-x 6 root root 4096 Feb 27 22:34 feed
-rw-r--r-- 1 root root 31725 Feb 27 22:34 http_tcp_udp.jsp
-rw-r--r-- 1 root root 23983 Feb 27 22:34 index.html
-rw-r--r-- 1 root root 27998 Feb 27 22:34 item.jsp
-rw-r--r-- 1 root root 21769 Feb 27 22:34 linux_dd.jsp
-rw-r--r-- 1 root root 46909 Feb 27 22:34 linux_directory.jsp
-rw-r--r-- 1 root root 29464 Feb 27 22:34 linux_maintain.jsp
-rw-r--r-- 1 root root 21370 Feb 27 22:34 linux_tail.jsp
-rw-r--r-- 1 root root 27968 Feb 27 22:34 linux_top.jsp
-rw-r--r-- 1 root root 19848 Feb 27 22:34 linux_wget.jsp
-rw-r--r-- 1 root root 20321 Feb 27 22:34 mobiledetect.jsp
```

熟练掌握wget命令,可以帮助你扒皮网站。

以上转载自: [https://wujunze.com/linux\\_wget.jsp](https://wujunze.com/linux_wget.jsp)


### 小紫注

当然咯,这个命令只能扒静态资源,你就别指望把源码扒下来了,除非对方服务器配置有问题不解析服务端动态语言

你所看到截图中的.jsp文件其实只是类似于伪静态的方式实现的,实际上都是JSP执行后输出的html而已

◀ 上一篇 · 下一篇 ▶

10 条评论



说点什么吧...

昵称 ▼ 发表评论



雲途科技信息发展有限公司

2017-5-26 18:18

刚测试了下，这个命令除了字体下载不到，其他的完全没问题，非常给力。



Z4HD

2017-5-21 23:04

这个可以用来抢救一些有珍贵静态资料的网站比如FL吧导航



霜酱

2016-8-13 20:40

我去下国外菊苣的博客群http://komkon.net/下的了，可是我下国内的网站CSS丢了、、、



小刘

2016-6-23 14:53


这tm就有点尴尬了 他是怎么递归的



静静

2016-3-16 18:25


HTTTrack表示笑了



行云流火

2016-3-2 09:53

我比较想知道他是怎么递归的。。。因为网站一般并不提供filelist。。。。如果说从sitemap上看的话，命令中并没有相关语句啊== P.S.不觉得多说评论框真的好丑好不搭么。。。



mokeyjay 博主

2016-3-2 10:13


我也不知道，就是感觉好厉害，就先转了再说评论框啊，我觉得都是淡灰色还挺搭网页背景色的。。。



asd

2016-4-7 04:01

从html中读取载入的资源？



行云流火

2016-4-7 17:21

类似于蜘蛛么？那完全不能从任何页面中找到的页面是不是没法扒。。。

热门文章

- 记人生中第一次被请到派出所喝茶
- 爱否氮化镓GaN 2C1A充电器简单开箱
- 腾讯QQ 1024程序员节重大更新！Linux版回归！

我的小伙伴

Kenxix	萝莉社
思起	四次元领域
Pi.1415926	Sonic853
LWL的自由天空	Server Not Found
RHW Home	繁星啊~
喵の空	流觞曲水
逃跑计划	HydricAcid
Waxxh's Blog	Ver4's Blog
FGHRSH	小霖
鹊湖居士	KK 的博客
唐霜的博客	Pch18's Blog
龙缘博客	更多大佬...

PHP	绝对灵域	wordpress
Linux	PhpStorm	yii
MySql	nginx	软件
简评	公告	Google
JavaScript	Codeigni...	百度
三坪房...	Ubuntu	二次元
硬件	开箱	脑残



