每个网页的构造都是有所不同的，所以需要在数据抓取之前分析网页。借助Google Chrome浏览器中的检查功能（Ctrl+Shift+I）就可查看网页结构，而且每个form、table等标签位置在网页中都是同步显示的，非常直观。

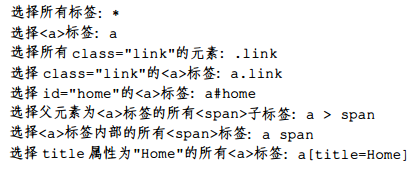
三种数据抓取的方法：

1. 正则表达式
2. Beautiful Soup
3. Lxml

正则表达式不可否认是一种强大的工具，但是它最大的问题就是不支持版本的变更——网页一变就不可用，所以我们一般不会使用它。示例：myRegex.py

使用Beautiful Soup的第一步是将已下载的html内容解析为soup文档，这个soup文档可以理解为完整的html。然后利用find或者find\_all去寻找所要的标签。使用起来比较方便快捷。示例：myBSoup.py

Lxml的使用和Beautiful Soup的使用都是很相似的，因为它们为我们提供了很好的封装。而在选择元素则会使用到CSS选择器，照猫画虎就行。



Lxml是最佳的选择，因为它的解析速度和正则表达式一样快，但是使用起来简单。而Beautiful Soup使用简单，但是很慢。个人认为爬虫最关键的就是速度。