Tedu.cn
达内教育

运维开发实战

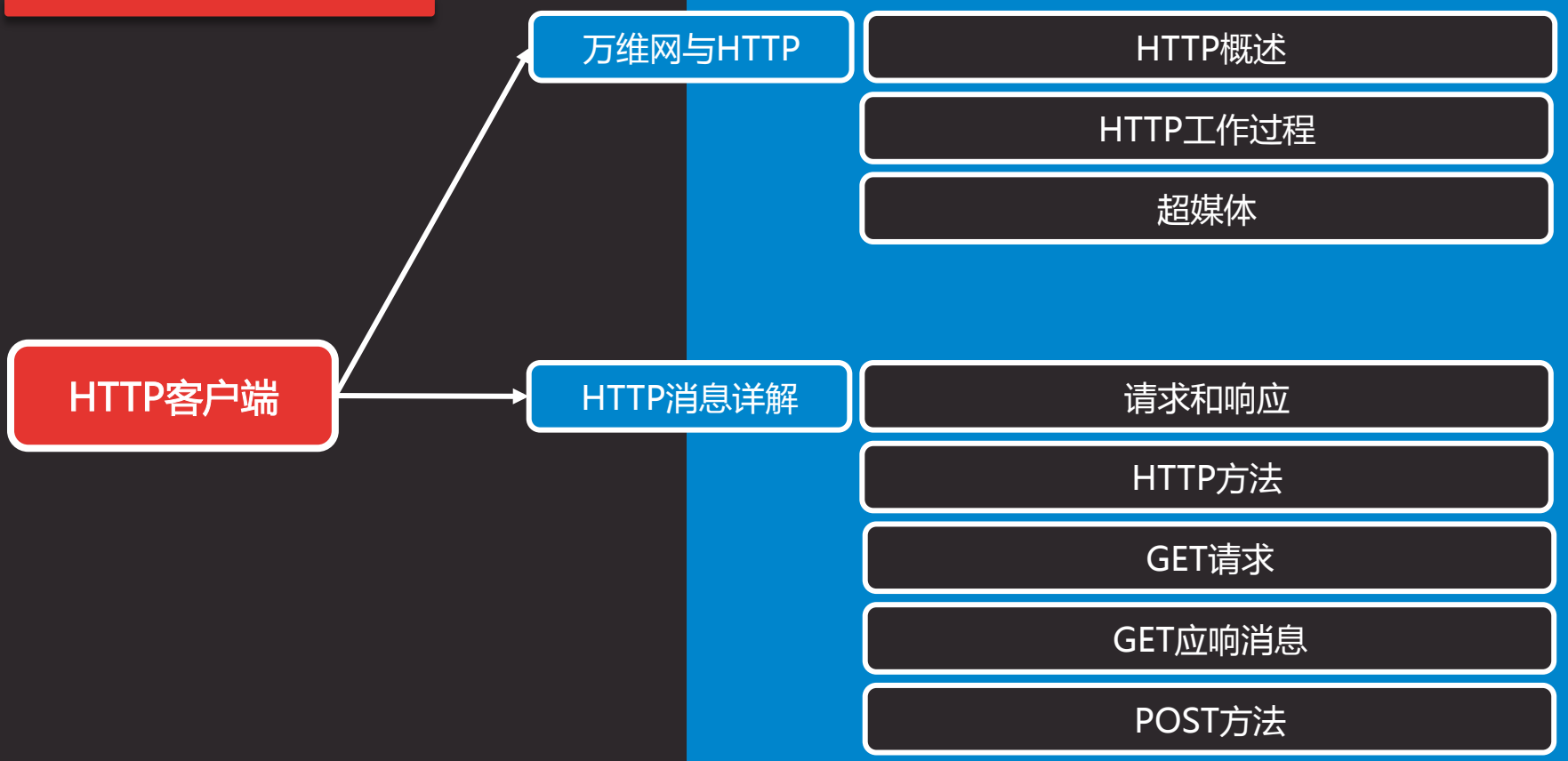
NSD DEVOPS

DAY03

内容

上午	09:00 ~ 09:30	作业讲解和回顾
	09:30 ~ 10:20	HTTP客户端
	10:30 ~ 11:20	
	11:30 ~ 12:00	urllib模块
下午	14:00 ~ 14:50	
	15:00 ~ 15:50	paramiko模块
	16:10 ~ 17:00	
	17:10 ~ 18:00	总结和答疑

HTTP客户端


HTTP客户端
万维网与HTTP
HTTP概述
HTTP工作过程
超媒体
HTTP消息详解
请求和响应
HTTP方法
GET请求
GET应响消息
POST方法

Tedu.cn
达内教育

万维网与HTTP

HTTP概述

知识讲解

- 超文本传输协议（HTTP，HyperText Transfer Protocol)是互联网上应用最为广泛的一种网络协议。所有的WWW文件都必须遵守这个标准。设计HTTP最初的目的是为了提供一种发布和接收HTML页面的方法。
- HTTP是一个客户端和服务器端请求和应答的标准。客户端是终端用户，服务器端是网站。通过使用Web浏览器、网络爬虫或者其它的工具，客户端发起一个到服务器上指定端口（默认端口为80）的HTTP请求。

HTTP工作过程

- 客户端发起一个请求，建立一个到服务器指定端口（默认是80端口）的连接
- 服务器则在那个端口监听客户端发送过来的请求。一旦收到请求，服务器（向客户端）发回一个状态行，比如"HTTP/1.1 200 OK"，和（响应的）消息
- 消息的消息体可能是请求的文件、错误消息、或者其它一些信息
- HTTP使用TCP协议

知识讲解

Tedu.cn
达内教育

HTTP消息详解

请求和响应

- 客户端向服务器发送获取文档的请求（request）
- 一旦发送完请求，客户端就会进行等待，直到从服务器接收到完整的响应（response）为止
- 当前最流行的HTTP 1.4版本的协议中，不允许客户端在尚未收到上一个请求前就在同一个套接字上开始发送第二个请求

知识讲解

请求和响应（续1）

```
GET /ip HTTP/1.1                          # 请求
User-Agent: curl/7.35.0
Host: localhost:8000
Accept: */*

HTTP/1.1 200 OK                           # 响应
Server: gunicorn/19.1.1
Date: Sat, 20 Sep 2014 00:18:00 GMT
Connection: close
Content-Type: application/json
Content-Length: 27
Access-Control-Allow-Origin: *
Access-Control-Allow-Credentials: true
```

知识讲解

请求和响应（续2）

知识讲解

- 第一行包含一个方法名和要请求的文档名；在响应消息中，第一行包含了返回码和描述消息。无论是在请求和响应消息中，第一行都以回车和换行（CR-LF）结尾
- 第二部分包含零或多个头信息，每个头信息由一个名称，一个冒号以及一个值组成。HTTP头的名称区分大小写。头信息之后要再跟上一个空行
- 第三部分是一个可选的消息体。消息体紧跟着头信息后面的空行

HTTP方法

- GET和POST这两种方法提供了HTTP基本的“读”和“写”操作
  - GET　请求获取Request-URI所标识的资源
  - POST　在Request-URI所标识的资源后附加新的数据
- 其余方法可以分为两大类：本质上类似于GET的方法和本质上类似于POST的方法

知识讲解

HTTP方法（续1）

- OPTIONS 请求与给定路径匹配的HTTP头的值
- HEAD 请求服务器做好一切发送资源的准备，但是只发送头信息
- DELETE 请求服务器删除Request-URI所标识的资源
- PUT 请求服务器存储一个资源，并用Request-URI作为其标识
- TRACE 请求服务器回送收到的请求信息，主要用于测试或诊断
- CONNECT 保留将来使用

知识讲解

GET请求

- 在浏览器的地址栏中输入网址的方式访问网页时，浏览器采用GET方法，现在默认使用的协议版本是1.1
- http : // www.tedu.cn /
- 协议部分 分隔符 目标域名 请求的资源

模拟访问：

```
telnet  www.tedu.cn  80

Trying 60.10.3.93...
Connected to www.tedu.cn.
Escape character is '^]'.
GET / HTTP/1.1
Host: www.tedu.cn
```

知识讲解

GET请求（续1）

- 常用的请求报头
  - METHOD 请求资源的方法，这个是必须的
  - Host 被请求资源的名子，这个是必须的
  - Accept 请求报头域用于指定客户端接受哪些类型的信息
  - Accept-Encoding 它是用于指定可接受的内容编码
  - User-Agent 客户端信息
  - Connection 是否关闭连接

知识讲解

GET应响消息

- HTTP/1.1 200　协议、版本和状态码
- Date　日期时间
- Server　服务器信息
- Content-Type　响应内容类型
- Content-Length　响应数据长度
- Last-Modified　资源最后更改时间
- Connection　连接方式

知识讲解

GET应响消息（续1）

知识讲解

HTTP/1.1 200 OK
Date: Thu, 18 Aug 2016 10:15:04 GMT
Server: tarena
Content-Type: text/html
Content-Length: 159232
Accept-Ranges: bytes
Age: 3921
Connection: keep-alive

<!DOCTYPE html PUBLIC 具体页面内容

POST方法

- 要求被请求服务器接受附在请求后面的数据
- 常用于提交表单
- 一般要在头部声明数据长度
- 信息头说明参见GET方法

知识讲解

案例1：熟悉HTTP工作流程

1. 为Firefox安装firebug插件
2. 打开Firefox的firebug或Chrome开发者工具
3. 访问http://www.tedu.cn
4. 在开发者工具的“网络”选项卡中查看请求和响应

课堂练习

urllib模块


urllib模块
urllib基础
urllib简介
爬取网页
下载网络资源
模拟客户端
urllib进阶
数据编码
HTTP异常处理

Tedu.cn
达内教育

urllib基础

urllib简介

- 在Python2版本中，有urllib和urlib2两个库可以用来实现request的发送。而在Python3中，已经不存在urllib2这个库了，统一为urllib
- urllib中包括了四个模块
  - urllib.request可以用来发送request和获取request的结果
  - urllib.error包含了urllib.request产生的异常
  - urllib.parse用来解析和处理URL
  - urllib.robotparse用来解析页面的robots.txt文件

知识讲解

爬取网页

知识讲解

- 先需要导入用到的模块：urllib.request
- 在导入了模块之后，我们需要使用urllib.request.urlopen打开并爬取一个网页
- 读取内容常见的有3种方式：
  - read()读取文件的全部内容，与readlines()不同的是，read()会把读取到的内容赋给一个字符串变量。
  - readlines()读取文件的全部内容，readlines()会把读取到的内容赋值给一个列表变量。
  - readline()读取文件的一行内容。

爬取网页（续1）

```
import urllib.request
html = urllib.request.urlopen('http://www.tedu.cn')
html.readline()
html.read(4096)
html.readlines()
```

知识讲解

案例2：爬取网页

1. 爬取的网页为http://www.tedu.cn
2. 保存的文件名为/tmp/tedu.html

课堂练习

下载网络资源

- urllib不仅可以下载网页，其他网络资源均可下载
- 有些文件比较大，需要像读取文件一样，每次读取一部分数据

```
import urllib.request
html = urllib.request.urlopen('http://172.40.50.116/python.pdf')
fobj = open('/tmp/python.pdf', 'ab')
while True:
    data = html.read(4096)
    if not data:
        break
    fobj.write(data)
fobj.close()
```

知识讲解

案例3：爬取图片

1. 将http://www.tedu.cn所有的图片下载到本地
2. 本地的目录为/tmp/images
3. 图片名与网站上图片名保持一致

课堂练习

模拟客户端

- 有些网页为了防止别人恶意采集其信息所以进行了一些反爬虫的设置，而我们又想进行爬取
- 可以设置一些Headers信息（User-Agent），模拟成浏览器去访问这些网站

```
import urllib.request

url='http://www.tedu.cn'
hearder={
   'User-Agent':'Mozilla/5.0 (X11; Fedora; Linux x86_64) AppleWebKit/
537.36 (KHTML, like Gecko) Chrome/58.0.3029.110 Safari/537.36'
}
html=urllib.request.Request(url,headers=header)
data=urllib.request.urlopen(request).read()
```

知识讲解

Tedu.cn
达内教育

urllib进阶

数据编码

知识讲解

- 一般来说，URL标准中只会允许一部分ASCII字符，比如数字、字母、部分符号等
- 而其他的一些字符，比如汉字等，>是不符合URL标准的。此时，我们需要编码。
- 如果要进行编码，可以使用urllib.request.quote()进行

```
>>> urllib.request.quote('hello world!')
'hello%20world%21'
>>> urllib.request.unquote('hello%20world%21')
'hello world!'
```

HTTP异常处理

- 如果访问的页面不存在或拒绝访问，程序将抛出异常
- 捕获异常需要导入urllib.error模块

知识讲解

```
>>> html = urllib.request.urlopen('http://172.40.50.116/a.html')
urllib.error.HTTPError: HTTP Error 404: Not Found
>>> html = urllib.request.urlopen('http://172.40.50.116/aaa')
urllib.error.HTTPError: HTTP Error 403: Forbidden
```

案例4：处理下载错误

1. 起动一个web服务
2. 在web服务器的文档目录下创建目录ban，权限设置为700
3. 编写python程序访问不存在的路径和ban目录，处理404和403错误
4. 404错误打印“无此页面”，403错误打印“无权访问”

课堂练习

paramiko模块

paramiko

安装paramiko模块

- 本地安装

```
# yum  install  -y  gcc  gcc-c++  python-devel
# tar  xzf  paramiko-1.15.4.tar.gz
# python  setup.py  install
```

- 网络安装

```
# pip install paramiko
```

知识讲解

基础使用介绍

- SSHClient
  - 创建用于连接ssh服务器的实例

```
>>> ssh = paramiko.SSHClient()
```

- paramiko.AutoAddPolicy
  - 设置自动添加主机密钥
- ssh.connect
  - 连接ssh服务器
- ssh.exec_comand
  - 在ssh服务器上执行指定命令

知识讲解

paramiko实例

- 编写用于实现ssh访问的脚本
  - 创建SSHClient实例
  - 设置添加主机密钥策略
  - 连接ssh服务器
  - 执行指定命令
  - 在shell命令行中接受用于连接远程服务器的密码以及在远程主机上执行的命令

知识讲解

案例5：利用多线程实现ssh并发访问

- 编写脚本程序
  1. 在文件中取出所有远程主机IP地址
  2. 在shell命令行中接受远程服务器IP地址文件、远程服务器密码以及在远程主机上执行的命令
  3. 通过多线程实现在所有的远程服务器上并发执行命令

课堂练习

Tedu.cn
达内教育

总结和答疑