第3讲 内容解析与提取

慧科集团 李皓

主要内容

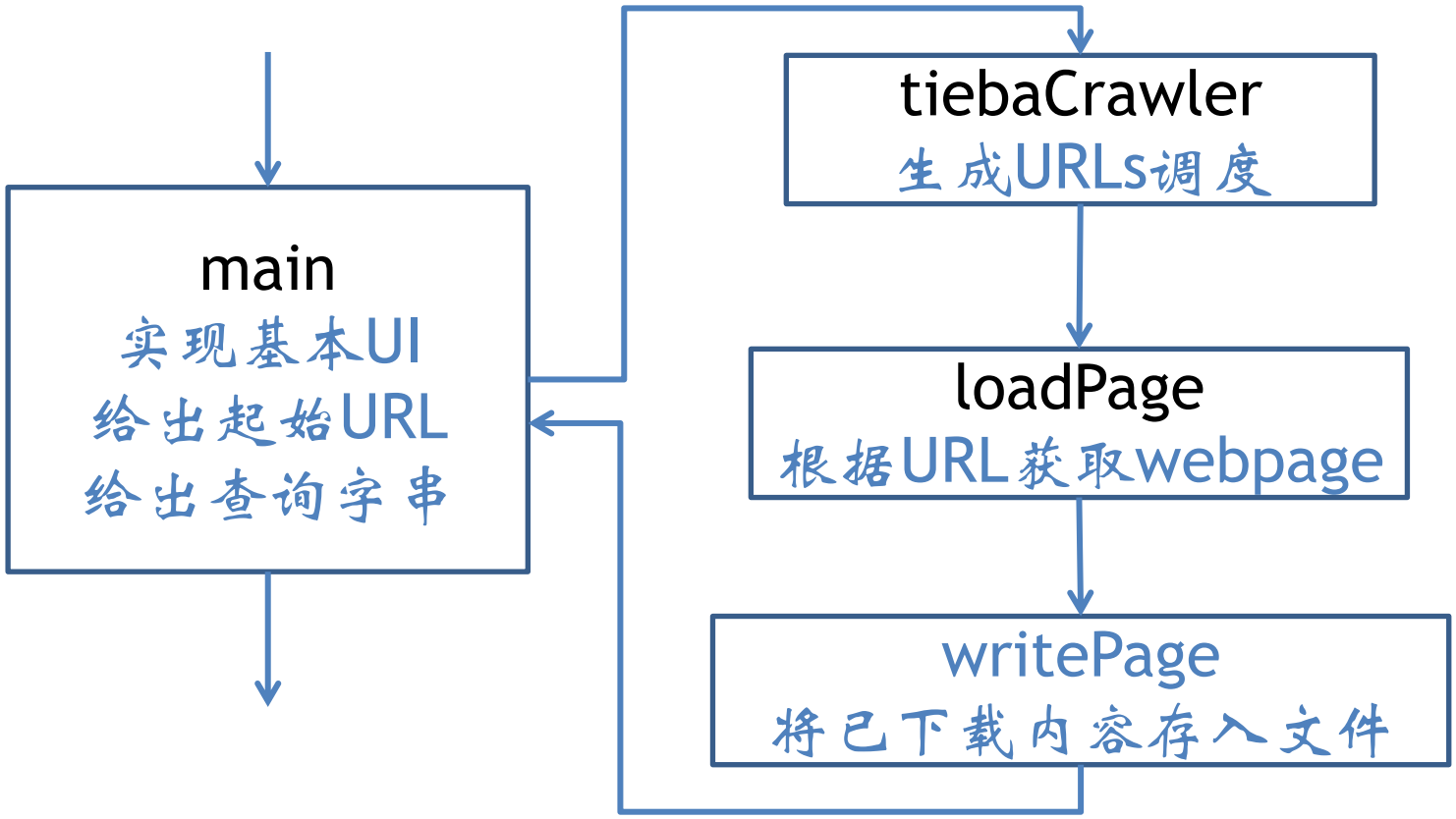
- 上周作业讲评
- 网页获取进阶
  - 如何通过HTTP认证?
  - 如何防止服务器封爬虫IP?
  - 如何获取ajax异步加载的页面?
  - 如何利用cookie访问登录后页面?
- 更简洁的Requests库
  - 基本方法
  - 响应对象
  - 爬取框架

1

Review

从百度贴吧下载多页话题内容

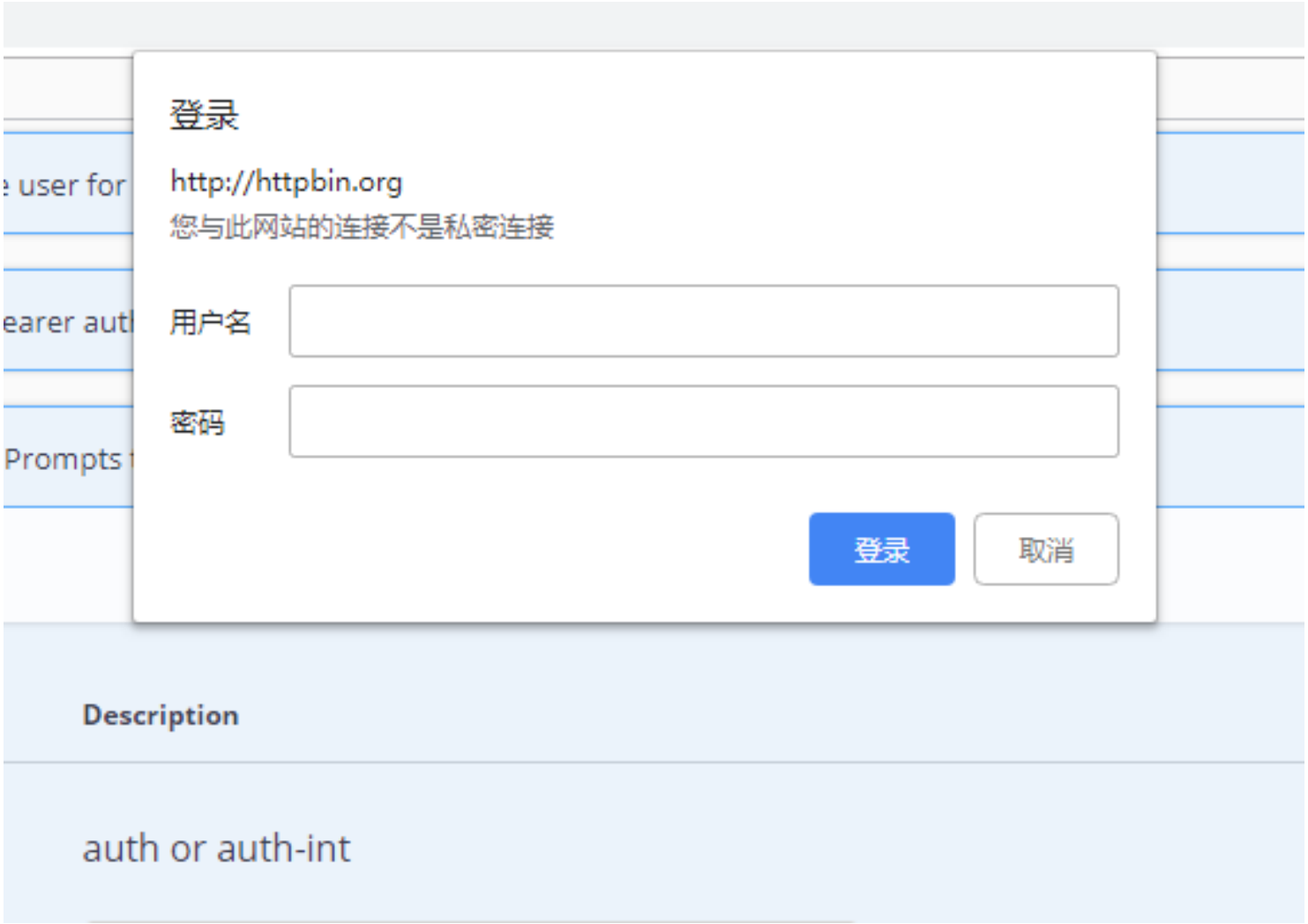
- http://tieba.baidu.com/f?



2

网页获取进阶

如何通过HTTP认证?


登录
http://httpbin.org
您与此网站的连接不是私密连接
用户名
密码
登录
取消
Description
auth or auth-int

老办法的问题

- Urllib.request.urlopen()
  - 调用了HTTPHandler来处理无错误的HTTP请求.
  - 功能有限，如何扩展?

```
        https_handler = HTTPSHandler(context=context)
        opener = build_opener(https_handler)
    elif context:
        https_handler = HTTPSHandler(context=context)
        opener = build_opener(https_handler)
    elif _opener is None:
        _opener = opener = build_opener()
    else:
        opener = _opener
    return opener.open(url, data, timeout)
```

扩展处理能力的方法

- 自定义Opener对象，调用特定Handler
  - HTTPHandler
  - HTTPSHandler
  - FileHandler
  - DataHandler
  - CacheFTPHandler
  - HTTPRedirectHandler
  - HTTPCookieProcessor
  - ProxyHandler
  - HTTPBasicAuthHandler
  - HTTPDigestAuthHandler
  - ProxyBasicAuthHandler
  - ProxyDigestAuthHandler
  - UnknownHandler

实例：自动实现简单用户认证

- HTTP规范中定义了两种认证模式
  - Basic Auth
  - Digest auth
- Basic Auth认证的基本过程是：
  1. 客户请求访问网页；
  2. 服务器端返回401错误，要求认证（401消息的头里面带了挑战信息，例如：认证头： WWW-Authenticate: Basic realm="zhouhh@mydomain.com"）；
  3. 客户端重新提交请求并附以认证信息，这部分信息将被编码；
  4. 服务器检查信息，通过则给以正常服务页面；否则返回401错误。

实例：自动实现简单用户认证

- HTTP Digest Auth
  - 使用MD5散列算法处理密钥。

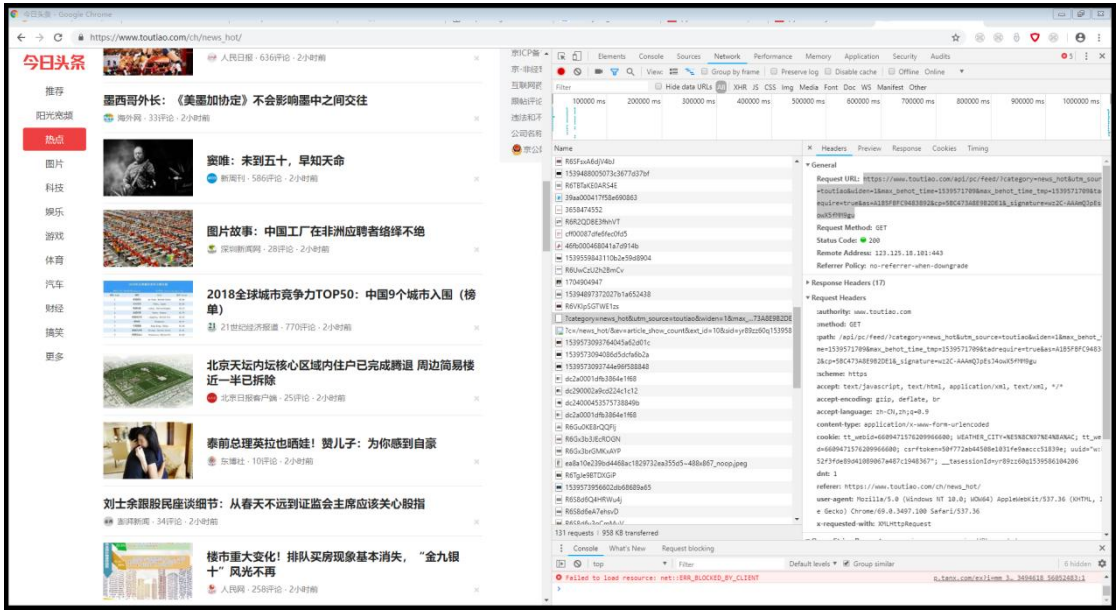
- 基本过程：
  1. 客户发送请求；
  2. 收到一个401消息，包含一个Challenge和一个唯一nonce（每次不一样）；
  3. 客户将用户名密码和401消息返回的挑战一起MD5加密后传给服务器；
  4. 服务器检查是否合法。

如何防止服务器封爬虫IP

- 很多server会检测外来IP的访问频次
- 使用proxy
  - 免费proxy
  - 收费proxy
- 使用特定Handler利用proxy访问服务器

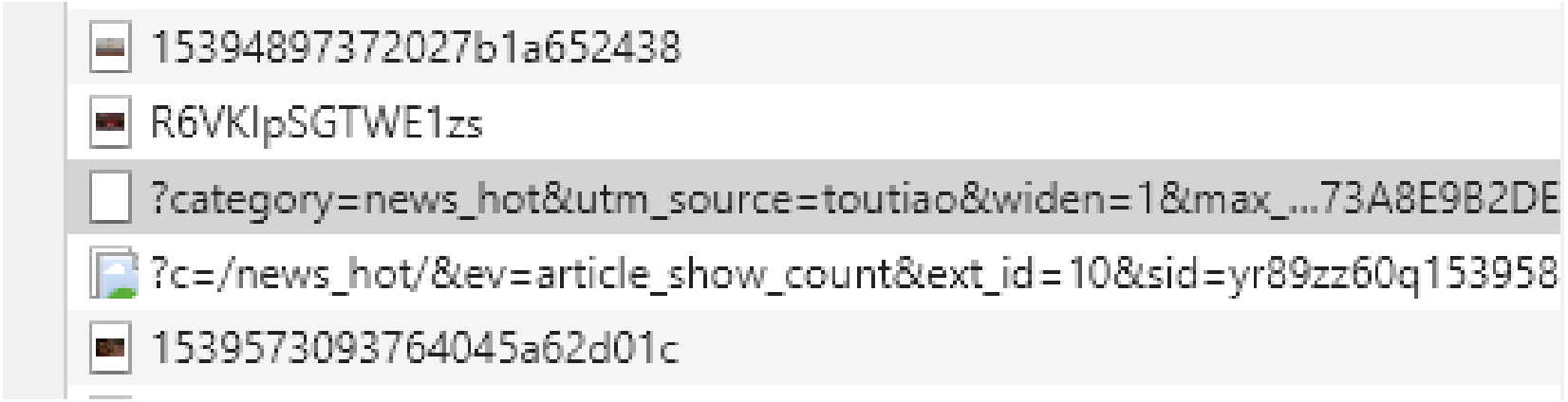
如何获取ajax异步加载的页面?

- AJAX
  - AJAX = Asynchronous JavaScript and XML
  - 不重新加载整个页面，就可以与服务器交换数据并更新部分网页内容
  - 不需要任何浏览器插件，但需要用户允许JavaScript在浏览器上执行。



如何获取ajax异步加载的页面?

- 手动找到关键URL



3

更简洁的Requests库

Requests库

- 安装试用
- 简单页面获取
  - 比urllib中的类型丰富
- 更易用的响应对象
- 异常处理
  - 构建自己的爬取框架
- 向server提交数据
- 二进制文件下载
- Json响应内容的解析
- 重定向历史的查看