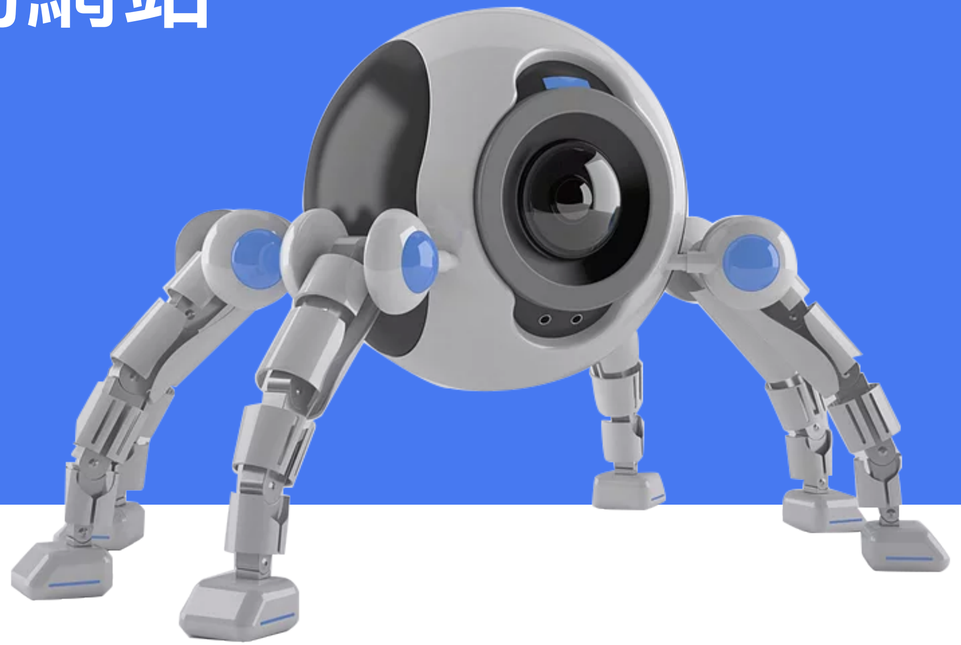
Day 32 如何克服反制爬蟲的網站

反爬：驗證碼處理



出題教練：張維元

python

本日知識點目標

- 了解「驗證碼機制」的反爬蟲機制
- 「驗證碼機制」反爬蟲的因應策略

常見的反爬蟲機制有哪些？

CUPOY

檢查 HTTP 標頭檔

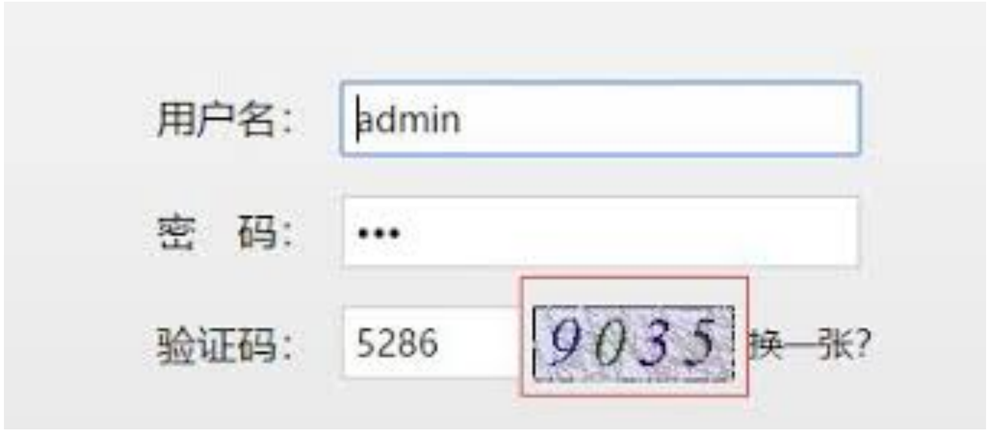
驗證碼機制

登入權限機制

IP 黑/白名單

驗證碼

驗證碼機制是許多網站再傳送資料的檢查機制，對於非人類操作與大量頻繁操作都有不錯的防範機制。



驗證碼是一種圖靈測試

CAPTCHA 的全名是「Completely Automated Public Turing test to tell Computers and Humans Apart」，或「全自動區分電腦與人類的圖靈測試」，實作的方式很簡單，就是問一個電腦答不出來，但人類答得出來的問題。

爬蟲該怎麼辦？

爬蟲在實作上遇到驗證碼的做法會是這樣，先把圖抓回來，

再搭配圖形識別工具找出圖中的內容。

環境工具準備

CUPOY

1. Tesseract

Tesseract 是一個OCR庫(OCR是英文Optical Character Recognition的縮寫)，它用來對文字資料進行掃描，然後對影像檔案進行分析處理，獲取文字及版面資訊的過程

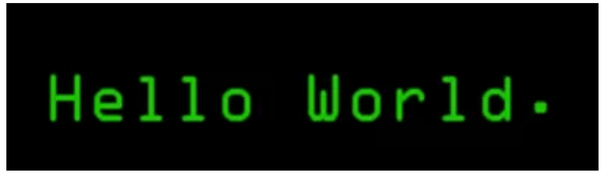
安裝方式：https://github.com/tesseract-ocr/tesseract/wiki

2. pytesseract

在 Python 中呼叫 Tesseract 的套件

安裝方式（利用 pip）：https://pypi.org/project/pytesseract/

來看個例子吧！




```
1  import requests
2  import pytesseract
3  from io import BytesIO

   response = requests.get('https://i0.wp.com/www.embhack.com/wp-content/uploads/
   2018/06/hello-world.png')
   img = Image.open(BytesIO(response.content))
   code = pytesseract.image_to_string(img)
   print(code)
```

重要知識點複習

- 了解「驗證碼機制」的反爬蟲機制
- 「驗證碼機制」反爬蟲的因應策略

參考資料

python識別驗證碼

奔跑中的兔子

網頁連結

此篇針對驗證碼的網站，提出了幾種不同的處理機制。

參考資料

Python 實現識別弱圖片驗證碼

猴哥yuri

網頁連結

圖片識別的精準度是一個麻煩的問題，容易受到圖形的模糊或是干擾而降低。本篇文章利用幾種常見的技巧來克服弱圖片的驗證碼。

解題時間
LET'S CRACK IT

請跳出 PDF 至官網 Sample Code & 作業

開始解題

START