Data Science & AI for Economists

Lecture 9: Web Scraping and Crawling(II)

Zhaopeng Qu
Business School, Nanjing University
November 20 2025

Roadmap

Today's Agenda

知识目标

- 掌握浏览器开发者工具的使用
- 学会分析网络请求和响应

技能目标

- 能够独立分析任意网站的API接口
- 提取关键的请求参数和响应数据
- 为编写爬虫做好技术准备

实战案例

- **百度指数** - 完整的API逆向工程实例

浏览器开发者工具界面介绍

如何打开开发者工具?

方法1：键盘快捷键（推荐）

- Windows: `F12` 或 `Ctrl + Shift + I` (Edge)
- Mac: `Cmd + Option + I` (Chrome)

方法2：右键菜单

- 网页任意位置右键
- 选择"检查"或"检查元素"

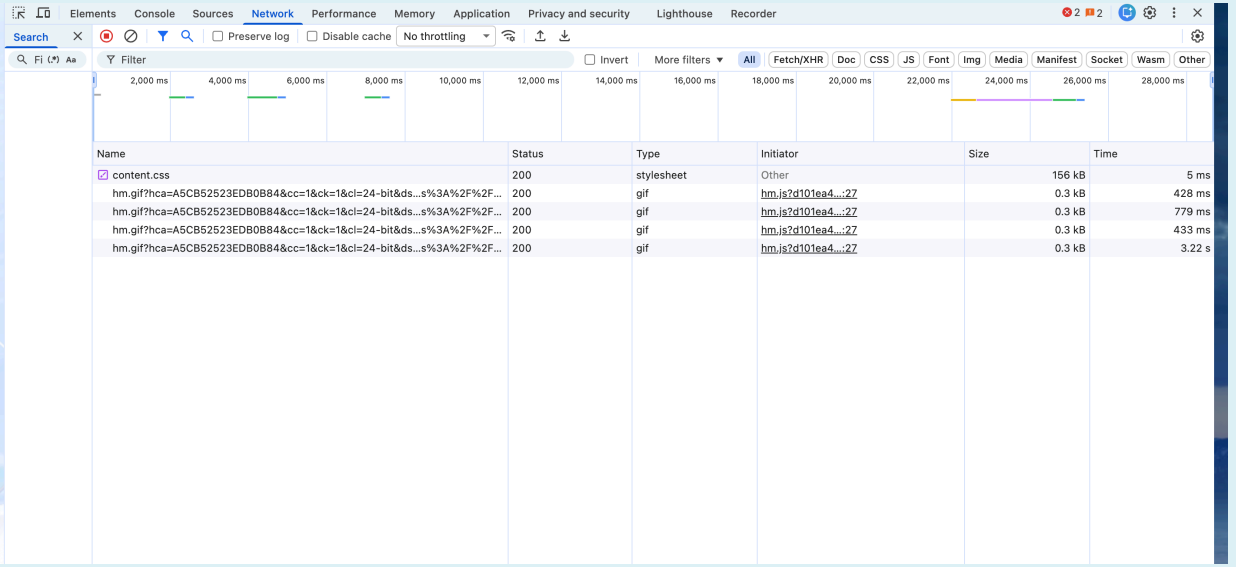
方法3：浏览器菜单

- Chrome: 更多工具 → 开发者工具
- Firefox: 更多工具 → Web开发者工具

Developer Tools Panels(Panel Overview)

1. Elements/元素 - 查看HTML结构
2. Console/控制台 - JavaScript调试
3. **Network/网络** ← 我们的重点！

1. Sources/源代码 - 查看源文件
2. Application/应用 - 存储、Cookie等
3. Performance/性能 - 性能分析
4. Memory/内存 - 内存分析



Network面板详解

- **Network面板**记录了浏览器与服务器之间的所有网络通信
  - 每一个请求（Request）
  - 每一个响应（Response）
  - 详细的时间线和性能数据

Network面板详解

- Network面板的核心组成

1. 工具栏

- 🔴 录制按钮（开始/停止记录）
- 🚫 清除按钮（清空记录）
- 🔍 过滤器（筛选请求类型）
- 🔎 搜索框（查找特定请求）

2. 请求列表

显示所有网络请求：

- Name（请求的URL）
- Status（状态码：200、404等）
- Type（类型：XHR、JS、CSS等）
- Size（数据大小）
- Time（耗时）

Network面板详解（续）

3. 请求详情面板:点击任一请求，右侧显示详细信息：

Headers（请求头）

- General（概要信息）
  - Request URL（请求地址）
  - Request Method（GET/POST等）
  - Status Code（状态码）
- Request Headers（请求头）
  - Cookie（身份凭证）
  - User-Agent（浏览器标识）
  - Referer（来源页面）
- Response Headers（响应头）
  - Content-Type（内容类型）

Preview/Response（响应内容）

- Preview：格式化显示（JSON、HTML等）
- Response：原始响应内容

Timing（时间线）

- 请求各阶段的耗时

Cookies

- 请求和响应的Cookie

Lab: Scraping a Dynamic Website: Baidu Index