

AI共學社群 > Python 網路爬蟲實戰研習馬拉松(舊) > D30: 爬蟲可能會遇到的問題

D30: 爬蟲可能會遇到的問題



簡報閱讀

範例與作業

問題討論

實務上的爬蟲與挑戰

本日知識點目標

>

實務上爬蟲有幾個優化的 角度

反爬是什麼?常見的反爬 蟲機制有哪些?

如何為爬蟲程式加速?

實務上的爬蟲與挑戰

>

 \square $\stackrel{\mathcal{H}}{\vdash}$ $\stackrel{\circ}{\cup}$ $\stackrel{\circ}{\circ}$

们用排任日勤儿史제

參考資料

解題時間 >

Day 3U 進陷爬蛾総見

實務上的爬蟲與挑戰



出題教練:張維元





本日知識點目標



本日知識點目標

- 實務上爬蟲可能遇到的問題有哪些
- 淺談常見防爬蟲機制與處理策略



實務上爬蟲有幾個優化的角度

在前面的課程中,我們討論了一個網頁從該如和思考和撰寫。接下來我們要討論的是「爬蟲可以順利拉到資料,然後呢?」我們針對這三個方向來做優化:

反爬

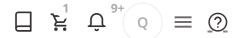
加速

自動化更新

反爬是什麼?常見的反爬蟲機制有哪些?

許多網站為了保護資料,避免網頁上的公開資訊被網頁爬蟲給抓取,因此有了「反 爬蟲」的機制出現。爬蟲工程師也發展了出一系列「反反爬蟲」的策略!







如何為爬蟲程式加速?

第二種實務爬蟲需要考慮的問題是加速,當資料量龐大或是更新速度較為頻繁的狀況下。依照正常的爬蟲程式,可以會因此受到應用上的限制。所以必須用程式的方法,來思考如何加速爬蟲的處理速度。

多線程 爬蟲加速

非同步爬蟲

真實世界中的資料是瞬息萬變的,也代表資料會有更新的需求。但爬蟲爬的資料只是一個片刻,所以必須要思考如何與資料源上的資料做同步或是更新,確保拿到的資料是最新的。常見的做法可以利用一個排程機制,週期性地重新抓取資料。

在迴圈中加上 Sleep

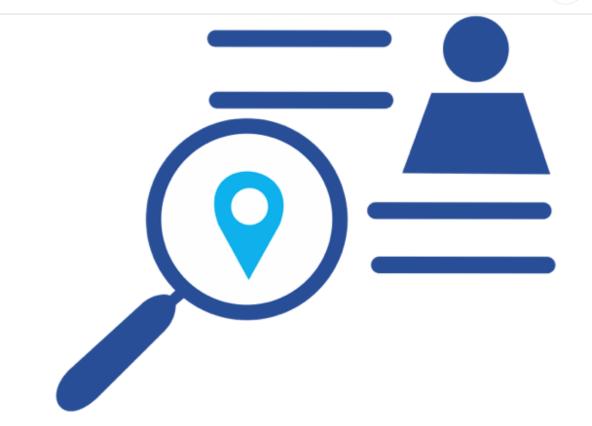
利用 threading 的 Timer

第三方套件 schedule

參考資料







Python爬蟲系统學習十一:常見反爬蟲機制與應對方法

guangyinglanshan/CSDN

網頁連結

作者簡介了他所看到的幾種反爬蟲的處理方法:標頭檔(User-Agent)反爬蟲機制解析、訪問頻率限制、代理IP或者分佈式爬蟲等等。

Python爬蟲筆記(六)— 應對反爬策略

菜到懷疑人生 / CSDN





主機策略、頻繁更改user-agent 等等。

解題時間



下一步:閱讀範例與完成作業

anboth