

Week 7 Regular Expression

用途

validation ex. email · phone number

clean up: 整理化混亂的數據

Extraction: ex 從 url 中獲取用戶名稱

import re

re.search(pattern, string, flags=0): 在字符串中找尋第一個匹配項目

ex. re.search(".+@.+", email, re.IGNORECASE) 忽略大小寫

re.match(pattern, string, flags=0): 從字符串開頭找尋匹配項目

re.fullmatch(pattern, string, flags=0): 嘗試匹配整個字符串

re.sub(pattern, string, count=0, flags=0): 替換所有字符串中的該項目

if find 返回匹配項目

else return None

ex. match = re.search(...)

if match is None:

return False

特殊字元規則範例說明

中國不需角括號，標題等皆

[]	括號內的任何字元	[a-zA-Z0-9]	大小寫字母或數字
[^]	不在括號內的任何字元	[^abc]+	Anything but abc.
.	換行之外的單一字元	..	中間有空白的兩個字
	表示所有可能的條件	a b	a or b , pick one!

Anchor 規則

Anchor	規則	範例	說明
^	輸入的開頭	^aaa	以aaa 開頭
\$	輸入的結尾	bbb\$	以bbb 結尾
()	小括弧內的字元形成群組	(dog)	
\	特殊符號		

ex. match = re.search("^.+@.+\\$")

→ match.group(1) * 從第1開始 i=1,2,3...

→ ex. \com

0 or 1

>= 0

>= 1

重複表示 規則

重複表示	規則	範例	說明
?	未出現或出現一次	ba?	b後面沒有a或一個a
*	未出現或出現多次	ba*	b後面沒有a或多個a
+	出現一次或多次	ba+	b後面有一個以上的a

重複表示 規則

重複表示	規則	範例	說明
{m,n}	最少出現 m 次, 最多出現 n 次	a{3,6}	3到6個a
{n}	出現 n 次	a{3}	剛好三個a
{n,}	至少出現 n 次以上	a{3,}	3個a以上

正規表示式 特別序列 (Special Sequences)

特別序列 說明

\b	單字的界線字元。
\B	字元的界線字元。
\d	數字，從 0 到 9 。
\D	非數字。
\s	各種空白符號，包含換行符號 \n 。 空格、制表符
\S	非空白符號。
\w	任意文字字元，包括數字。 = [a-zA-Z0-9_]
\W	非文字字元，包括空白符號。

原始字符串 (raw strings)

- 加上 r 前綴 ex: r"pattern" 可創建原始字符串
- 將反斜線 \" 當為單面單元

Flags

- re.IGNORECASE : 忽大小寫
- re.MULTILINE : 使 ^ 匹配每行開頭, \$ 每行結尾

將數量算符 :=

- if match := re.search(pattern, string): 可將 re.search 的結果賦值給 matches,並 check if match == True.