Python职场实用技能

——网络爬虫入门

rennywang(王瑞刚)

课程大纲

技能一：如何爬取网页的内容？

```
import requests

r = requests.get("https://www.baidu.com/")

print (r.text)
```

课程大纲

技能一：如何爬取网页的内容？

技能二：如何解决爬取结果中有乱码的问题？

网页源文件格式　　IDE解析格式

课程大纲

技能一：如何爬取网页的内容？

技能二：如何解决爬取结果中有乱码的问题？

技能三：如何巧妙躲避“反爬措施”？

反爬措施一：来源审查

来源审查：判断User-Agent进行限制

检查来访HTTP协议头的User-Agent域，只响应浏览器或友好爬虫的访问

反爬措施二：数量统计

反爬措施三：robots协议

Robots协议

Robots Exclusion Standard，网络爬虫排除标准

作用：

网站告知网络爬虫哪些页面可以抓取，哪些不行

形式：

在网站根目录下的robots.txt文件

课程大纲

技能一：如何爬取网页的内容？

技能二：如何解决爬取结果中有乱码的问题？

技能三：如何巧妙躲避“反爬措施”？

技能四：如何从爬取内容中过滤出我们想要的信息？

技能五：如何批量爬取信息？

技能六：如何获得标题对应链接？

技能七：如何爬取动态加载的内容？

技能四到七可以参考对应代码哦