小象学院

HTTP

主　讲：杨真

小象学院

Part 1 基础

- 环境搭建
- HTML 基础
- 第一个10行代码的爬虫
- 内容抽取及解析
- HTTP 协议
- POSTMAN 工具详解

Part 2 爬虫

- 网站结构分析
- 抓取方案
- 多线程并行及排重
- 用 MySQL 信息存储

Part 3 进阶

- 网站服务结构
- Cookie 及 登录处理
- 控制抓取的节奏
- 日志
- 守护进程

Part 4 实战

- 网站结构分析
- 网页抓取方案
- 数据提取
- 存储方案


应用层
表示层
会话层
应用层
传输层
网络层
数据链路层
物理层
互联网络层
网络接口层
TSI 四层模型
OSI 七层模型

OSI

- 物理层：电器连接
- 数据链路层：交换机，STP，帧中继
- 网络层：路由器，IP 协议
- 传输层：TCP、UDP 协议
- 会话层：建立通信连接，网络拨号
- 表示层：每次连接只处理一个请求
- 应用层：HTTP、FTP

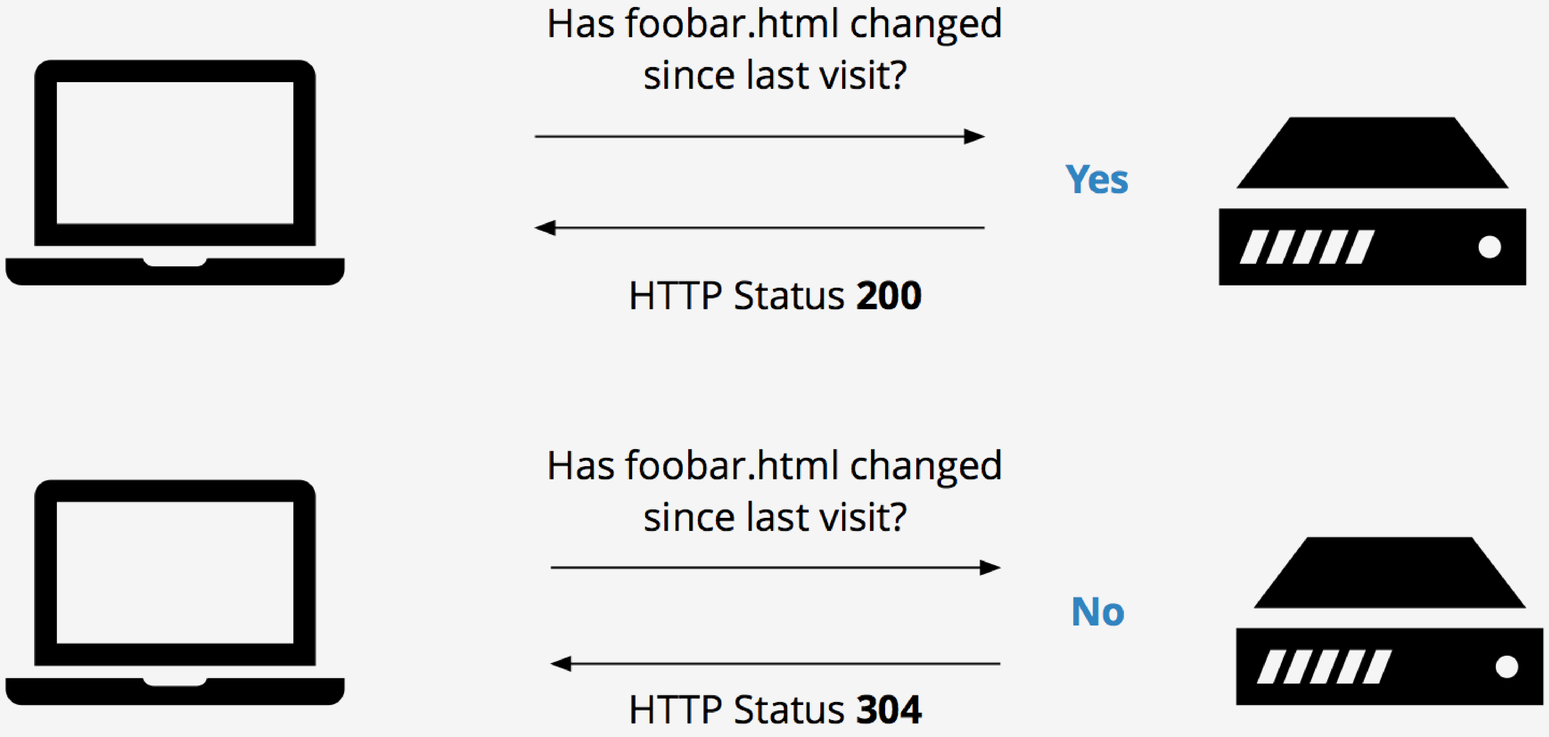
HTTP

- 应用层的协议
- 无连接：每次连接只处理一个请求
- 无状态：每次连接、传输都是独立的

Request Header

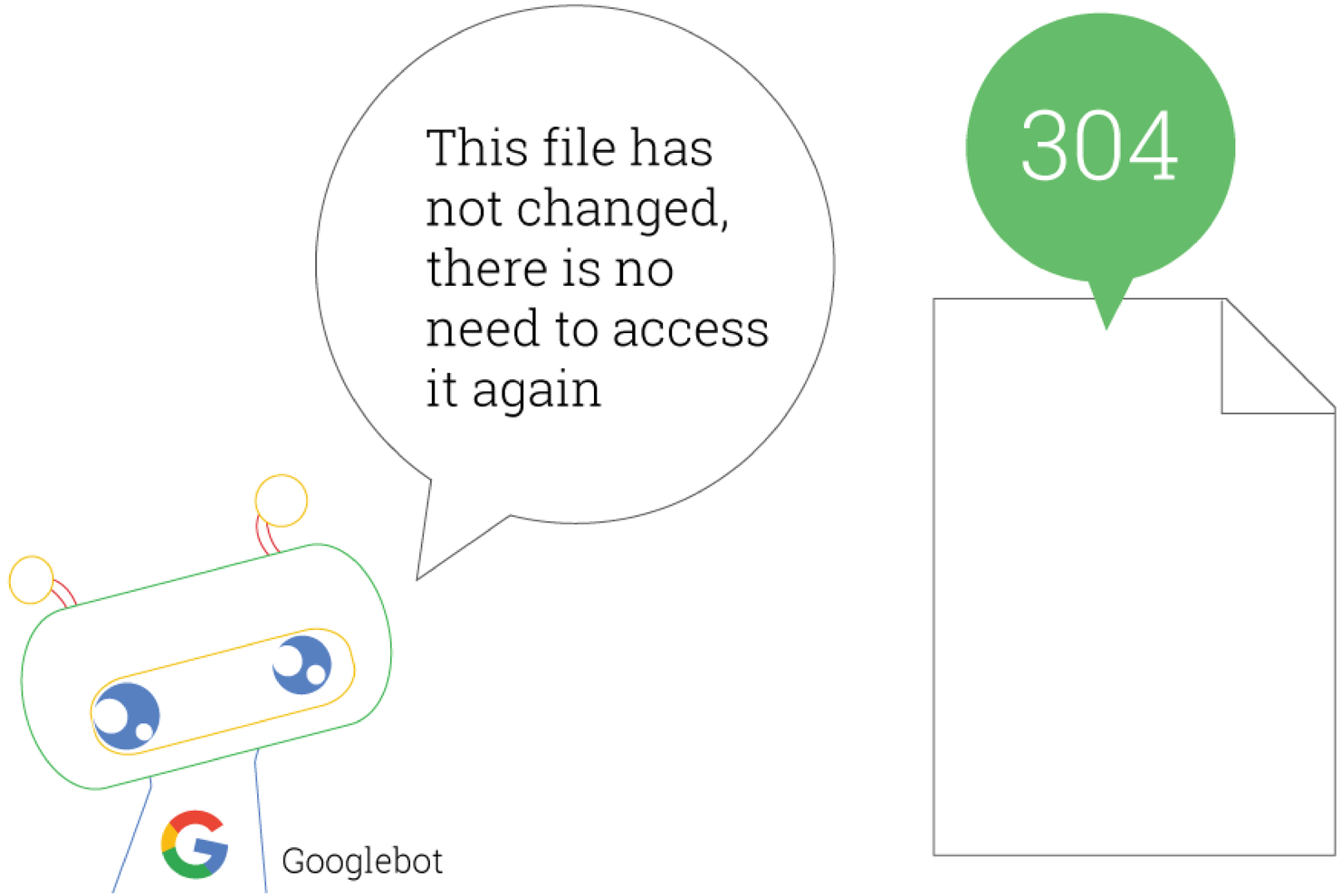
- Accept: text/plain
- Accept-Charset: utf-8
- Accept-Encoding: gzip, deflate
- Accept-Language: en-US
- Connection: keep-alive
- Content-Length: 348
- If-Modified-Since
- Content-Type: application/x-www-form-urlencoded
- Date: Tue, 15 Nov 1994 08:12:31 GMT
- Host: en.wikipedia.org:80
- User-Agent: Mozilla/5.0 (X11; Linux x86_64; rv:12.0) Gecko/20100101 Firefox/21.0
- Cookie: $Version=1; Skin=new;

If-Modified-Since



If Modified Since HTTP Header

If-Modified-Since



Keep Alive

小象学院

HTTP/1.1

默认情况下所在HTTP1.1中所有连接都被保持，除非在请求头或响应头中指明要关闭：Connection: Close

Keep-Alive功能使客户端到服务器端的连接持续有效，当出现对服务器的后继请求时，Keep-Alive功能避免了建立或者重新建立连接。

Response Header

- Accept-Patch: text/example;charset=utf-8
- Cache-Control: max-age=3600
- Content-Encoding: gzip
- Last-Modified: Tue, 15 Nov 1994 12:45:26 GMT
- Content-Language: da
- Content-Length: 348
- ETag: "737060cd8c284d8af7ad3082f209582d “
- Expires: Thu, 01 Dec 1994 16:00:00 GMT
- Location: http://www.w3.org/pub/WWW/People.html
- Set-Cookie: UserID=JohnDoe; Max-Age=3600; Version=1
- Status: 200 OK

Request Method

HTTP Method	RFC	Request Has Body	Response Has Body	Safe	Idempotent	Cacheable
GET	RFC 7231	No	Yes	Yes	Yes	Yes
HEAD	RFC 7231	No	No	Yes	Yes	Yes
POST	RFC 7231	Yes	Yes	No	No	Yes
PUT	RFC 7231	Yes	Yes	No	Yes	No
DELETE	RFC 7231	No	Yes	No	Yes	No
CONNECT	RFC 7231	Yes	Yes	No	No	No
OPTIONS	RFC 7231	Optional	Yes	Yes	Yes	No
TRACE	RFC 7231	No	Yes	Yes	Yes	No
PATCH	RFC 5789	Yes	Yes	No	No	Yes

Status Code

- 2XX 成功
- 3XX 跳转
- 4XX 客户端错误
- 500 服务器错误

300

- 300 Multiple Choices 存在多个可用的资源，可处理或丢弃
- 301 Moved Permanetly 重定向
- 302 Found 重定向
- 304 Not Modified 请求的资源未更新，丢弃

400 500

- 400 Bad Request 客户端请求有语法错误，不能被服务器所理解
- 401 Unauthorized 请求未经授权，这个状态代码必须和WWW-Authenticate报头域一起使用
- 403 Forbidden 服务器收到请求，但是拒绝提供服务
- 404 Not Found 请求资源不存在，eg：输入了错误的URL
- 500 Internal Server Error 服务器发生不可预期的错误
- 503 Server Unavailable 服务器当前不能处理客户端的请求，一段时间后可能恢复正常

错误处理

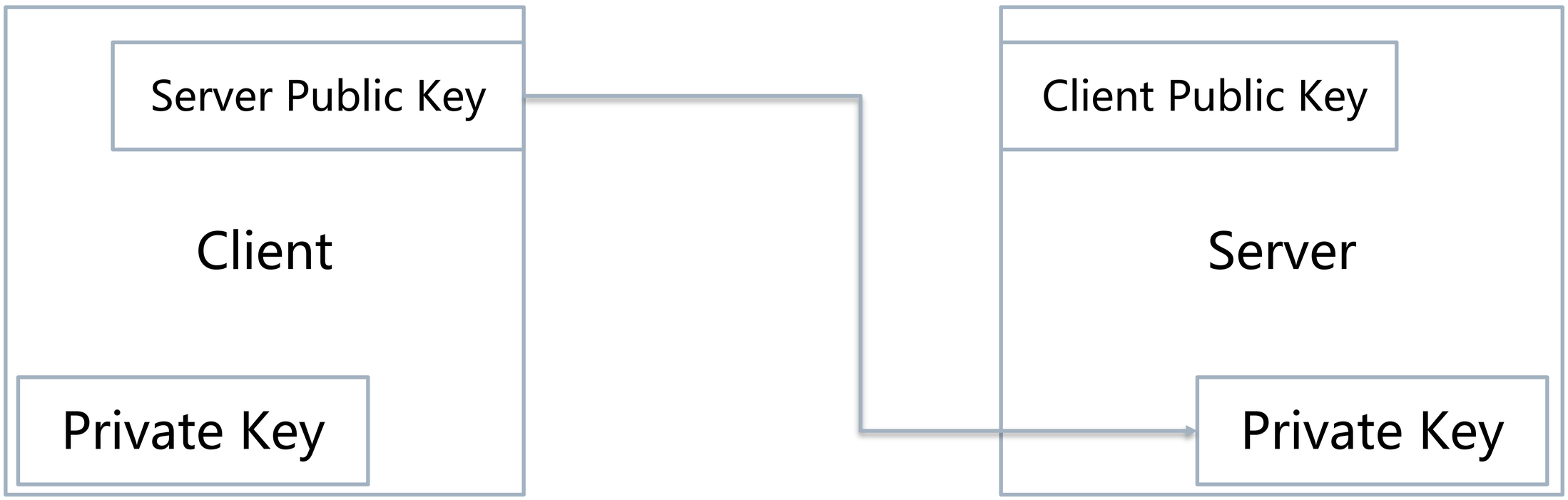
- 400 Bad Request 检查请求的参数或者路径
- 401 Unauthorized 如果需要授权的网页，尝试重新登录
- 403 Forbidden

    1. 如果是需要登录的网站，尝试重新登录
    2. IP被封，暂停爬取，并增加爬虫的等待时间，如果拨号网络，尝试重新联网更改IP

- 404 Not Found 直接丢弃
- 5XX 服务器错误，直接丢弃，并计数，如果连续不成功，WARNING 并停止爬取

HTTPS

HTTPS


Server Public Key
Client
Private Key
Client Public Key
Server
Private Key

小象学院

只为遇见明天更优秀的你！