首先，什么是有用的内容？我是想要什么类型的内容呢，我是想要以图片形式还是文本形式的内容？对于这些内容，在这个网站中我可能能从哪里获得呢，是网站本身的内容，还是评论诸如好差评，还是浏览量等等。因此首先要做的就是界定好我需要的有用内容到底是什么。

确定目标之后，即有用的内容是什么之后。据我了解，网站上页面的显示都是将处理后的HTML发送到用户的浏览器，这些处理后的HTML都是我们可以看得懂的网页内容。所以如果给定一个网站的话，那么最笨的方式就是右击查看页面源代码，将这些源代码保存下来，然后对内容进行提取（用python），需要内容的话我保存文字，需要图片我就保存图片。在课上学习之后，我们可以通过实现爬虫程序，搜索某一网站站点的URLS（网页地址），然后通过页面采集模块，完成对网页数据的采集，最后通过页面分析模块，对上一步采集的数据进行分析，提取出我们需要的网页数据和信息。

获取到的内容不一定就是有用的内容，可能更多成分是无用的，所以我们需要对爬取得到的数据进行处理和清洗。数据清洗是指在数据集中发现不准确、不完整或不合理数据，并对这些数据进行修补或移除以提高数据质量的过程。如果是图片类型的数据的话，需要自行对图片进行观察，找到含有有用信息的图片。对于数据类型的，首先是对得到的数据进行整体的浏览找到缺失值，通过插值的方法填充异常值。其次通过箱型图或者3sigma原则找到异常数据的值，对该数据进行判断，是剔除异常值还是通过插值等方法重新填充值。最后对数据进行描述性统计等操作，最终得到我们想要的有用内容。