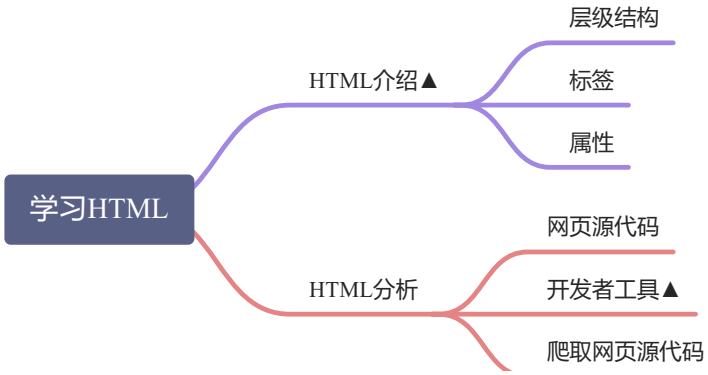
第1课 学习HTML

一、课程结构导图



注：▲为重点知识点。

二、知识点讲解

2.1 HTML 介绍

概念： HTML是用来描述网页的一种语言，英文全称是Hyper Text Markup Language，也称为超文本标记语言。HTML 文档主要由HTML 元素组成。HTML 元素由标签以及标签中的内容构成。

示例： （<!-- 注释内容-->为HTML 注释）

```
<!DOCTYPE html>                          <!-- 声明为HTML5文档-->
<html>                                   <!-- HTML文档（起点）-->
    <head>                               <!-- 文档头（起点）-->
        <meta charset="UTF-8">           <!-- 设置编码-->
        <title>风变一下，就能学到</title><!-- 设置网页标题-->
    </head>                              <!-- 文档头（终点）-->
    <body>                               <!-- 文档体（起点）-->
        <p>风变一下</p>                  <!-- <p>标签定义文字段落-->
    </body>                              <!-- 文档体（终点）-->
</html>                                  <!-- HTML文档（终点）-->
```

2.1.1 HTML 层级结构

```
# 根据包含关系进行判断。
<!DOCTYPE html>
<html>
    <head>
        <meta charset="UTF-8">
        <title>风变一下，就能学到</title>
    </head>
    <body>
        <p>风变一下</p>
    </body>
</html>
```

代码解释：最外层是一个<html></html>标签，把所有的代码包含其中，在<head></head>标签中，又包含了两个标签<meta>和<title>。在<body></body>标签中，包含了<p>标签。

2.1.2 标签

概念： 标签用于标记文本信息，指用尖括号（<>和</>）括起来的字母和英文，一般标签形式有两大类：闭合标签和自闭合签。闭合标签，常见于HTML 代码中，一般成对出现（有开始标签<>,也有结束标签</>），如：<title>...</title>是标题标签， ... 是块标签。

自闭合标签，只有一个尖括号</>(斜杠/可省略)，标签的内容都在尖括号中展示，比如：<img />是图片标签，<input />是input标签。

示例：

```
<!DOCTYPE html>
<html>
<head>
    <title>风变一下，就能学到</title>                    <!-- 闭合标签-->
</head>                <!-- 闭合标签-->
<body>
    <div style="text-align:center;margin-top:80px;">
        <img width="150px" src="https://res.pandateacher.com/横版:中文组合:风变蓝.png"><!-- 半闭合标签-->
        <form id="search_form">
```

```
            <input type="text" class="input_box">     <!-- 半闭合标签-->
            <button class="btn_submit" disabled="true">风变一下</button>  <!-- 闭合标签-->
        </form>       <!-- 闭合标签-->
    </div>            <!-- 闭合标签-->
</body>               <!-- 闭合标签-->
</html>               <!-- 闭合标签-->
```

常见的HTML标签：

2.1.3 属性

概念：在HTML 文档中，属性存放在尖括号<>中，都是以赋值形式存在，即：属性=属性值。常见的属性有：class、id、href、style。

示例：

```
<!DOCTYPE html>
<head>
    <meta charset="UTF-8">  <!-- charset是一个属性 -->
</head>
<style>
    /*定义了class属性为style_1的样式*/
    .style_1 {  width: 100px;   }
    /*定义了class属性为style_2的样式*/
    .style_2 {  width: 50px;    }
</style>
<body style="background:#a8c7e2"> <!-- style是一个属性 -->
    <img class='style_1' src="https://res.pandateacher.com/风变科技logo1.png" ><!-- class是
    <img class='style_2' src="https://res.pandateacher.com/forchange_blue_logo.png" >
    <img class='style_2' src="https://res.pandateacher.com/forchange_black_logo.png" >
    <br>
    <img class='style_1' src="https://res.pandateacher.com/风变科技logo2.png" >
    <img class='style_2' src="https://res.pandateacher.com/panda_black_logo.png" >
    <img class='style_2' src="https://res.pandateacher.com/panda_white_logo.png" >
</body>
</html>
```

注意：为了方便反复调用，部分内容的尺寸、颜色等设置会放到<head>标签中的<style>标签里。在<style>标签中，class属性的样式用点'.'，id属性用井号'#'。

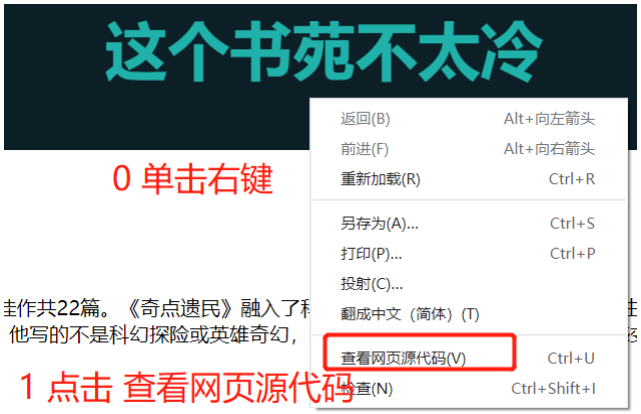
2.2 HTML分析

我们可以通过两种方式查看网页源代码（注：所有演示都是在谷歌浏览器下进行。），一种是直接打开网页源代码查看，另一种是打开开发者工具进行查看。

注意两者的一个区别，当网页是静态网页的时候，二者基本相同；当网页是动态网页时，开发者工具中的代码是浏览器渲染之后显示的内容，包括网页源代码，还有藏在Network 面板中的一些数据（第4~5课会介绍动态网页）。

2.2.1 网页源代码

1. 随机打开一个网站，在网页任意地方点击鼠标右键，然后点击“查看网页源代码”。

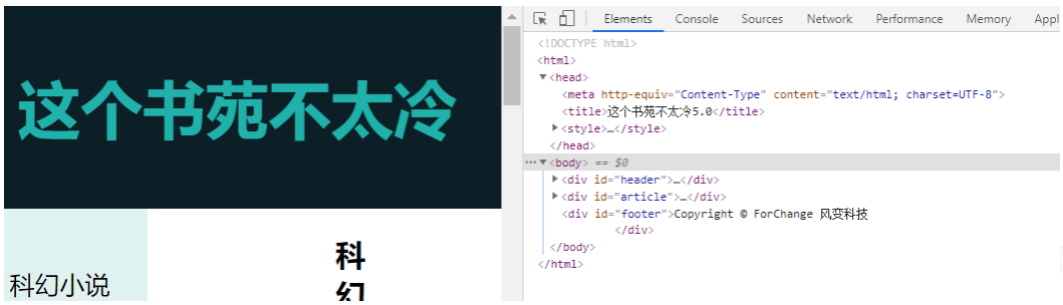


2. 点击“查看网页源代码”之后，会弹出一个新窗口（如下图），该代码便是网页源代码。

2.2.2 开发者工具

1. 随机打开一个网页，右键点击【检查】选项，Windows电脑的同学可直接按F12或Fn+F12。

2. 随即，网页下方（或右方）弹出一个子窗口，这便是浏览器的开发者工具。



HTML的层级关系可通过这些小三角形进行查阅，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，就像电脑中一层一层的文件夹。

2.2.3 爬取网页源代码

示例：

```
import requests # 调用requests 库
res = requests.get('https://localprod.pandateacher.com/python-manuscript/crawler-html/spider-men5.0.html')   # 获取
code = res.text # 将返回的Response 对象转化为字符串格式
print(code)     # 打印Response 对象转化的字符串格式
```

注意：使用requests.get() 爬取的是网页源代码的内容，当网页是动态网页时，爬取到的代码和开发者工具中的代码是不一样的。