【零基础学爬虫】网页的基本构成

爬虫程序之所以可以抓取数据，是因为爬虫能够对网页进行分析，并在网页中提取出想要的数据。在学习 Python 爬虫模块前，我们有必要先熟悉网页的基本结构，这是编写爬虫程序的必备知识。

## 网页的基本结构

> 关于 Web 初步教程：Here

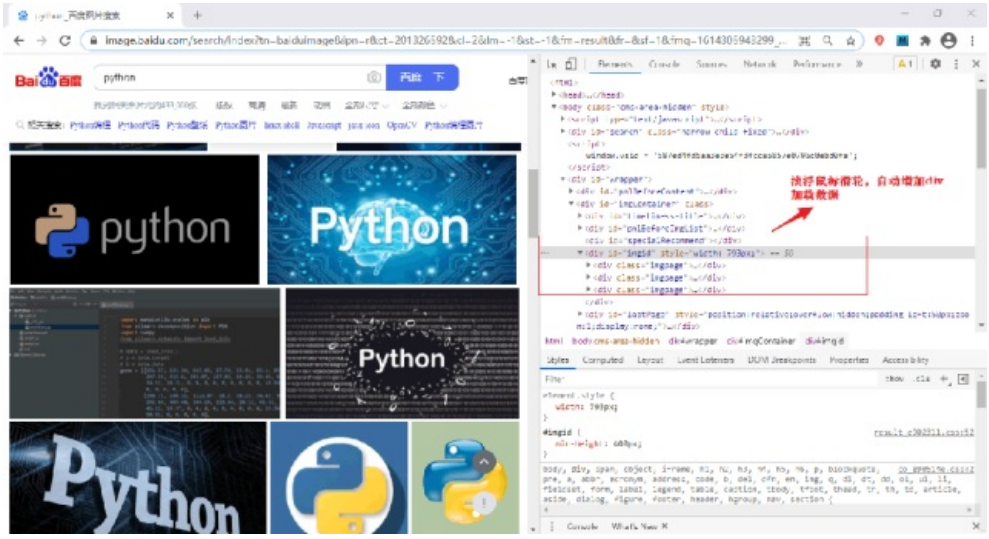
- 网页在组成上一般由三部分组成，分别是
- HTML（超文本标记语言）负责定义网页的内容、
- CSS（层叠样式表）负责描述网页的布局、
  JavaScript（简称“JS”动态脚本语言）负责网页的行为，它们三者在网页中分别承担着不同的任务。
- 网页从类型分为静态与动态，
- 静态网页是标准的 HTML 文件，通过 GET 请求方法可以直接获取，文件的扩展名是`.html`、`.htm`等（静态并非静止不动，它也包含一些动画效果，这一点不要误解）
- 动态网页指的是采用了动态网页技术的页面，比如 AJAX（是指一种创建交互式、快速动态网页应用的网页开发技术）和JSP(是 Java 语言创建动态网页的技术标准) 等技术，它不需要重新加载整个页面内容，就可以实现网页的局部更新。

针对静/动态的一个重要区别在于，有无需要连接后台数据库，

由于静态网页的内容相对固定，且不需要连接后台数据库，因此响应速度非常快。但静态网页更新比较麻烦，每次更新都需要重新加载整个网页。

动态页面使用“动态页面技术”与服务器进行少量的数据交换，从而实现了网页的异步加载。下面看一个具体的实例：

打开百度图片（https://image.baidu.com/）并搜索 Python，当滚动鼠标滑轮时，网页会从服务器数据库自动加载数据并渲染页面，这是动态网页和静态网页最基本的区别。如下所示：

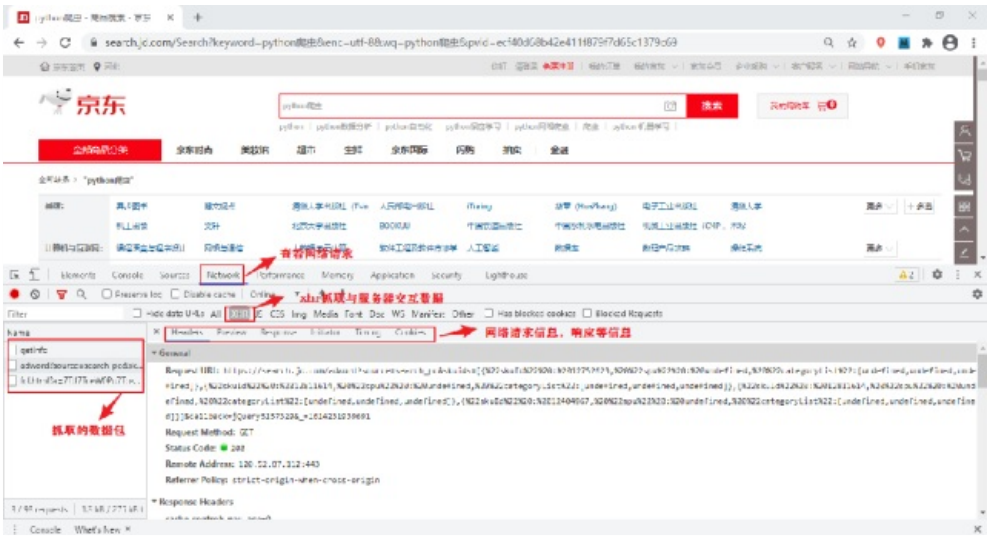



动态网页中除了有 HTML 标记语言外，还包含了一些特定功能的代码。这些代码使得浏览器和服务器可以交互，服务器端会根据客户端的不同请求来生成网页，其中涉及到数据库的连接、访问、查询等一系列 IO 操作，所以其响应速度略差于静态网页。

> **注意：**一般网站通常会使用动静相结合的方式，使其达到一种平衡的状态。

当然动态网页也可以是纯文字的，页面中也可以包含各种动画效果，这些都只是网页内容的表现形式，其实无论网页是否具有动态效果，**只要采用了动态网站技术，那这个网页就称为动态网页。**

抓取动态网页的过程较为复杂，需要通过动态抓包来获取客户端与服务器交互的 JSON 数据。抓包时，可以使用 Chrome 的开发者模式（快捷键：F12）`Network` 选项，然后点击 XHR，找到获取 JSON 数据的 URL，如下所示：




或者您也可以使用专业的抓包工具 Fiddler 。关于动态网页的数据抓取，在后续内容会做详细讲解。

## 审查网页元素

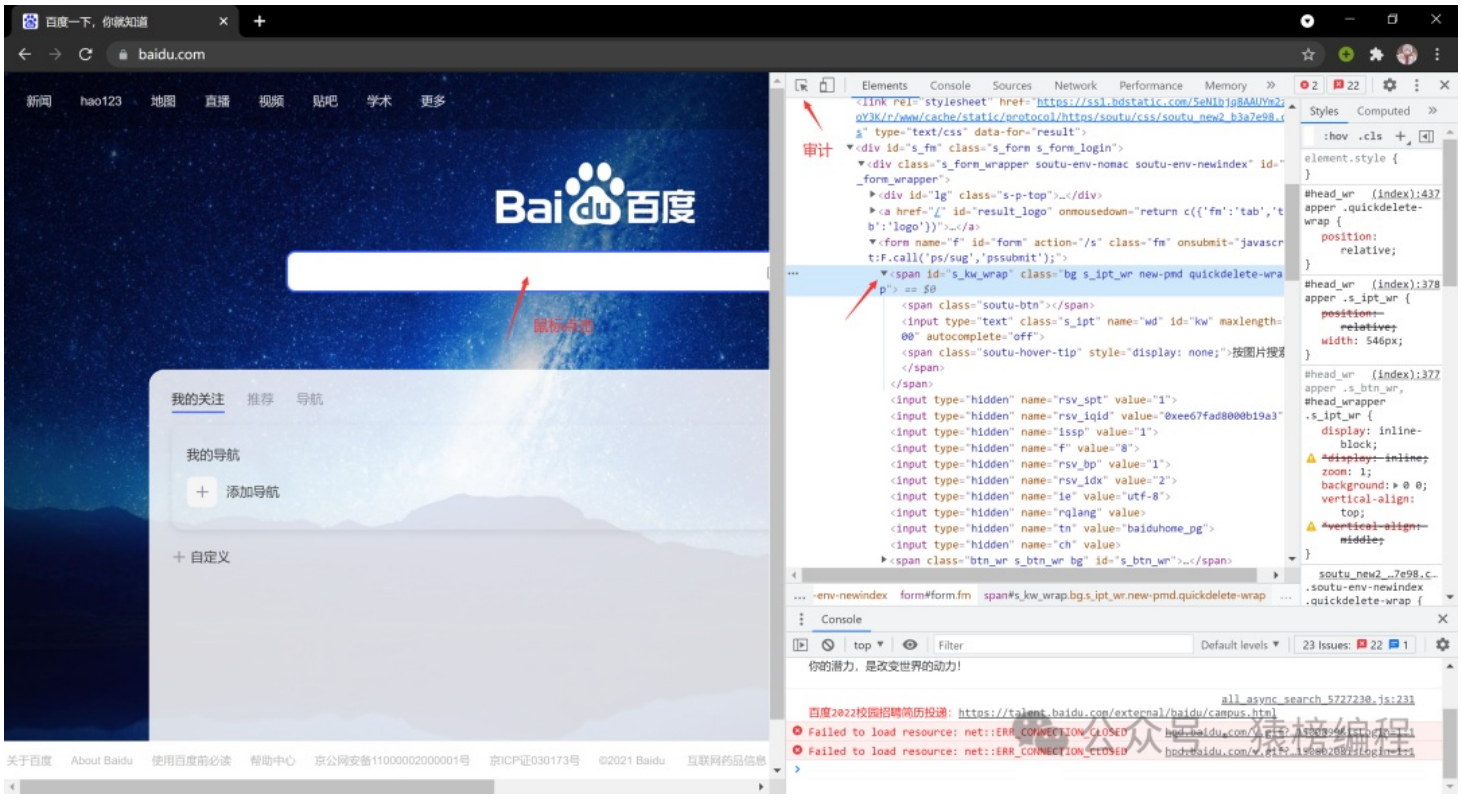
> 前面说了这么多关于网页的构成，其实是要引出本部分内容。

浏览器都自带检查元素的功能，不同的浏览器对该功能的叫法不同， Chrome 称为“检查”，而 Firefox 则称“查看元素”，尽管如此，但它们的功却是相同的。（接下来以 Chrome 进行操作）

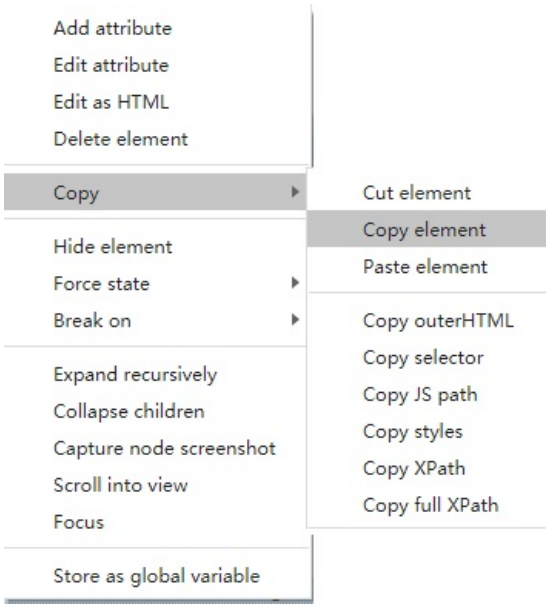
在动手编写爬虫程序前，必须要对网页元素进行审查、并且能从中提炼出有效的信息。

这就要求我们能善于发现网页元素的规律。

## 检查百度首页




点击审查元素按钮，然后将鼠标移动至您想检查的位置，比如百度的输入框，然后单击，此时就会将该位置的代码段显示出来（如图 1 所示）。最后在该代码段处点击右键，在出现的会话框中选择 Copy 选项卡，并在二级会话框内选择“`Copy element`”，如下所示：



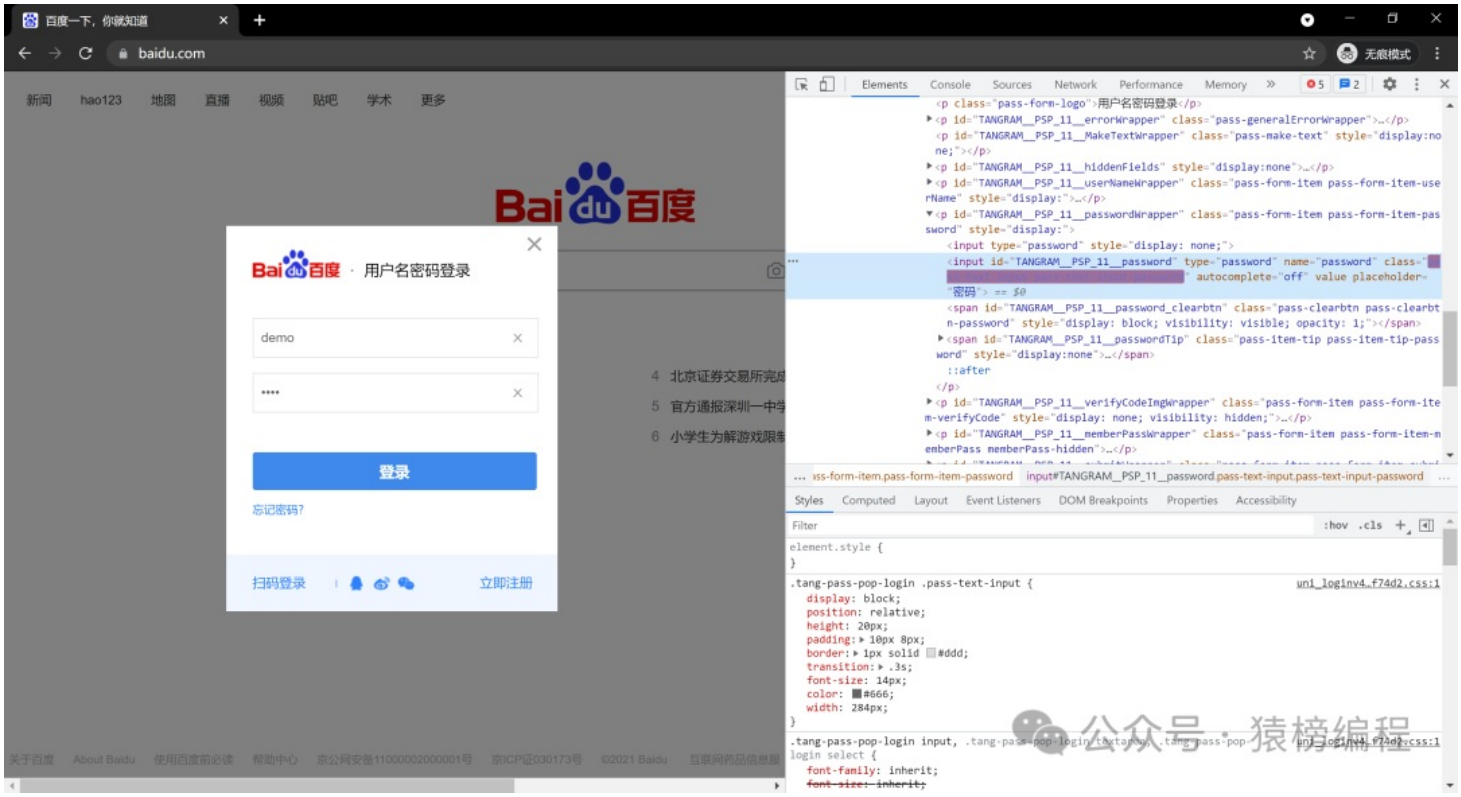

百度输入框的代码如下所示：

```
<input type="text" class="s_ipt" name="wd" id="kw" maxlength="100" autocomplete="off">
```

依照上述方法，您可以检查页面内的所有元素。

## 编辑网页代码

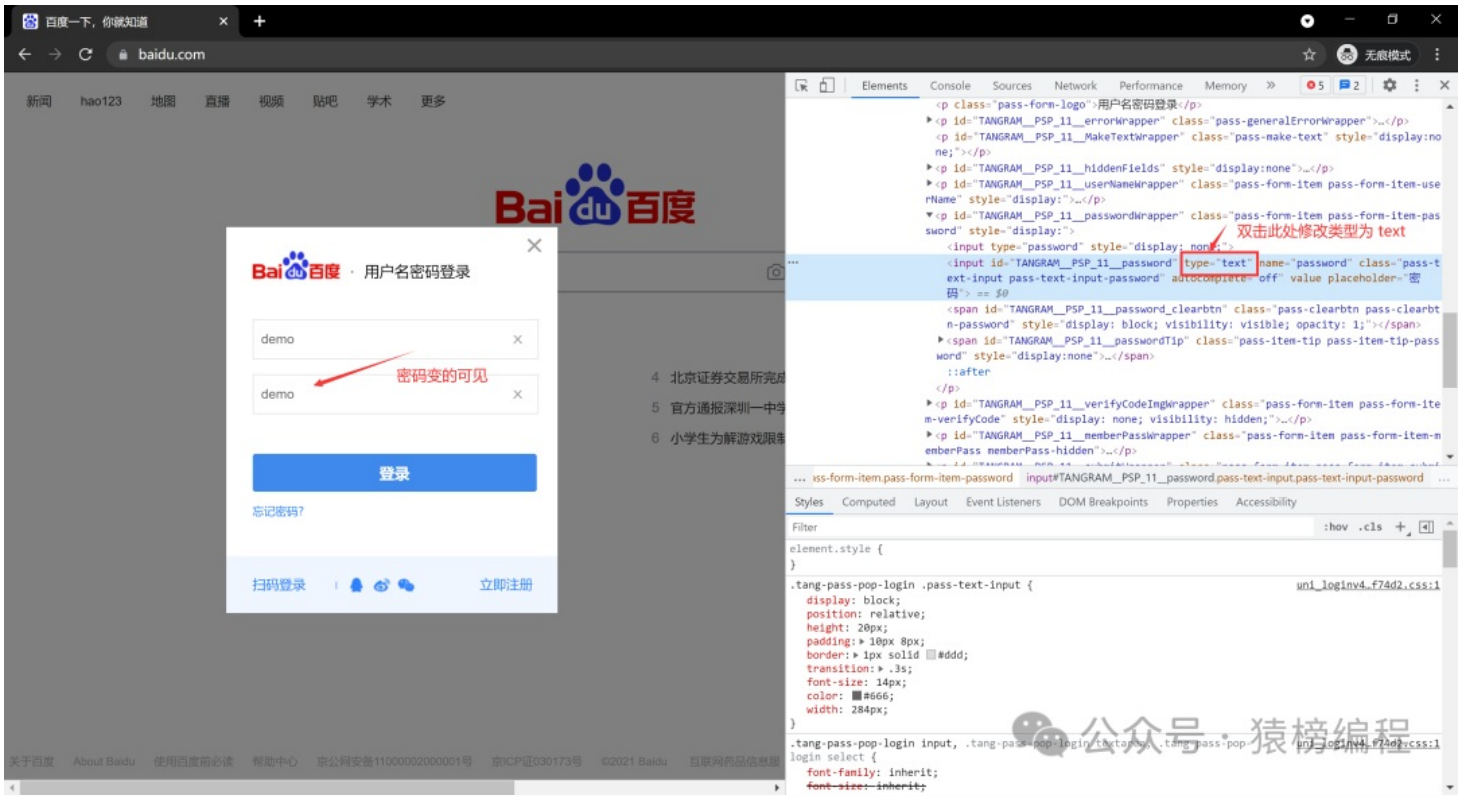
通过检查元素也可以更改网页代码，下面仍通过 百度 页面进行简单演示




检查密码框的 HTML 代码，代码如下所示：

```
<input id="TANGRAM__PSP_11__password" type="password" name="password" class="pass-text-input pass-tex
```

只要在显示出的代码段上稍微做一下更改，密码就会变为可见状态。如下图所示：




双击 type="password" 将输入框类型更改为 text，此类操作适用于所有网站的登录界面。但是需要注意，您做的更改仅限本次有效，当关闭网页后，会自动恢复为原来的状态。

## 检查网页结构

对于爬虫而言，检查网页结构是最为关键的一步，需要对网页进行分析，并找出信息元素的相似性。下面以猫眼电影网为例，检查每部影片的 HTML 元素结构。如下所示


第一部影片的代码段如下所示：

```
<div class="board-item-main">
      <div class="board-item-content">
              <div class="movie-item-info">
        <p class="name"><a href="/films/1299372" title="你好，李焕英" data-act="boarditem-click" data-
        <p class="star">
                主演：贾玲,张小斐,沈腾
        </p>
<p class="releasetime">上映时间：2021-02-12</p>    </div>
    <div class="movie-item-number score-num">
<p class="score"><i class="integer">9.</i><i class="fraction">5</i></p>
    </div>

      </div>
    </div>
```

接下来检查第二部影片的代码，如下所示：

```
<div class="board-item-main">
      <div class="board-item-content">
              <div class="movie-item-info">
        <p class="name"><a href="/films/553231" title="心灵奇旅" data-act="boarditem-click" data-val=
        <p class="star">
                主演：杰米·福克斯,蒂娜·菲,菲利西亚·拉斯海德
        </p>
<p class="releasetime">上映时间：2020-12-25</p>    </div>
    <div class="movie-item-number score-num">
<p class="score"><i class="integer">9.</i><i class="fraction">3</i></p>
    </div>

      </div>
    </div>
```

经过对比发现，除了每部影片的信息不同之外，它们的 HTML 结构是相同的，比如每部影片都使用 `<dd></dd>` 标签包裹起来。这里我们只检查了两部影片，在实际编写时，你可以多检查几部，从而确定它们的 HTML 结构是相同的。

> PS：通过检查网页结构，然后发现规律，这是编写爬虫程序最为重要的一步。