我期末報告爬取的網站是 IT 幫幫忙，這個網站主要是跟科技還有資訊相關的問答網站，有點類似 Dcard，只是在這個網站中更著重於資訊相關的問題。在這個網站中有涵蓋的問題像是程式、軟體、資安、硬體等的問題。

我這次要分析的內容是在這個網站中大家最常對哪種類型的內容做提問，所以我就觀察了一下可以當作分析的指標結果發現每個文章標題下面都有「標籤」，因此我就打算以此為指標做分析，然後我原本打算使用數位人文研究平台做分析，但後來遇到了一些問題就改用成其他的分析平台。

在點進去內文時也可以看的到「標籤」，因此我打算爬出來的東西有包括「標題」、 「標籤」以及「內文」，然後我總共爬了 500 頁，大約是三年的資料量。

圖一及圖二是我爬出來的成果。

圖一是內文(程式:FinalProject.py、文本:finalproject.txt)，圖二是所有出現過的標籤名稱(程式:analysis.py 原本只有爬出標籤名稱，後來才加入正規表示式等的分析程式碼、文本:tag_key.txt)，我打算把它們當作權威詞做使用。

標題:在ESXi Server複製VM結果變大了
標籤:vmware esxi vm

大家好
我在VMware ESXi 6.5安裝了數個VM
因為要升級成 VMware ESXi 7.0，所以我就拿了一顆SSD 裝了ESXI7.0 然後把原來裝有VM的硬碟裝上去。再把VM整個搬過去但是發現....
原本的VM整個變大了
原本一開始做VM的時候是規劃150G但是一開始沒用那麼多，所以佔整個硬碟只有幾十G
但是搬到新硬碟後發現它變大到150G
請問這是為什麼？有什麼辦法可以讓它變小嗎？

標題:問答程式跑不出題目
標籤:python入門

```
from untitled5 import Question

choosel=[
    "有沒有想我啊?\n (a)當然摟 想死你了\n (b)怎麼可能 我已經忘記你勒\n",
]
questions = [
    Question(choosel[0],"a"),
]
def test(questions):
    for question in questions:
        answer = input(question.description)
        if answer == question.answer:
            print("太棒了","看你疑惑的眼神,我還以為你忘了我(","給予","一個大大的擁抱")

        if answer != question.answer:
            print("琳達:真沒想到你竟然忘了我,我真的好難過")
            print("琳達跑走了")

class Question:
    def __init__(self,description,answer):
        self.description=description
        self.answer=answer
```

根據影片做了個問答程式 有小改過 但run的時候什麼都沒發生 很好奇怎麼沒跑出題目
請大神解惑

標題:Windows加入AD後，指令詢問(NTP Server、密碼更換為何時?)
標籤:ad

我的作業系統是Windows 10，然後有加入AD，想請教如下問題：
1. 如何用下指令方式，知道我NTP Server是對應到哪一台?

圖一

windows
server
devc++
map
兩片網卡
,一片是adsl用,一片是vpn用,route如何設定
samba
網域控制站
#python
#文字雲
tradingview
pine
oracle
進銷存
paloalto
防火牆
檔案權限
php
mail
return
value
工作排程
notebook
word
mysql
資料管理
dell
r430
bios更新
google
chrome瀏覽器
網路印表機
gmail
ipad2
itunes
網頁伺服器
瀏覽限制
qnap
nas
excel
自動回傳
it大神
@大神幫幫忙
python系列文章
深度學習
#bat

yolov4
cuda
cudnn
vdi
rdp
python
xls
xlsm
win32com
vba
vmware
bridged
dns
網域
svm
esxi
vm
python入門
ad
for
loop
groupby
pandas
python3
dataframe
erp規劃師認證
c#
大括號
dao
程式設計
框架
database
ms
sql
android
app
十全大補
cache
react
前端
line
messaging
api
linebot
login
active

#電腦
mss
server2016
gcb
lgpo
react.js
odoo
sheet
自動編號
vcenter
converter
centos
p2v
unable
to
query
the
live
source
machine
函數
vbi
大神幫幫我
arma
資料結構與演算法
演算法
postfix
dovecot
vue.js
javascript
cpu使用率
cpu製程
cpu演算法
主機板通道
電子簽核整合平台建議
系統更換
swith
router
ap
#伺服器
#excel
#vba
seo
網站地圖
sitemap
proxy

圖二

當我用權威詞做分析後就發現分析的結果有問題，如圖三。因為有些人在為自己的文章取標籤名稱時，有可能會用到一些奇奇怪怪的名稱，同時一字詞出現一定很容易就會跟許多相關的詞做搭配，所以我當時就想要去刪掉一字詞這樣結果可能會比較準確。

權威詞	頻率	再次查詢按鈕	查詢關鍵字按鈕
.	144075	再限縮查詢結果	查詢關鍵字:.
0	75682	再限縮查詢結果	查詢關鍵字:0
t	72487	再限縮查詢結果	查詢關鍵字:t
,	63883	再限縮查詢結果	查詢關鍵字:,
的	61043	再限縮查詢結果	查詢關鍵字:的
2	59764	再限縮查詢結果	查詢關鍵字:2
1	54067	再限縮查詢結果	查詢關鍵字:1
-	49843	再限縮查詢結果	查詢關鍵字:-
a	42991	再限縮查詢結果	查詢關鍵字:a
s	40673	再限縮查詢結果	查詢關鍵字:s

圖三

改善方式 ： 我回去程式中加入了可以過濾掉一字詞的程式。圖四中的 CountWord 就是我用來計算標籤名稱會有多少個字，超過一個字的才會被寫入我用來做分析的權威檔裡。

```
#將dic字典中的key寫入tag_key.txt中
with open("tag_key.txt",mode="a",encoding = "utf-8") as tagname:
    #將dic中的key取出
    for key in dic:
        CountWord = 0 #避免抓取一個字的標籤名稱
        #確認標籤名稱的字數
        for i in key:
            CountWord += 1
        #將大於1個字的標籤名稱寫入檔案中
        if CountWord != 1:
            tagname.write(key+"\n")
```

圖四

這是我重新抓下來後再做一次分析的結果，結果還是會發現它仍有問題，如圖五。因為還是會有字詞容易跟其他字詞做搭配，後來我就開始手動刪除他列出來的這些高頻率的字詞，總共刪了三次，每次看到結果都還是覺得有問題，所以我後來甚至在重抓一遍，把二字詞也排除掉。但後來想到好像這麼做也不是解決問題的好方法，所以我就改用其他工具做分析。

圖五

我改成用 python 做分析，在用 python 分析的時候有用到老師上課教過的正規表示式。(程式:analysis.py、文本:tag.txt)

圖六是我用來分析的程式碼，當中有兩個比較重要的部分，第一個式這個 target 變數，它是用來抓取「標籤:」這整段的文字，在當初我抓文章時就有讓標題標籤跟內容用換行符號做區隔。

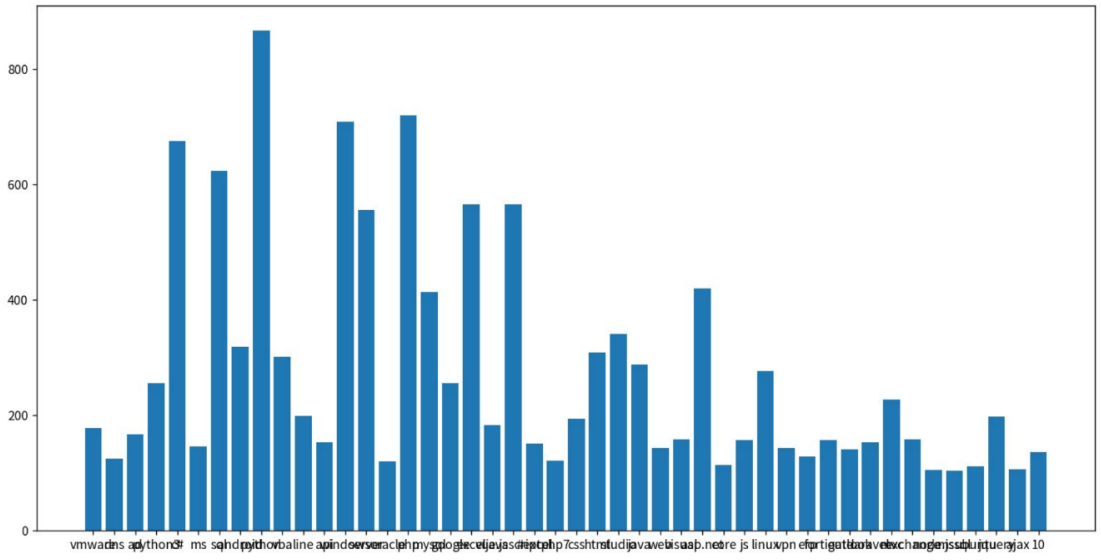
```
pattern = {} #創立空字典，後續作圖使用
#將dic字典中每個key及value寫入tag.txt中
with open("tag.txt",mode="a",encoding = "utf-8") as ftext:
    #將dic中的key取出
    for key in dic:
        ftext.write(key+":") #寫入字典中的key
        ftext.write(str(dic[key])+"\n") #寫入字典中的value
        #只將字典value的值大於100的資料加進pattern裡
        if dic[key] > 100:
            pattern[key] = dic[key] #將pattern這個空字典加入符合條件的資料
```

圖六

第二個比較重要的就是最後一行的程式碼。它可以用來判斷我上面建立的空字典中有沒有出現過當下取得的標籤名稱。如果沒出現過就把它加入字典裡，並把它的 value 設成 1。反之，如果有出現過則直接將該名稱的 value+1。

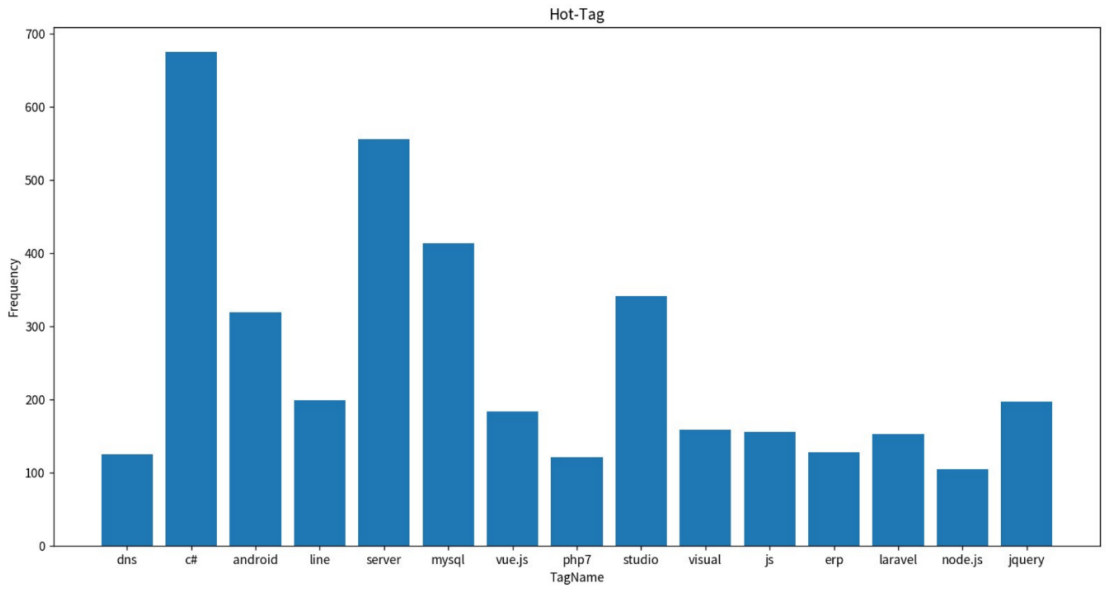
因為抓出來後字典內會有很多筆資料，但有些資料它的 value 值很小我後面打算把它做圖表分析會沒有甚麼意義，所以我就以 100 這個量去做篩選。

我用「matplotlib」這個模組幫我將資料做視覺化的分析，做出來後長得像圖七這樣，X 軸代表標籤的名稱，Y 軸代表每個標籤出現過的數量，因為下面的文字都擠在一起，所以我把他分三份圖。



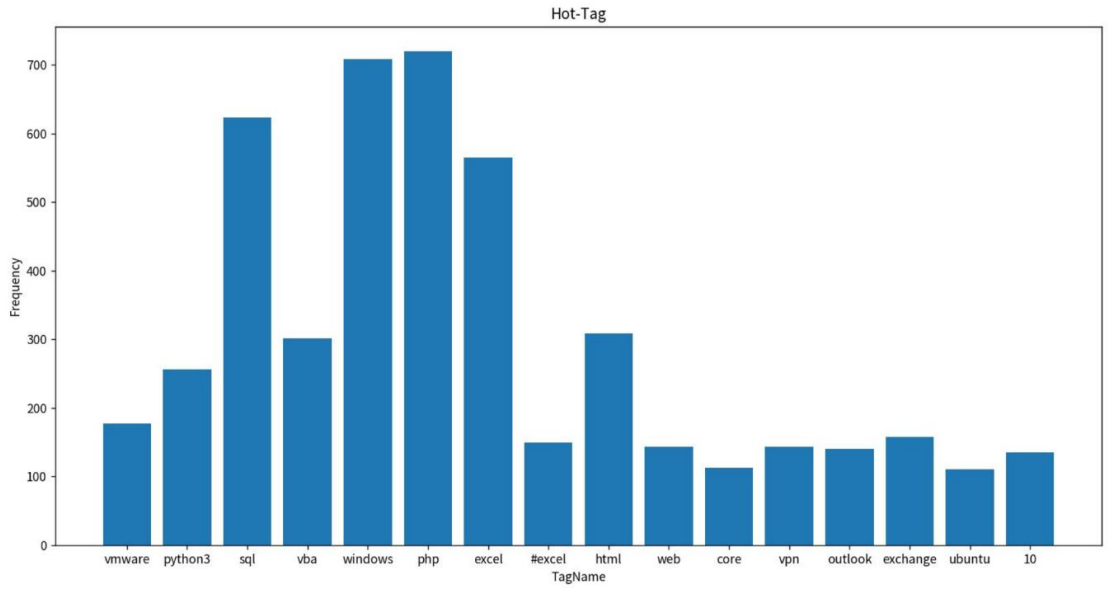
圖七

圖八是第一份拆分後的圖表，從左邊的 Y 軸看來她的最大值是 600 多，標籤名稱是 C#。



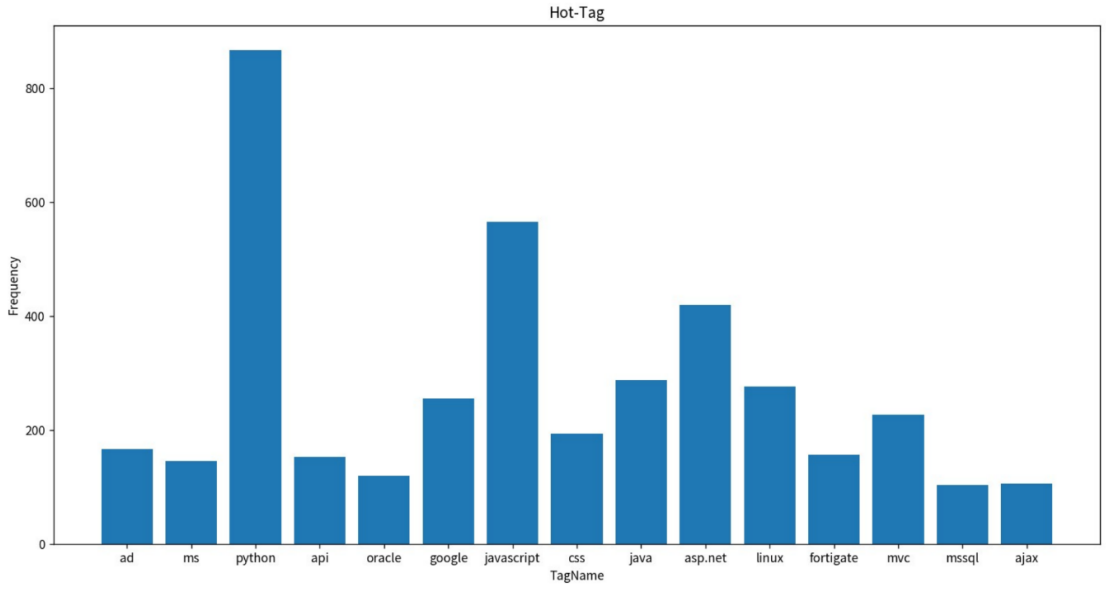
圖八

圖九是第二份拆分後的圖表，從左邊的 Y 軸看來她的最大值是 700 多，標籤名稱是 php。



圖九

圖十是最後一份拆分後的圖表，從左邊的 Y 軸看來她的最大值是 800 多，標籤名稱是 python。



圖十

因此可以從這三張圖中得到結論，就是大家最常詢問的問題是跟 python 有關的問題。

後來我有再多分析 N 字詞(分析的是整個文本的，也就是上面用爬蟲抓下來的文本)。一字詞因為常需要和其他詞作搭配，所以分析 1 字詞比較沒什麼意義，因此我就只有列出 2 到 4 字詞。

我有用黃色標起來我覺得比較有意義的字詞，如圖十一。接下來我每個 N 字詞都舉一個我有做標記的來稍微說明可能會出現這些詞的情況。

2字詞	頻率		3字詞	頻率		4字詞	頻率
標題	15317		各位大	2644		問題標籤	2426
標籤	15245		題標籤	2438		各位大大	1932
問題	10968		問題標	2426		請問各位	1060
資料	10481		的問題	2096		各位前輩	1016
使用	9222		想請問	1969		請問一下	637
可以	8536		位大大	1963		請教各位	635
請問	8421		不知道	1923		問各位大	602
一個	6709		的資料	1874		各位大神	578
各位	6363		資料庫	1775		錯誤訊息	525
如何	5516		程式碼	1663		各位先進	472
設定	5248		使用者	1538		的程式碼	453
沒有	5114		有沒有	1521		的問題標	428
程式	4936		資料夾	1285		一個問題	421
目前	4706		問各位	1237		這個問題	419
無法	3904		伺服器	1225		方法可以	386
謝謝	3601		該如何	1166		想請問各	386
檔案	3455		個問題	1137		可以正常	386
網路	3307		請問各	1079		位大大好	376
電腦	3302		想請教	1043		如何解決	353
顯示	3223		請問有	1042		請教一下	347

圖十一

二字詞我以"設定"為例，可能大家會問像密碼的設定，安裝資料的設定，網卡設定之類的，如 Proof 1、Proof 2。

SERVER 兩片網卡 ,一片是ADSL用,一片是VPN用,ROUTE如何設定

server 兩片網卡 ,一片是adsl用,一片是vpn用,route如何設定

hcg1256 2022-06-09 18:25:23 · 251 瀏覽

讚 0

一片網卡(ADSL上網用) - IP:192.168.2.90 子網路遮罩:255.255.255.0 預設閘道:192.168.2.1
另一網卡(實體VPN中華電信) - IP:192.168.9.90 子網路遮罩:255.255.255.0 預設閘道:192.168.9.1 , 高雄VPN閘道:192.168.10.1
如何設定ROUTE?又可上網,高雄的Client 又可以使用這台server

Proof 1

Windows加入AD後，指令詢問(NTP Server、密碼更換為何時?)

ad

klm2242 2022-06-10 13:58:59 · 218 瀏覽

讚 0

我的作業系統是Windows 10，然後有加入AD，想請教如下問題：
1、如何用下指令方式, 知道我NTP Server是對應到哪一台?
2、如何在我這台電腦前下指令，知道我上一次密碼更換為何時?及一些密碼原則設定?
3、若無法下指令狀況之下的話，在我PC前，可以看的我自己本身帳號的一些相關設定嗎?
以上幾點，再請各位前輩回覆，謝謝!

Proof 2

三字詞我以"伺服器"為例，可能就是詢問某種伺服器無法登入的問題，或是伺服器更新之類的，如 Proof 3、Proof 4。

在網域控制站的事件檢視器出現一堆事件識別碼5774的事件

active directory　samba　網域控制站　dns

vinix 2022-06-09 18:19:36 · 184 瀏覽

讚 0

在Windows Server 2012 R2網域控制站(dc2)的事件檢視器中，每個幾個小時就會出現類似如下的錯誤：

> 在下列 DNS 伺服器上動態登錄 DNS 記錄 '_ldap._tcp.Default-First-Site-Name._sites.gc._msdcs.domainname.com. 600 IN SRV 0 100 3268 dc2.edtung-internal.com.' 失敗:
>
> DNS 伺服器 IP 位址: 192.168.1.10
> 傳回的回應碼 (RCODE): 0
> 傳回的狀態碼: 9016
>
> 對於要用來尋找這個網域控制站的電腦及使用者，這個記錄必須 在 DNS 中登錄。
>
> 使用者動作
> 判定造成這個失敗的可能原因、解決問題、並由網域控制站初始 DNS 記錄的登錄。如

Proof 3

Dell R430 BIOS更新問題

dell　r430　bios更新

wind2124 2022-06-08 18:16:43 · 285 瀏覽

讚 0

各位前輩們好

小弟公司有一台R430因為使用VMware6.5想要升級到7.0u3

但這台伺服器的bios還停留在2017的版本看了下官網的相容性需要以下版本
Dell Inc. 2.11.0 (Boot Mode:Legacy BIOS)
Dell Inc. 2.10.5 (Boot Mode:Legacy BIOS)
Dell Inc. 2.2.5
Dell Inc. 2.2.5

目前看官網最新為2.13.0版，但問題來了

官網提供的bios版本檔案副檔名為XXX.BIN

我放到USB內後進入Lifecycle Controller選擇固件更新/從USB/打上檔案

Proof 4

四字詞以錯誤訊息來說，問問題的人可能會給一段錯誤訊息的程式碼問大家要怎麼解決，如 Proof 5、Proof 6。

使用Python網路爬蟲輸出CSV時出現I/O operation on closed file.錯誤

pythonm 網路爬蟲

baltic 2022-05-30 21:38:10 · 226 瀏覽

讚 0

我最近在做網路爬蟲，我爬蟲的網址是這個https://www.jkf.net/lady/2021/rank.php ，我有成功把資料爬下來，但是我要輸出成CSV檔案，卻發生了錯誤，錯誤訊息是 ValueError: I/O operation on closed file.

我的程式如下:

```
import csv
import requests
from bs4 import BeautifulSoup

url = 'https://www.jkf.net/lady/2021/rank.php'
response = requests.get(url=url)

soup = BeautifulSoup(response.text, 'lxml')

info_items = soup.find_all('li', 'in_fade')
```

Proof 5

CentOS實體機轉虛擬機

vcenter converter centos p2v unable to query the live source machine

acoldliu 2022-06-08 08:53:01 · 454 瀏覽

讚 0

跪求指點迷津？

我有三台CentOS 7.9的實體機，因為維護需求所以必須轉成虛擬機
爬文找到的方法是用VMware vCenter Converter工具去作線上實體機轉換
1.轉換機(安裝Converter)
2.來源機(CentOS)
3.目的機(Esxi)
均在同網段

透過轉換機進行轉換時，會發生unable to query the live source machine之錯誤訊息
關於此錯誤也有到VMware的KnowlegeBase爬文過
https://kb.vmware.com/s/article/1009153
此篇文章的作法也有嘗試過(但當我列出mount | grep /tmp時，/dev/裡的路徑卻沒有類似此篇文章的情形，不過我還是有進行文中提到的掛載步驟)
仍然會出現上方敘述的錯誤訊息

Proof 6

以上是我簡單列舉的幾個例子，實際上發問的內容還是要依據各自的需求。

與作業 2 之相關性與差異:

爬蟲的部分多加了以下程式用來取得標籤的名稱，當然不只這一行程式碼，還有與其相關要整理並寫入檔案中的程式。

(程式:FinalProject.py、文本:finalproject.txt)

```
tags = root2.find("div",class_="qa-header__tagGroup") #以列表形式找出所有 class_=qa-header__tagGroup 的 div
```

報告後增修:

1. 權威詞的部分可以找出利用程式抓出英文的 N 字詞:

這個部份我確實有利用程式抓出只有英文的字詞，但利用權威詞統計出來的結果仍不佳。此原因在於我抓出來的字詞還是會出現許多單詞(例如:a,b,c,d 這種詞)，即使沒有單詞兩個英文組成的詞或更多英文字母組成的詞仍會因為有囊括到其他字詞的可能(像是 in 這個由 2 個英文字母組成的詞，會因為像是 print, define, line, insert, running 等的英文字而被重複計算)分析出來的效果仍不佳。(程式:power.py、文本:TagContent.txt、權威詞:power.txt)

2. 除了分析最熱門的問題外，可以再做進一步的分析:

這個部份我有再多寫一個程式去抓出另外的文章，而抓出來的新文章是跟某個領域相關。舉例來說像是我以關鍵字「python」為例，抓出來的文章都會跟 python 相關，這樣就可以進一步分析與其相關的資料。(程式:FindTagContent.py、文本: TagContent.txt)

與原本混雜著許多領域的文本相比，在新的文本中利用「一般查詢」功能可以較容易找出相關連性。

數位人文研究平台帳號:41071102H，密碼:tony20030201