

by etng via cheatography.com/51777/cs/14117/

辅助函数	
修剪	清理两端的空白

辅助函数 (cont)
取整 转化为整数



By **etng** cheatography.com/etng/

Published 28th December, 2017. Last updated 28th December, 2017. Page 1 of 100.



by etng via cheatography.com/51777/cs/14117/

辅助函数 (cont)

取浮点数

转化为小数

辅助函数 (cont)

↑清除 html 标签 清理掉 html 只保留文字

辅助函数 (cont)

提取图片

提取文中所有的图片地址



By **etng** cheatography.com/etng/

Published 28th December, 2017. Last updated 28th December, 2017. Page 2 of 100.



by etng via cheatography.com/51777/cs/14117/

爬虫配置	
4年 11	<b>化</b> 理 服 冬 哭 劢 丰

爬虫配置 (cont) 浏览器 浏览器列表



By **etng** cheatography.com/etng/

Published 28th December, 2017. Last updated 28th December, 2017. Page 3 of 100.



by etng via cheatography.com/51777/cs/14117/

	(cont)

超时 链接超时,避免服务器假死

爬虫配置 (cont)

ookie 登录信息

爬虫配置 (cont)

http 头 避免识别成非浏览器

以上都是为了防止认为是非浏览器



By **etng** cheatography.com/etng/

Published 28th December, 2017. Last updated 28th December, 2017. Page 4 of 100.



by etng via cheatography.com/51777/cs/14117/

#### 常见问题

爬取速度怎样

分布式爬取,一般5分钟爬完

#### 常见问题 (cont)

爬取频率如何

可以自由设定,爬取完一个再取得下一个

#### 常见问题 (cont)

可以增加 worker 数量不

只要主机数量增加,即可分配更多的 worker



By etng cheatography.com/etng/

Published 28th December, 2017. Last updated 28th December, 2017. Page 5 of 100.



by etng via cheatography.com/51777/cs/14117/

页面规则	
类型	CSS/XPATH

页面规则 (cont)	
内容	article.content>.main

各地区主机分布	
美国	12
荷兰	25
印度	30
中国	60
澳大利亚	22

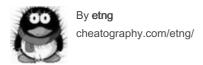


By **etng** cheatography.com/etng/

Published 28th December, 2017. Last updated 28th December, 2017. Page 6 of 100.



### SpiderMan Cheat Sheet by etng via cheatography.com/51777/cs/14117/



Published 28th December, 2017. Last updated 28th December, 2017. Page 7 of 100.