



學習使用正規表達式(Regular expression) · 過濾及擷取資料


[簡報閱讀](#)[範例與作業](#)[問題討論](#)[學習心得\(完成\)](#)

學習使用正規表達式(Regular expression) · 過濾及擷取資料

**Day 10** 靜態網頁資料爬蟲

學習使用正規表達式(Regular expression)
過濾及擷取資料




出題教練：張齊文

本日知識點目標

本日知識點目標

- 何謂正規表達式(Regular Expression)
- 正規表達式運作及基本語法
- Python正規表達式用法

- 正規表達式是使用一段語法，來描述符合該語法規則的一系列文本。常用簡稱：regex, regexp。
- 正規表達式常用來處理文本資料。例如搜尋、過濾、新增、移除、隔離等功能。

應用場景

假設有一段網頁原始碼如下圖，如何用一種表達式，同時滿足這些有規則的類似字串？

```
<div class="pc-movie-schedule-form"> == $0
  <div class="area_timebox">
    <div class="area_title">台北市</div>
    <ul id="theater_id_29" class="area_time_c jq_area_time" data-theater_name="國賓影城(台北長春廣場)" data-theater_url="http://www.ambassador.com.tw/" data-theater_schedules="https://movies.yahoo.com.tw/theater_result.html/id=29">...</ul>
    <ul id="theater_id_30" class="area_time_c jq_area_time" data-theater_name="欣欣秀泰影城" data-theater_url="http://www.showtimes.com.tw/" data-theater_schedules="https://movies.yahoo.com.tw/theater_result.html/id=30">...</ul>
    <ul id="theater_id_32" class="area_time_c jq_area_time" data-theater_name="台北美麗華大直影城" data-theater_url="http://www.miramarcinemas.tw/" data-theater_schedules="https://movies.yahoo.com.tw/theater_result.html/id=32">...</ul>
    <ul id="theater_id_33" class="area_time_c jq_area_time" data-theater_name="華威天母影城" data-theater_url="http://www.woviecinemas.com.tw/" data-theater_schedules="https://movies.yahoo.com.tw/theater_result.html/id=33">...</ul>
    <ul id="theater_id_34" class="area_time_c jq_area_time" data-theater_name="士林陽明戲院" data-theater_url="http://www.silinmingsheng.com.tw/" data-theater_schedules="https://movies.yahoo.com.tw/theater_result.html/id=34">...</ul>
```

正規表達式運作及基本語法

正規表達式(regex)由兩種字元所組成：

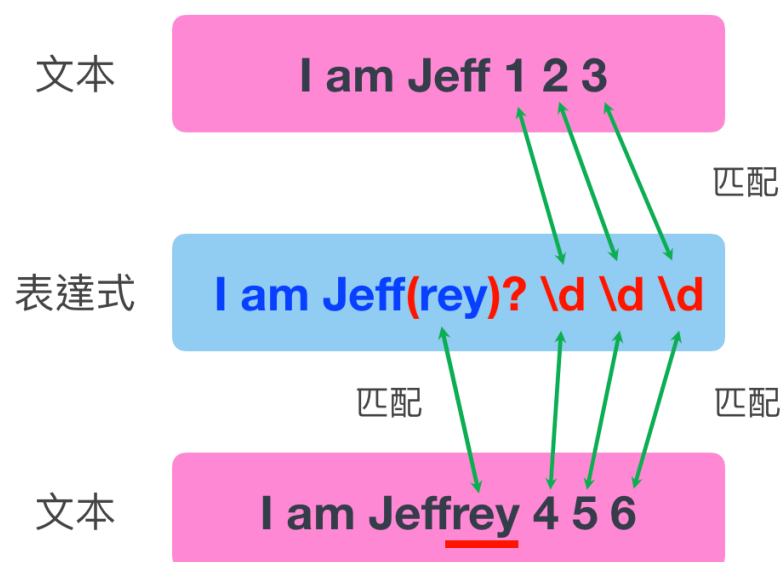
- 詮釋字元(metacharacters)。擁有特殊意義的字元。
- 字面文字(literal)，或稱為一般文字。

可以把 regex 比喻成一段句子。詮釋字元是語法，字面文字是單字。單字+語法=>句子，這句子就是一種表達式。此表達式可用來尋找匹配(matching)此規則的一系列文字。

下圖

藍色為字面文字，regex引擎會逐字比對文本是否匹配。

紅色為詮釋字元，regex引擎會根據其代表的意義，來匹配相對應的字元。



JavaScript regex quick reference (hide):

[abx-z]	One character of: a, b, or the range x-z	^	Beginning of the string
[^abx-z]	One character except: a, b, or the range x-z	\$	End of the string
a b	a or b	\d	A digit (same as [0-9])
a?	Zero or one a's (greedy)	\D	A non-digit (same as [^0-9])
a??	Zero or one a's (lazy)	\w	A word character (same as [_a-zA-Z0-9_])
a*	Zero or more a's (greedy)	\W	A non-word character (same as [^_a-zA-Z0-9_])
a+?	Zero or more a's (lazy)	\s	A whitespace character
a+	One or more a's (greedy)	\S	A non-whitespace character
a+?	One or more a's (lazy)	\b	A word boundary
a{4}	Exactly 4 a's	\B	A non-word boundary
a{4,8}	Between (inclusive) 4 and 8 a's	\n	A newline
a{9,}	9 or more a's	\t	A tab
(?=...)	A positive lookahead	\cY	The control character with the hex code Y
(?!...)	A negative lookahead	\xYY	The character with the hex code YY
(?!. . .)	A non-capturing group	\uYYYY	The character with the hex code YYYY
(...)	A capturing group	.	Any character
		\Y	The Y'th captured group

Python 正規表達式用法

Python 正規表達式模組(re)，其運作流程如下：



編譯



正規表達式字串

正規表達式對象

匹配



所要匹配的文本

匹配結果

```
In [1]: import re #載入re模組

In [2]: regex = 'abcde\s\d+' # 定義正規表達式字串
        pattern = re.compile(regex) # 編譯正規表達式字串，轉換成pattern

In [3]: test_string = "abcde 12345" # 定義所要比對的字串，或是從檔案中讀出字串

# 呼叫re模組所提供的函數來匹配，以下是一些常用的函數
re.match(pattern, test_string) # match函數會從test_string的開頭開始比對
re.search(pattern, test_string) # search與match類似，但search會掃描整個test_string來比對
re.split(pattern, test_string) # 比對test_string並將結果分割
re.findall(pattern, test_string) # 搜尋整個test_string，並將結果以list的形式輸出
re.sub(pattern, '*', test_string) # 替換掉匹配的字串
```

重要知識點複習

- 何謂正規表達式(Regular Expression)
正規表達式是使用一段語法，來描述符合該語法規則的一系列文本。
- 正規表達式運作及基本語法
參考前面講義所列出的詮釋字元用法
- Python正規表達式用法
先將表達式轉換成pattern，再用搜尋函式來比對。

參考資料

目，讓學員們練習regex。匹配的結果會即時顯示，相當適合練習建構regex。

Pythex

- 線上建構regex，並測試結果是否能匹配文本。

常用 Regular Expression 範例

- 常用的regex patterns參考。

解題時間



學習使用正規表達式
(Regular expression)...



本日知識點目標



何謂正規表達式



正規表達式運作及基本語法



Python 正規表達式用法



[下一步：閱讀範例與完成作業](#)

