Python HTTP

mainly talk about Requests

by laike9m
laike9m@gmail.com
https://github.com/laike9m

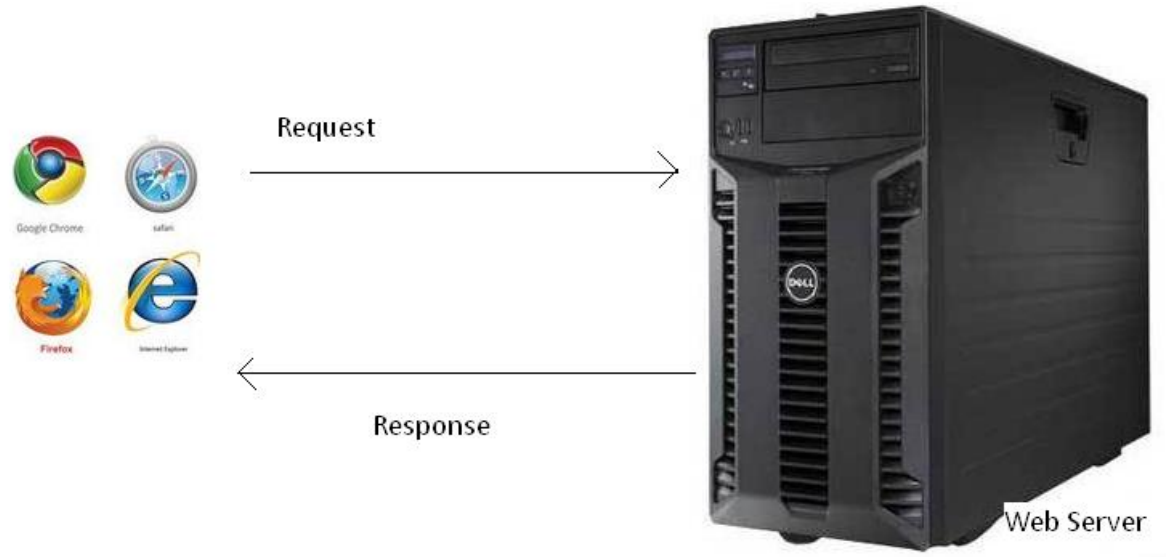
Topics

- HTTP 基础知识
- Requests 库介绍
- 实战
  - ChinaUnicom模拟登陆
  - RenRen模拟登陆，抓取自己的日志
- More...

What is HTTP

- HTTP = Hyper Text Transfer Protocol
- 协议是指计算机通信网络中两台计算机之间进行通信所必须共同遵守的规定或规则，超文本传输协议(HTTP)是一种通信协议，它允许将超文本标记语言(HTML)文档从Web服务器传送到客户端的浏览器
- 目前我们使用的是HTTP/1.1 版本

Web Server, Browser

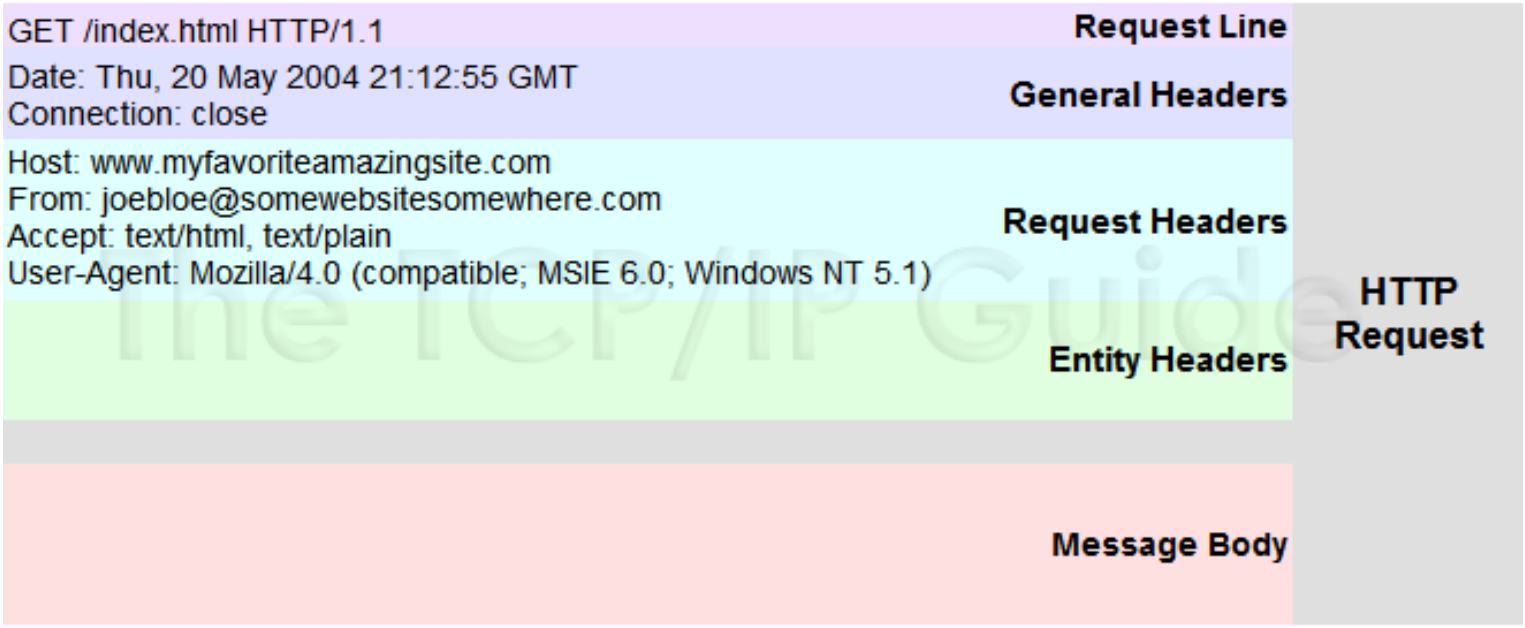


当我们打开浏览器，在地址栏中输入URL，然后我们就看到了网页。原理是怎样的呢？
实际上我们输入URL后，我们的浏览器给Web服务器发送了一个Request, Web服务器接到Request后进行处理，生成相应的Response，然后发送给浏览器， 浏览器解析Response中的HTML,这样我们就看到了网页，过程如图所示

open a webpage

- 打开一个网页需要浏览器发送很多次Request
  - 当你在浏览器输入URL http://www.cnblogs.com 的时候，浏览器发送一个Request去获取 http://www.cnblogs.com 的html. 服务器把Response发送回给浏览器.
  - 浏览器分析Response中的 HTML，发现其中引用了很多其他文件，比如图片，CSS文件，JS文件。
  - 浏览器会自动再次发送Request去获取图片，CSS文件，或者Js文件。
  - 等所有的文件都下载成功后。 网页就被显示出来了。

HTTP Request



HTTP Request = Request Line + Headers + Body

HTTP Methods

- GET - 不向服务器发送数据，Body是空的
- POST - 向服务器发送数据，包含在Body中
- PUT
- DELETE
- OPTIONS
- HEAD

GET vs POST

- GET提交的数据会放在URL之后，以?分割URL和传输数据，参数之间以&相连，如EditPosts.aspx?name=test1&id=123456. POST方法是把提交的数据放在HTTP包的Body中.
- GET提交的数据大小有限制（因为浏览器对URL的长度有限制），而POST方法提交的数据没有限制.
- GET方式提交数据，会带来安全问题，比如一个登录页面，通过GET方式提交数据时，用户名和密码将出现在URL上，如果页面可以被缓存或者其他人可以访问这台机器，就可以从历史记录获得该用户的账号和密码.
- GET一般用于获取/查询资源信息，而POST一般用于更新资源信息.

HTTP Response

- 和 Request 区别不大，但是一般会返回数据
- 示例

```
********************response line**********************
HTTP/1.1 200 OK

********************response header********************
Content-Type: text/xml; charset=UTF-8

********************response body**********************
<?xml version="1.0" encoding="utf-8"?>
<string xmlns="http://www.codecademy.com/">Accepted
</string>
```

演示

- F12
- www.baidu.com
- 观察request和response

What is Requests

- Requests是一个Python第三方库
- 最好的Python HTTP library
- http://docs.python-requests.org/en/latest/index.html

Requests tutorial

- http://cn.python-requests.org/en/latest/user/quickstart.html
  - 完整讲解

- http://cn.python-requests.org/en/latest/user/advanced.html
  - Session Object讲解

Get images

```
import shutil   # shell utilities
import requests
import os

url = ‘http://www.baidu.com/img/bdlogo.gif’
r = requests.get(url, stream=True)
os.chdir(‘Desktop’)
with open(‘baidu icon’, ‘wb’) as f:
        shutil.copyfileobj(r.raw, f)
```

ChinaUnicom模拟登陆

- 关于前期步骤，参见

  ChinaUnicom模拟登陆.pdf

代码讲解：

https://github.com/laike9m/CU_login/tree/master/src

人人日志导出

- 代码
- https://github.com/laike9m/DumpRenrenPosts2Markdown/blob/master/renren_get_posts.py

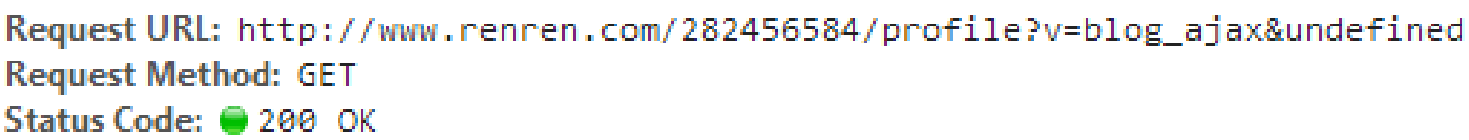
step 1 - 登陆


www.renren.com/282456584
人人网
首页
个人主页
好友
应用
左遥
VIP 1
310天
发状态
传照片
发分
你正在干嘛？
今日话题：在临沂大学作弊的各种成本，看完还敢作
我常用的
新鲜事
新鲜事
特别关注 new
关注内容

step2 - 访问profile页面(个人主页)

step 3 – 进入日志tab

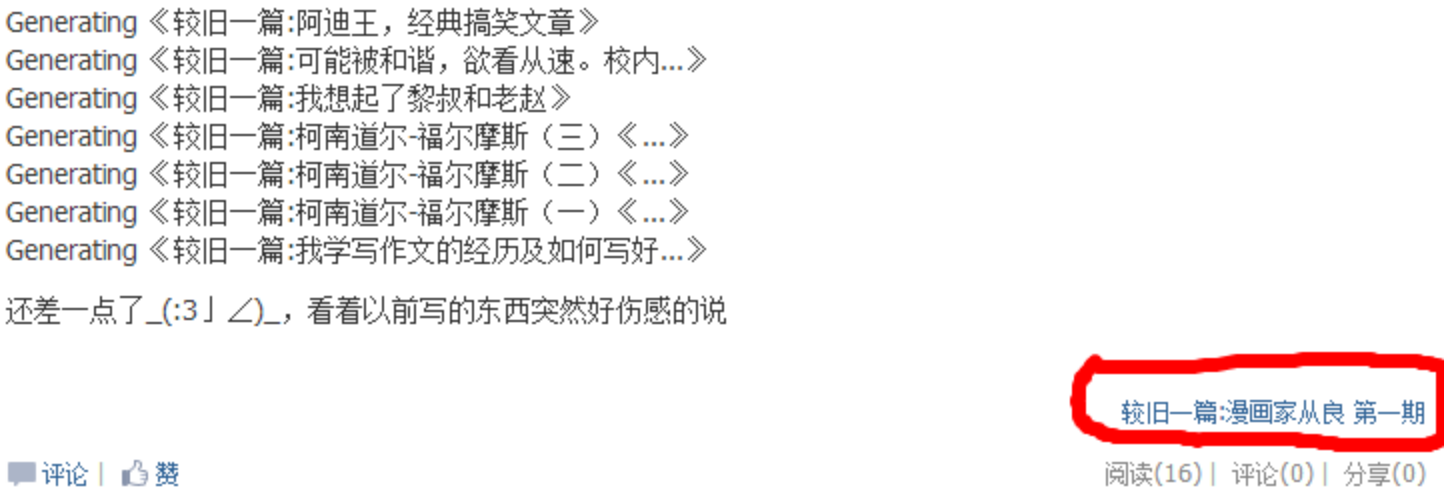
- 通过调试工具查看在个人主页点击“日志”tab时，浏览器发出了什么请求



至此，我们进入了日志页面

step 4？

- 请大家自己对照代码研究
- 需要一点css知识和lxml这个库
- 本质上，就是不停地访问红圈部分



More

- 这一讲的内容是比较精确的HTTP请求，如何获取特定的数据，和完整抓站不一样。
- 抓取大量数据，需要多线程
- 遇到动态内容，需要根据Js代码追溯其来源
- Cookie往往也要追溯来源
- 网站或许会拒绝访问

Q&A