爬虫概念、工具和HTTP

1.什么爬虫

- 爬虫就是`模拟客户端(浏览器)发送网络请求`，获取响应，按照规则提取数据的程序
- `模拟客户端(浏览器)发送网络请求`：照着浏览器发送一模一样的请求，获取和浏览器一模一样的数据

2.爬虫的数据去哪了

- 呈现出来：展示在网页上，或者是展示在app上
- 进行分析：从数据中寻找一些规律

3.需要的软件和环境

- python3
  - 黑马python基础班15天视屏：http://yun.itheima.com/course/214.html
  - 基础语法（字符串，列表，字典，判断和循环）
  - 函数（函数的创建和调用）
  - 面向对象（如何创建一个类，如何使用这个类）
- pycharm
  - python编辑器
- chrome浏览器
  - 分析网络请求用的

4.浏览器的请求

- url
  - 在chrome中点击检查，点到network，
  - url = 请求的协议+网站的域名+资源的路径+参数
- 浏览器请求url地址
  - 当前url对应的响应+js+css+图片 ---》elements中的内容
- 爬虫请求url地址
  - 当前url对应的响应
- elements的内容和爬虫获取到的url地址的响应不同，爬虫中需要以当前url地址对应的响应为准提取数据
- 当前url地址对应的响应在哪里
  - 从network中找到当前的url地址，点击response
  - 在页面上右键显示网页源码

5.认识HTTP、HTTPS

- HTTP:超文本传输协议
  - 以明文的形式传输
  - 效率更高，但是不安全
- HTTPS:HTTP + SSL(安全套接字层)
  - 传输之前数据先加密，之后解密获取内容
  - 效率较低，但是安全
- get请求和post请求的区别
  - get请求没有请求体，post有，get请求把数据放到url地址中
  - post请求常用于登录注册，
  - post请求携带的数据量比get请求大，多，常用于传输大文本的时候
- HTTP协议之请求
  - 1.请求行
  - 2.请求头
    - User-Agent:用户代理：对方服务器能够通过user_agent知道当前请求对方资源的是什么浏览器

        - 如果我们需要模拟手机版的浏览器发送请求，对应的，就需要把user_agent改成手机版
    - Cookie：用来存储用户信息的，每次请求会被携带上发送给对方的浏览器
        - 要获取登录后才能访问的页面
        - 对方的服务器会通过cookie来判断是我们是一个爬虫
  - 3.请求体
    - 携带数据
    - get请求没有请求体
    - post请求有请求体
- HTTP协议之响应
  - 1.响应头
    - Set-Cookie：对方该字段设置cookie到本地
  - 2.响应体
    - url地址对应的响应