Python爬虫3-request库与爬虫设计的基本思路

2020年3月12日　　12:56

对于没有WEB经验的同学，预习推荐：
Web的基本工作原理、HTTP协议和URL说明：https://www.cnblogs.com/fwnboke/p/9114381.html
Web服务器的基本原理：https://blog.csdn.net/qq_36359022/article/details/81666221
Web服务器/Web容器/Web应用程序服务器/反向代理：
https://blog.csdn.net/a3192048/article/details/89792586
Web服务器的工作原理（概括）及其相关协议（HTTP、TCP/IP与FTP）：
https://blog.csdn.net/lz233333/article/details/51210554

域名www.baidu.com（URL） -> 电信DNS：223.5.5.5->转换成IP
HTTP：超文本传输协议　****P 协议
TCP的漏洞就造成了DDOS

request（请求）获取获取网页的时候，拿到的是一段文本（若干个字符串）（即response）
浏览器拿到response的时候，解释渲染成我们看到的样子（假象，只是当时服务器的镜像，是可以在本地进行修改的，本地修改的不是改动服务器上的内容）

这段文本：
1、HTML：网页的基础，标记语言（有标签组成，eg. 天气不错！ ）
　　天气不错！　整体叫元素， 叫标签
　　H5 <!DOCTYPE html>　HTML 5.0
　　HTML4的年代，HTML只负责表示文字的意义
　　学HTML有个非常大的误区，就是：HTML标签应该表示的是意义，而不是样子
2、有哪些标签呢？（第一部分）
　　<h1><h2><h3>...<h6>　标题
　　　段落，正文
　　　无序列表　 子项　在HTML制作过程中，还经常被用来做菜单（下拉菜单）
　　　有序列表　 子项
　　<a>　超链接标签　实现了多个网页之间的关联
　　[1]　href、target是属性，分别代表超链接的目标地址和跳转方式
3、以一篇英文论文来制作我们的简单HTML demo
https://www.sciencedirect.com/science/article/pii/S0898122111009084#br000005

Eg. 爬取油气学院新闻

原则：不给人家添麻烦，盗亦有道

发现所有新闻都藏在了

2020-01-14油气工程学院2020年寒假值班安排

\r\n(时间)(新闻的标题)

如何去找这种“有规则的文字”呢？ A：正则表达式
参考资料：百度“python 正则表达式”或者“python re”

具体代码参见附件