1. 什么是爬虫

可以理解为在网络上进行爬行的一个蜘蛛，在网络上遇到一些网站(网络上的结点)k可以将其爬取下来。

正式定义:请求网站并提取数据的自动化程序

请求:向服务器发送请求获取到网络资源

我们的任务:从请求到网页文本中提取到我们需要的数据存储到数据库中。

1. 基本流程
2. 向目标站点发送请求(向服务器发送Request)
3. 服务器正常响应则返回Response
4. 解析获取到的内容
5. 保存数据
6. Request和Response

我们向服务器发送一个代码请求访问一些内容，服务器获取到这串代码后向我们返回一个结果。我们向服务器发送的那一串代码(请求头部)就是Request，服务器返回的结果就是响应(Response)。

1. Request中包含的信息
2. 请求方式:GET和POST这两种常用请求，在GET请求中将请求的参数是包含在URL中，POST请求中请求的参数放在FormData中，GET请求可以直接输入URL访问，POST请求不行。(HEAD,PLT,
3. 请求URL
4. 请求头信息(包含浏览器信息等,是一个请求的配置信息)
5. 请求体（post请求时才有内容）
6. Response中的信息
7. 响应状态码(200为正常，300以上为跳转，400一些错误如资源不存在,500y以上服务器处理异常)
8. 响应头(内容类型、内容长度、服务器信息、设置cookie等)
9. 相应体(服务器反馈过来的资源)
10. Request库

Request.get(url) 向指定URL发送get请求如果正常返回一个响应

Response.headers 响应头

Response.content 获取到响应体

Response..status.code 响应的状态码 //进行实验时没有这个属性

1. 爬虫可以抓取怎么样的数据
2. 网页文本
3. 图片，使用二进制保存
4. 视频
5. 其它，能够请求到的
6. 解析方式
7. 直接处理
8. Json解析
9. 正则表达式
10. BeautifulSoup
11. PyQuery
12. XPath
13. 如何解决JavaScript渲染的问题
14. 分析Ajax请求
15. Selenium /WebDriver
16. Splash
17. PyV8、Ghost.py
18. 怎样保存数据
19. 文本
20. 关系型数据库
21. 非关系型数据库(key-value形式)
22. 二进制文件