

# python用正则表达式将网页中所有 形式中的XXX的字符串提取出

```
import urllib
import re
def GetWebContent(url):
    '''
    获取网页源码
    '''
    try:
        wp = urllib.urlopen(url)
        content = wp.read()
    except:
        print u('获取失败, 请重试...')
        return ""
    return content
def sect(matchstr, rawstr):
    '''
    正则表达式截取 并返回匹配结果列表
    '''
    if len(matchstr) == 0 : return ;
    compile_obj = re.compile(rawstr, re.I|re.S)
    match_obj = compile_obj.findall(matchstr)
    #all_groups = match_obj.groups()
    return match_obj
if __name__ == '__main__':
    matchstr= GetWebContent("http://blog.ask3.cn/")
```

```
#rawstr=r' [\s\S].<img.*src="([\s\S]*?)".*>[\s\S].'  
rawstr=r'<img.*?>'  
groups=sect(matchstr,rawstr)  
for group in groups:  
    match_obj=re.search('src="(.*?)"',group)  
    group=match_obj.groups()  
    print group
```