**如何去掉网页中的广告?**

当成功爬取到的网页数据中有很多不相干的html广告标签时, 你是否会感到无可奈何, 有时候即使将XPath的效果发挥到极致, 也无法去掉顽固的html广告标签, 咋整呢?

本节给你介绍通过selector类的remove方法去除html广告标签, 可提取有用数据或清理无用数据.

举个栗子:

在爬取某论坛问答帖时, 发现有很多html广告标签以及一些无用数据, 就需要在on\_extract\_field回调函数中调用selector的remove方法了

$configs = array(

// configs的其他成员

...

'fields' => array(

array(

'name' => "question\_detail",

'selector' => "XXX",

),

),

);

$spider->on\_extract\_field = function($fieldname, $data, $page)

{

if ($fieldname == 'question\_detail')

{

// 将data中符合XPath: "//div[contains(@class,'a\_pr')]"的数据去掉

$data = selector::remove($data, "//div[contains(@class,'a\_pr')]");

return $data;

}

};

有时, 如果无用数据太多, 最好调用selector的select方法直接将有用的数据提取出来, 这么做会比调用remove方法更加方便.