അ AI共學社群 我的

AI共學社群 > Python 網路爬蟲實戰研習馬拉松 > D5:修改爬蟲程式中的 Headers 以成功存取第三方網站

☐ Ä Û₀ II → ふ (ive)

D5:修改爬蟲程式中的 Headers 以成功存取第三方網站

囯 E 學習心得(完成) 課程閱讀 本日作業 問題討論

API 資料串接 - 以 知乎 API 實作範例 本日知識點目標 以知乎 API 作為範例 思考流程與使用套件 **Request Library** 為什麼要加 Headers? 在 Request 上加上 Headers

存取回傳的 JSON 格式

如何知道要加上哪些

Headers ?

API 資料串接 - 以 知乎 API 實作範例 entoch Day 05 使用 API 存取網路資料 修改爬蟲程式中的 Headers 以成功存取第三方網站 出題教練:張維元

本日知識點目標

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

www.zhihu.com

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



હ

載入函式庫

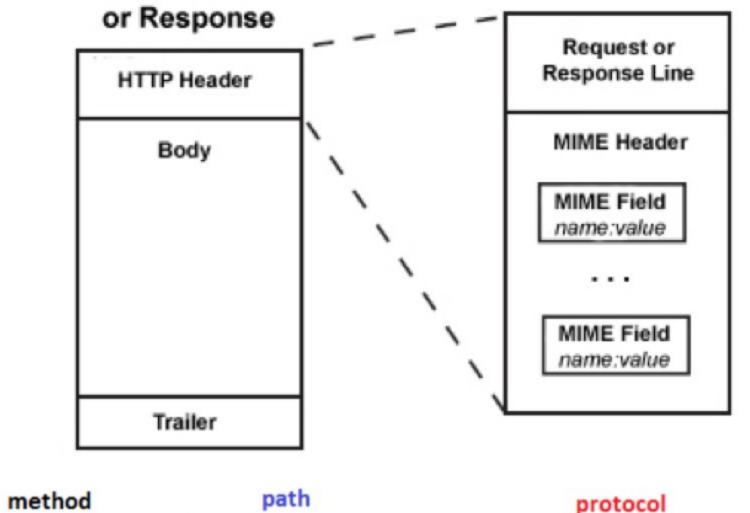
```
2 import requests
     # 發送請求
     r = requests.get('https://www.zhihu.com/api/v4/questions/55493026/answers')
   7 # 取回請求的回覆
 Out[4]: '<html>\r\n<head><title>400 Bad Request</title></head>\r\n<body bgcolor="white">\r\n<center><h1>400 Bad Request</h1>
       </center>\r\n<hr><center>openresty</center>\r\n</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。」

Host: net.tutsplus.com

HTTP Header HTTP Request



User-Agent: Mozilla/5.0 (Windows; U; Windows NT 6.1; en-US; rv:1.9.1 Accept: text/html,application/xhtml+xml,application/xml;q=0.9,*/*;q= Accept-Language: en-us, en; q=0.5 Accept-Encoding: gzip, deflate Accept-Charset: ISO-8859-1, utf-8; q=0.7, *; q=0.7 Keep-Alive: 300 Connection: keep-alive Cookie: PHPSESSID=r2t5uvjq435r4q7ib3vtdjq120 Pragma: no-cache Cache-Control: no-cache HTTP headers as Name: Value • 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 两部分。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。

GET /tutorials/other/top-20-mysql-best-practices/ HTTP/1.1

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。

• 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

1 import requests 3 # 定義標頭檔內容

headers = {'user-agent': 'my-app/0.0.1'}

在 Request 上加上 Headers

requests.get('https://www.zhihu.com/api/v4/questions/55493026/answers',hea ders=headers) response = r.text ('data":[{"id":683070334,"type": 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_ten": "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": "normal", "question": "duestion", "id": 55493026, "title": "https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_{ size}.jpg", "is_org":true, "type": "normal", "question": "duestion", "id": 55493026, "title": "https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_{ size}.jpg", "is_org":true, "type": "normal", "duestion", "id": 55493026, "title": "https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_{ size}.jpg", "is_org":true, "type": "normal", "duestion", "id": 55493026, "title": "https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_{ size}.jpg", "is_org":true, "type": "normal", "duestion", "id": "duestion", "id": "https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_{ size}.jpg", "is_org":true, "type": "normal", "duestion", "id": "duestion", "id": "https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_{ size}.jpg", "is_org":true, "type": "normal", "duestion", "duestion" e", "people", "url": "https://www.zhihu.com/api/v4/people/36f69162230003d316d0b8a6d8da20ba", "user type": "organization", "headline": "有趣的前沿科技——公众号: ObitAI", "badge":[{"type": "identity", "description": "已认证的官方被号", "topics": 在 requests 當中加上 headers 參數就可以正常拉回資料:!(683070334",

'{"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_tok en":"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, "type": "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":55493026,"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/v2-03afe38ldbea789c0f918d6aac15556c_{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。

2 data = json.loads(response) # 利用 json 解析網頁內容 3 # 取出資料

import json

for d in data['data']: print(d)

{'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://www.zhihu.com/api/v4/questions/55493026', 'relationship': {}}, 'author': {'id': '36f69162230003d316d0b8a6d8d a20ba', 'url_token': 'liang-zi-wei-48', 'name': '量子位', 'avatar_url': 'https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edba e635cee82d007_is.jpg', 'avatar_url_template': 'https://pic4.zhimg.com/v2-ca6e7ffc10a0d10edbae635cee82d007_{size}.jp g', '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

- data: [id: 683070334, type: "answer", answer_type: "normal",

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",
                                                                                       重新载入
            relationship: { }
                                                                                       另存為...
       - author: {
           id: "36f69162230003d316d0b8a6d8da20ba",
                                                                                       翻譯成中文 (繁體)
            url_token: "liang-zi-wei-48",
                                                                                       新捉網頁截蓋 - FireShot的
           (2) JSONView
                                                                                       检视纲真原始碼
           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: [ ]
● 🛇 📟 🔻 Q. View: 🚟 🐾 🔲 Group by frame : □ Preserve log □ Disable cache : □ Offline No throttling 🔻
                  ☐ Hide data URLs ⚠ XHR JS CSS Img Media Font Doc WS Manifest Other
  50 ms 100 ms 150 ms 200 ms
                                  250 ma 300 ma 350 ma 400 ma 450 ma 500 ma 550 ma 600 ma 650 ma 700 ma
```

Elements Console Sources Network Performance Memory Application Security Audits Redux x Headers Preview Response Cookies Tirring answers options.png :authority: www.zhihu.com favicon.ico :method: GET accept-encoding: gzip, deflate, br accept-language: zh-TW, zh; q=0.9, en-US; q=0.8, en; q=0.7, zh-CN; q=0.6, ja; q=0.5 cookie: _zap=a5634a47-a806-4828-8ab4-2cbc03734bda; d_c0="AFBoGjc1fg6PTus2_H76YgMM3xvztVHRnCs=|1541767835"; __gads=ID=18e1196642fc5994:T=154497508 1:S=ALNI_MZ2l0jiLHPlLxjVqGGH6o-Eil2lu0; z_c0="2|1:0|10:1551204934|4:z_c0|92:Mi4xM1FkMER:nQUFBQUFBVUdnYU56Vi1EaVlBQUFCZ0FsVkSSdEJpWFFDUXZ6REVHVWwGQW $tXb \ line with the windows and the windows$ 00|1543510922000; _utmv=51854390.100--|2=registration_date=20190226=1^3=entry_date=20181138=1; _utma=51854390.567487929.1560948518.1560948518.15 3 requests 9.1 KB transf-在 Request 上加上 Headers import requests

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

5 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!

 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

重要知識點複習



https://getfireshot.com

下一步:完成作業