(圖解) 網路爬蟲 API 常見的 3 種「翻頁」方式

2021年07月10日 網路爬蟲 #API

目錄

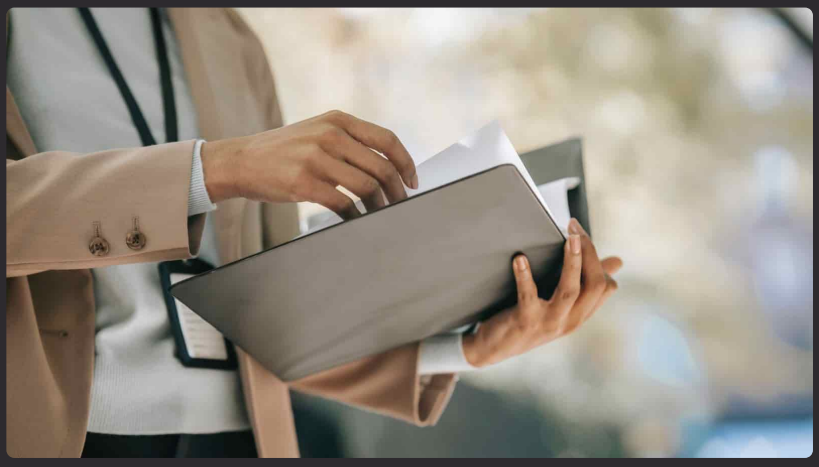
前言
說明
頁數 (page)
偏移 (limit & offset)
指定ID (pid)
三者方式比較
結語

前言

之前時不時有網友在詢問有關網路爬蟲"翻頁"的問題：

- 我該如何抓取下一頁的文章呢？
- 使用 limit 最多只能抓到前 100 筆留言，那之後的該怎麼取得？

我發覺可能之前 Python 網路爬蟲實例系列內沒有說明清楚。
因此這篇文章，將整理目前我遇過的網路爬蟲 API 中，常遇見的三種「翻頁」方式，並且搭配簡易圖示，希望讓剛進此領域的網友能更容易理解。



圖片來源：Pexels

想要爬取某個網站，也順利找出網頁是使用動態載入請求 API 的方式，但遇到像 Dcard 這種的文章列表或留言列表，它是往下滾，就會送出新請求來取得下一頁資料。

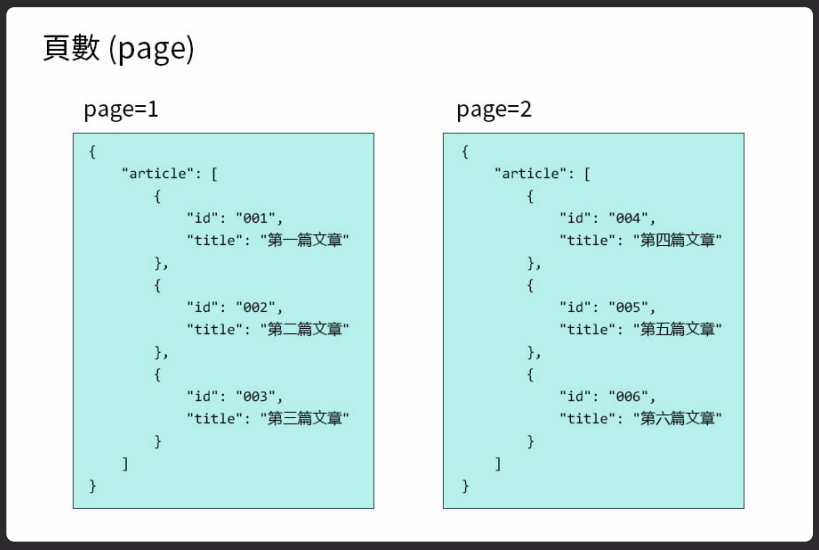
那麼，它是如何達成「翻頁」的呢？

底下會依照這三種常遇到的「翻頁」方式來分別說明：

1. 頁數 (page)
2. 偏移 (limit & offset)
3. 指定ID (pid)

目錄
前言
說明
頁數 (page)
偏移 (limit & offset)
指定ID (pid)
三者方式比較
結語

頁數 (page)



翻頁方式 - 頁數

第一種最容易理解、最直覺的是頁數，這就跟我們一般瀏覽網頁一樣，你想要看第幾頁，就給它第幾頁的頁數即可。

- `page` 代表資料的頁數。

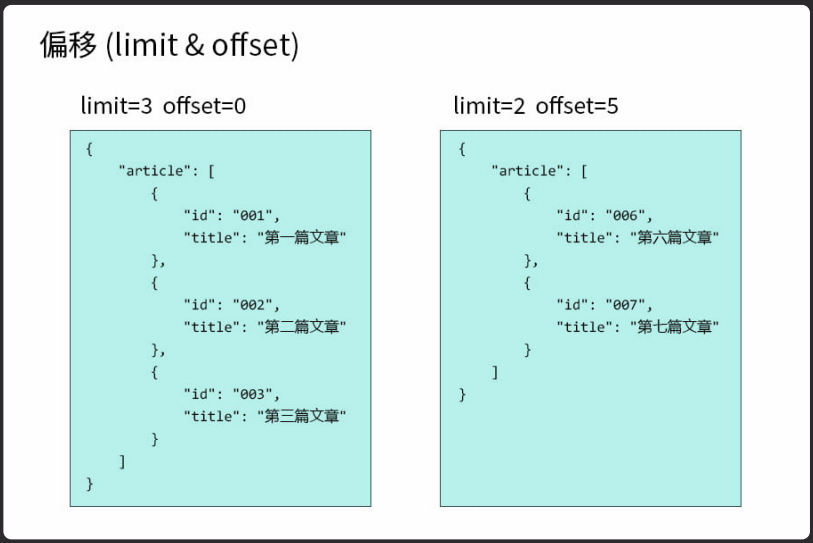
不過缺點是不能彈性調整每次抓取的量，假如此 API 一頁是 30 則留言，就算我只想取前 5 則留言，一樣一次請求還是會抓到 30 則留言，除了會占用較多流量，也可能花費較多時間。

舉例來說：
想抓第一頁 `page=1`，想抓第二頁是 `page=2`，同理第99頁就是 `page=99`。

實際網站範例：
「PChome 線上購物」"商品搜尋"中的 `page`。
「聯合新聞網」"文章列表"中的 `page`。

目錄
前言
說明
頁數 (page)
偏移 (limit & offset)
指定ID (pid)
三者方式比較
結語

偏移 (limit & offset)



翻頁方式 - 偏移

第二種翻頁方式就解決了第一種的問題，變成可「彈性調整抓取量」。
它藉由兩個參數來達成，分別為"limit"與"offset" (不同 API 參數名稱可能不同)

- `limit` 代表一次請求最大資料筆數。
- `offset` 代表資料的偏移值。

* `offset` (foodpanda)參數在不同網站的 API 有不同名稱，例如 `newest` (蝦皮購物)、`after` (Dcard)。

舉例來說：
抓前三筆資料是 `limit=3 offset=0`。
想取得第六、七筆資料，將資料偏移 5 (從第一筆資料開始往下加五筆的意思)、限制一次 2 筆，就是 `limit=2 offset=5`。

實際網站範例：

「foodpanda」"搜尋餐廳"中的 `limit` 和 `offset` 。

「蝦皮購物」"搜尋商品"中的 `limit` 和 `newest` 。

「Dcard API」"留言列表"中的 `limit` 和 `after` 。

目錄

前言
說明
頁數 (page)
偏移 (limit & offset)
指定ID (pid)
三者方式比較
結語

指定ID (pid)

翻頁方式 - 指定ID

第三種翻頁方式感覺像是第二種的改版，有些網站一樣有"limit"來限制最大資料筆數，但"offset"換成了"pid"參數，"pid"參數需要帶入上一頁最後一筆資料的數值。
但它就限制了你，不能直接跳到後面的頁數，例如我想看第 100 筆資料，你還是要請求一頁才知道下一頁的網址。

- `limit` 代表一次請求最大資料筆數。
- `pid` 上一頁最後一筆的 ID。

* `pid` (NOWnews)參數在不同網站的 API 有不同名稱，例如 `before` (Dcard)、`pageToken` (YouTube Data API)。而 NOWnews 不需要 `limit` 參數，它 API 已經有限制一次的資料量了。

舉例來說：
抓前三筆資料是 `limit=3 pid=` ，因為這是最前面的資料，pid 就不需要給值。
而取得第四、五筆資料，因為只要兩筆，limit 帶入 2，前一頁最後一筆資料 ID 為 003，因此 pid 帶入 003，結果就

實際網站範例：

「NOWnews今日新聞」"新聞列表"中的 `pid` 。

「Dcard API」"文章列表"中的 `limit` 和 `before` 。

「YouTube Data API」"留言列表"中的 `maxResults` 和 `pageToken` 。

注意：YouTube Data API 的 `pageToken` 有點不太一樣，它是帶入上一頁回傳的的 `nextPageToken` ，而不是最後一筆資料的 ID。

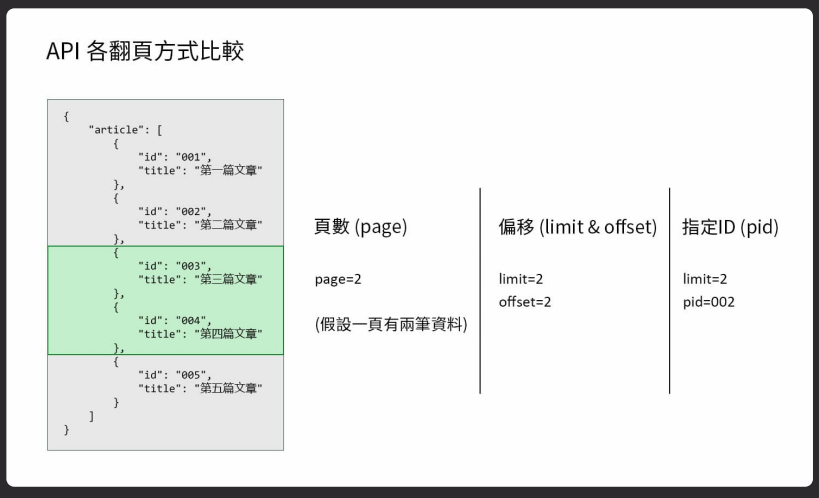
目錄

前言
說明
頁數 (page)
偏移 (limit & offset)
指定ID (pid)
三者方式比較
結語

三者方式比較

我進一步將以上三種翻頁方式放在一起比較，如下圖範例所示。

假如左邊是此 API 可以獲取的全部資料，可以看到共有五篇文章，各自有 ID(id) 和 標題(title) 欄位，同樣要取得綠色區塊(第三和第四篇文章)，以上三種翻頁方式實際會需要帶入這些數值。



三種翻頁方式比較

- 頁數：假設它一頁是兩筆資料，那我們要帶入 `pega=2` 來取到第二頁資料。
- 偏移：一次想取兩筆資料，因此 `limit=2` ；我們要從第三筆開始取，因此 `offset=2` 。
- 指定ID：一次想取兩筆資料，因此 `limit=2` ；上一頁最後一筆資料的 ID 是 002，因此 `pid=002` 。

結語

目錄
前言
說明
頁數 (page)
偏移 (limit & offset)
指定ID (pid)
三者方式比較
結語

不知道透過上面的說明與比較，是不是讓你對於 API 常見的幾種翻頁方式更了解了呢？
提醒各網站 API 本身可能有些微差異，詳細規則還是要查看 API 文件。

歡迎追蹤『IT空間』FB 粉專，取得最新發文通知🔔

參考：

Dcard API | IT空間
蝦皮購物 爬蟲 | IT空間
foodpanda 爬蟲 | IT空間
PChome 線上購物 爬蟲 | IT空間
YouTube Data API | IT空間
NOWnews今日新聞 爬蟲 | IT空間
聯合新聞網 爬蟲 | IT空間

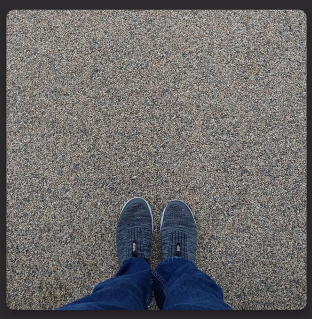
> Stay hungry. Stay foolish
> 求知若飢，虛心若愚。
>
> —— 史蒂夫・賈伯斯

🔻 如果覺得喜歡，歡迎在下方獎勵我 5 個讚~

分享

目錄

前言
說明
頁數 (page)
偏移 (limit & offset)
指定ID (pid)
三者方式比較
結語



作者
Jia
軟體工程師

相關內容

- Google 搜尋結果 API — Aves API 完整教學
- [Python爬蟲實例] YouTube-使用 YouTube Data API
- 爬蟲 Dcard API 2.0 版本？！

[Python爬蟲實例] 教你爬取"foodpanda"餐...
YouTube「剪輯片段」新功能 - 完整說明與教學

覺得有幫助就給個心情吧~

0 Responses

讚讚 爆笑 喜歡 驚訝

目錄

前言
說明
頁數 (page)
偏移 (limit & offset)
指定ID (pid)
三者方式比較
結語