HTMLParser是python用来解析html的模块。它可以分析出html里面的标签、数据等等，是一种处理html的简便途径。 HTMLParser采用的是一种事件驱动的模式，当HTMLParser找到一个特定的标记时，它会去调用一个用户定义的函数，以此来通知程序处理。它主要的用户回调函数的命名都是以handler_开头的，都是HTMLParser的成员函数。当我们使用时，就从HTMLParser派生出新的类，然后重新定义这几个以handler_开头的函数即可。这几个函数包括：

handle_startendtag 处理开始标签和结束标签

handle_starttag 处理开始标签，比如<xx>

handle_endtag 处理结束标签，比如</xx>

handle_charref 处理特殊字符串，就是以&#开头的，一般是内码表示的字符

handle_entityref 处理一些特殊字符，以&开头的，比如

handle_data 处理数据，就是<xx>data</xx>中间的那些数据

handle_comment 处理注释

handle_decl 处理<!开头的，比如<!DOCTYPE html PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN"

handle_pi 处理形如<?instruction>的东西

这里我以从网页中获取到url为例，介绍一下。要想获取到url，肯定是要分析<a>标签，然后取到它的href属性的值。下面是代码：

```
# -*- encoding: gb2312 -*-
from html.parser import HTMLParser

class MyParser(HTMLParser.HTMLParser):
    def __init__(self):
        HTMLParser.HTMLParser.__init__(self)

    def handle_starttag(self, tag, attrs):
        # 这里重新定义了处理开始标签的函数
        if tag == 'a':
            # 判断标签<a>的属性
            for name,value in attrs:
                if name == 'href':
                    print value
```

```
if __name__ == '__main__':
    a = '<html><head><title>test</title><body><a href="http: //www.163.com">链接到163</a></body></html>'

    my = MyParser()
    # 传入要分析的数据，是html的。
    my.feed(a)
```

说明：3.2 以后的版本需要

from html.parser import HTMLParser 导入 模块