

辅助工具

主讲：杨真

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
- 辅助工具

Part 2 爬虫

- 网站结构及常见爬虫策略
- BBS网站结构分析及方案
- 控制节奏

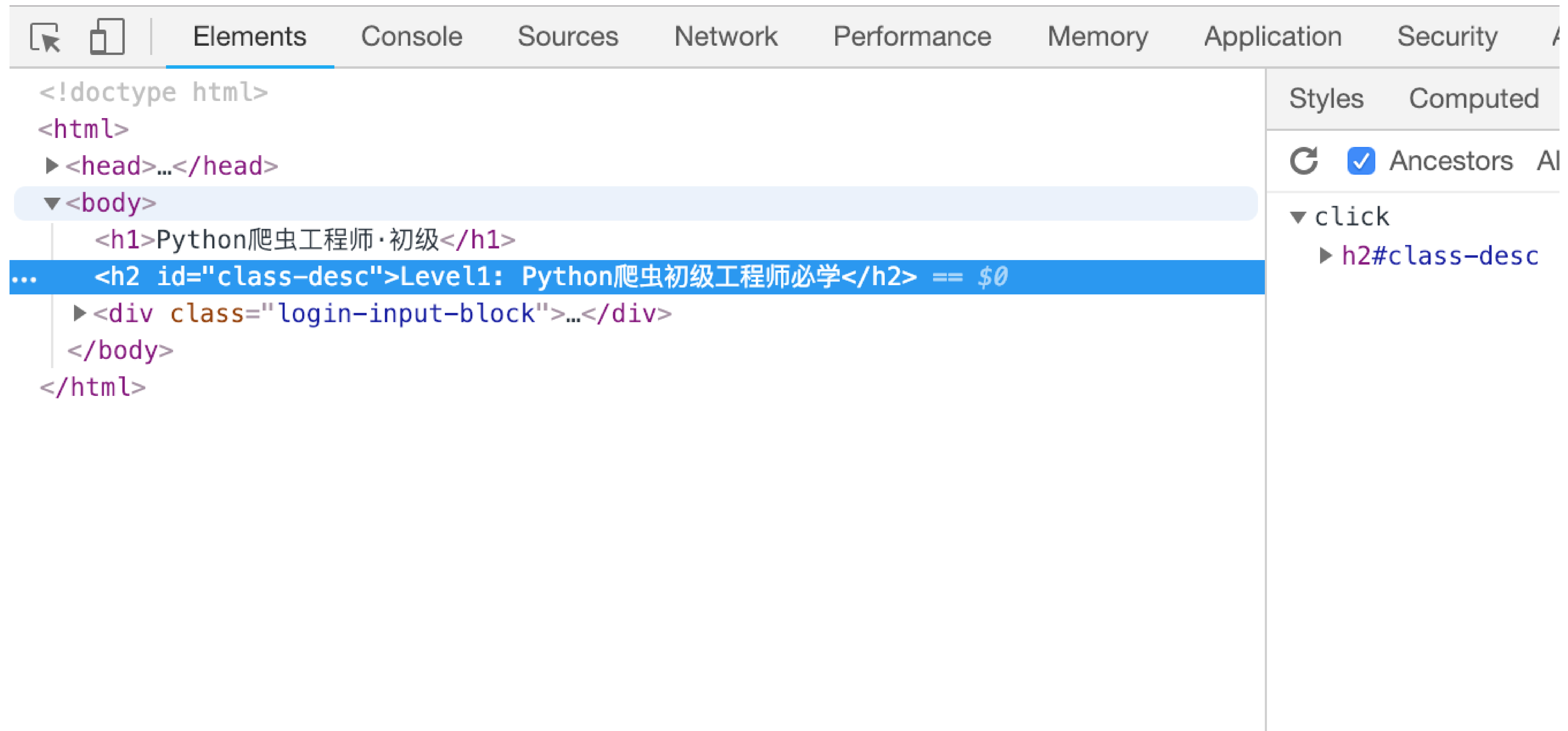
Part 3 进阶

- 多线程
- 并行抓取
- MySQL 数据库
- 网站服务架构
- 表单、登录及Cookie处理

Part 4 实战

- 网站结构分析
- 网页抓取方案
- 数据提取
- 存储方案

Inspector - Elements





The screenshot displays the Chrome DevTools 'Elements' panel. The DOM tree on the left shows the following structure:

- `<!doctype html>`
- `<html>`
- `<head>...</head>`
- `<body>`
 - `<h1>Python爬虫工程师·初级</h1>`
 - `<h2 id="class-desc">Level1: Python爬虫初级工程师必学</h2> == $0` (Selected)
 - `<div class="login-input-block">...</div>`

The right sidebar shows the 'click' event listener for the selected element, with the following details:

- Event: `click`
- Target: `h2#class-desc`

Inspector - Console



Elements


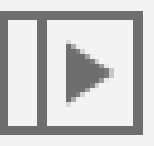
Console

Sources

Network


Performance

Memory



top

▼



Filter

De

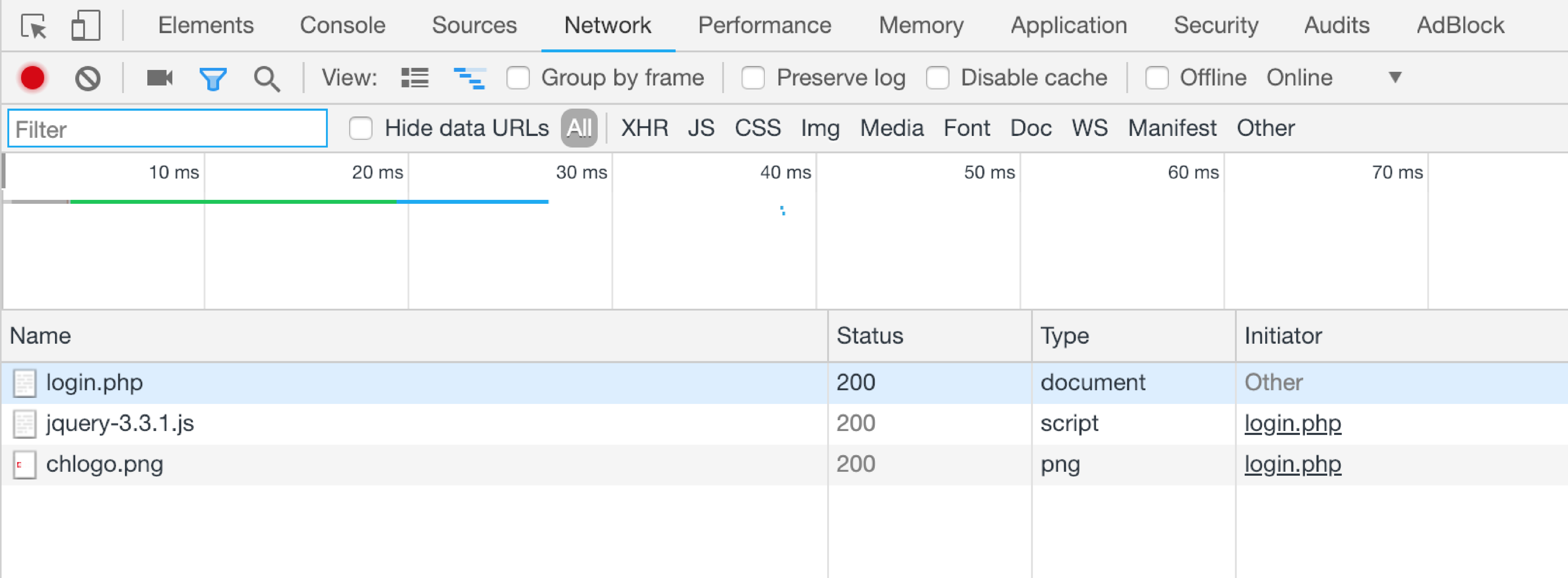
-----init----- <http://localhost/login/login.php>

> document.getElementById('class-desc')

< <h2 id="class-desc">Level1: Python爬虫初级工程师必学</h2>

> |

Inspector – Network



Inspector – Headers

× Headers Preview Response Timing

▼ General

Request URL: http://localhost/login/login.php

Request Method: GET

Status Code: ● 200 OK

Remote Address: [::1]:80

Referrer Policy: no-referrer-when-downgrade

▼ Response Headers view source

Connection: Keep-Alive

Content-Length: 1086

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

Date: Sun, 30 Dec 2018 14:32:51 GMT

Keep-Alive: timeout=5, max=100

Server: Apache/2.4.34 (Unix) PHP/7.1.19

X-Powered-By: PHP/7.1.19

▼ Request Headers view source

Accept: text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,image/apng,*/*;q=0.8

Accept-Encoding: gzip, deflate, br

Accept-Language: zh-CN,zh;q=0.9,en-US;q=0.8,en;q=0.7

Cache-Control: max-age=0

Connection: keep-alive

Host: localhost

Upgrade-Insecure-Requests: 1

User-Agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_2) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/70.0.3538.102 Safari/537.36

Inspector – Preview

× Headers Preview Response Timing

Python爬虫工程师·初级

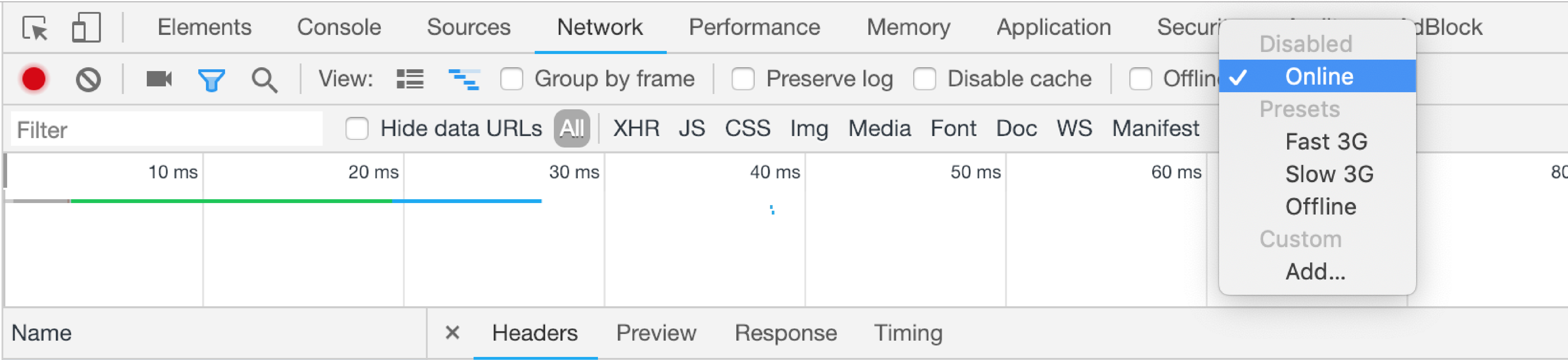
Level1: Python爬虫初级工程师必学

Inspector – Response

```

x  Headers  Preview  Response  Timing
1  <!DOCTYPE html>
2  <html>
3
4  <head>
5    <link rel="stylesheet" type="text/css" href="style.css" />
6
7    <script src="jquery-3.3.1.js"></script>
8
9    <script>
10      $(document).ready(function(){
11        $("#class-desc").click(function(){
12          $("#class-desc").hide();
13        });
14      });
15    </script>
16  </head>
17
18  <body>
19
20  <h1>Python爬虫工程师·初级</h1>
21  <h2 id="class-desc">Level1: Python爬虫初级工程师必学</h2>
22
23
24    <div class='login-input-block'>
25
26      
27
28      <form action="login.php" method="POST">
29        <div class='input-box'>
30          <!--<div style="width: 100px;display: inline-block;">Name:</div> -->
31          <input class='input-field' type="text" name="name" placeholder="User Name">
32        </div>
33        <div class='input-box'>
34          <!-- <div style="width: 100px; display: inline-block;">Password</div> -->
35          <input class='input-field' type="password" name="password" placeholder="Password">
36        </div>
37        <div class="ed-login">
38          <input type="submit" value="Login">
39        </div>
40      </form>
41    </div>
42
43  </body>
44
45  </html>
```


Inspector – Offline



Inspector – Source

🔍 📄

Elements

Console

Sources

Network

Performance

Page

Filesystem

>>

⋮

🔍

login.php ×

▼ 📁 top

▼ ☁ localhost

▼ 📁 login

📄 login.php

📄 jquery-3.3.1.js

📄 style.css

📄 chlogo.png

1 <!DOCTYPE html>

2 <html>

3

4 <head>

5 <link rel="stylesheet" type="text/css" href="style.css">

6

7 <script src="jquery-3.3.1.js">

8

9 <script>

10 \$(document).ready(function(){

11 \$("#class-desc").click(function(){

12 \$("#class-desc").text("Hello World!");

13 });

14 });

15 </script>

16 </head>

17

18 <body>

19

Inspector – Cookie

ElementsConsoleSourcesNetworkPerformanceMemoryApplication

>>

Application

Manifest

Service Workers

Clear storage

Storage

Local Storage

https://m.weibo.cn

Session Storage

https://m.weibo.cn

IndexedDB

Web SQL

Cookies

https://m.weibo.cn

Cache

Cache Storage

Application Cache

Frames

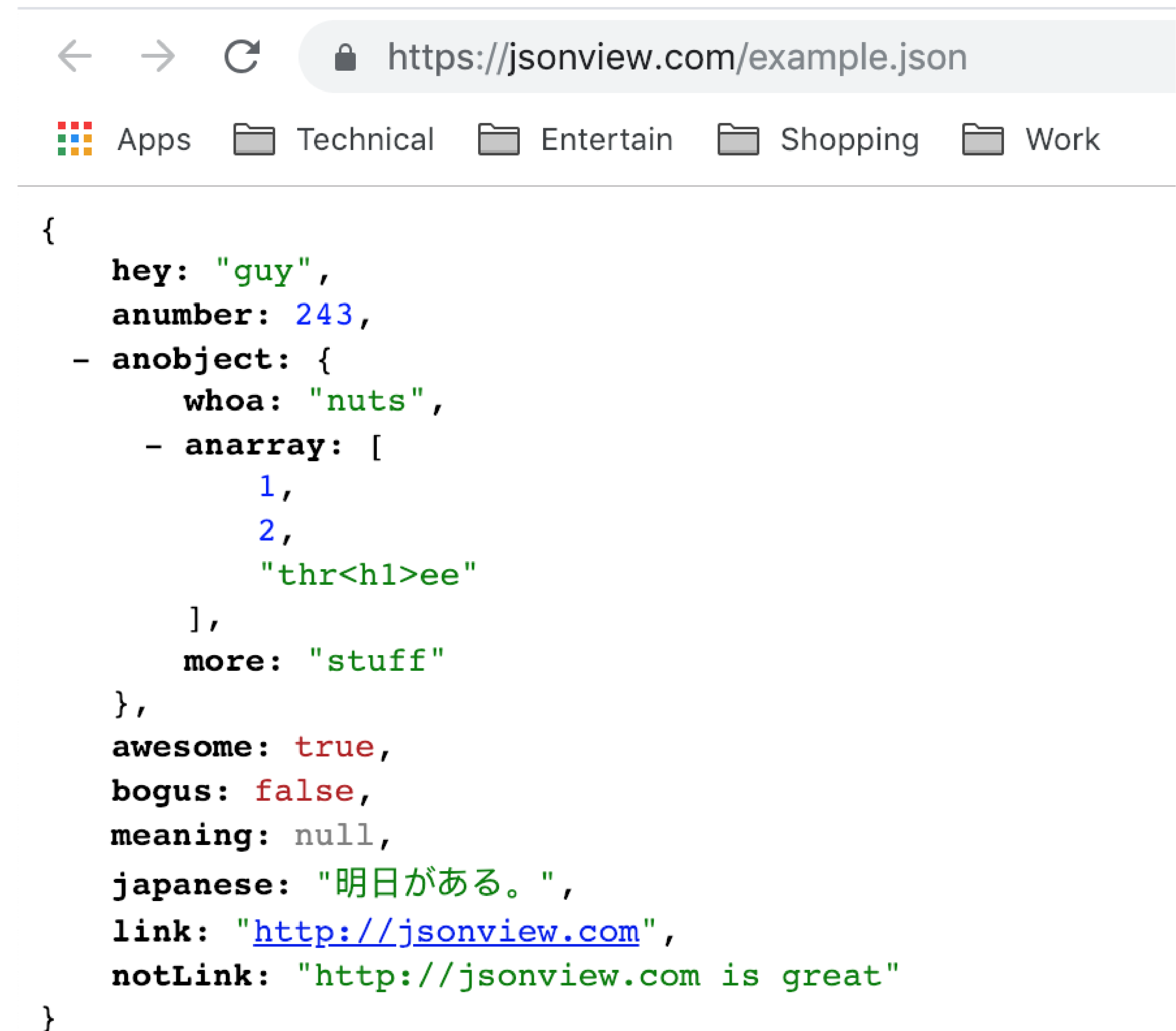
top

↻ⓧ✕Filter

Name	Value	Domain	Path	Ex...	Si
ALF	1547631724	.weibo....	/	20...	
MLOGIN	0	.weibo....	/	20...	
M_WEIBOCN_PARAMS	luicode%3D10000...	.weibo....	/	20...	
WEIBOCN_FROM	1110006030	.weibo....	/	19...	
_T_WM	e58e0fa0a021c98...	.weibo....	/	20...	

JSONView

```
{hey: "guy",anumber: 243,anobject:
{whoa: "nuts",anarray:
[1,2,"thr<h1>ee"],more:
"stuff"},awesome: true,bogus:
false,meaning: null,japanese: "明日が
ある。",link:
"http://jsonview.com",notLink:
"http://jsonview.com is great"}
```



试试看: <https://m.weibo.cn/api/container/getIndex?containerid=102803&openApp=0>

JSONView

安装方法：

1. Chrome Store 在线安装

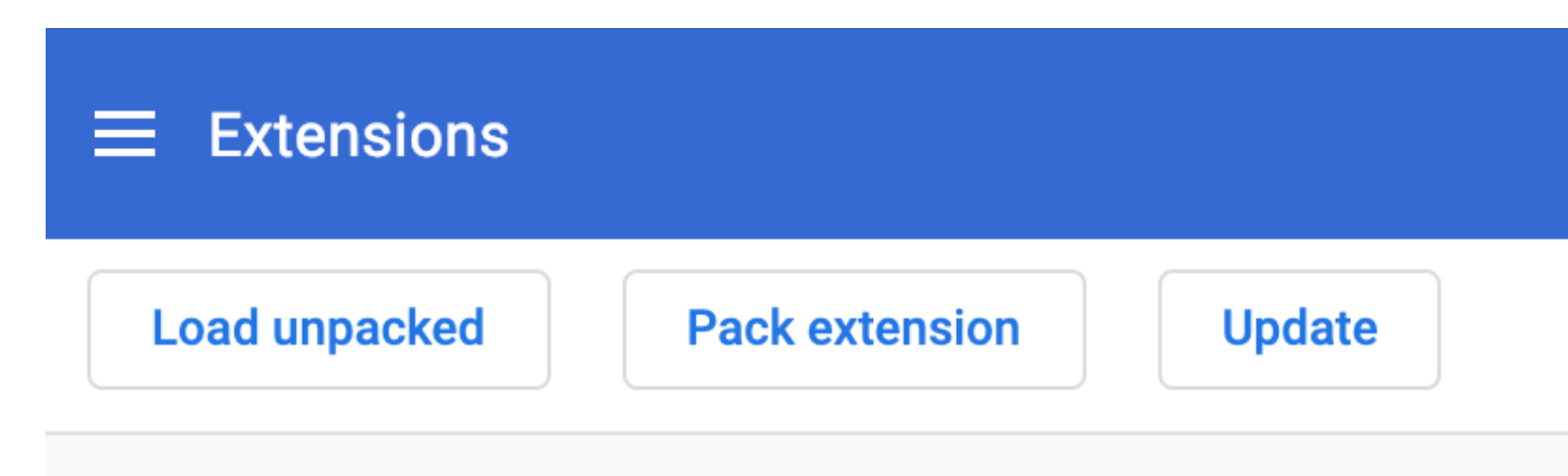
2. 离线安装：

2.1 下载压缩包并解压

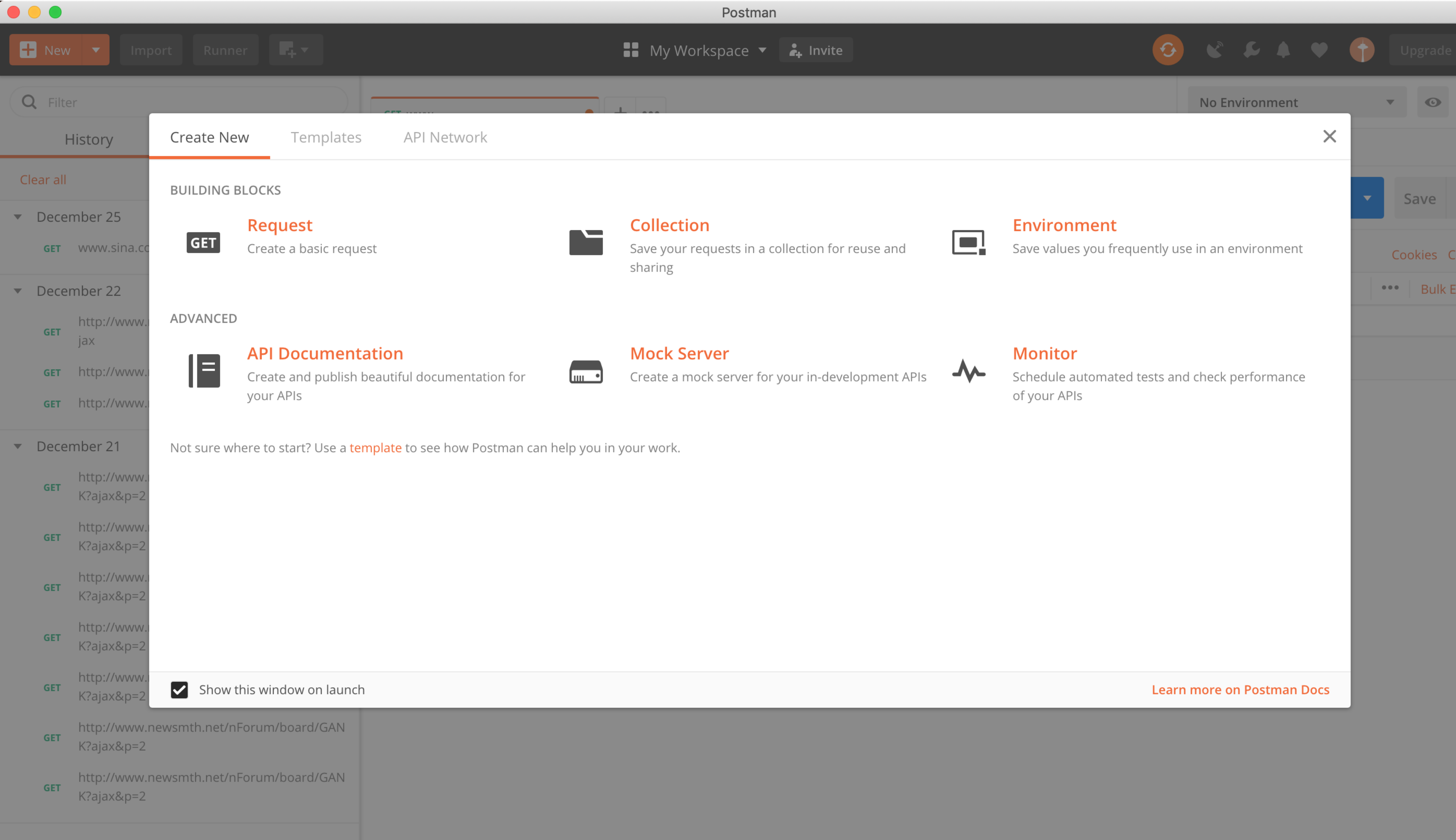
<https://github.com/gildas-lormeau/JSONView-for-Chrome>

2.2 在谷歌浏览器地址栏中输入：chrome://extensions/

2.3 加载已解压的扩展程序



POSTMAN



POSTMAN

离线安装插件：

POSTMAN:

<https://github.com/postmanlabs/postman-app-support/releases>

POSTMAN Interceptor:

<https://github.com/postmanlabs/postman-chrome-interceptor/releases>

POSTMAN

- 模拟网络请求
- 捕获 Chrome 的所有网络流量
- 生成 Python Header 及 Requests 代码

只为遇见明天更优秀的你！