一，了解网页结构

对任何网页点右键点击检查会出来代码

Div部分是html写的。。Class等等标签是规定的CSS样式。。有的前缀是script，这就是javascript写的

绝大多数网页三者构成

Html好像是房间的结构部分，区别每一个部分是干什么用的，比如卧室和客厅

CSS是样式，比如房间的颜色

Js是具体设置，比如电视

二，<div> </div>

表示网页中的固定区域

两个div中加一个p标签加文字

还可以加上class修饰区域



P标签是文字内容

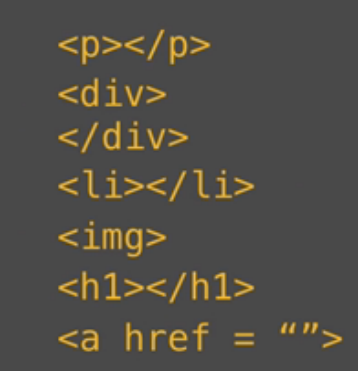
Div是文件框架

Li是列表

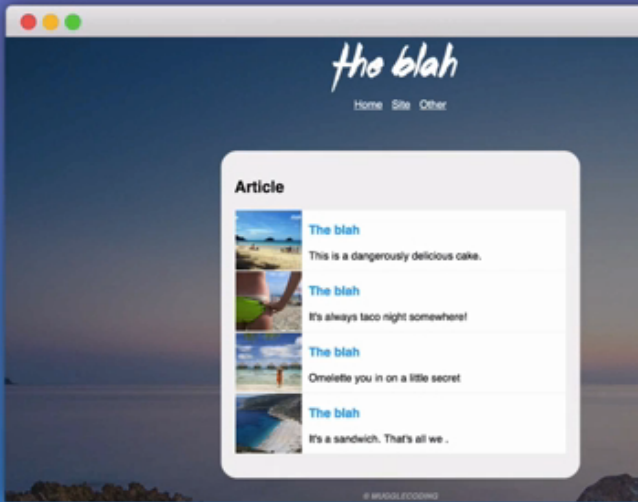
Img插入图片

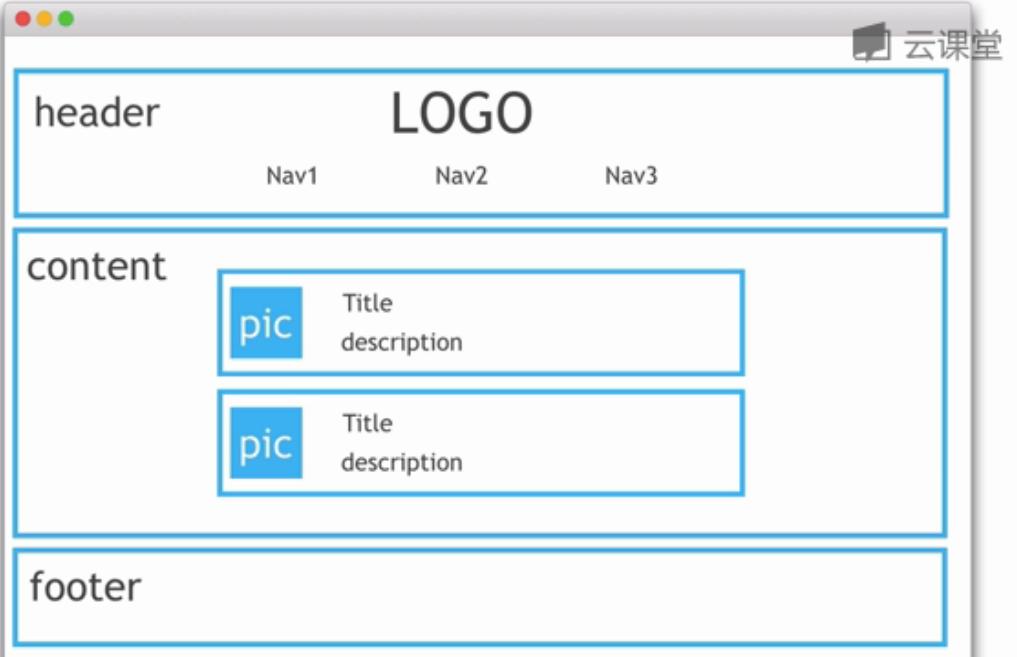
H1代表不同字号标题

A href =’’ 加入链接

\

三 做一个简单界面



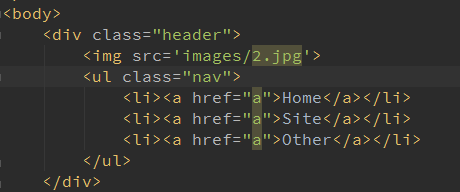


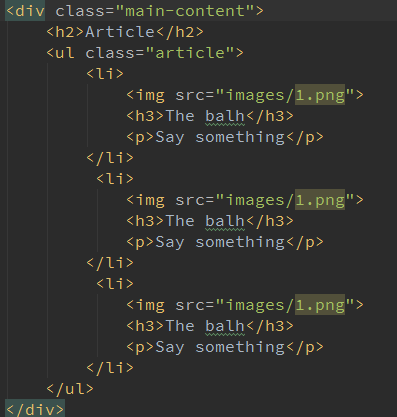
Body是网页中可见的部分

Head是给浏览器看的，我们不可间

Class定义的CSS样式要有文件定义，不然你是显示不出来的

Html自动支持可以显示

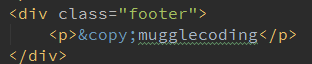




不出效果是因为没有他定义的CSS样式

图片调整大小

最后填写body中的footer部分

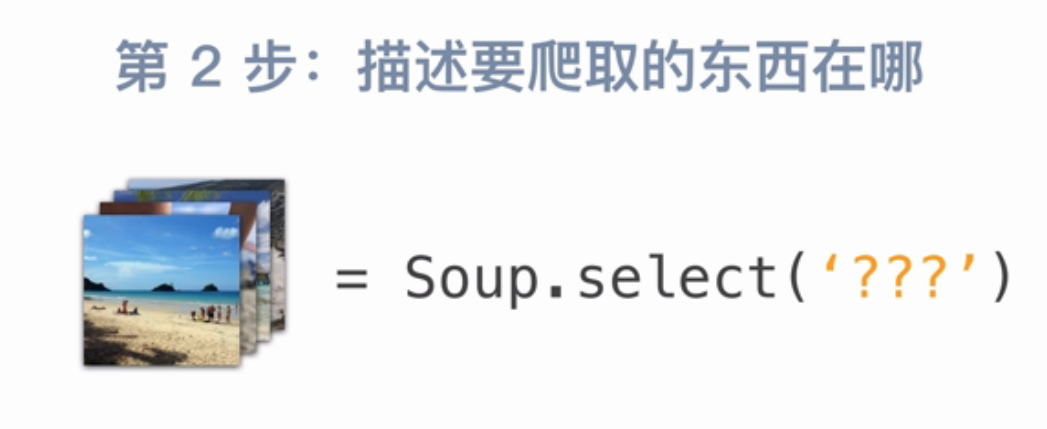


&copy就是那个C的标志

四 爬取表单界面

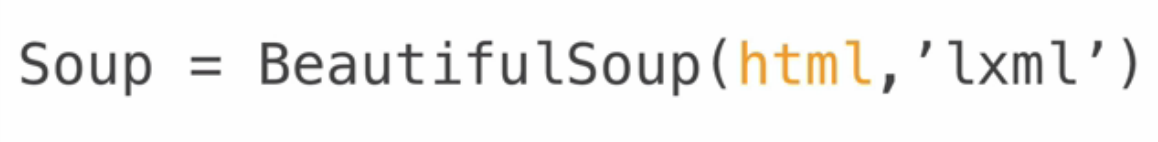








把信息按照一定格式装载数据容器中

1. 

解析网页，有lxml，这是一种解析库，还有其他库可以选择

对于一行描述的html可以通过右键copy把选择色了selector或者XPath来获取信息

CSS selector

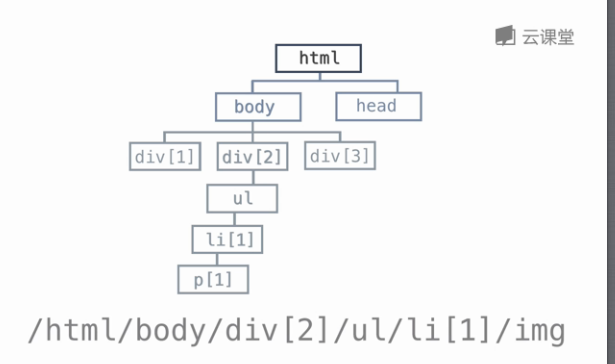
body > div.main-content > ul > li:nth-child(1) > img

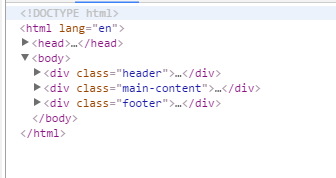
XPath:

/html/body/div[2]/ul/li[1]/img

Beautifulsoup指认第一种写法，去爬信息

1. XPAth与CSS Se介绍





Xpath通过序号选择



按照网页中的样式去进行选择，li后面的表示第一个

3.导入beautifulsoup库