⾸先分析将要爬取的新闻⽹⻚。在⽹站上右击“检查 和“查看页面源代码” 可以查看要爬取的⽹⻚的源码，利⽤代码栏左上⾓的“查看器”按钮，在⽹⻚上选择想要审视的要素，就可以看到该要素对应的 html 源代码。利 ⽤这个⽅式，我们可以确定要爬取的内容的格式。

在种子页面，我们需要爬取的就是新闻网页链接地址，供我们之后爬取新闻内容使用。网页链接一般在标签内。新浪新闻的主⻚是我们本次爬取的种⼦⻚⾯，我们将解析出这个⻚⾯上的新闻⻚⾯超链接并对这些新闻⻚⾯进⾏爬取，根据需要爬取的数据修改format内容，以及用正则表达式匹配新闻url。

在内容页面，我们需要爬取的就是新闻具体内容，内容包括 title、author、keywords等。因为不同的内容位于不同的标签内，并且有不同的name，在之后我们要根据我们所需要爬取的内容，使用选择器选择我们要爬取的元素。

按网页搜索的内容进行分页管理