D5：修改爬蟲程式中的 Headers 以成功存取第三方網站

API 資料串接 - 以 知乎 API 實作範例

本日知識點目標

- 了解知乎 API 使用方式與回傳內容
- 撰寫程式存取 API 且添加標頭

以知乎 API 作為範例

知乎是一家創立於2011年1月26日的中國大陸社會化問答網站，產品形態與美國同類網站Quora類似。「知乎」在文言文中意為「知道嗎」。我們這個例子會利用其提供的 API 接口進行練習。

思考流程與使用套件

Request Library

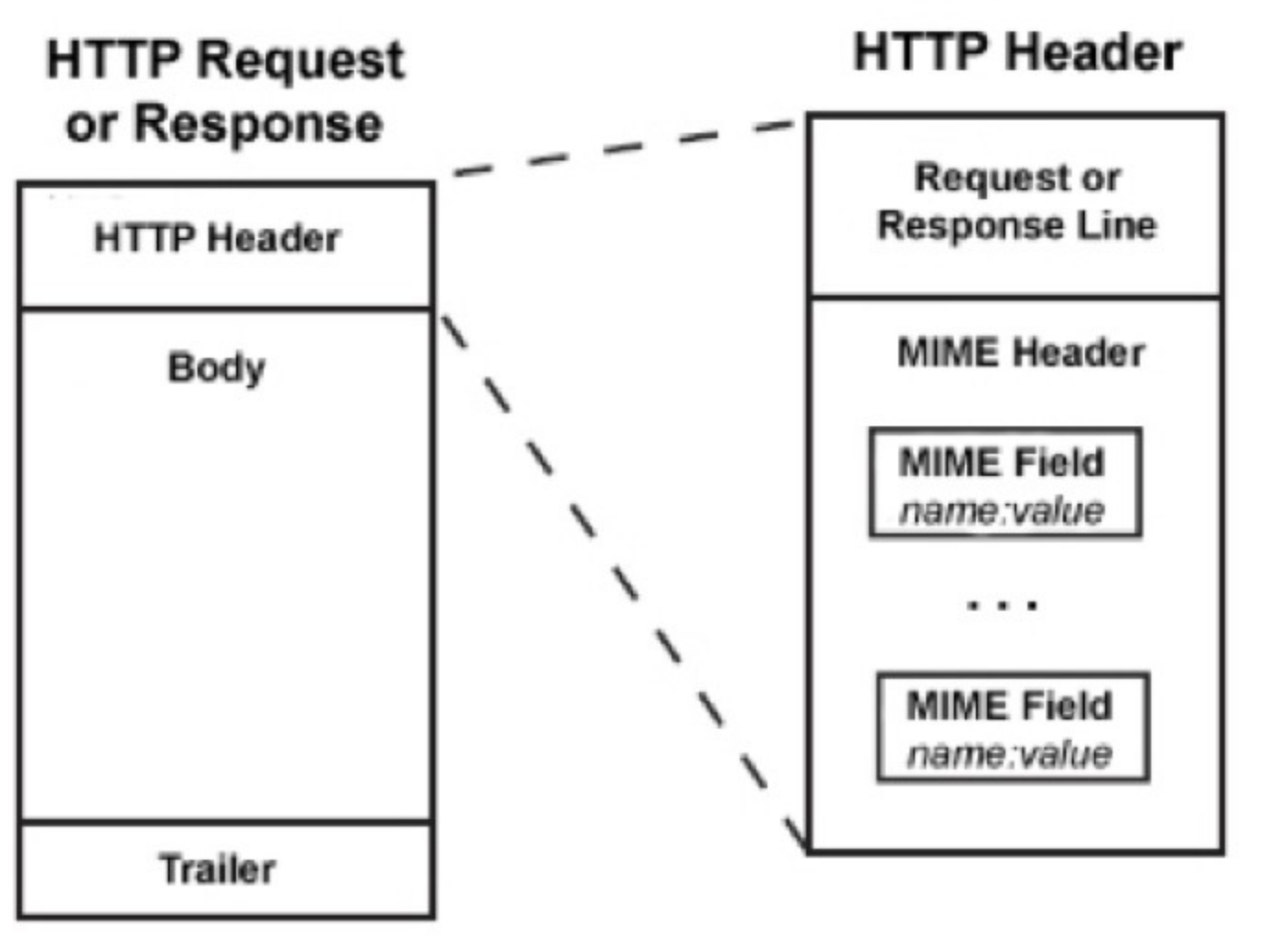
```
# 載入函式庫
import requests

# 發送請求
r = requests.get('https://www.zhihu.com/api/v4/questions/55493026/answers')

# 取回請求的回覆
response = r.text
```

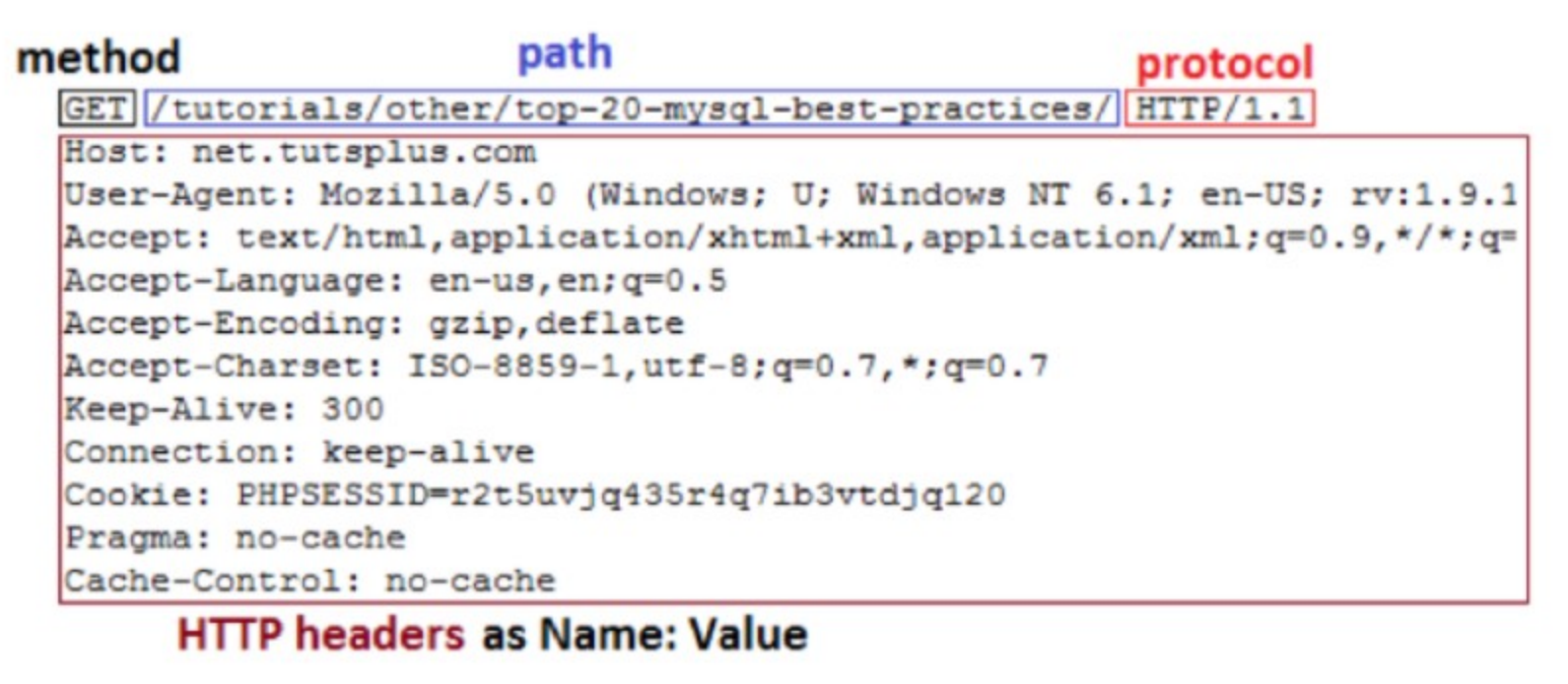
```
Out[4]: '<html>\r\n<head><title>400 Bad Request</title></head>\r\n<body bgcolor="white">\r\n<center><h1>400 Bad Request</h1></center>\r\n<hr><center>openresty</center>\r\n</body>\r\n</html>\r\n'
```

當我們直接對 API 發送請求時會發生錯誤，原因是知乎的伺服器有做檢查的機制，會透過 Headers 檢查「發送方」是否合法。

為什麼要加 Headers ?

「Header 是 HTTP 協定中規範資料傳輸的一種機制，會記載跟「發送方」有關的資訊。」





- Headers 是指從這個 API 的哪個部分來獲知請求中的內容是使用何種編碼方式，需輸入的欄位包含了 Key 和 Value 兩部分。通常會包含發送資料方的資訊，例如「時間」、「權限」、「瀏覽器資訊」等等的。
- 所以 Headers 可以視為是最基本的檢查機器，可以用來判斷發出 Request 的那一方是否為一個正常的來源。因此我們在這邊可以加上一些資訊，讓我們利用 Python 發出的 Request 更像是一個正常的使用者行為。

在 Request 上加上 Headers

```
import requests

# 定義標頭檔內容
headers = {'user-agent': 'my-app/0.0.1'}

r =
requests.get('https://www.zhihu.com/api/v4/questions/55493026/answers',hea
ders=headers)
response = r.text
```

'{"data":[{"id":683070334,"type":"answer","answer_type":"normal","question":{"type":"question","id":55493026,"title":"你们都是怎么学 Python 的？","question_type":"normal","created":1486390229,"updated_time":1543075931,"url":"https://www.zhihu.com/api/v4/questions/55493026","relationship":{}},"author":{"id":"36f69162230003d316d0b8a6d8da20ba","url_token":"liang-zi-wei-48","name":"量子位","avatar_url":"https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_is.jpg","avatar_url_template":"https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_{size}.jpg","is_org":true,"type":"people","url":"https://www.zhihu.com/api/v4/people/36f69162230003d316d0b8a6d8da20ba","user_type":"organization","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":55493026,"title":"你们都是怎么学 Python 的？","question_type":"normal","created":1486390229,"updated_time":1543075931,"url":"https://www.zhihu.com/api/v4/questions/55493026","relationship":{}},"author":{"id":"788f207a6bf8f66c5bad79bd0f011065","url_token":"simonlearn","name":"赛门喵Simon","avatar_url":"https://pic2.zhimg.com/v2-03afe381dbea789c0f918d6aac15556c_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

存取回傳的 JSON 格式

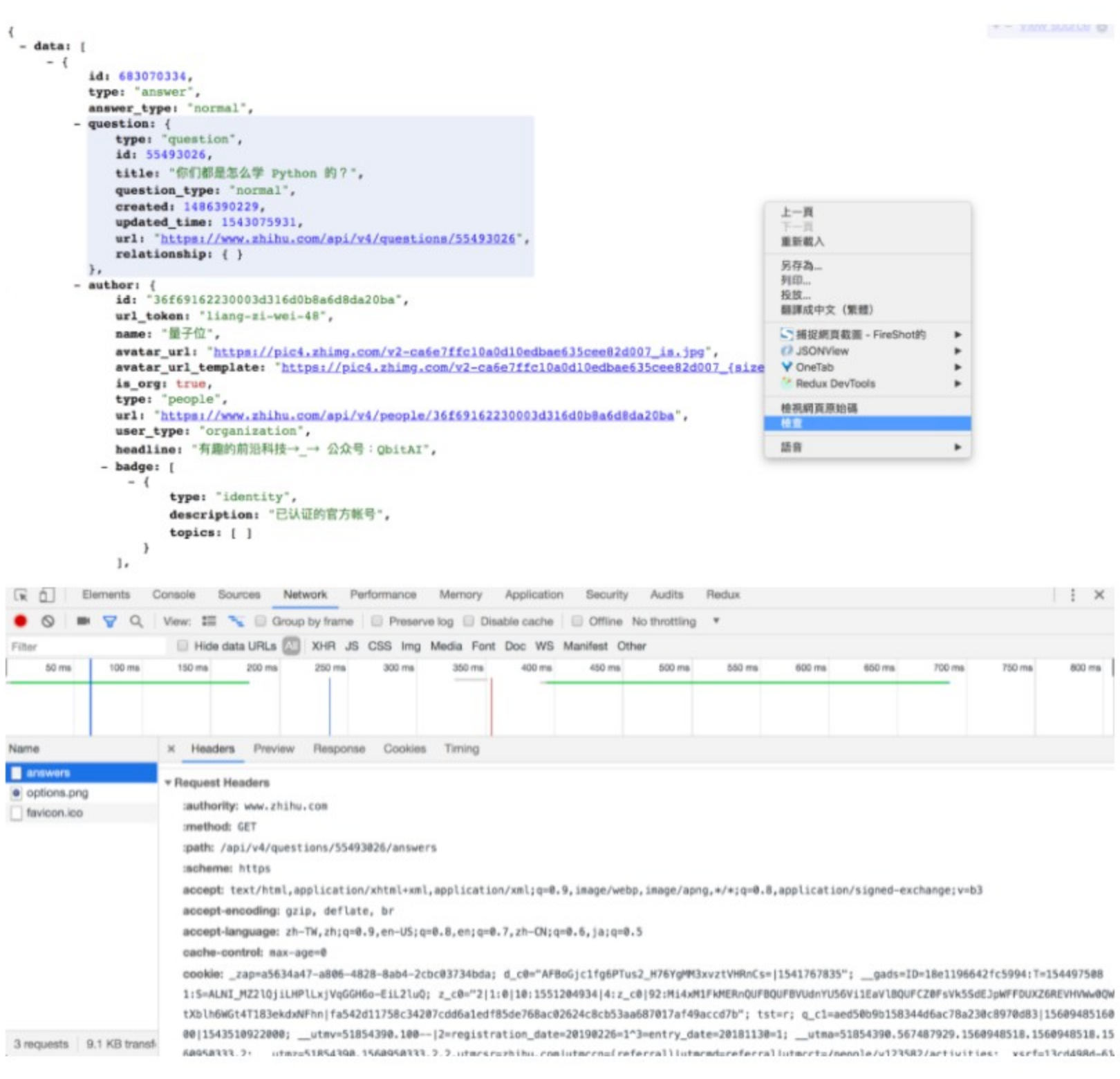
這是一個以 JSON 格式的字串，因此我們可以利用 json 函式庫解析成 Python 內的資料結構。

```
import json
data = json.loads(response) # 利用 json 解析網頁內容

# 取出資料
for d in data['data']:
    print(d)
```

{'id': 683070334, 'type': 'answer', 'answer_type': 'normal', 'question': {'type': 'question', 'id': 55493026, 'title': '你们都是怎么学 Python 的？', 'question_type': 'normal', 'created': 1486390229, 'updated_time': 1543075931, 'url': 'https://www.zhihu.com/api/v4/questions/55493026', 'relationship': {}}, 'author': {'id': '36f69162230003d316d0b8a6d8da20ba', 'url_token': 'liang-zi-wei-48', 'name': '量子位', 'avatar_url': 'https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_is.jpg', 'avatar_url_template': 'https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_{size}.jpg', 'is_org': True, 'type': 'people', 'url': 'https://www.zhihu.com/api/v4/people/36f69162230003d316d0b8a6d8da20ba', 'user_type': 'organization', 'headline': '有趣的前沿科技→_→ 公众号：QbitAI', 'badge': [{'type': 'identity', 'descriptio

如何知道要加上哪些 Headers ？

1. 先從網路進去該 API
2. 右鍵檢查，下方 Console 選 Network
3. 左方資源選擇第一個
4. 查看右方中的 Request Header 區塊



在 Request 上加上 Headers

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

從瀏覽器的開發者工具中可以在這裡看出瀏覽器帶了一堆 Headers 在請求中，不過我們的範例卻只加了 user-agent 而已？

因為經過測試之後，發現知乎伺服器只會檢查 user-agent 而已，如果在一開始不確定的情況下，會建議所有參數都加上去！

重要知識點複習

- 了解知乎 API 使用方式與回傳內容
- 撰寫程式存取 API 且添加標頭

解題時間

下一步：完成作業