**任务1，爬取新闻**

1. 新浪网和搜狐网的新闻比较好爬取，它并没有使用JS异步加载，我是用BeautifulSoup和re直接解析的，
2. 网易新闻的标题及内容是使用js异步加载的，单纯的下载网页源代码是没有标题及内容的，我在Network的js中找到了需要的内容，然后用正则表达式获取了需要的标题及其链接，并使用了BeautifulSoup来获取相应标题的内容比较困难。
3. 除了上面三个新闻网站外，我还尝试了今日头条。但是今日头条的新闻跟前两个也都不一样，它的标题和链接是封装到json文件中的，但是他json文件的url参数是通过一个js随机算法变化的，所以需要模拟json文件的参数。该网站自带反爬机制，所以我添加了cookie。关于新闻的内容，我用了正则表达式提取了中文。具体URL的获取我参考了一些网上的方法，但是一直出现各种各样不同的问题，没有能够调出来。