

Python 进阶训练营

尹会生

② 简单爬虫

目录 CONTENTS

- 01 编写爬虫需要掌握的 HTTP 、 HTML 基础知识
- 02 urllib、Requests 库的深入讲解

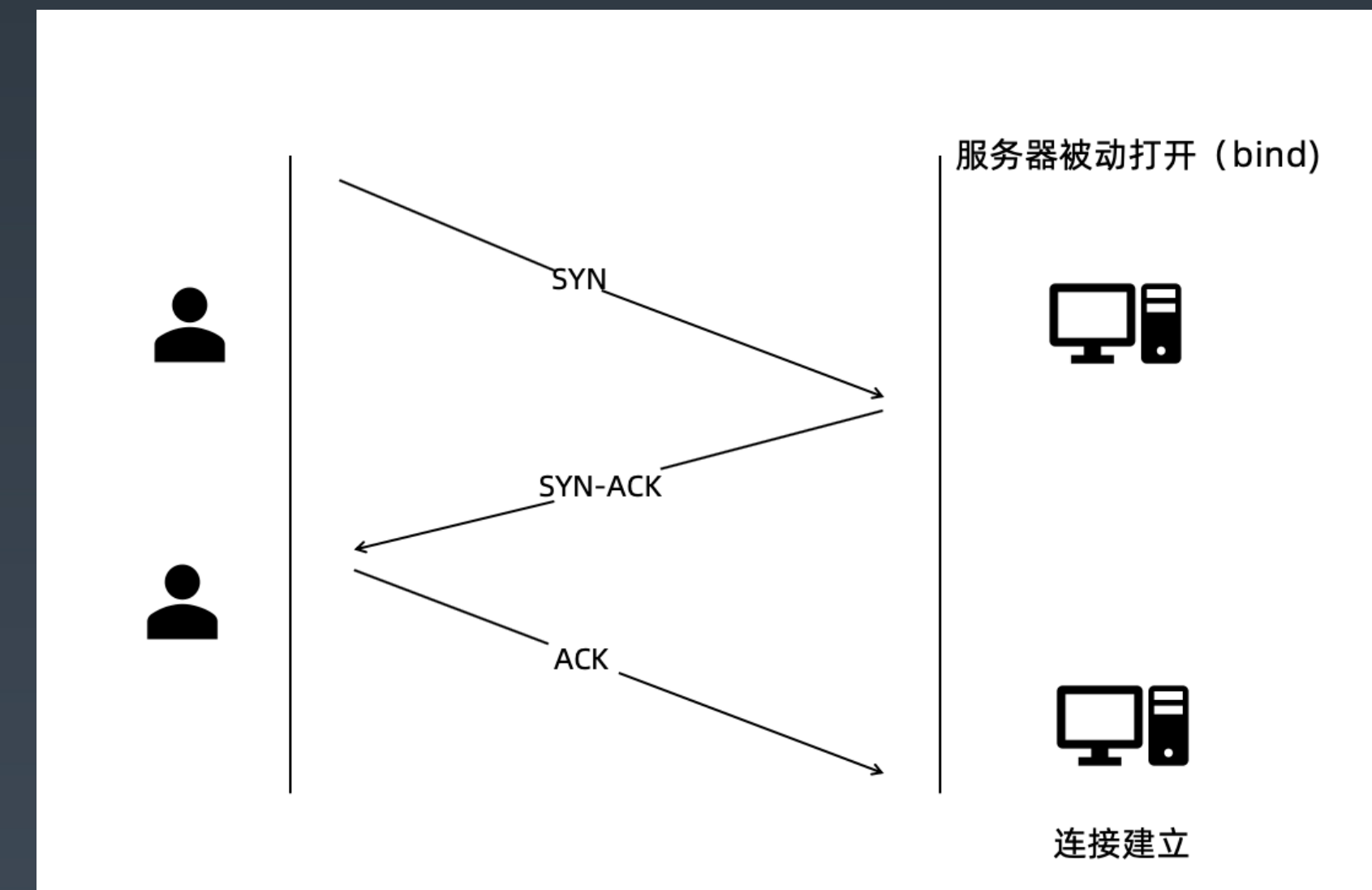
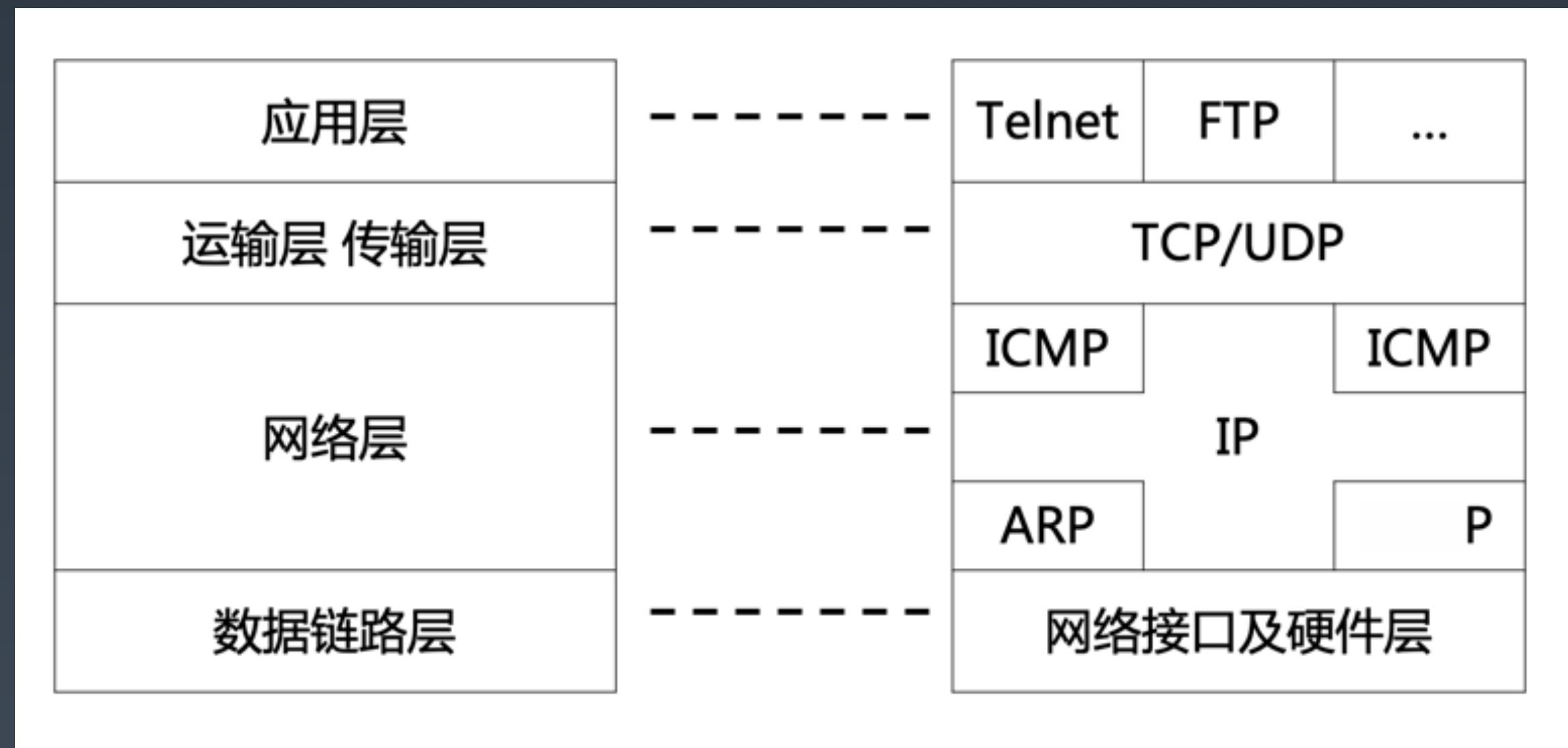
学习目标

掌握编写爬虫需要掌握的 HTTP、HTML 基础知识

掌握 Requests 库的用法

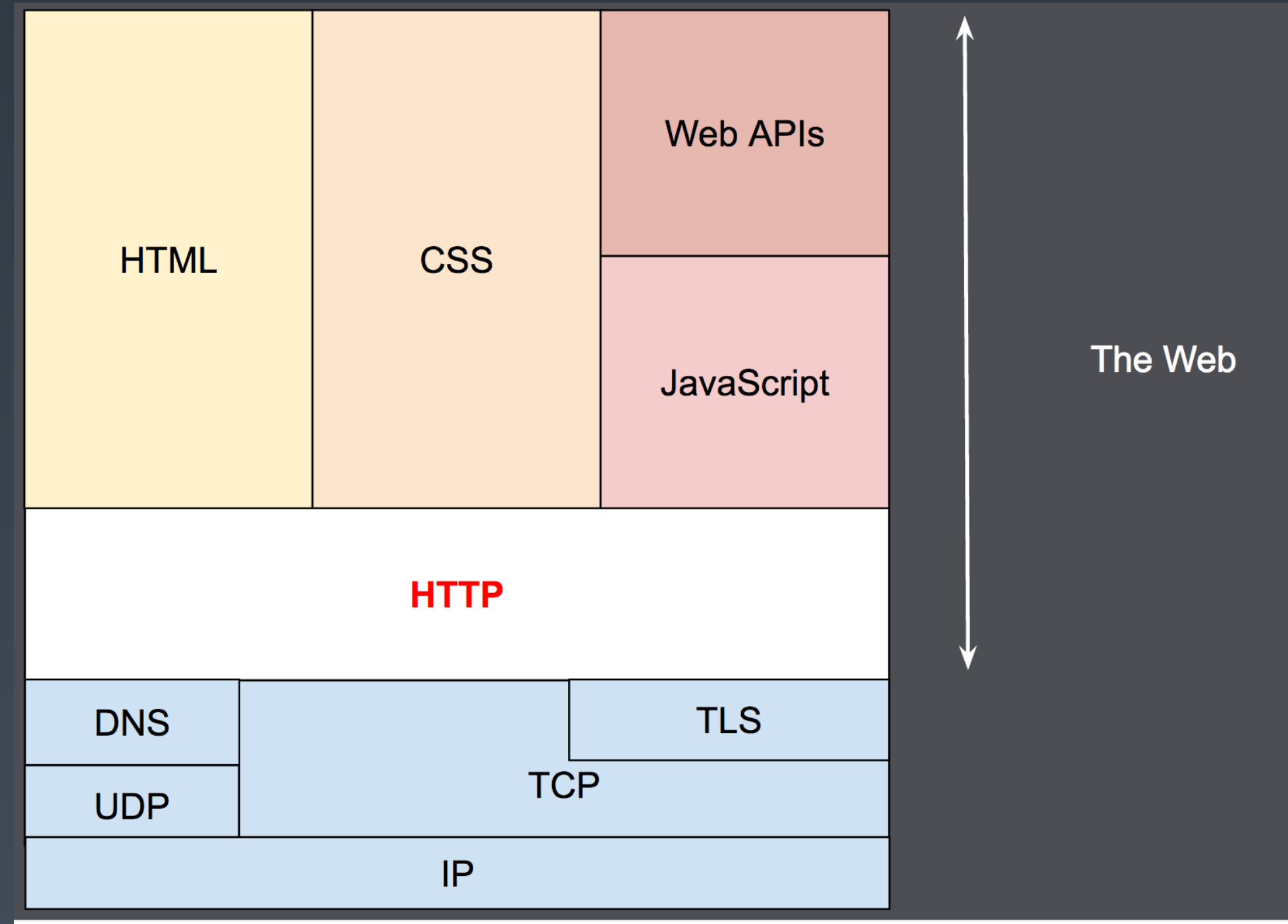
掌握 XPath 的用法

TCP 协议简介



TCP 协议的分层概念

HTTP 协议



HTTP 协议与浏览器

HTTP 协议

▼ General

Request URL: https://www.douban.com/

Request Method: GET

Status Code: 🟢 200 OK

Remote Address: 154.8.131.172:443

Referrer Policy: no-referrer-when-downgrade

× Headers Preview Response Initiator Timing Cookies

Remote Address: 154.8.131.172:443

Referrer Policy: no-referrer-when-downgrade

▼ Response Headers [view source](#)

Cache-Control: must-revalidate, no-cache, private

Connection: keep-alive

Content-Encoding: br

Content-Type: text/html; charset=utf-8

Date: Tue, 18 Feb 2020 09:10:08 GMT

Expires: Sun, 1 Jan 2006 01:00:00 GMT

Keep-Alive: timeout=30

Pragma: no-cache

Server: dae

Strict-Transport-Security: max-age=15552000;

Transfer-Encoding: chunked

Vary: Accept-Encoding

Vary: Accept-Encoding

X-DAE-App: sns

X-DAE-Instance: home

HTTP 协议的请求过程

HTTP 协议请求与返回头部

HTTP 协议

HTTP 请求方式	
GET	获取数据
POST	将实体提交到指定的资源
DELETE	删除指定资源
HEAD	获取响应头部，没有响应体
PUT	替换目标资源
.....

登录是怎么实现的

帐号登录

安全登录

邮箱/会员帐号/手机号

请输入密码

☐

记住我

忘记密码

登录

还没有微博? [立即注册!](#)

其它登录: [淘](#) [QQ](#) [微信](#) [微博](#) [+](#)

HTTP 协议

HTTP 状态码 (响应代码)	
1xx	信息响应
2xx	成功响应
3xx	重定向
4xx	客户端响应
5xx	服务端响应

HTTP 协议与 HTML 的关系

- W3C 标准
- HTML 常用的标签和属性
- 网页的三大组成部分：结构、表现、行为

HTML

HTML 不是编程语言，是一套标记，使用标记来描述网页的结构。

对于数据获取有用的标记：

`<html>内容</html>`

`<head>内容</head>`

`<body>内容</body>`，在此标记之间可以包含如`<p></p>`标记

``

`链接文字或者图片`

`<div>`：内容划分元素

``：`` 与 `<div>` 元素很相似，但 `<div>` 是一个块元素，而 `` 则是行内元素。

CSS

层叠样式表 (Cascading Style Sheets)

用途：解决内容与表现分离的问题

内联样式表

`<body style="background-color:red; "></body>`

嵌入式样式表

写在 `<style type= " text/css" ></style>` 标记之间

外部样式

写为 .css 文件

`<link rel="StyleSheet" type="text/css" href="style.css">`

JavaScript

轻量级的脚本语言，由浏览器进行解释执行。

直接引用

在 `<script></script>` 标记中编写代码

外部引用

使用单独 .js 文件

JSON

JavaScript 对象表示法 (JavaScript Object Notation)

- 多用于存储和交换文本信息
- Web 前端中运用非常广泛
- JSON 使用 JavaScript 语法来描述数据对象，但独立于语言 and 平台
- JSON 解析器和 JSON 库支持许多不同的编程语言

XPath

七种类型的节点：元素、属性、文本、命名空间、处理指令、注释以及文档（根）节点。

节点关系：

- 父 (Parent)
- 子 (Child)
- 同胞 (Sibling)
- 先辈 (Ancestor)
- 后代 (Descendant)

XPath

XPath 是沿着路径来选取节点的

XPath 路径表达式	
nodeName	节点下所有的子节点
/	根节点选取
//	任意位置选取
.	当前节点
..	当前节点的父节点
@	选取属性

XPath

XPath 谓语句 (选取位置)	
第一个元素	/root/path/to/node[1]
最后一个元素	node[last()]
倒数第二个元素	node[last() - 1]
选择属性	//nodep[@name] //nodep[@name='value'] //nodep[@name>100
可以使用通配符	* 和

urllib

历史上出现的 urllib、urllib2、urllib3 库的区别和联系：

- Python2 中，urllib 和 urllib2 是 python 的标准库，urllib2 是 urllib 的增强版。
- Python3 中，urllib2 合并到 urllib 中。
- urllib3 提供现场安全连接池和文件 post 支持。

Requests

Requests 基于 urllib3 开发

Requests: HTTP for Humans

简单且强大

Requests 示例

Requests 库实现 HTTP 协议的常见操作

Requests 下载文件

Requests 库处理 cookies

cookies 和 session

动态网页与 Selenium 的使用简介

使用 Selenium 来进行模拟登录微博

```
pip install selenium
```

下载 Chrome 浏览器的 webdriver 到 Python 的 scripts 目录下

THANKS! |  极客大学