# 實作題目:圖形驗證碼CAPTCHA

名字:陳曉龍

學號:409416145

背景:

圖形驗證碼CAPTCH是由路易斯馮安（Luis von Ahn）所發明的。2000年的時雅虎當時有在提供免費電郵服務。受當時技術限制，每個email地址每天只能寄出500封信，所以引起駭客利用程式自動化，不斷申請新的帳號，然後用這些帳號發出大量垃圾郵件。

馮安經過一段時間的思考後，做出圖形驗證碼CAPTCH，利用一張扭曲文字的圖片，請使用者輸入，讓機器人程度難以分別圖中文字，達到能夠區分使用者和機器人。[[1]](#footnote-1)

製作動機:

我本身對資安和爬蟲這個領域感到興趣，雖然captcha這個技術已經快被google推出的re-captcha取代，但還是有部分的公司或網站會摘用captcha來擋掉自動程式。

這個題目比較簡單就可以做出，所以我會嘗實作一個captcha，並且常試去破解自己的captcha。

功能介紹:

圖形驗證碼CAPTCHA設計出來的目的是防機器人自動化。現時資訊發達，程式開發員會嘗試利用程式作出爬蟲，截取和收集網站資訊，用作搶奪門票，或是

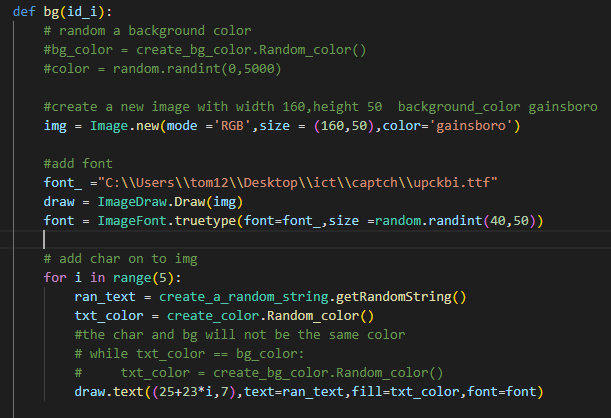
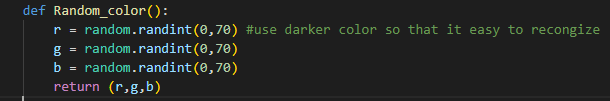
駭客利用程式惡意註冊大量帳號，或是暴力破解密碼，或是在網站或社交媒體上發大量的帖文。圖形驗證碼CAPTCH是用來區分使用者和機器人。防止駭客的惡意行為，或是讓人很難以得到網站資源。

技術:

**文字型驗證碼由數字、英文大小寫字母**

第一步驟:

我產生了5個隨機的數的數字或英文字母



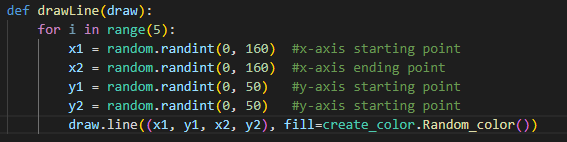
輪出結果:



第二步驟:

**在圖像加入干擾線:**

目是為了讓影像處理不容易進行特徵分析。



輪出結果:



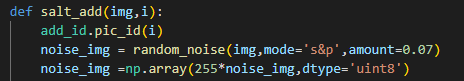
**小總結:**

我們人類之所以可以辨識出各種文字或圖案是因為固定型態的辨識，我們可以透過語言的邏輯，大腦可以在短時間內辨識出同一個字的變化，而這個數量可以高達無限，但是相反電腦卻需要透過資料庫的新增來有辦法做判讀。[[2]](#footnote-2)

所以當加入干擾線或對文字進行扭曲後電腦對特徵的辨識便會大幅下降。

最後一個步驟:

背景加入噪音: 椒鹽噪音

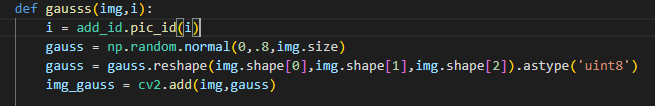


輪出結果:



或者

背景加入噪音: 高斯噪音



輪出結果:

一張含有 線條 的圖片

自動產生的描述

**基本上一張圖形驗證碼便完成了**

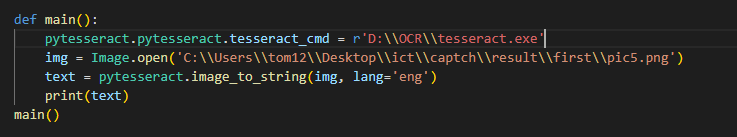
我們可以用google推出的光學字元辨識[[3]](#footnote-3)tesseract來進行測試:

我們先以一張清晰無任何雜音的圖像來進行文字辨識，確保他是能準確地辨識。

我用這張圖來作範例:

一張含有 文字 的圖片

自動產生的描述



執行結果:

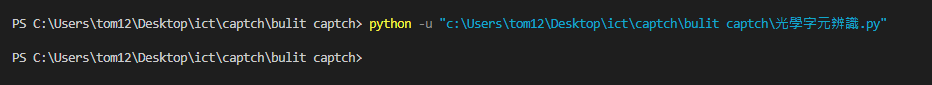


他可以清楚解讀出4tx2x的數字組合

接下來用我剛做好的圖像進行測試:

一張含有 線條 的圖片

自動產生的描述



執行結果:

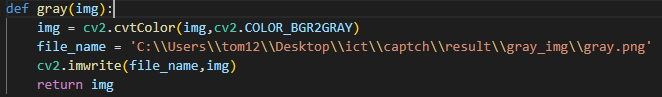
沒有顯示任何內容，表示不能解讀出任何一個數字或字母

**破解方法:**

(本範例用含椒鹽噪音的圖像)

**步驟一:**

把需要破解的圖案轉為灰階:



執行結果:

一張含有 文字, 差異, 數個 的圖片

自動產生的描述

**步驟二:**

把圖像二值化一張含有 文字 的圖片

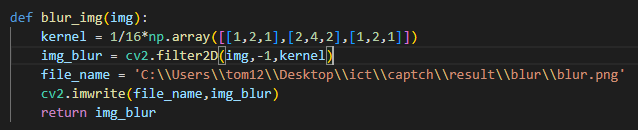
自動產生的描述

執行結果:



**步驟三:**

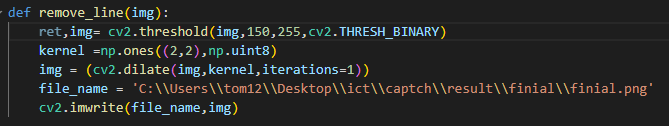
把圖像進行模糊化，降噪



執行結果:



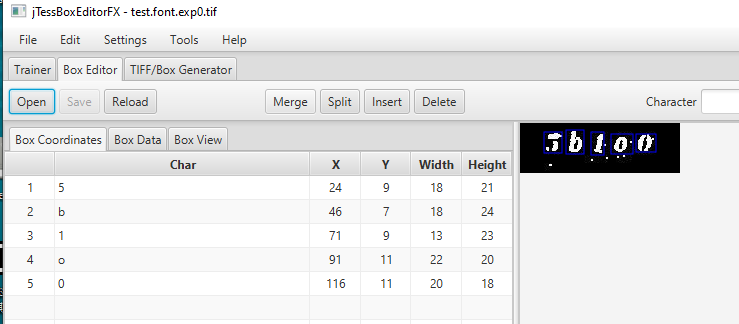
**步驟四:**

再把圖像二值化 找到最接近的閥值除去干擾線

執行結果:



最後放進使用tesseract裏做分割與標籤建立新的文字檔案，提高識別率。



**結論:**因為圖像的干擾線(圖一)，沒辦法找到一個合適的閥值，能完全清理掉以及其他字元的分別難道例如1和l的分別(圖二)，大幅增加了標籤的難道。結果只做到8成的辨識。

圖一



圖二



報告心得:

起初在想報告題目時，是希望結合影像和ai的技術做出一樣東西。例如人臉識別，識別特定物件之類的。網路上也有很多的資源，我發現如果要在半個月內，從零到一，完全做出一個相關的技術要花7成左右的時間。但因為接近期未考，又受到疫情影響，很多的課堂都改做要交報告當期未成績，而且不知為何還把時間提早了1-2星期，所以時間上是沒太鬆動。

後來，因為我想起以前在學爬蟲的時候多數會面對的問題是圖像驗證碼的問題，所以不能單一靠程式自動輸入資料。會卡在不停轉變的圖像驗證碼，當時對程式和影像處理還不是很了解，因此那時便暫時放棄了。直到今次的報告才想起我可以重新嘗試再次圖像驗證碼挑戰，所以這就是我選擇做這次主題的原因之一。

製作一張圖像驗證碼並沒有太大難度，應用的影像的技術也只有在加入噪音這一項。現時用PIL的python套件便可以輕鬆做出一張圖像和在圖像上添加文字，matplotlib.pyplot也有提供draw這個function讓我可以輕鬆的在圖像上加上線條。

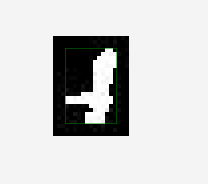
在破解的時候也有運用到一些技術，例如先要把圖像轉灰再二值化分清楚特徵和輪廓，再者要把圖像模糊，再用卷積除噪音，最後再一次的二值化。

而真正困難點是標籤，有時因為二值化不能完全處理干擾線，有的把原有的特徵也除去了，就例如4(圖一)，連我沒看沒二值化前的圖像都不知道他是4，誤以為他是1或l橫線是沒處理好的干擾線。，結果花最多的時間是在標籤這一部分。

這次雖然運用到技術沒有太多，但起碼讓我了解多驗證碼。但其實隨著ai的成熟，其實這一類型的驗證碼，只要有大量的sample和花多一點時間，他1實在太容易破解了，對於bots來說就是一個擺設。所以google的re-captcha其做到很好，而且很難去破解，他分成二個case，第一是檢查你的滑鼠指標在點擊I’m not robot 先利用其「風險分析引擎」做一次安全檢測，如果直接檢驗通過的話，無須進入圖像辨別的環節。

所以剛好可以在這個技術淘汰前嘗試自己做一做。也是一個不錯的成果。

圖一



YouTube連結: https://youtu.be/S4T0hvVh\_c8

1. https://www.businessweekly.com.tw/business/blog/3002826 [↑](#footnote-ref-1)
2. https://progressbar.tw/posts/78 [↑](#footnote-ref-2)
3. https://zh.wikipedia.org/zh-tw/%E5%85%89%E5%AD%A6%E5%AD%97%E7%AC%A6%E8%AF%86%E5%88%AB [↑](#footnote-ref-3)