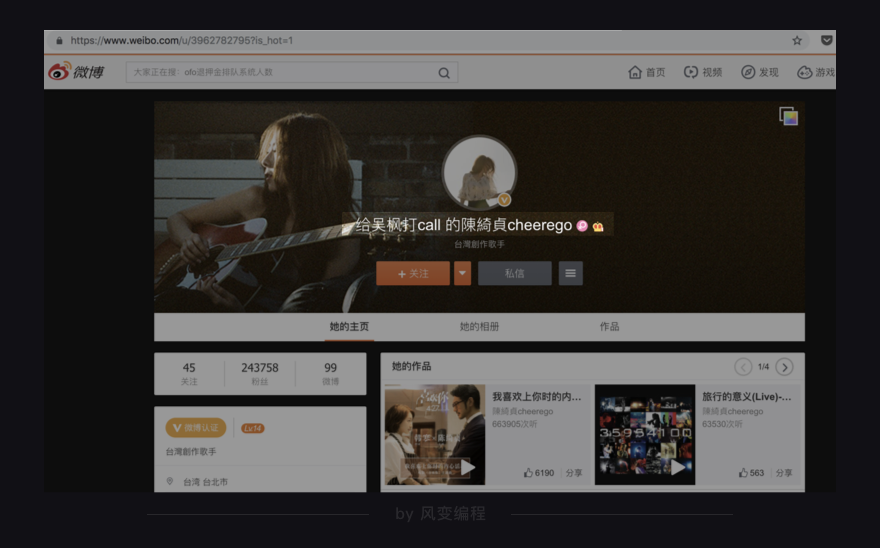
**爬虫第一关 我也可以写个网页**

而在今天这一关，我们要学的是——HTML基础。

如果把HTML的学习依序分为三个层次的话，应该是读懂、修改、编写。

【读懂】：只有读懂了HTML，我们才能看得懂网页结构，才有可能运用Python的其他模块去解析数据和提取数据。所以想写爬虫程序的话，一定要先学好HTML基础。

【修改】：在读懂HTML文档的基础上，学会修改HTML代码，是可以做些有趣的事情的，比如修改我女神的微博。看下图：



【编写】：如果达到了这个水平，那就可以去应聘前端工程师了，这是专业的程序员水平了。

当然，我们这一关的学习目标只要达到前两个——读懂HTML、能够修改HTML文档即可。

好，那马上开始今天的学习吧！

HTML是什么？

HTML（Hyper Text Markup Language）是用来描述网页的一种语言，也叫超文本标记语言 。

打个比方就更好理解了，HTML之于网页就好比建筑图纸之于建筑。



建筑图纸是建筑师设计房子时使用的语言，工匠会根据图纸内容，修建出它所描述的房子。

而HTML文档就是前端工程师设计网页时使用的语言，浏览器会根据HTML文档的描述，解析出它所描述的网页。

上一关，我们讲到向浏览器中输入某个网址后，浏览器会向服务器发出请求，然后服务器就会作出响应。其实，服务器返回给浏览器的这个结果就是HTML代码了。

而紧接着，浏览器会根据这个HTML代码，解析成我们所能看见的漂亮的网页。

查看网页的HTML代码

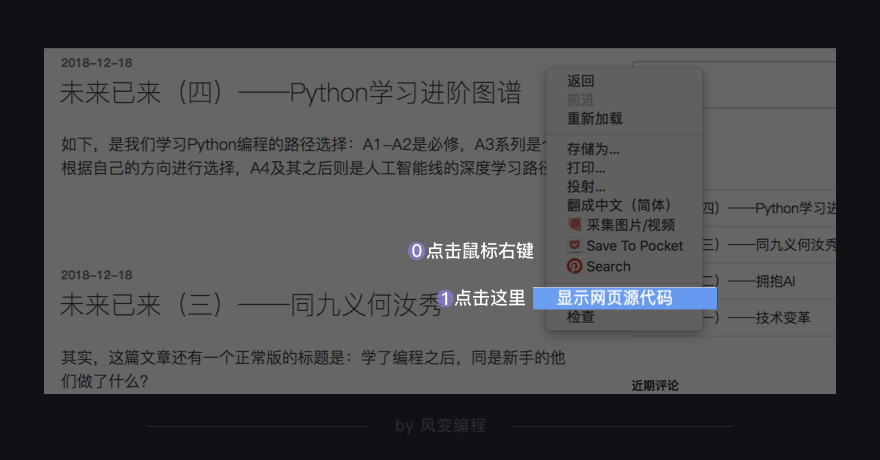
接下来，我们就来看看，每个漂亮的网页背后的HTML代码是怎样的，请一步一步跟着我操作。

【注：下面我们的示范，会用谷歌浏览器（Chrome）进行演示，火狐浏览器（Firefox）的操作方式是一样的。推荐你也使用这二者之一。】

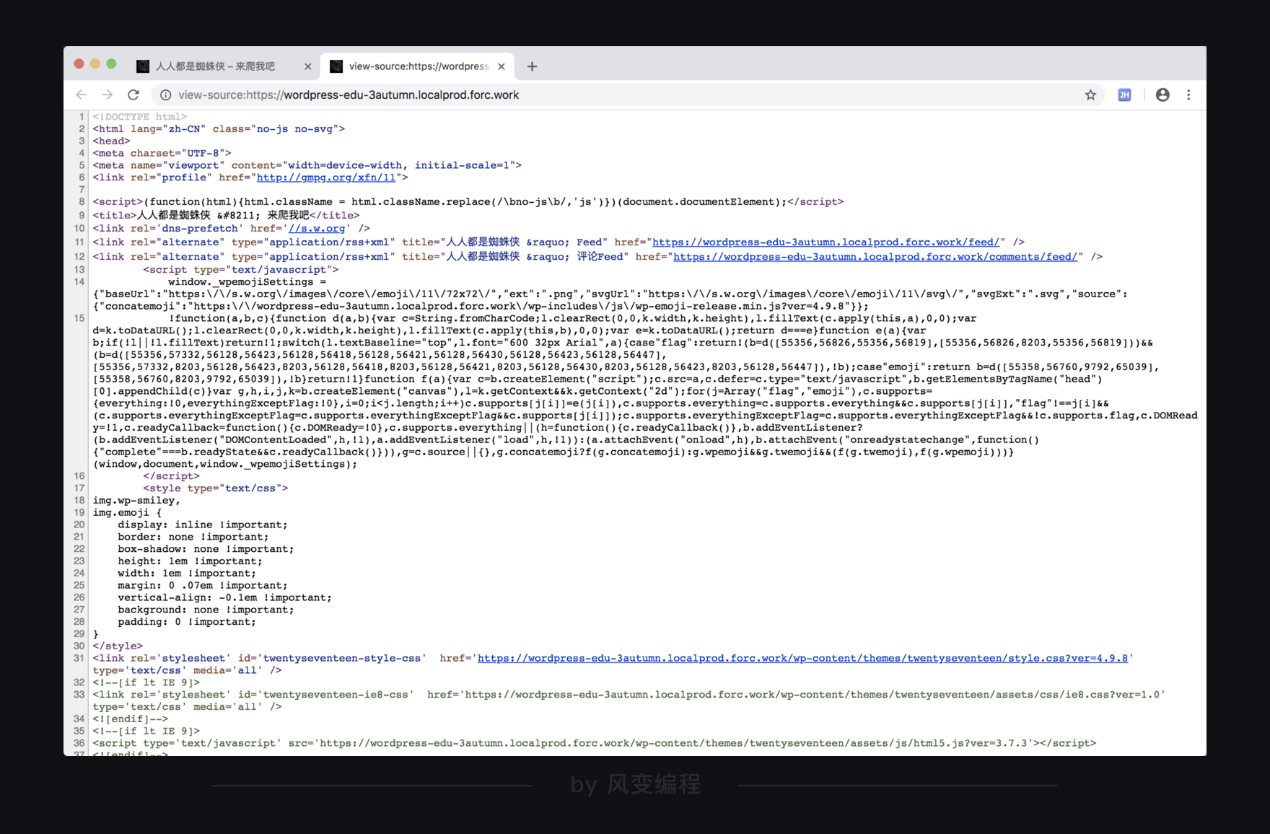
先打开我们的教学网站：人人都是蜘蛛侠。

<https://spidermen.cn/>

在网页任意地方点击鼠标右键，然后点击“显示网页源代码”。（Windows系统的电脑还可以使用快捷键ctrl+u来查看网页源代码）



你会看到，浏览器弹出了一个新的标签页：



没错，这就是HTML源代码了。

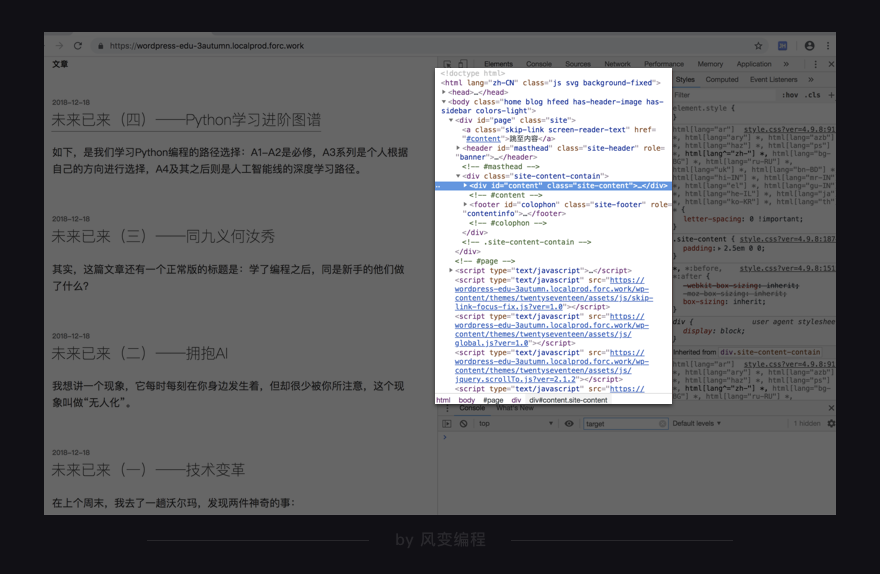
这样查看的好处是，整个网页的源代码都完整地呈现在你面前。坏处是，在大部分情况下，它都会经过压缩，导致结构不够清晰，你不太容易懂每行代码的含义。而且，源代码和网页分开在两个页面展示。

所以更多时候，我们会用这样一种方法：

在网页的空白处点击右键，然后选择“检查”（快捷方式是ctrl+shift+i）。

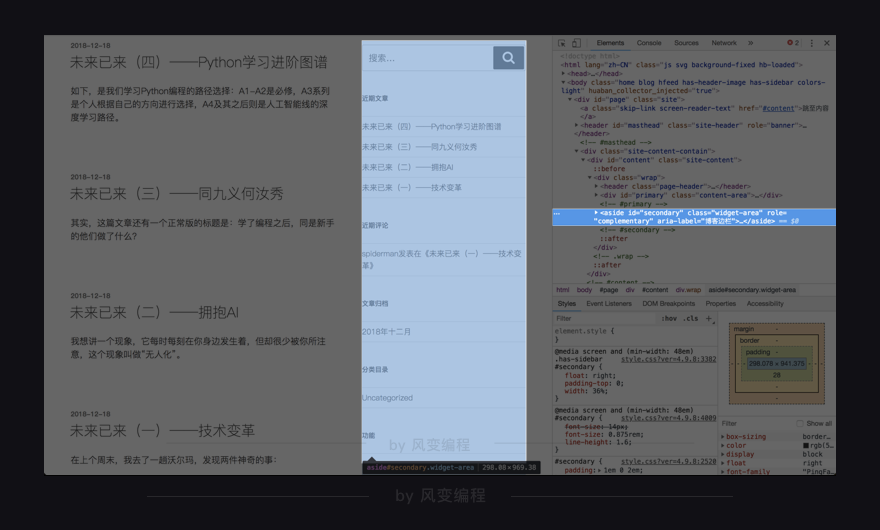


接着，你会看到一个新的界面——开发者工具栏：



上图中标亮的部分就是网页的HTML代码。

将鼠标放在HTML源代码上，你会发现，左边网页上有一些内容会被标亮。这其实就是这行代码所描述的网页内容，它们一左一右，相互对应。

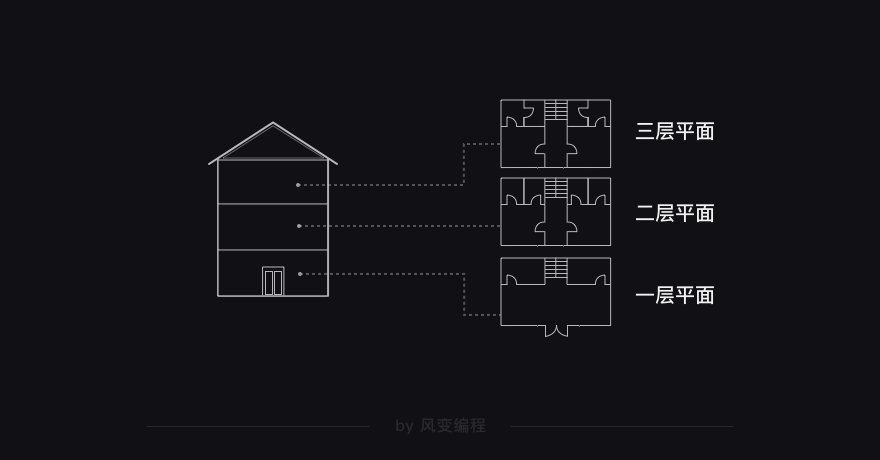


HTML的层级

尽管在HTML里面有一些陌生的符号和文字，但不要担心噢，HTML很好入门。

因为，万事万物都是有条理的。

先以房子为例：房子由不同楼层所组成，每一层中，都会包含一些房间，一个房间还可能划分为几个更小的房间，每个房间又是由门、窗、墙壁、地板等等构建组成的。

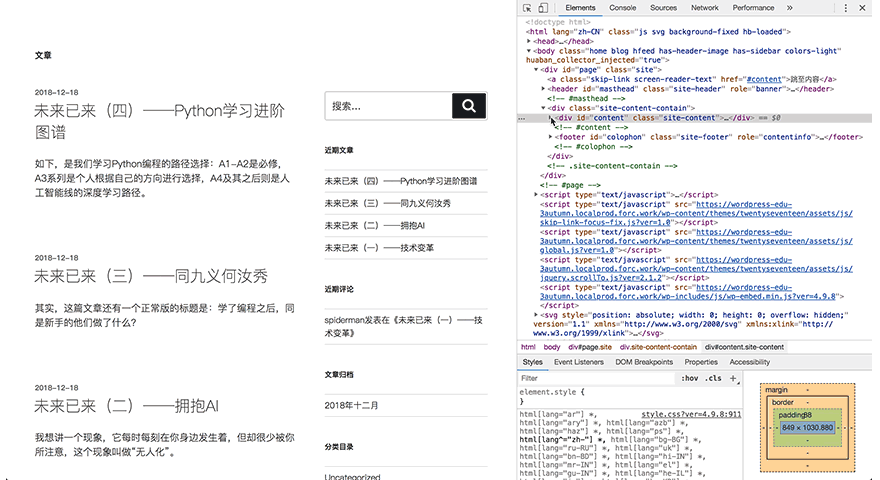


那么相对应的，建筑图纸也是由不同楼层的设计图所组成，而在每层楼的图纸中，又都详细描述了每个房间大小，以及门、窗的位置、尺寸、样式等等。

生活中是如此，在数据的世界中，这种组织的规则会更加明显。

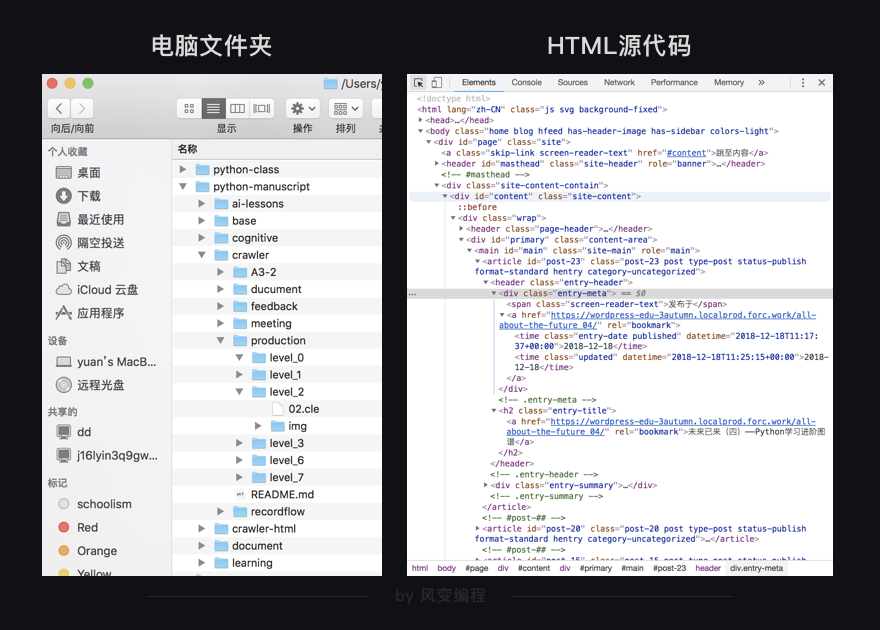
所以HTML源代码之于网页，就像建筑图纸之于房子，会有鲜明的层级结构，以及互相对应的关系。

我们再回到刚才网页，仔细看开发者工具栏：



可以看到，HTML源代码中有一些小三角形，每一个三角形都可以展开或合上。尖角向下代表展开，向右代表合上了，这就是HTML的层级关系。

这每一个可以展开和合上的小三角形里包含的内容，都是一个层级，它很像电脑中一层一层的文件夹。



好，从整体上感受了HTML文档的特点后，我们再具体地看看HTML的组成。

HTML的组成

标签和元素

首先，我们一起看个最简单的HTML文档，感受一下。（提示：HTML代码在当前学习系统中可以直接同步显示渲染结果，点击运行之后会退出代码，进入下一条信息。）

好，回过头来先看HTML文档，可以看到很多夹在尖括号<>中间的字母，它们叫做【标签】。

标签通常是成对出现的：前面的是【开始标签】，比如<body>；后面的是【结束标签】，如</body>。

不过，也有标签是形单影只地出现，比如HTML代码的第四行<meta charset="utf-8">（定义网页编码格式为 utf-8），就是此类。这些你知道就好，大部分情况下用的都是成双成对出现的标签。



事实上，开始标签+结束标签+中间的所有内容，它们在一起就组成了【元素】。



下面的表格列出了几个常见元素：



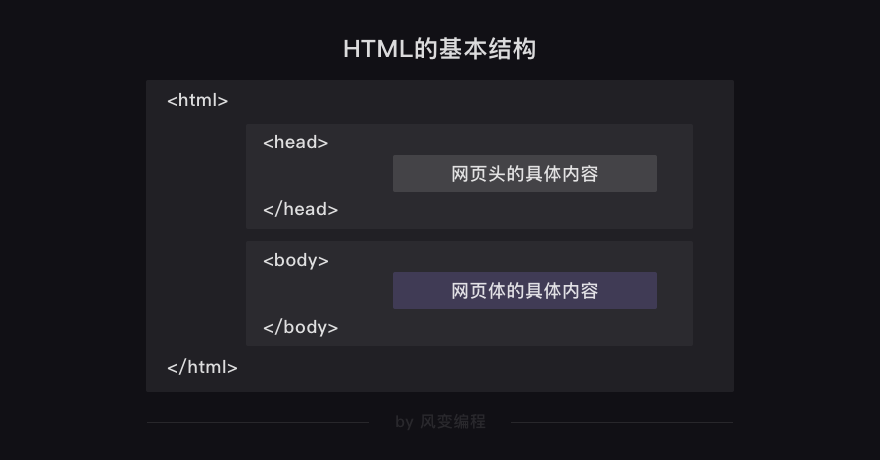
根据表格，请你回看一下上面那段HTML代码，里面就有<h1>，<h2>和<p>。你再对照看代码的显示结果，<h1>是一级标题，<h2>是二级标题，<p>是段落文本，它们一一对应。

你会在后面反复见到这些标签，我也会反复讲，最终，你肯定会对它们熟悉起来。

注意一下：HTML标签是可以嵌套标签的，而且可以多层嵌套；这就像是在电脑中，一个硬盘可以包含数个文件夹，文件夹中还可以嵌套文件夹。

网页头和网页体

事实上，HTML文档的基本是由【网页头】和【网页体】组成的：



HTML文档的最外层标签一定是<html>，里面嵌套着<head>元素与<body>元素。<head>元素代表了【网页头】，<body>元素代表了【网页体】，这是最基本的网页结构。

HTML文档和网页的内容一定是一一对应的。只是，【网页头】的内容不会被直接呈现在浏览器里的网页正文中，而【网页体】的内容是会直接显示在网页正文中的。来具体看看：

先来看<head>元素，也就是【网页头】，它里面一般会有哪些内容呢。

<head>

<meta charset="utf-8">

<title>我是网页的名字</title>

</head>

第2行的<meta charset="utf-8">定义了HTML文档的字符编码。

第3行的<title>元素用来定义网页的标题，这个标题就是显示在浏览器的标签页中的内容，比如我们的网站——人人都是蜘蛛侠：



【网页头】中的编码是没办法在网页中直接被看到的，标签页的内容也不属于网页的正文。

而<body>元素中，即【网页体】，就是那些你能看到的显示在网页中的内容了。

我写了一个简单的小网页，请对照HTML看效果：

<html>

<head>

<meta charset="utf-8">

<title>这个书苑不太冷1.0</title>

</head>

<body>

<h1>这个书苑不太冷</h1>

<h3>吴枫推荐的书:</h3>

<h2>《奇点遗民》</h2>

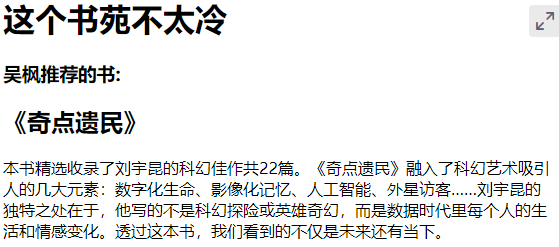
<p>本书精选收录了刘宇昆的科幻佳作共22篇。《奇点遗民》融入了科幻艺术吸引人的几大元素：数字化生命、影像化记忆、人工智能、外星访客……刘宇昆的独特之处在于，他写的不是科幻探险或英雄奇幻，而是数据时代里每个人的生活和情感变化。透过这本书，我们看到的不仅是未来还有当下。

</p>

</body>

</html>

**代码效果**



此时，我们能看到这个HTML文档是由网页头和网页体组成的，网页头不显示在网页中。

而网页体中依次有四个内容：<h1>元素代表一级标题，对应网页中的“这个书苑不太冷”；<h3>元素代表三级标题，对应网页中的“吴枫推荐的书”；<h2>元素代表二级标题，对应网页中的“奇点移民”；然后是<p>元素，对应网页中“本书精选......还有当下”这一整段文本。

恭喜！现在你已经基本掌握了HTML的标签与元素，以及由网页头和网页体组成的基本结构。

属性

现在，我们来玩一个找不同的游戏吧！我在刚才代码的基础上，添加了几行代码，试着找出和刚才1.0版本的不同。

<html>

<head>

<meta charset="utf-8">

<title>这个书苑不太冷2.0</title>

</head>

<body>

<h1 style="color:#20b2aa;">这个书苑不太冷</h1>

<h3>吴枫推荐的书:</h3>

<a href="https://spidermen.cn" target="\_blank">点这里看看</a>

<br>

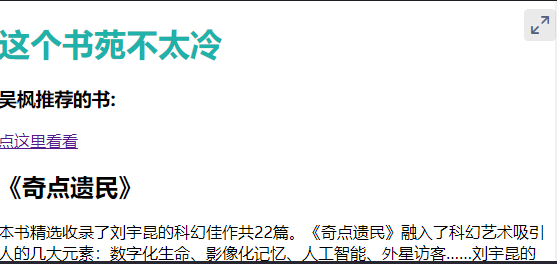
<h2>《奇点遗民》</h2>

<p>本书精选收录了刘宇昆的科幻佳作共22篇。《奇点遗民》融入了科幻艺术吸引人的几大元素：数字化生命、影像化记忆、人工智能、外星访客……刘宇昆的独特之处在于，他写的不是科幻探险或英雄奇幻，而是数据时代里每个人的生活和情感变化。透过这本书，我们看到的不仅是未来还有当下。

</p>

</body>

</html>



2.0和1.0相比，一共有两个不同：<h1>元素中的文字添加了颜色；多增加了一个链接。其实啊，这些都是HTML【属性】的功劳。

（注意：HTML的属性和Python中的属性不是一个东西，不要搞混。）

HTML标签可以通过设置【属性】来为HTML元素描述更多的信息。比如上面用到的例子：

<h1 style="color:#20b2aa;">这个书苑不太冷</h1>

这行代码给<h1>元素添加了一个style属性，属性中的内容规定了这行文字的颜色。

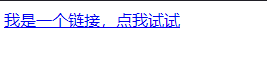


style属性可以用来定义网页文本的样式，比如字体大小、颜色、间距、对齐方式等等。

在上面的代码中，style属性添加在了<h1>的开始标签中，因为属性通常都是在HTML元素的开始标签中设置的。

说完了style属性，看看href属性的用法——添加链接。

<a href="https://wordpress-edu-3autumn.localprod.forc.work/">我是一个链接，点我试试</a>



在HTML中，链接一般都由<a>标签定义，href属性用于规定指向页面的URL，所以刚刚2.0的网页中就多了一个超链接。

除了style和href，HTML中还有两个很常用的属性，即class与id。

先看class的用法。我使用class属性，把刚才的代码2.0又升级为了3.0，下面是两个版本的HTML源代码，再玩一次找不同吧：

这个书苑不太冷3.0：

<html>

<head>

<meta charset="utf-8">

<title>这个书苑不太冷3.0</title>

<style>

.book {

float: left;

margin: 5px;

padding: 15px;

width: 350px;

height: 240px;

border: 3px solid #20b2aa;

}

</style>

</head>

<body>

<h1 style="color:#20b2aa;">这个书苑不太冷</h1>

<h3>吴枫喜欢的书:</h3>

<a href="https://spidermen.cn">点这里看看</a>

<div class="book">

<h2>《奇点遗民》</h2>

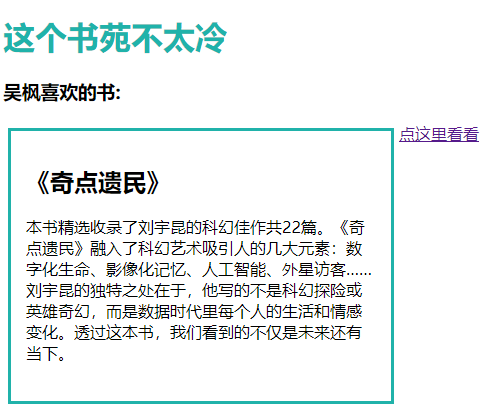
<p>本书精选收录了刘宇昆的科幻佳作共22篇。《奇点遗民》融入了科幻艺术吸引人的几大元素：数字化生命、影像化记忆、人工智能、外星访客……刘宇昆的独特之处在于，他写的不是科幻探险或英雄奇幻，而是数据时代里每个人的生活和情感变化。透过这本书，我们看到的不仅是未来还有当下。

</p>

</div>

</body>

</html>



细心的你会看见，网页头中多了一个很长的<style>元素。

<style>

.book {

/\*以下是.book的具体样式规定\*/

float: left; /\*控制元素浮动\*/

margin: 5px; /\*外边距为5像素\*/

padding: 15px; /\*内边距为15像素\*/

width: 350px; /\*宽度为350像素\*/

height: 240px; /\*高度为240像素\*/

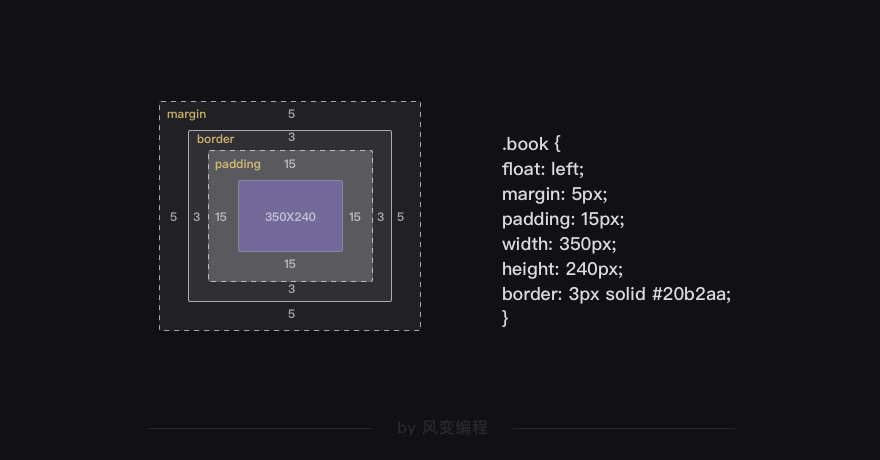
border: 3px solid #20b2aa; /\*边框为3像素\*/

}

</style>

<style>元素的内容，是对.book的具体描述，比如float: left;，而/\*控制元素浮动\*/是对代码的注释。

其实这是一段对网页布局的描述。你看上面代码，在大括号内部写的就是一条条的样式规定。代码的注释中有每个具体样式的意思，你可以和下面的图片对照着看：



这段HTML对照在网页中，就是"吴枫喜欢的书"下面的那个绿框。

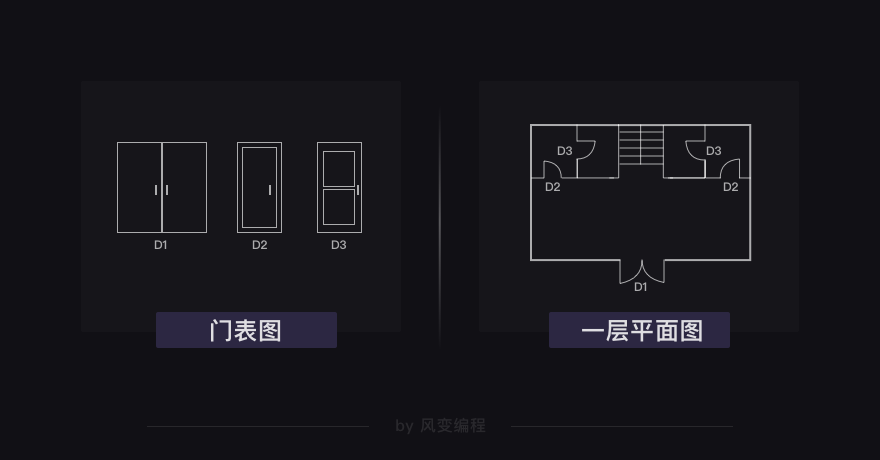
那刚提到的.book又是什么呢？其实，在网页体中，3.0比2.0版本多了这样两个东西：

一个是网页头里的.book；一个是网页体里的<div class="book">。

其实.对应class，所以.book代表class book。因此，网页头中的.book和网页体中的class="book"是有联系的。

在网页头里面，定义了class属性，属性值为"book"，然后下面一长串代码是对这个class属性的描述；接着再在网页体中调用，所以看到了<div class="book">。

这种方法，在建筑设计中同样也会用到。比如，建筑师会做门表图，规定好每一种门的详细尺寸和样式，然后给每个门都起个名字，比如 D1，D2，D3……



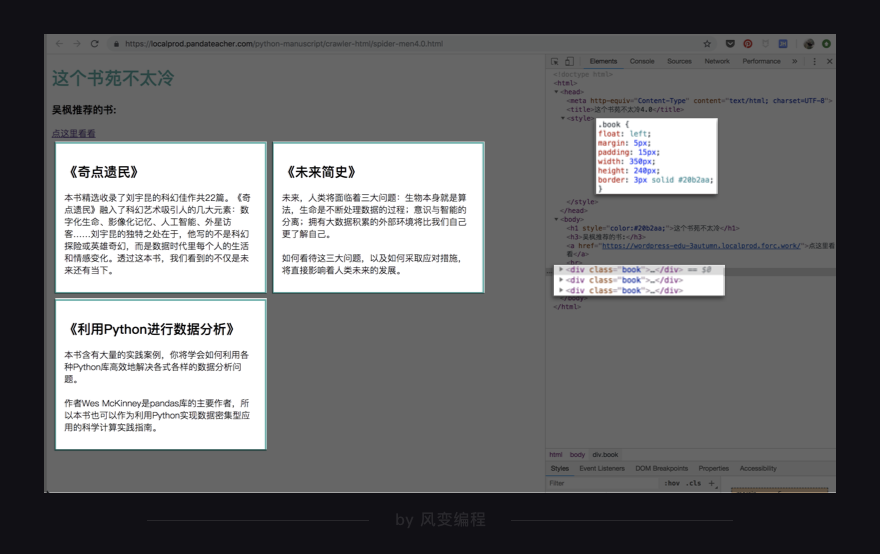
然后，在其它图纸中，只需要用一个简单的符号代表门，在它旁边写上D1，就知道了这个门的具体样式，而不用详细画出每一个门的样子。

在建筑图纸中，一个门可以在平面中使用多次。同样的，在HTML中，class属性也可以被多次利用。

如何用呢？我做了一个网站，也就是【这个书苑不太冷4.0】，URL是：[https://localprod.pandateacher.com/python-manuscript/crawler-html/spider-men4.0.html](https://localprod.pandateacher.com/python-manuscript/crawler-html/spider-men4.0.html" \t "_blank)

打开网页后，用我教你的右键-检查的办法，在开发者工具查看网页的HTML源代码。

你会看到，<div class="book"> 在源代码中出现了三次，与此对应，网页中也有三个一样的块。



网页头的<style>元素中定义了.book的样式，因此，凡是class="book"的元素都会继承它的样式。

聪明如你，应该能猜到，假如想在这个页面中多加一些书籍的介绍，只要在网页体中增加一些class="book" 的div元素就好了。

以上就是class属性的用法，接下来是id属性。

id属性和class属性的用法类似，给元素定义id和class的目的都是为了查找、定位元素，或者为元素设置样式。

但id属性用于标识唯一的元素，而class用于标识一系列的元素。id就像是学生的学生证号码，每个人都是唯一的；而学生们可以属于同一个班级，班级就像class。



好啦，再总结一下最常用的几个HTML属性：



至此，我们就讲完了HTML的组成：标签、元素、结构（网页头和网页体）、属性。

掌握这些知识后，我们就可以通过阅读HTML文档，来对照理解网页了。一起来做个复习：

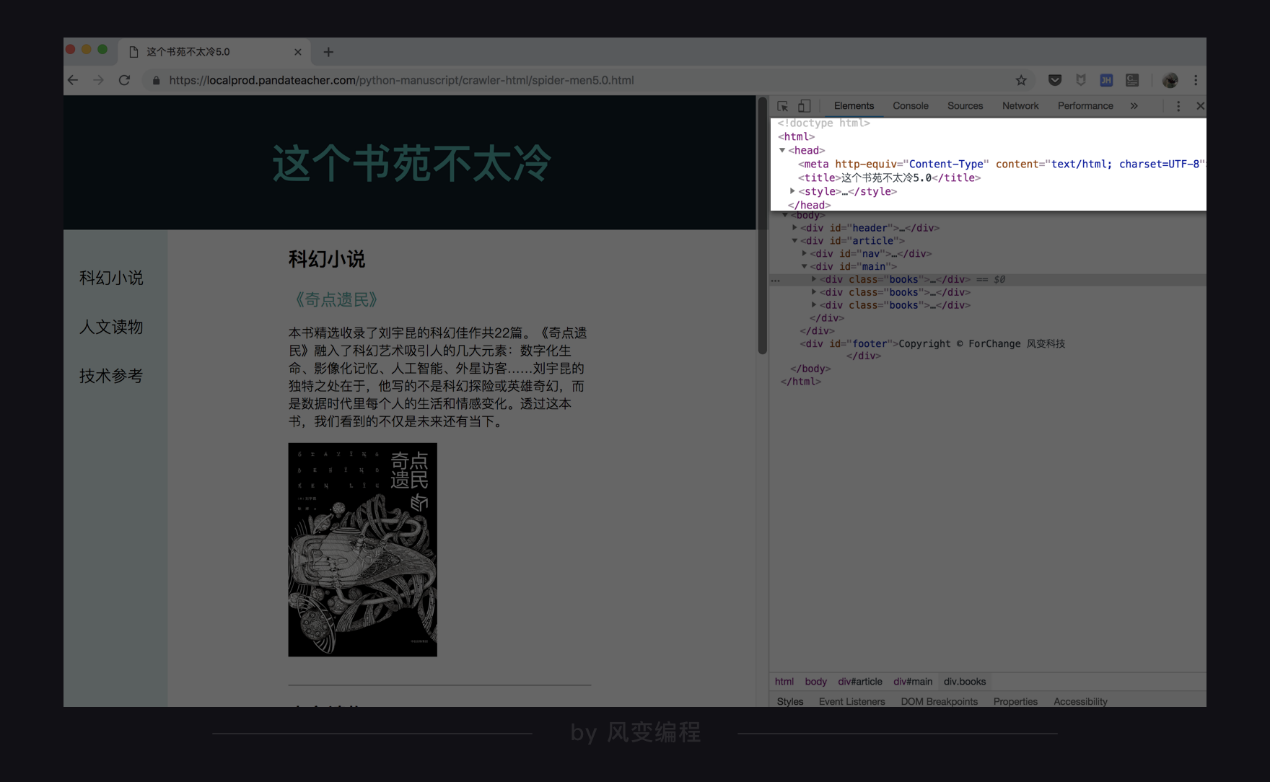
一起来读HTML

我把刚刚的4.0版本又升级成了5.0，刚才讲的所有知识都可以在这个网页看到运用，我们边看HTML源代码，边看它的呈现。

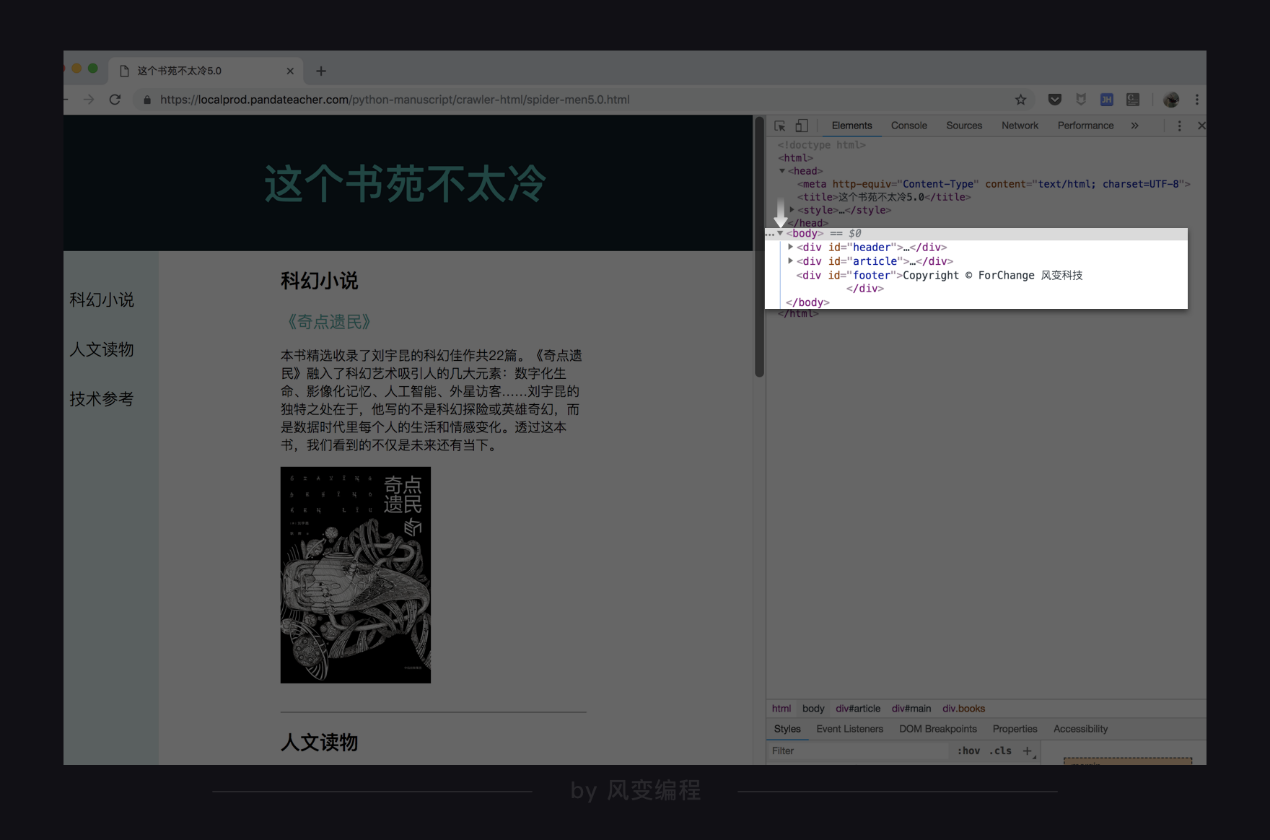
点击URL：[https://localprod.pandateacher.com/python-manuscript/crawler-html/spider-men5.0.html](https://localprod.pandateacher.com/python-manuscript/crawler-html/spider-men5.0.html" \t "_blank)

然后在网页中点击右键——检查，我们从上往下，一起来读读这个HTML文档：

首先，网页头中定义了编码、选项卡之类的内容，里面的<style>元素也定义了一些样式。



然后看网页体，点击三角符号折叠，让我们更清晰地看到内部的结构是这样：

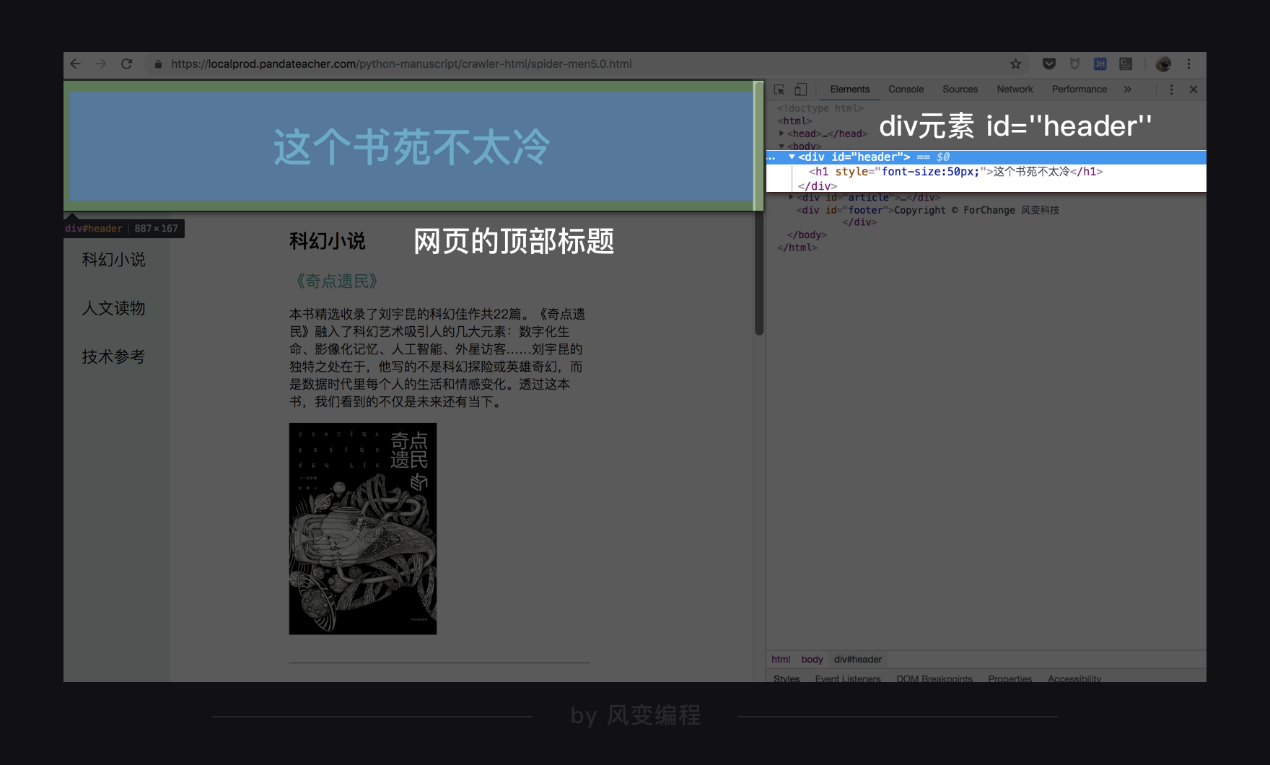


网页体有三大部分，<div id="header">元素的部分，<div id="article">元素的部分，和<div id="footer">元素的部分。HTML中的三部分，对应的也是网页布局中的三部分：



三部分分别是：顶部的标题、中间的内容、以及底部，三个区域。

首个<div>元素是网页的顶部标题：id="header"：

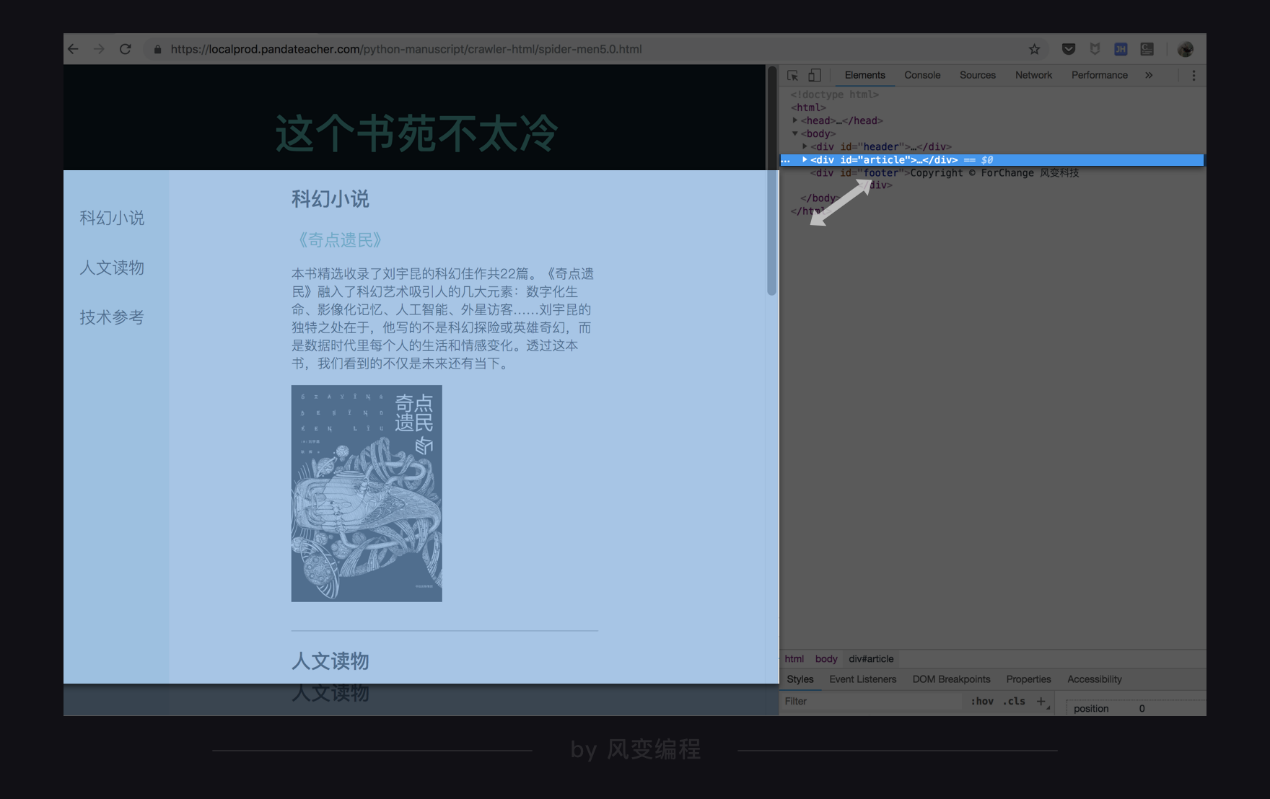


<div id="header">

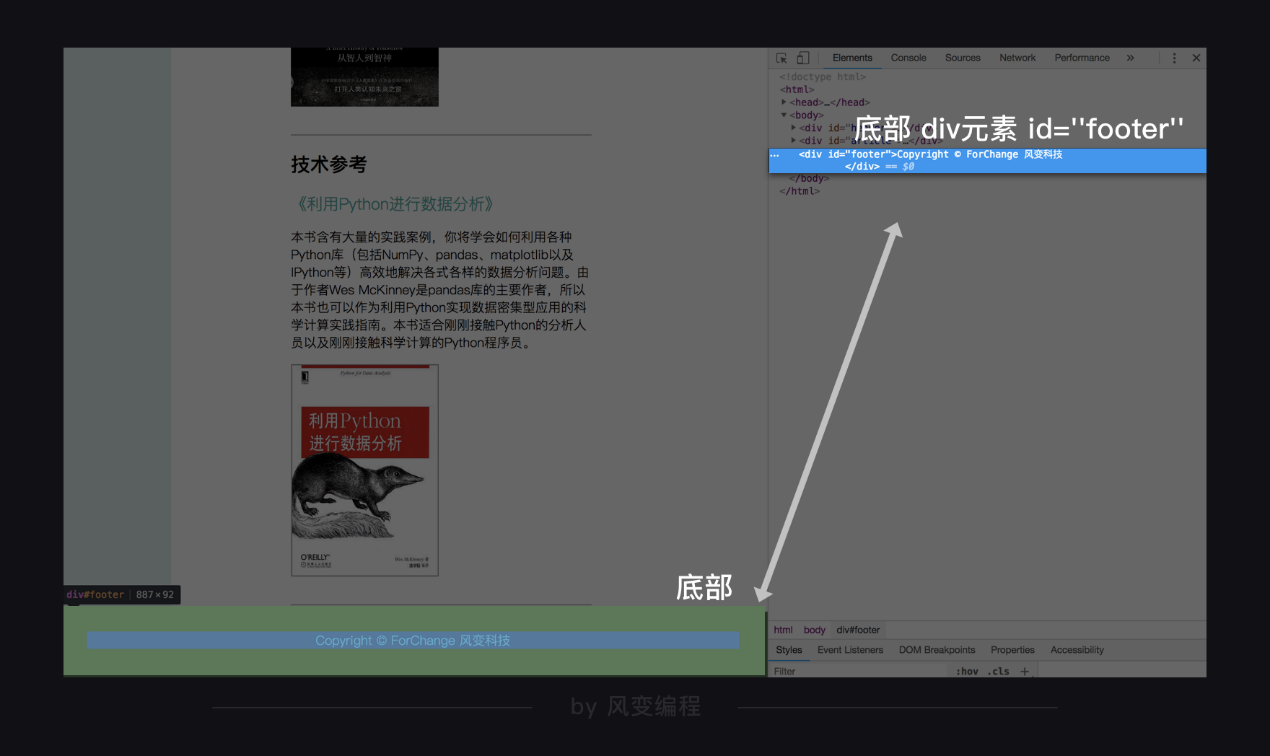
<h1 style="font-size:50px;">这个书苑不太冷</h1>

</div>

然后是<div id="article">元素，它对应的是中间的正文部分。



底部对应着的<div id="footer">，这部分的内容也比较简单：



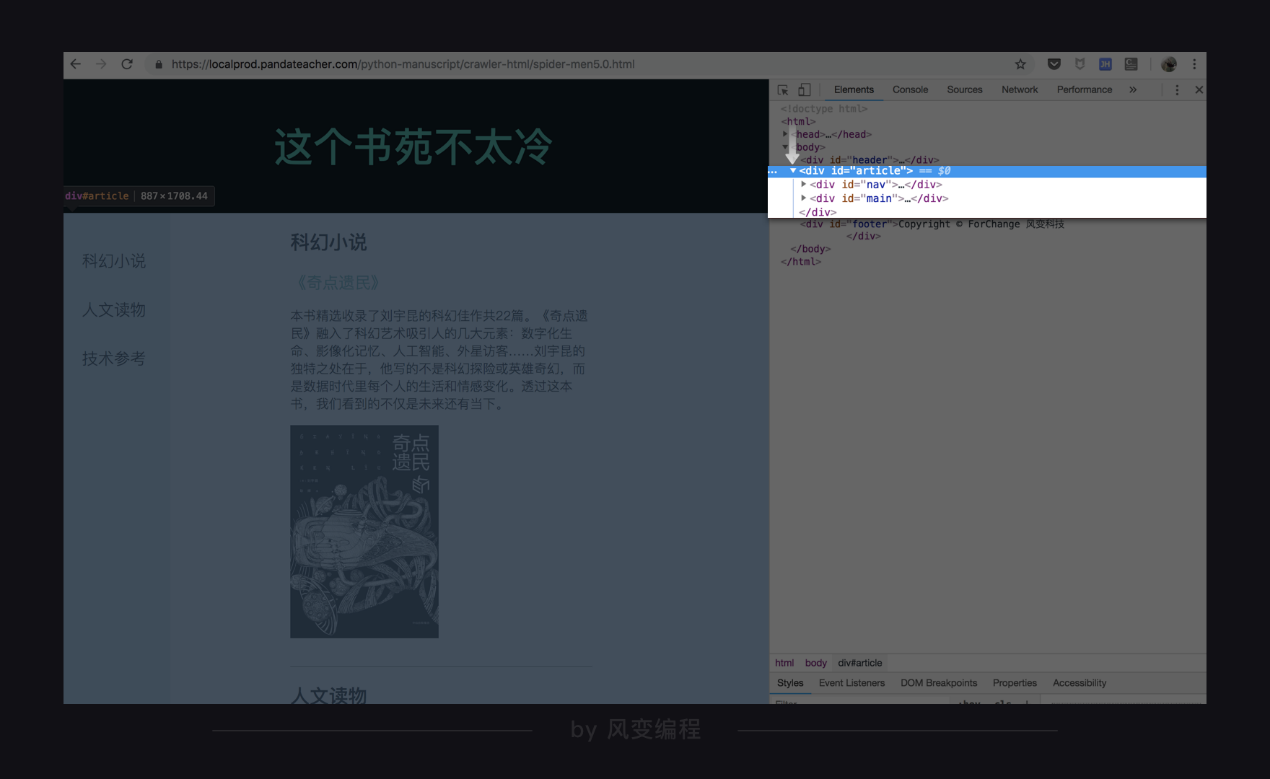
<div id="footer">

Copyright © ForChange 风变科技

</div>

接着我们细看中间正文部分。

继续点开HTML的<div id="article">元素，折叠三角符号，以便清晰地看到它的结构：



<div id="article">元素分为两部分，它包含着两个<div>元素，分别对应着网页中间的左边栏和正文部分。HTML源代码和网页的结构仍然是一一对应的：



第一个<div>元素， id="nav"，它对应着侧边栏：



点进去看里面的代码：

<div id="nav">

<a href="#type1" class="catlog">科幻小说</a><br>

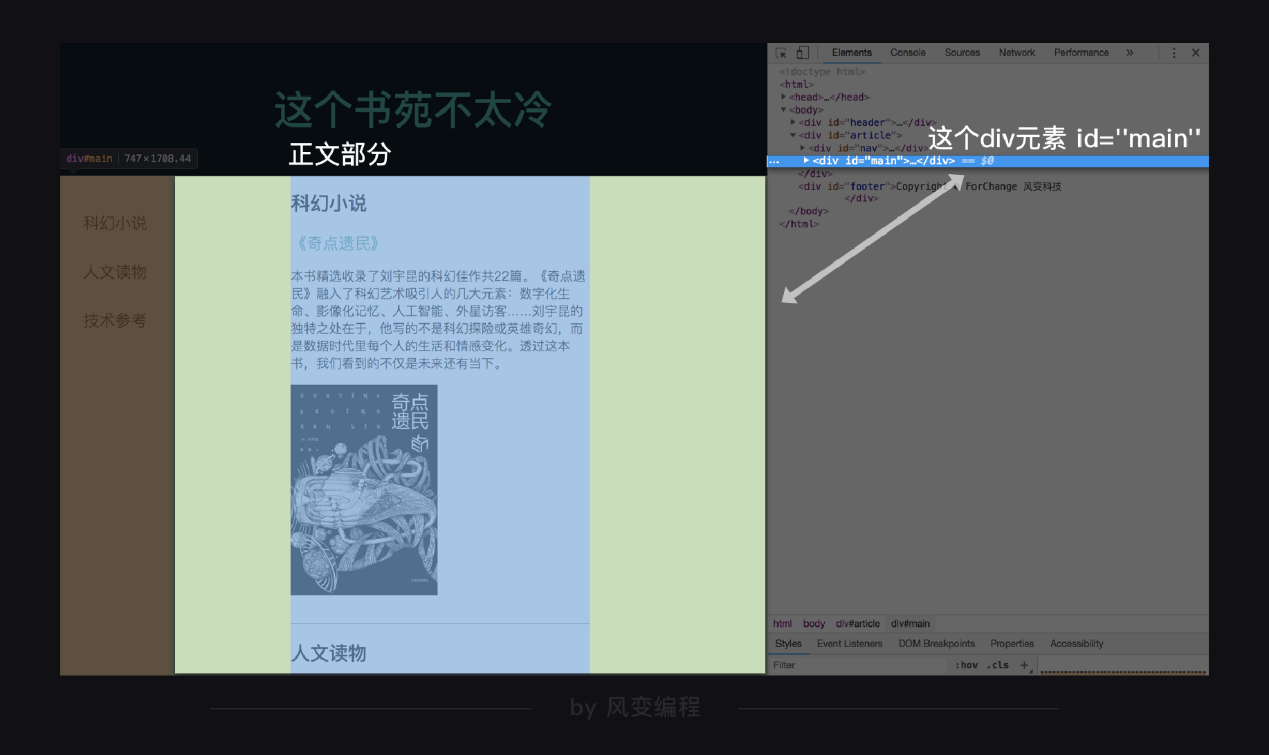
<a href="#type2" class="catlog">人文读物</a><br>

<a href="#type3" class="catlog">技术参考</a><br>

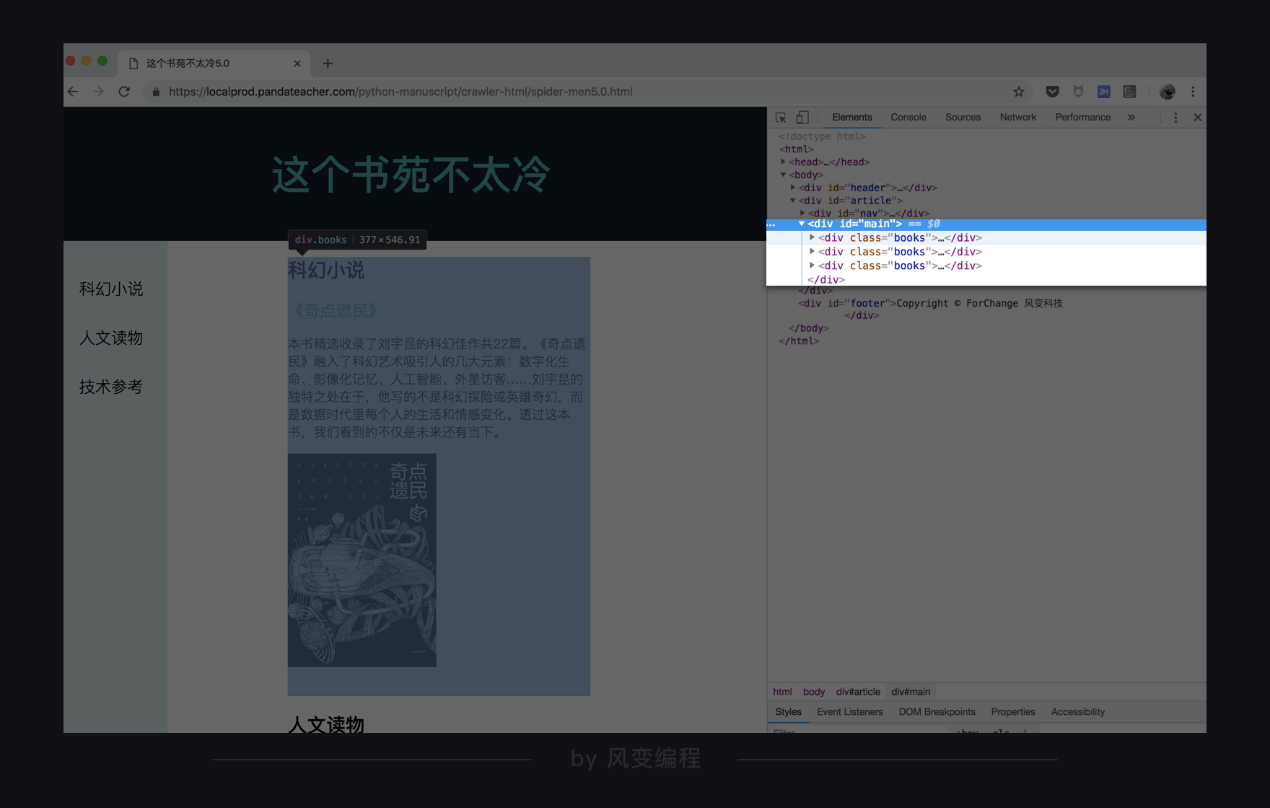
</div>

这个<div>元素中，包含了三个<a>元素，内容是你所看到的：科幻小说、人文读物、技术参考。它们三个都由href属性和定义了内部链接，试着点击看看，你会看到，它跳转到了正文中对应的位置。

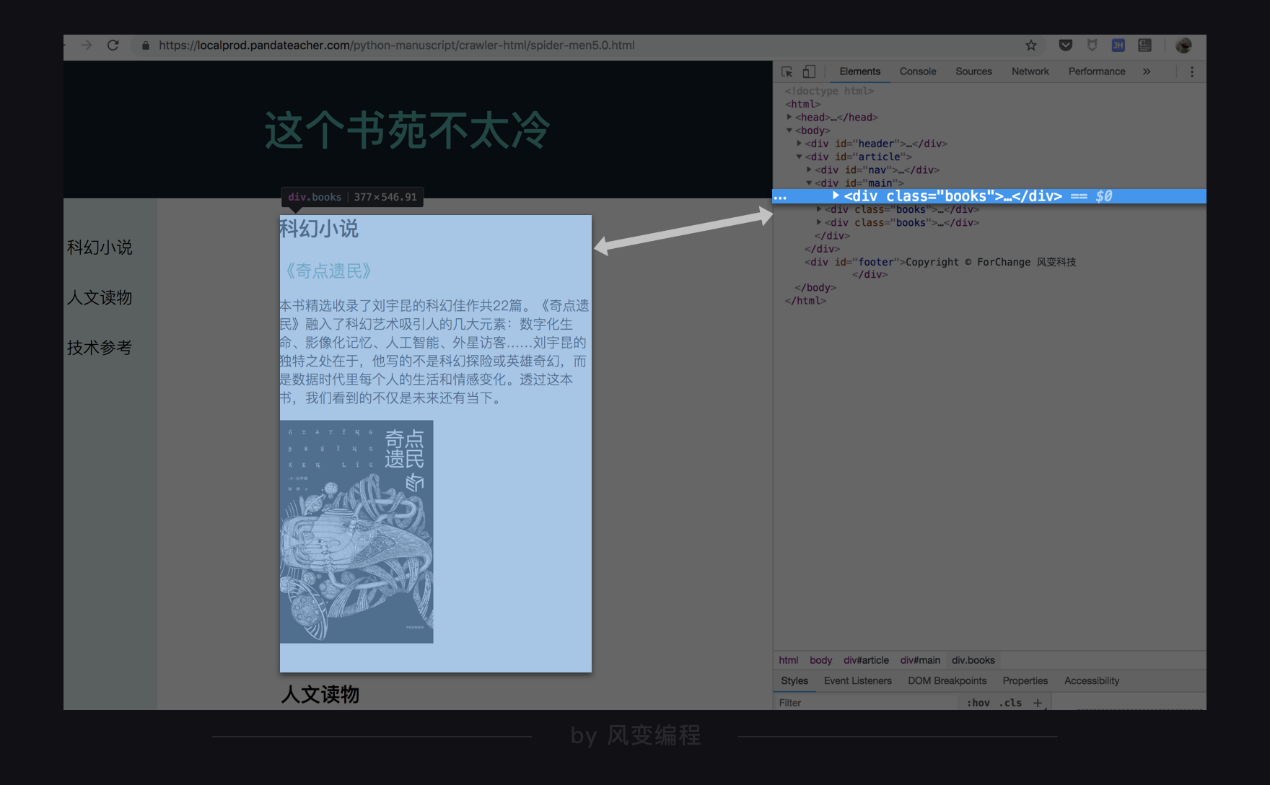
第二个<div>元素对应着右边正文部分， id="main"：

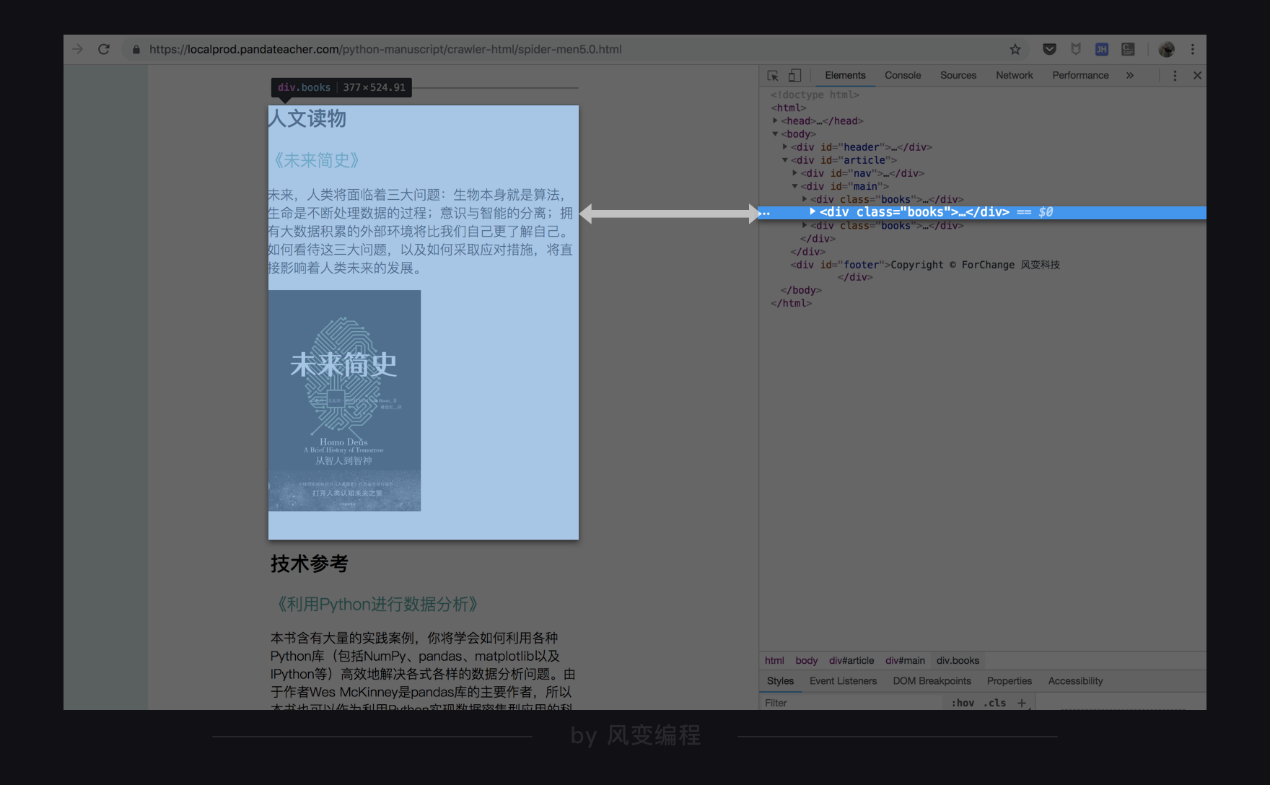


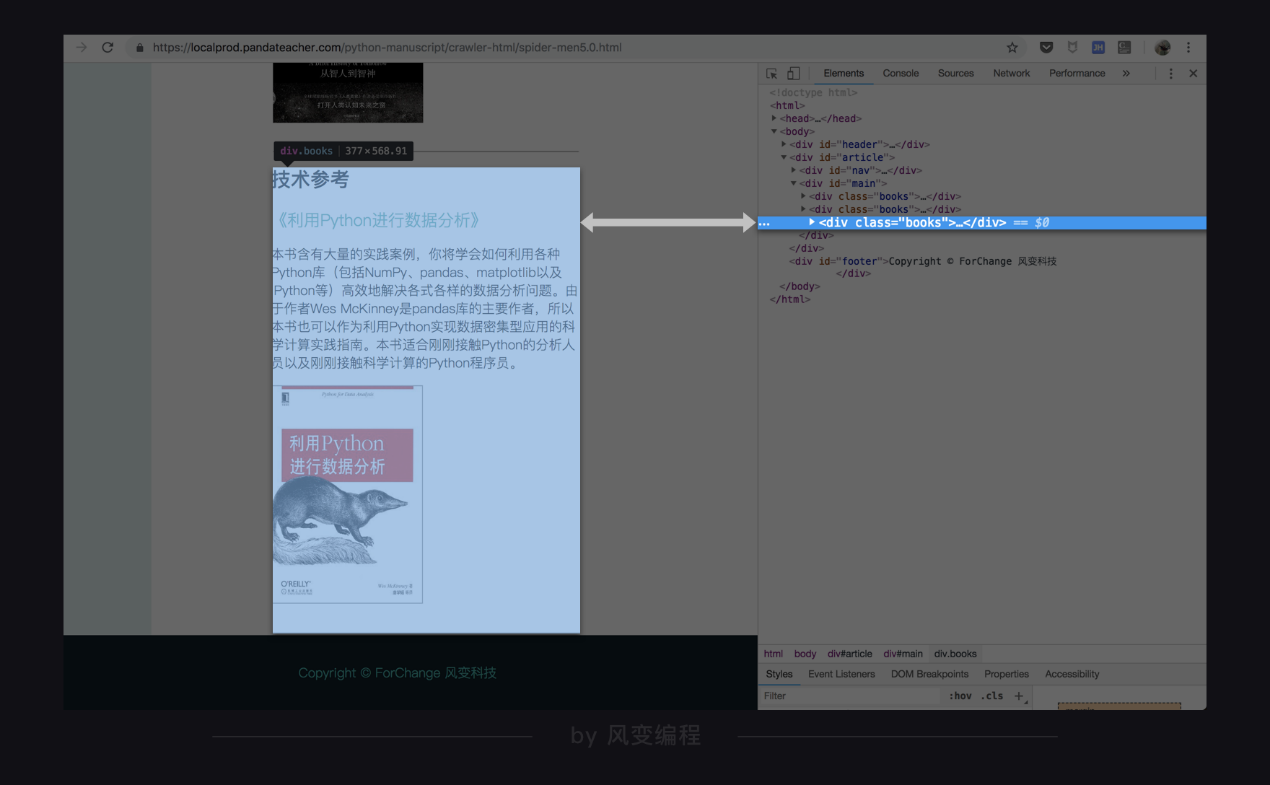
点进去看里面的结构：



在<div id="main">中，又包含了三个<div>元素，它们都有同样的属性：<class="books">。每个<div>元素分别介绍了一本书的内容，这个部分和网页结构是基本一样的：







现在，我们完全展开<section>元素来看看更详细的代码：

<div id="article">

<div id="nav">

<a href="#type1" class="catlog">科幻小说</a><br>

<a href="#type2" class="catlog">人文读物</a><br>

<a href="#type3" class="catlog">技术参考</a><br>

</div>

<div id="main">

<div class="books">

<h2><a name="type1">科幻小说</a></h2>

<a href="https://book.douban.com/subject/27077140/" class="title">《奇点遗民》</a>

<p class="info">本书精选收录了刘宇昆的科幻佳作共22篇。《奇点遗民》融入了科幻艺术吸引人的几大元素：数字化生命、影像化记忆、人工智能、外星访客……刘宇昆的独特之处在于，他写的不是科幻探险或英雄奇幻，而是数据时代里每个人的生活和情感变化。透过这本书，我们看到的不仅是未来还有当下。</p>

<img class="img" src="./spider-men5.0\_files/s29492583.jpg">

<br/>

<br/>

<hr size="1">

</div>

<div class="books">

<h2><a name="type2">人文读物</a></h2>

<a href="https://book.douban.com/subject/26943161/" class="title">《未来简史》</a>

<p class="info">未来，人类将面临着三大问题：生物本身就是算法，生命是不断处理数据的过程；意识与智能的分离；拥有大数据积累的外部环境将比我们自己更了解自己。如何看待这三大问题，以及如何采取应对措施，将直接影响着人类未来的发展。</p>

<img class="img" src="./spider-men5.0\_files/s29287103.jpg">

<br/>

<br/>

<hr size="1">

</div>

<div class="books">

<h2><a name="type3">技术参考</a></h2>

<a href="https://book.douban.com/subject/25779298/" class="title">《利用Python进行数据分析》</a>

<p class="info">本书含有大量的实践案例，你将学会如何利用各种Python库（包括NumPy、pandas、matplotlib以及IPython等）高效地解决各式各样的数据分析问题。由于作者Wes McKinney是pandas库的主要作者，所以本书也可以作为利用Python实现数据密集型应用的科学计算实践指南。本书适合刚刚接触Python的分析人员以及刚刚接触科学计算的Python程序员。</p>

<img class="img" src="./spider-men5.0\_files/s27275372.jpg">

<br/>

<br/>

<hr size="1">

</div>

</div>

</div>

在这里，5.0对比4.0版本变化的部分是：（第9、19、29行）给h2元素标注了name属性，< section id='nav'>中的超链接标签(第3、4、5行)以这个name属性为标识，设置了跳转到这个标题的锚点；（第10、20、30行）给书名添加了超链接，可以链接到这本书的豆瓣主页；以及，（第12、22、32行）用<img>标签添加了书的封面图片。

这样，我们就读完网页的HTML源代码了。

从【这个书苑不太冷】1.0到5.0，你能看到，随着标签和属性逐步增加，网页也在逐步变得更加复杂。你在这个过程中，循序渐进地学会了HTML基础。

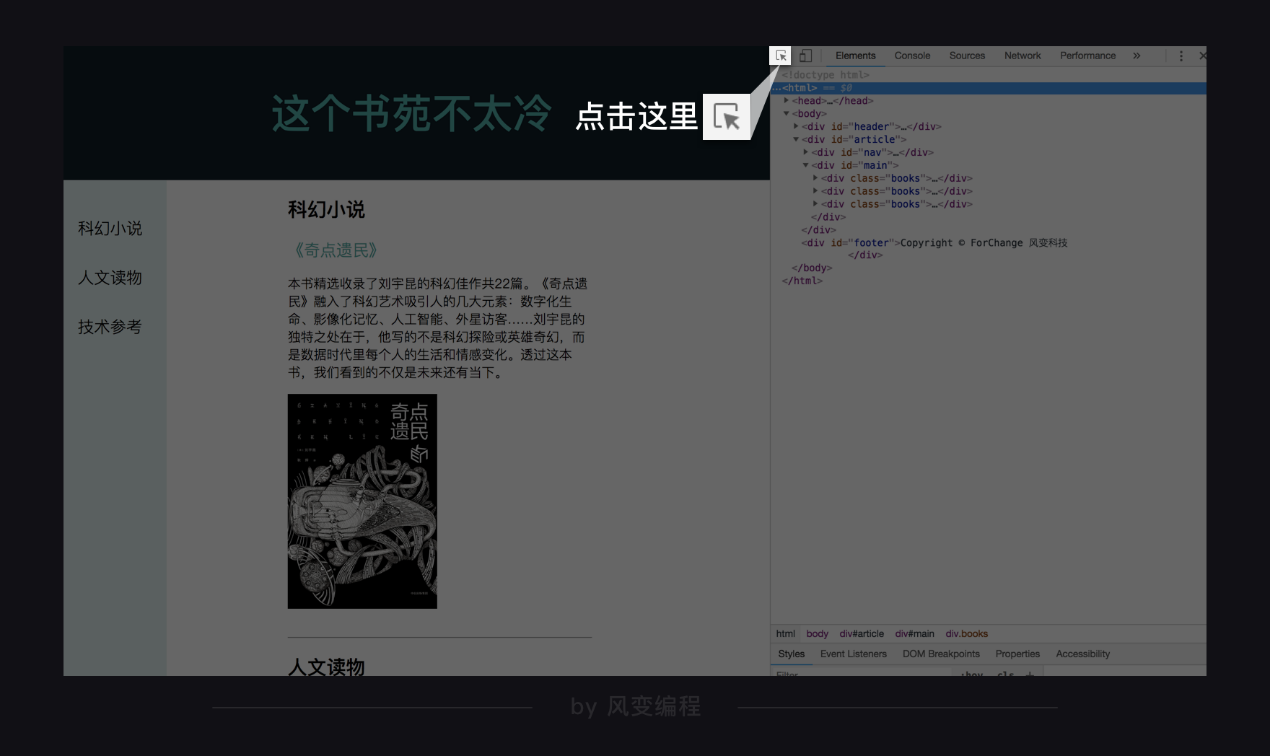
说了这么多，接下来我们就来做点有趣的事吧。

实操一下

修改网页

首先，我来教你一个网页偷天换日法：

打开【这个书苑不太冷5.0】，你能看到开发者工具的左上角，有一个图标长这样吗？



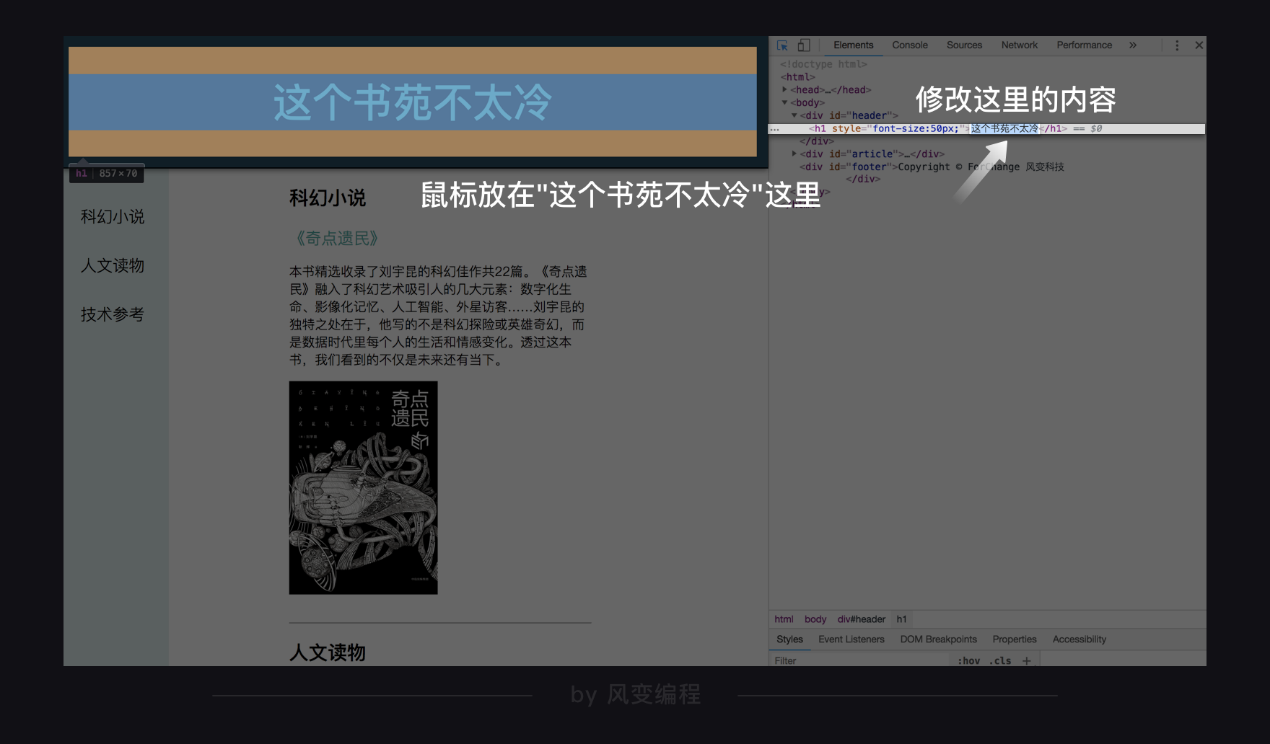
点击它，然后再把鼠标放在网页中，你会发现很好玩的事情：和点击源代码的情景恰恰相反，当鼠标放在网页上，右边代码区中描述它的代码会被标亮出来：



这是一个你以后会经常用到的功能，用来定位你所要查找的网页内容的源代码。

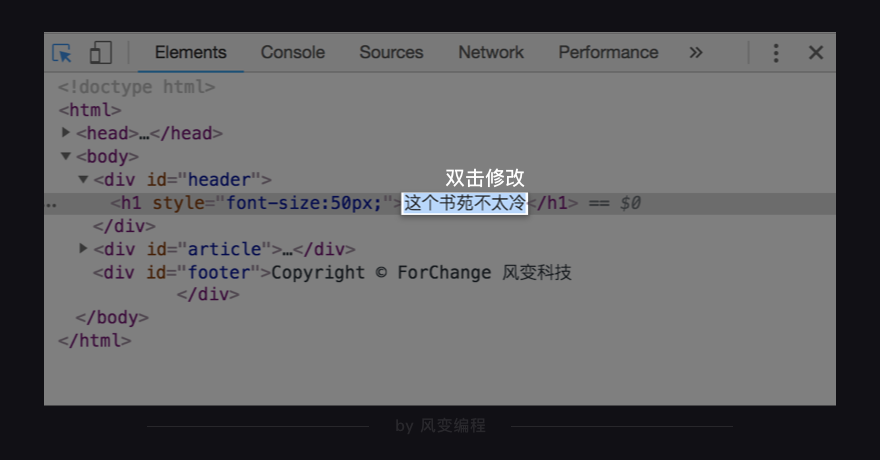
而我们已经知道，可以通过修改HTML源代码，从而修改网页。其实，我们也可以在网页的开发者工具这里修改HTML文件。

比如，我们想修改网页的标题，那就把鼠标放在网页"这个书苑不太冷"这里。



你看到网页源代码中，标题所在的这个元素被标亮了，那接下来要做的就是，修改这里的内容。

把鼠标点上去，双击，就和修改word文档一样。



改完之后，按下enter确认。你会发现左边的文字就变成了你刚才输入的样子。

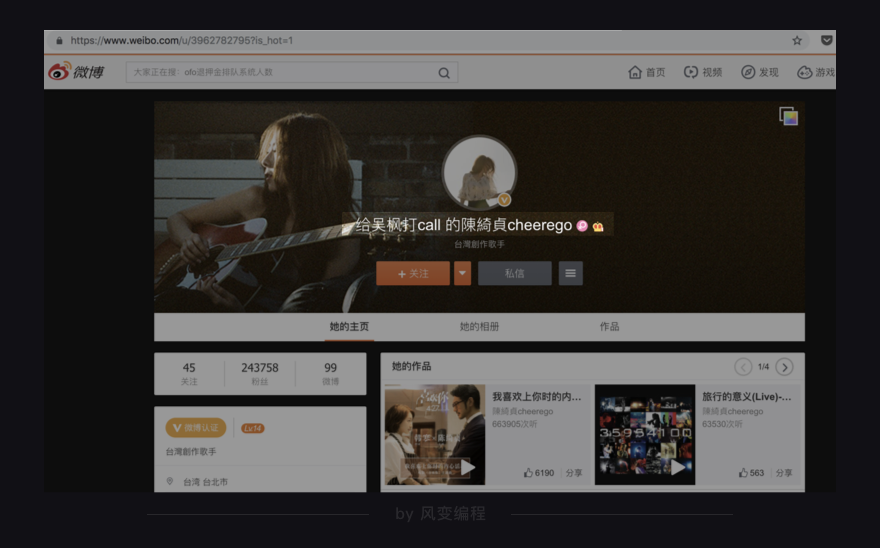


然后点击开发者工具右上角的叉号，关掉它，大功告成！



用同样的方法，可以修改网页的其它部分。

学会了这招，你还可以用来改偶像微博，比如修改女神的微博，发朋友圈炫耀。



当然，这样的修改只是在你本地的修改，而服务器上的源文件你是修改不了的，所以，这些改动仅供你自娱自乐。而程序员们会使用这个方法，在开发者工具这里，调试HTML代码。

到这里，我认为你已经足够了解HTML，可以继续学习爬虫啦。

获取数据

上一关，你已经学习了使用requests.get()获取数据，体验了用它来获取电子书和图片。



最后再教你一招，做一个本地的网页：

把刚才请求到的HTML源文件复制粘贴，在vscode中保存为后缀为.html 的文档，它就是一个保存在你本地中的网页了。

.html这种文件格式，可以用浏览器打开。这就好比MP3文件用音乐播放器打开，txt用记事本打开。

你现在应该更明白浏览器的工作原理了，简单来说：浏览器从服务器上接收一个HTML文档，然后拿去做解析，最后呈现给你。因此，它也可以把你电脑的HTML文档解析成漂亮的网页。

HTML其实很好入门，学会它，从此，你就可以开始睁开眼睛上网。