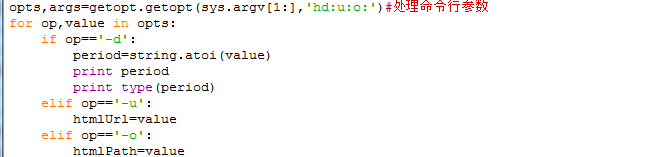
加入了处理命令行参数的程序：



为了增强程序的阅读性：

将主循环中的的代码封装成了函数，js，css，html，image的抓取代码都封装成了相对应的函数，将复用的代码也封装成了函数。函数的具体作用main.py里有相对应的注释。

补漏了关于css内联样式的抓取代码。

为了增加效率尝试下列的正则表达式先抓取所以所需的代码部分，然后在所抓取的buf里再次抓取对应的代码，以减少对原网页的重复抓取，试图减少每次循环的时间，亲测并没有什么用。。。所以还是之前的正则表达式。buf=re.findall('<script.+?</script>|<style.+?</style>|http.+css|style=".+?/"|http://.+?\.jpg|http://.+?\.png',buf1,re.S)

关于image部分：<img src 是插入图片的格式，但是一些背景图片和超链接图片并不满足这一格式。在对<img src进行抓取的时候发现，搜狐上所有的插入图片都是jpg和png格式。因为图片的格式较多，为了正确性，我们先对网页预处理一下

import urllib2

import re

f=urllib2.urlopen('http://m.sohu.com')

buf=f.read()

print buf

out = open('souhu.html','w')

out.write(buf)

out.close()

addrjs=re.findall('\.[a-zA-Z]{3,5}',buf,re.S)

print list(set(addrjs))

得到的结果是：



可以看出所有的图片都是jpg和png格式的（jpeg是jpg的另一种同名），所以用下面的正则表达式

