



Day 31

如何克服反制爬蟲的網站

反爬：瀏覽器標頭與基本資訊



出題教練：張維元



python

本日知識點目標

- 了解「檢查 HTTP 標頭檔」的反爬蟲機制
- 「檢查 HTTP 標頭檔」反爬蟲的因應策略

常見的反爬蟲機制有哪些？

檢查 HTTP
標頭檔

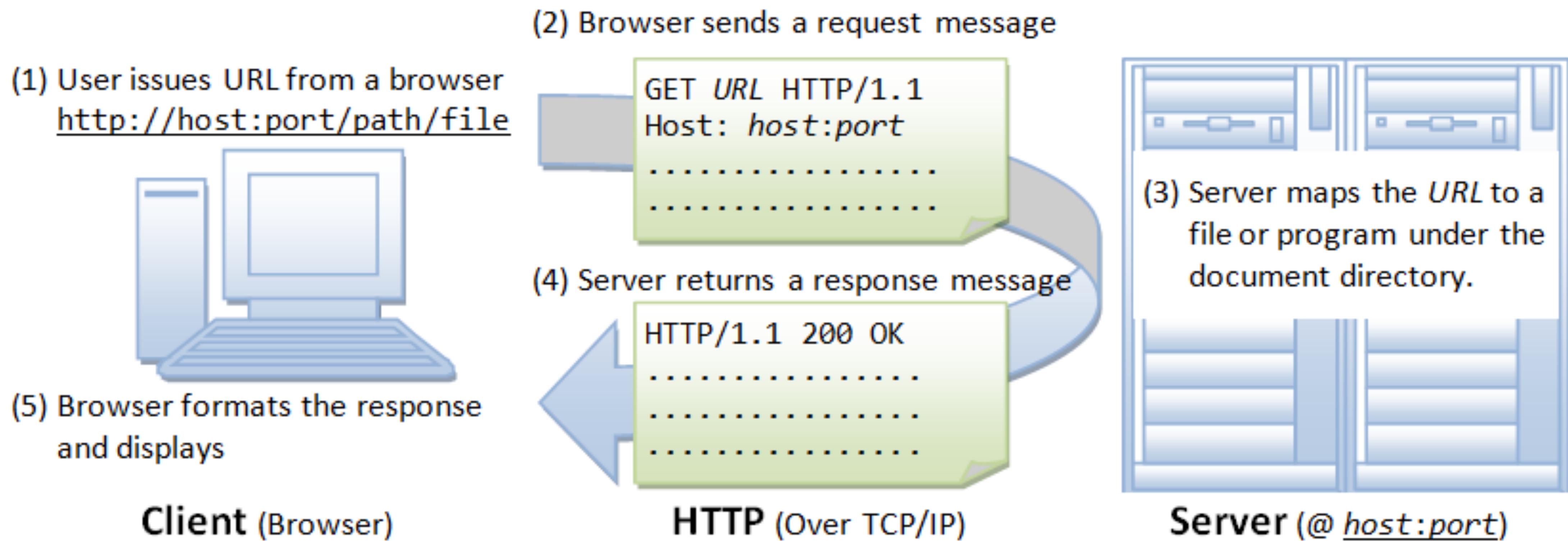
驗證碼機制

登入權限機制

IP 黑/白名單

檢查 HTTP 的發送請求方是否合法

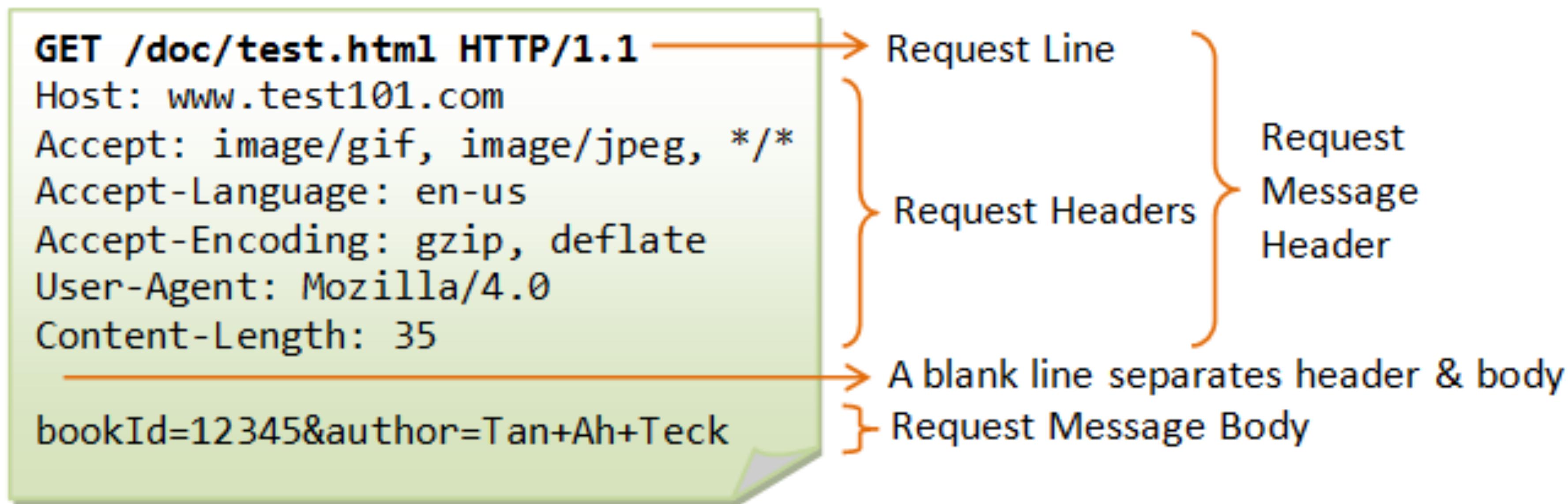
前面我們在提到網頁的傳輸有講到 HTTP 協定，HTTP 會將網路的傳輸分為「Request」和「Response」兩種角色。



Request 中的 Header

其中 Request 又可以分為幾個部分：

- Header：瀏覽器自動產生，包含跟發送方有關的資訊。
- Body：網頁服務真正要傳送的資料



Header 包含發送方的資訊

一般來說，Header 可能會包含：

- 發送方的位址 (Host)
- 發送方的瀏覽器版本 (User-Agent)
- 發送方的語言/格式

... 等等



讓爬蟲程式也加上 Header

因為 Header 是由瀏覽器自動產生，因此如果透過程式發出的請求預設是沒有 Header 的。透過檢查 Header 是最基本的反爬機制。

解法：在爬蟲程式的 Request 加上 Header !

如果沒有加上 Headers

```
1 import requests  
2 requests.get('https://www.zhihu.com/api/v4/questions/55493026/answers')  
3 response = r.text
```

```
'<!DOCTYPE html>\n<html style="height:100%;width:100%">\n    <head>\n        <meta http-equiv="Content-Type"\ncontent="text/html; charset=utf-8" />\n        <meta http-equiv="Server" content="HuaweiCloudWAF" />\n    <title id="title">\n        è®¿é\x97®è¢«æ\x8b|æ\x88¤í\x81\n    </title>\n    <script>\n        function bindall() {var requestid = "30-30-14-20190729225616-d49c4c8c";if(requestid==null |\n        | requestid == ""){return;}document.getElementById("a").innerText = requestid;}</script></head><body onload="bindall\n()" style="height:100%;width:100%;margin:0px;font-family:Microsoft yahei"><div></div><div style="min-height:13.125re\nm;height:17%;width:100%"><div style="margin-top:3.75rem;margin-left:18.4375rem;height:52.38;width:83.04%;"><svg xmlns\n="http://www.w3.org/2000/svg" style="width: 3.125rem;height: 3.125rem" viewBox="0 0 50 50"><path fill="#e84e4c" d="M2\n5,0A25,25,0,1,0,50,25,25,0,0,0,25,0Zm1.6,37.16H22.85V33.41H26.6Zm0-6.63H22.85L22.35,13H27.1Z"/></svg><font style\n="font-family:MicrosoftYaHei;font-size:4.375rem;color:#e94d4c;margin-left: 0.75rem;font-weight: bold;">418</font></di\nv><div style="margin-left:18.4375rem;height:47.62%;width:83.04%;"><font style="font-family:MicrosoftYaHei;font-size:\n1.875rem;color:#999999;word-wrap:break-word;">æ\x82“ç\x9a\x84è”·æ†\x82ç\x96\x91ä\x94»å\x87»è\x8cä,ºí\x81</font><\n    p style="font-family:MicrosoftYaHei;font-size:0.9rem;color:#999999;word-wrap:break-word;"><span>äº\x8bä»¶IDí\x9a</sp\nan><span style="color:#499df2" id="a">False alarm ID</p></div></div><div style="height:15.625rem;width:100%;min-widt\nh:105rem;"><div style="margin-left:18.4375rem;float:left;width:50rem"><p style="margin-top:10px">å\x82æ\x9cæ\x82\n“æ\x98“ç“\x95¤í\x8cæ\x82“å\x8f“ä»¥å\x89\x8då\x80WAFæ\x8e§å\x88¶å\x8f°è\x9bè\x8cè““æ\x8a¥å\x8fè\x94½è@¾ç½®í\x8cè@æ\x82“ç\x9a\x84è@é\x97®ä,\x8då\x86\x8dè¢«æ\x8b|æ\x88¤</p></div></div></body>\n'
```



如果沒有加上 Headers

```
1 import requests  
2 requests.get('https://www.zhihu.com/api/v4/questions/55493026/answers')  
3 response = r.text
```

從資料的數量上，好像比預期中的少。內容也看起來怪怪的！

```
'<!DOCTYPE html>\n<html style="height:100%;width:100%">\n    <head>\n        <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />\n        <meta http-equiv="Server" content="HuaweiCloudWAF" />\n    <title id="title">\n        è®¿é\x97®è¢«æ\x8b|æ\x88¤í\x81\n    </title>\n<script>\n    function bindall() {var requestid = "30-30-14-20190729225616-d49c4c8c";if(requestid==null || requestid == ""){return;}document.getElementById("a").innerText = requestid;}</script></head><body onload="bindall()" style="height:100%;width:100%;margin:0px;font-family:Microsoft yahei"><div></div><div style="min-height:13.125rem; height:17%;width:100%"><div style="margin-top:3.75rem; margin-left:18.4375rem; height:52.38; width:83.04%;"><svg xmlns="http://www.w3.org/2000/svg" style="width: 3.125rem; height: 3.125rem" viewBox="0 0 50 50"><path fill="#e84e4c" d="M25,0A25,25,0,1,0,25,0,0,25,0,0,1,0,25,0m0-6.63H25.85L27.55,27.5H27.5V27.5L25.85,25.85H25v-6.63ZM0,0z" style="font-size:4.375rem; color:#e94d4c; margin-left: 0.75rem; font-weight: bold;">418</font></div><div style="margin-left:18.4375rem; height:47.62%; width:83.04%;"><font style="font-family:MicrosoftYaHei; font-size:1.875rem; color:#999999; word-wrap:break-word;">æ\x82“ç\x9a\x84è·æ±\x82ç\x96\x91ä¼å\x94»å\x87»è¡\x8cä,ºí\x81</font><p style="font-family:MicrosoftYaHei; font-size:0.9rem; color:#999999; word-wrap:break-word;"><span>äº\x8ba»¶IDí\x9a</span><span style="color:#499df2" id="a">False alarm ID</p></div></div><div style="height:15.625rem; width:100%; min-width:105rem;"><div style="margin-left:18.4375rem; float:left; width:50rem"><p style="margin-top:10px">å\x82æ\x9e\x9cæ\x82“æ\x98“ç«\x99é\x95¿í\x8cæ\x82“å\x8f“ä»¥å\x89\x8då\x80WAFæ\x8e§å\x88¶å\x8f°è\x9bè\x8cè--æ\x8a¥å±\x8fè\x94½è®¾ç½®í\x8cè®©æ\x82“ç\x9a\x84è®¿é\x97®ä\x8då\x86\x8dè¢«æ\x8b|æ\x88¤</p></div></div></body>\n'
```

從資料的數量上，好像比預期中的少。內容也看起來怪怪的！

在 Request 上加上 Headers

```
1 import requests  
2 headers = {'user-agent': 'my-app/0.0.1'}  
3 r = requests.get('https://www.zhihu.com/api/v4/questions/55493026/  
answers',headers=headers)  
response = r.text
```

```
' {"data": [{"id": 683070334, "type": "answer", "answer_type": "normal", "question": {"type": "question", "id": 55493026, "titl  
e": "你们都是怎么学 Python 的?", "question_type": "normal", "created": 1486390229, "updated_time": 1543075931, "url": "https://w  
ww.zhihu.com/api/v4/questions/55493026", "relationship": {}, "author": {"id": "36f69162230003d316d0b8a6d8da20ba", "url_t  
oken": "liang-zi-wei-48", "name": "量子位", "avatar_url": "https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_is.jp  
g", "avatar_url_template": "https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_{size}.jpg", "is_org": true, "typ  
e": "people", "url": "https://www.zhihu.com/api/v4/people/36f69162230003d316d0b8a6d8da20ba", "user_type": "organizatio  
n", "headline": "有趣的前沿科技→_→ 公众号：QbitAI", "badge": [{"type": "identity", "description": "已认证的官方帐号", "topics":  
[]}], "gender": -1, "is_advertiser": false, "is_privacy": false}, "url": "https://www.zhihu.com/api/v4/answers/683070334", "is  
_collapsed": false, "created_time": 1557824412, "updated_time": 1557824412, "extras": "", "is_copyable": true, "relationship":  
{"upvoted_followees": []}, {"id": 163642949, "type": "answer", "answer_type": "normal", "question": {"type": "question", "id": 5  
5493026, "title": "你们都是怎么学 Python 的?", "question_type": "normal", "created": 1486390229, "updated_time": 1543075931, "ur  
l": "https://www.zhihu.com/api/v4/questions/55493026", "relationship": {}, "author": {"id": "788f207a6bf8f66c5bad79bd0f011  
065", "url_token": "simonlearn", "name": "赛门喵Simon", "avatar_url": "https://pic2.zhimg.com/v2-03afe381dbea789c0f918d6aac1  
5556c_is.jpg", "avatar_url_template": "https://pic2.zhimg.com/v2-03afe381dbea789c0f918d6aac15556c_{size}.jpg", "is_org":  
false, "type": "people", "url": "https://www.zhihu.com/api/v4/people/788f207a6bf8f66c5bad79bd0f011065"}, "user_type": "peopl
```

怎麼檢查 Request 要帶哪些 Header？

知乎 首页 发现 等你来答 暴雪嘉年华 提问

最新专题



1. 右鍵點選檢查

双十一大促，有哪些电子产品值得入手？
2 天前更新 · 1,110,653 浏览

Redux DevTools
檢視網頁原始碼
檢查
語言

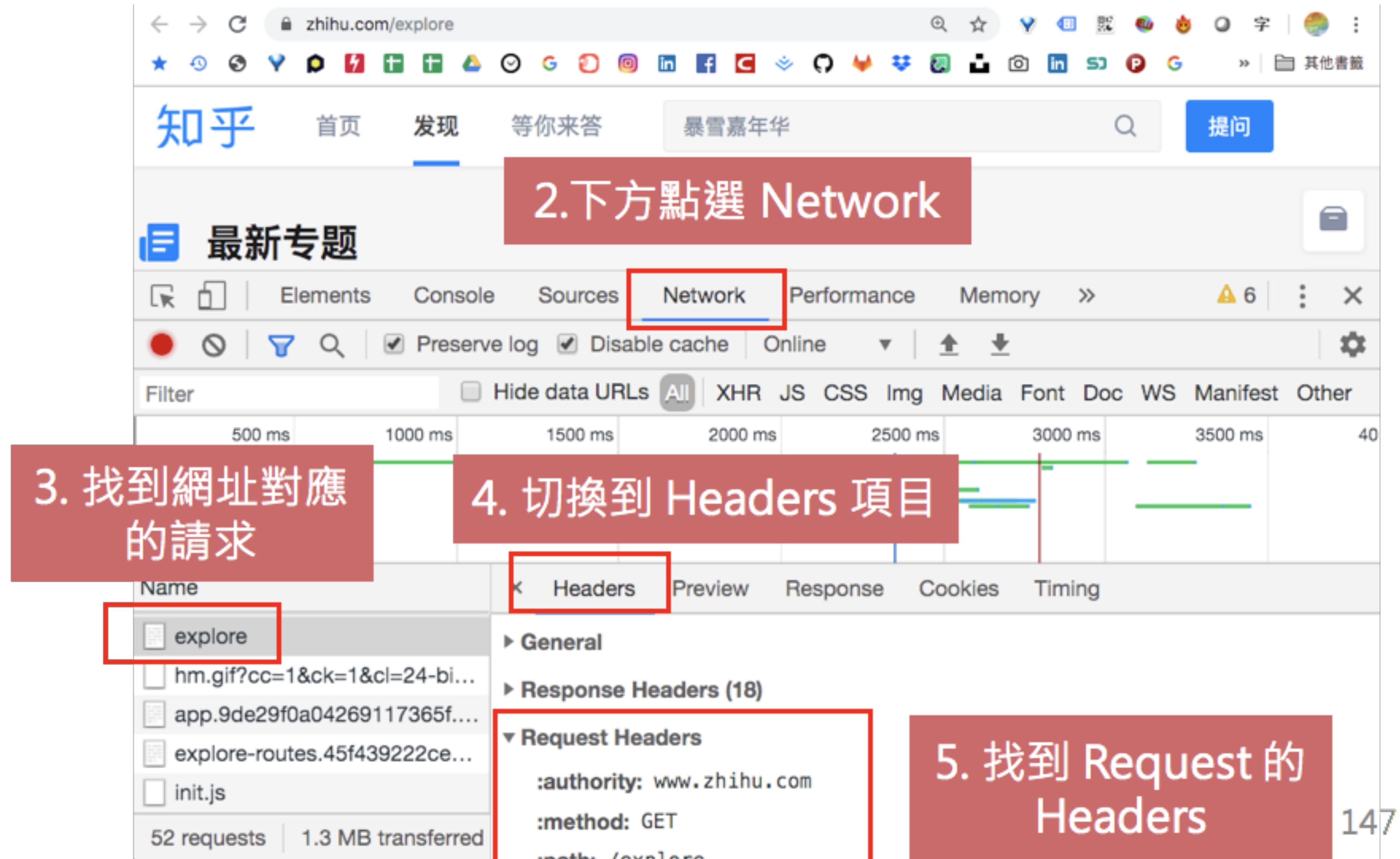
随身要带的 2019 年真无线蓝牙耳机有哪些推荐?
工作学习用得上 有哪些实用的平板电脑推荐?
好玩易上手 2019 年双十一，有哪些适合学生党的微单/单反?

会说话的博物馆



天津博物馆 | 甲骨文知多少 甲骨文是如何研究的
博物馆里知多少 博物馆存在的意义和价值
文物鉴赏知多少 如何分辨书法大家

怎麼檢查 Request 要帶哪些 Header？





在 Request 上加上 Headers

實際上的 Headers 應該參考瀏覽器的。但範例為了方便，我們這邊是先自己定義一的比較基本的。但不是每一個網站都可以通過，比較保險的方式建議模仿瀏覽器所帶出的標頭且整理成 dict 的型態（如下）。

```
1 headers = {  
2     'accept': '...',  
3     'accept-encoding': '...',  
4     'accept-language': '...',  
5     ...  
6     'user-agent': '...'  
7 }
```



在 Request 上加上 Headers

為什麼我們的範例中，只加上 user-agent 就可以了呢？原因是 user-agent 通常是最基本的篩選條件，他代表的是發出請求的瀏覽器版本，所以我們可以大膽假設用這個欄位。但實際上會檢查哪些欄位是由對方的伺服器決定，比較保險的方式還是把所有欄位補上去。

重要知識點複習

- 了解「檢查 HTTP 標頭檔」的反爬蟲機制
- 「檢查 HTTP 標頭檔」反爬蟲的因應策略



解題時間

LET'S CRACK IT

請跳出 PDF 至官網 Sample Code & 作業

開始解題

START

