**科學計算軟體實習期末專題**

**題目: Wootalk 之聊天過濾機器人**

第十二組 組長: 黃泰源 組員: 黃柏儒、吳文欽

一、摘要

此專題主要是在為wootalk聊天網站設計一款過濾器，主要能夠過濾男性並在女性出現的時候能通知使用者。我們將利用python所提供的正則表達式函式庫進行字串特徵比對，tkinter函式庫製作GUI介面，並利用Python提供的資料結構dictionary製作狀態轉移表。

1. 動機

通常在wootalk聊天軟體上，男性使用者通常會先直接留言:”男”，這時我們不想浪費時間，按下離開尋找下一個對象，但做這些動作可能會讓使用者感到無聊或煩悶。所以我們希望利用Python強大的爬蟲和字串剖析函式庫與上課所學的知識來製作一個wootalk聊天軟體的過濾軟體，來節省時間。

三、挑戰

Wootalk網站本身為了防止機器人所以採取了一些防護措施，以下是網站常用防止機器人的方式。

1.通過User-Agent來控制訪問:

很多網站都會創建 user-agent白名單，只有屬於正常範圍的user-agent才能夠正常訪問。

2.通過IP限制來反爬蟲

如果一個固定的ip在短暫的時間內，快速大量的訪問一個網站，那自然會引起注意，管理員可以通過一些手段把這個ip給封了。

3.通過JS腳本來防止爬蟲

如果想爬取某個網站，但是在進入網站之前，它會有一個驗證頁面來驗證你是不是機器。

4.通過robots.txt來限制爬蟲

搜索引擎本身就是一個超級大的爬蟲，Google開發出來爬蟲24h不間斷的在網上爬取着新的信息，並返回給數據庫，但是這些搜索引擎的爬蟲都遵守着一個協議：robots.txt

根據實際的測試wootalk網站是抓進出聊天室太多次的IP來封鎖，wootalk網站會要求該IP必須通過測試(如圖1)才能繼續使用。

我們的解決方法有以下兩種:

1. 利用代理IP池:

使用代理IP池來刷新IP，但此方法使用免費的代理IP，由於免費的代理IP品質不穩定，會導致程式執行的時候會有很大的延遲。

1. wootalk要求測試時，主動通知使用者:

此方法會使程式執行起來順暢，但缺點是IP被wootalk要求測試時，使用者必須主動去測試。

四、實作

此程式的資料流程圖(圖2)主要為三個部分:

UI程式:

UI介面(圖3)負責提供友善介面與使用者互動，提供輸入密語欄位，按下啟動按鈕開啟進入聊天室並開始過濾，按下停止按鈕離開聊天室並停止過濾。

爬蟲程式:

製作成woospider類別(圖4)負責提供一系列相關的方法以供操作，表1是方法功能描述。

Strategy程式:

設計成狀態機(圖5)的形式，在不同的狀態(表2)有不同的處理，並利用狀態轉移表(表3)進行狀態轉移。

五、 結論

在實作(圖6)後得到了以下結論:

* 狀態機的中斷並不是即時的，而是要等到一個狀態做完才能中斷
* 雖然有實作前有一些設計，但在實作過程中有些功能是製作中東加西改的，所以之後維護起來應該很痛苦。
* 在擷取對方的訊息還是有點延遲，有三種可能:

1. 網路太慢
2. Selenium函式庫的問題
3. Cpython的threading函式庫設計上的問題

之後可能會針對第三點改善，不使用threading函式庫而改multiprocessing函式庫來提升程式效能。



圖1、wootalk防機器人頁面



圖2、wootalk之聊天過濾機器人資料流程圖

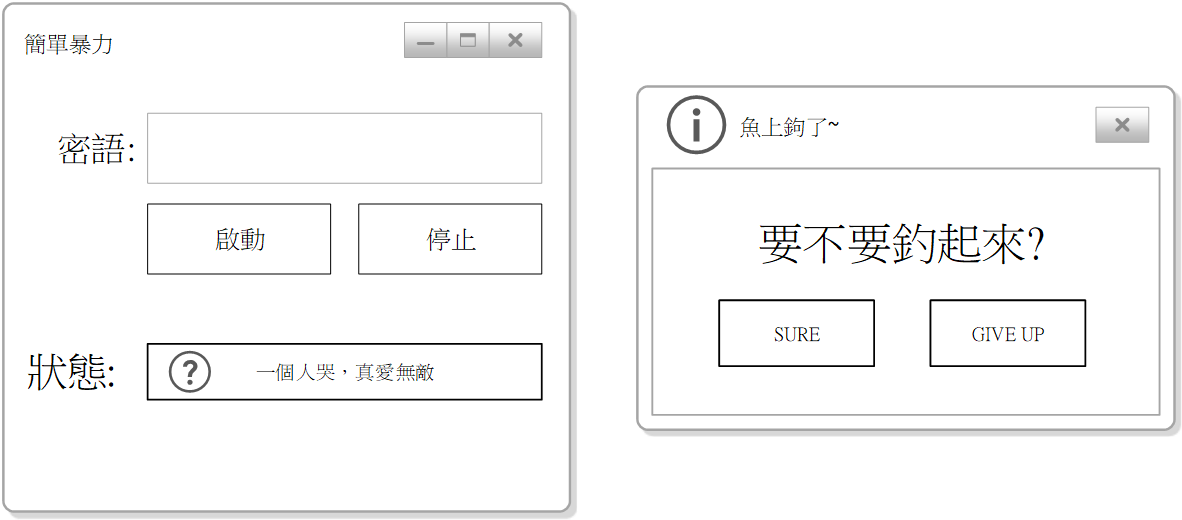


圖3、 使用者介面



圖4、woospider UML類別圖

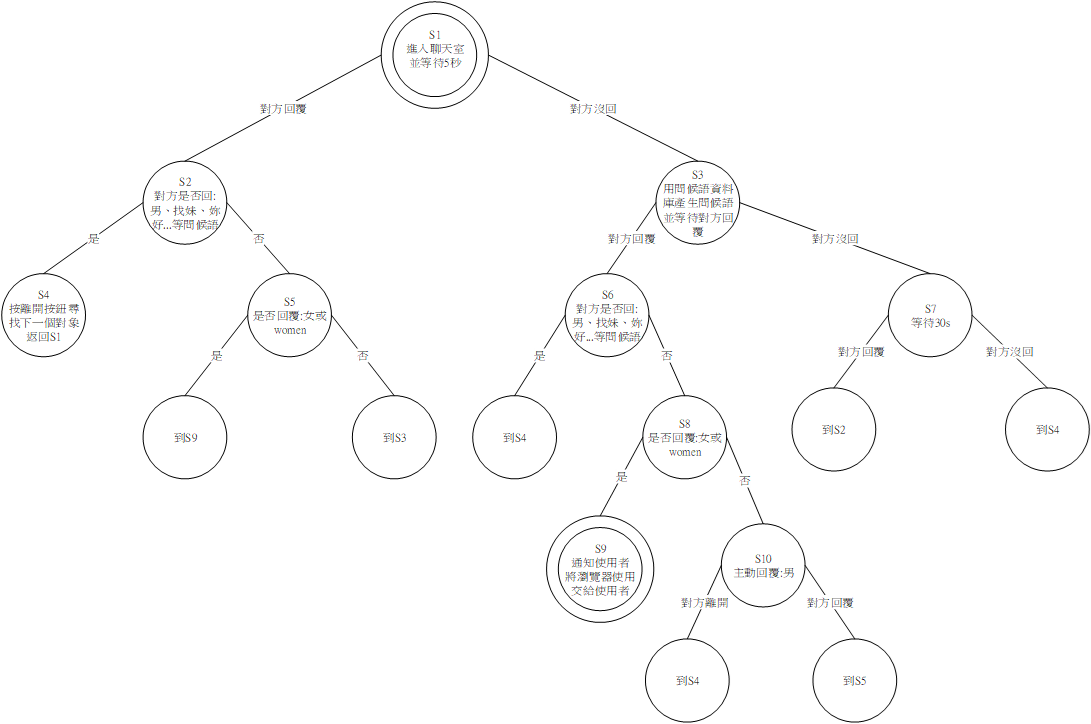


圖5、Strategy狀態機圖

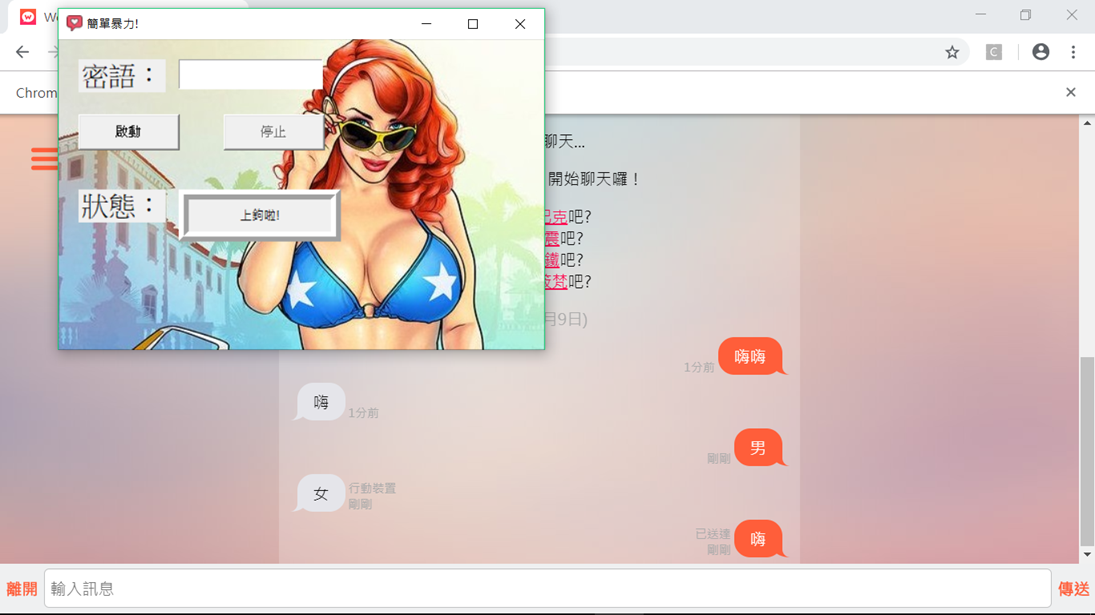


圖6、wootalk之過濾機器人成果畫面

|  |  |
| --- | --- |
| 方法 | 描述 |
| \_\_init\_\_ | 初始化woospider的各個屬性。 |
| update\_proxy | 更新代理伺服器IP以供使用。 |
| print\_proxy | 印出代理伺服器IP。 |
| set\_driver | 設定Chrome driver的設定。 |
| close\_driver | 關閉Chrome driver。 |
| connect | 打開wootalk網站。 |
| set\_secret | 設定wootalk網站密語欄位。 |
| start | 進入聊天室。 |
| leave | 離開聊天室。 |
| send\_message | 傳送訊息。 |
| get\_message | 獲得對方訊息，並可設定等待秒數。 |
| change\_secret | 更改密語。 |
| check\_stranger | 檢查對方是否主動離開聊天室。 |
| check\_enter | 檢查是否正常進入聊天室。 |
| maximize | 將Chrome瀏覽器最大化。 |

表1、woospider方法功能說明

|  |  |
| --- | --- |
| 狀態 | 狀態描述 |
| S1 | 進入聊天室，並等待對方回覆；等超過5秒，對方回覆到S2，對方沒回到S3。 |
| S2 | 對方是否回男性會回覆的訊息，是到S4，否到S5。 |
| S3 | 用問候語產生主動回覆並等待對方回覆，對方有回覆到S6，沒有到S7。 |
| S4 | 離開聊天室並回到S1。 |
| S5 | 對方是否回女性會回覆的訊息，是到S9，否到S3。 |
| S6 | 對方是否回男性會回覆的訊息，是到S4，否到S8。 |
| S7 | 等待對方回覆，最多30秒，有回覆到S2，沒回覆到S4。 |
| S8 | 對方是否回女性會回覆的訊息，是到S9，否到S10。 |
| S9 | 結束狀態，將瀏覽器交給使用者。 |
| S10 | 主動回覆”男”，對方回覆到S5，對方離開到S4。 |

表2、各狀態描述

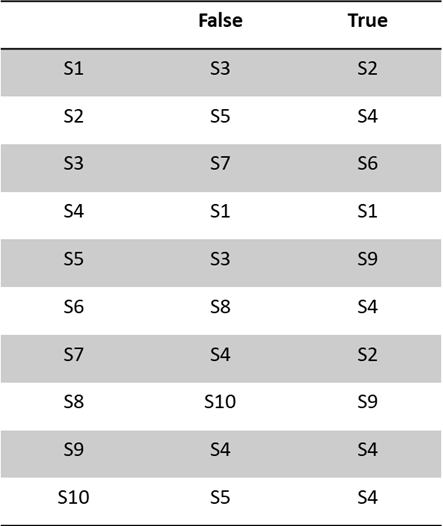


表3、 Strategy狀態轉移表