O.请问URL 是“统一资源标识符”还是“统一资源定位符”?

答:统一资源标识符

正确答案:URI是统一资源标识符 URL是统一资源定位符 一句话概括:URI是用字符串来标识某一互联网资源 而URL则是标识资源的地址(如我们现在常说的网址就URL)因此URI属于父类 而URL属于URI的子类

1.什么是爬虫?

答:如果说现实中的话就是恶心的昆虫 计算机领域中的爬虫就是可以在互联网上爬取自己想要的数据的虫

2.设想一下，如果你是负责开发百度蜘蛛的攻城狮，你在设计爬虫时应该特别注意什么问题?

答:编码问题和一些资源信息处理的问题

正确答案:不要重复的爬取同一个URL的资源 如果没做这个防范 如果一个URL内容中包含该URL本身 那么就会陷入无限循环

3.设想一下，如果你是网站的开发者，你应该如何禁止百度爬虫访问你网站中的敏感内容? (课堂上没讲，可以自行百度答案)

答:在防火墙上添加策略 如果定义1分钟内并发连接数超过10ip为非法

正确答案:在网站的根目录下创建并编辑robots.txt文件 用于表明你不希望搜索引擎抓取工具访问你网站上的哪些内容

4.urlib.request.urlope 返回的是什么类型的数据?

答:返回是一个HTTPResponse实例对象 它属于htto.client模块 调用read()方法才能读出URL的内容

5.如果访问的网址不存在，会产生哪类异常? (虽然课堂没讲过，但你可以动手试试)

答:HTTPError异常

6.鱼C工作室([http://www.fishc.com](http://www.fishc.com/)) 的主页采用什么编码传输的?

答:gbk

7.为了解决ASCII 编码的不足，什么编码应运而生?

答:中国的GB2312和utf-8