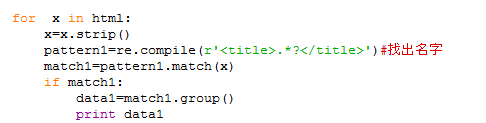
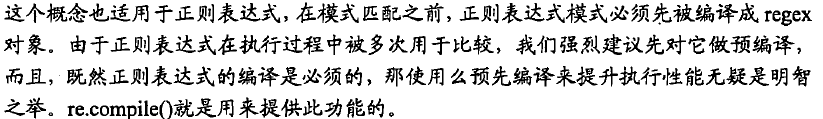
正则表达式：使用模块re

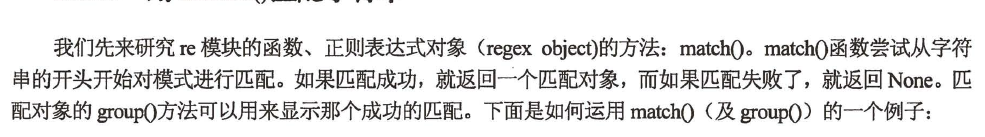


1、使用正则表达式之前应该先对正则表达式进行编译来提高效率：re.compile（），上面的pattern其实就是regex对象。

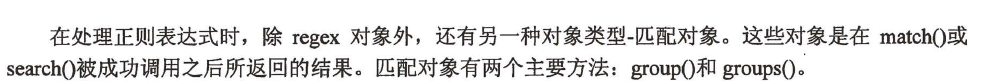


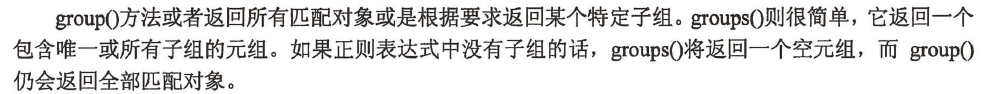


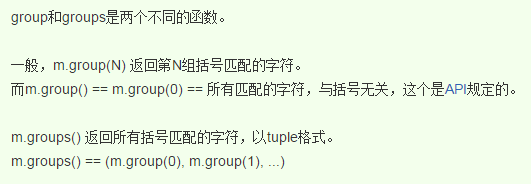
2、这个regex对象可以调用re的其他方法，使用re模块中的match方法在指定的对象中进行正则表达式的匹配。



3、对match或者search返回结果对象使用的方法：group（） or groups（）。



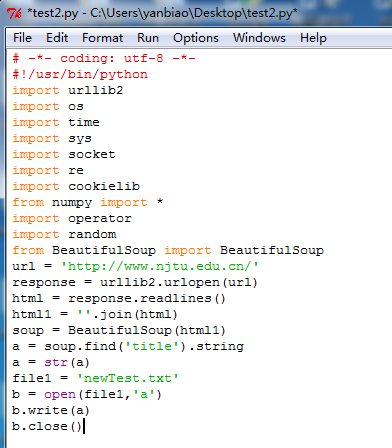




也就是说，如果上面使用的是groups（），那么这个对象的返回值就应该是对应字符串，其中每一个字符作为一个元组的一个元素的元组。

显然，在上面的情况下应该使用的函数应该是group（）

美丽汤BeautifulSoup：使用模块BeautifulSoup



具体使用方法：参见pyspider学习笔记

这个搜索方法跟上面的类似，都需要返回一个对应搜索模式的对象，这里面，这个对象就是上面的soup = BeautifulSoup（html）

之后，在这个对象的前提下，可以调用一系列的方法，比如findall（），find（）

find是返回第一个对象，findall则是返回所有的对象，以列表的形式。

Xpath提取信息：使用的模块：lxml

当想使用xpath提取信息的时候，需要html.urllib2.urlopen（url）.read（）

当想使用正则匹配提取信息的时候，需要把上面的read（）改成readlines（）

doc2=etree.HTML(res.decode('utf8','ignore'))

music=doc2.xpath('//title/text()')

提取其中一个属性的信息的格式，记住后面不需要加上/text（）：

comment=doc3.xpath('//li[@class="comment-item"]/div/a//@title')

提取其中一个子类信息的格式，在最后面需要加上/text（）：

comment2=doc3.xpath('//li[@class="comment-item"]/p[@class="comment-content"]/text()')

但是在这两句提取信息之后都需要去掉天然带着的回车，也就是‘\n’（反斜杠n，电脑屏幕的左下方是正斜杠n），使用strip函数