

爬蟲應用相關模組套件

Web Crawler Related Packages



Python Web Crawler

webbrowser 模組

◆ Python提供webbrowser模組，用來瀏覽網頁，我們呼叫這個模組的open()方法，就可開啟指定網頁。

```
import webbrowser

webbrowser.open('http://tw.yahoo.com')

#chrome_path = 'C:/Program Files(x86)/Google/Chrome/Application/chrome.exe %s'

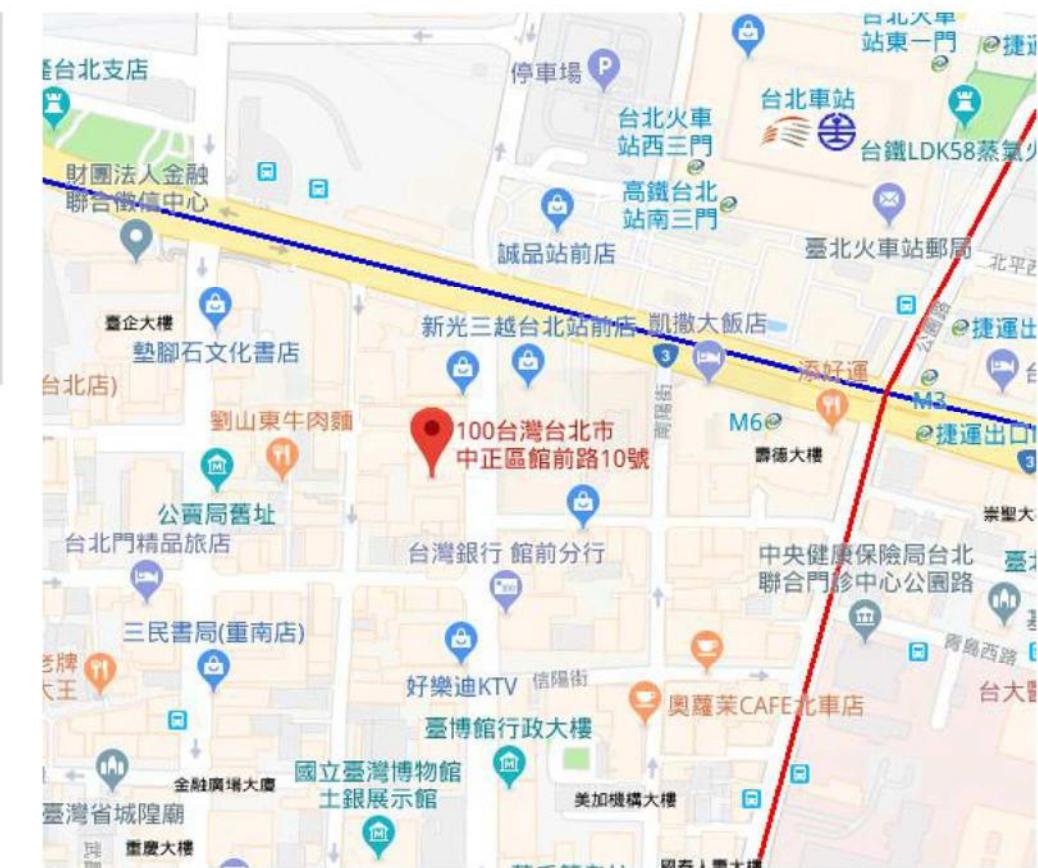
#webbrowser.get(chrome_path).open('http://tw.yahoo.com')
```



webbrowser 模組

- ◆ 我們可以進一步利用webbrowser模組，配合Google地圖服務，來進行地址查詢地圖的功能。
 - ◆ 只要讀取地址資訊，然後放在open()參數內與Google 地圖服務的網址連接即可。

```
import webbrowser  
  
address = input("請輸入地址:")  
  
webbrowser.open('http://www.google.com.tw/maps/place/' + address)
```



常見的爬蟲套件

常見的爬蟲套件有以下幾種，以下我們將會介紹urllib、request與BeautifulSoup。

- ◆ urllib
- ◆ requests
- ◆ BeautifulSoup
- ◆ Selenium



Urllib Request



Requests



urllib

urllib 套件簡介

- ◆ 當我們想要從網站抓取網頁並解析資料時，就要向伺服器發出HTTP請求 (request)。
- ◆ 發出HTTP請求與伺服器溝通是程式語言不可或缺的能力，在Python可使用urllib以及requests套件。
- ◆ urllib套件包含四個主要模組：`request` 用於模擬發送請求；`parse` 提供url處理；`error` 處理異常模組；`robotparser` 處理網站的reboot.txt。
- ◆ urllib套件底下有兩個主要函式，可以進行網頁擷取與網址解析：
 - `urlopen`：進行網頁擷取
 - `urlparse()`：進行網址解析

urllib module

urllib.request.urlopen() 取得網頁資料

- ◆ urllib套件的urlopen()函式可以針對網頁進行擷取，其語法如下：

urlopen(網址)

- ◆ 執行 urlopen() 函式，會回傳一個urllib.response物件。假設其物件名稱為x，則其屬性/函式資料如下：

屬性/函式	說明	使用方式
read()	網址的通訊協定	x.read()
geturl()	網站網址	x.geturl()
getheader()	網站路徑	x.getheader()
status	查詢url的參數字串	x.status

urllib.request.urlopen() 取得網頁資料

- ◆ urllib套件中的urlopen()函式可以將網頁內容擷取至程式。
- ◆ 以下例子就是讀取http://python.org 的網頁資料(HTML)並輸出。

```
import urllib.request  
  
response=urllib.request.urlopen('http://python.org/')  
  
html=response.read()  
  
print(html)
```

```
In [1]: runfile('C:/Users/Alex/Desktop/untitled0.py', wdir='C:/Users/Alex/Desktop')  
b'<!doctype html>\n<!--[if lt IE 7]>    <html class="no-js ie6 lt-ie7 lt-ie8 lt-ie9">  
<![endif]-->\n<!--[if IE 7]>        <html class="no-js ie7 lt-ie8 lt-ie9">      <!  
[endif]-->\n<!--[if IE 8]>        <html class="no-js ie8 lt-ie9">          <!  
[endif]-->\n<!--[if gt IE 8]><!--><html class="no-js" lang="en" dir="ltr">  <!--<!  
[endif]-->\n\n<head>\n    <meta charset="utf-8">\n    <meta http-equiv="X-UA-Compatible"  
content="IE=edge">\n    <link rel="prefetch" href="//ajax.googleapis.com/ajax/libs/  
jquery/1.8.2/jquery.min.js">\n    <meta name="application-name" content="Python.org">  
\n    <meta name="msapplication-tooltip" content="The official home of the Python  
Programming Language">\n    <meta name="apple-mobile-web-app-title"  
content="Python.org">\n    <meta name="apple-mobile-web-app-capable" content="yes">\n    <meta name="apple-mobile-web-app-status-bar-style" content="black">\n    <meta  
name="viewport" content="width=device-width, initial-scale=1.0">\n    <meta  
name="HandheldFriendly" content="True">\n    <meta name="format-detection"  
content="telephone=no">\n    <meta http-equiv="cleartype" content="on">\n    <meta http-  
equiv="imagetoolbar" content="false">\n    <script src="/static/js/libs/  
modernizr.js"></script>\n    <link href="/static/stylesheets/style.css"  
rel="stylesheet" type="text/css" title="default" />\n    <link href="/static/  
stylesheets/mq.css" rel="stylesheet" type="text/css" media="not print, braille,  
embossed, speech, tty" />\n    <!--[if (lte IE 8)&(!IEMobile)]>\n    <link  
href="/static/stylesheets/no-mq.css" rel="stylesheet" type="text/css" media="screen" />  
\n    <!--[endif]-->\n    <link rel="icon" type="image/x-icon" href="/  
static/favicon.ico">\n    <link rel="apple-touch-icon-precomposed" sizes="144x144"  
href="/static/apple-touch-icon-144x144-precomposed.png">\n    <link rel="apple-touch-  
icon-precomposed" sizes="114x114" href="/static/apple-touch-icon-114x114-  
precomposed.png">\n    <link rel="apple-touch-icon-precomposed" sizes="72x72" href="/  
static/apple-touch-icon-72x72-precomposed.png">\n    <link rel="apple-touch-icon-  
precomposed" href="/static/apple-touch-icon-precomposed.png">\n    <link rel="apple-  
touch-icon" href="/static/apple-touch-icon-precomposed.png">\n\n    <meta
```

urllib.request.urlopen() 取得網頁資料_2

- ◆ 使用urlopen() 函式，連結中央大學網頁擷取網頁資訊。網址及網頁如下

<https://www.ncu.edu.tw/>

The screenshot shows the homepage of the National Central University (NCU) website. At the top left is the university's logo and name. To the right is a weather widget showing current conditions: 30.3 °C and 0.0 mm/hr. Navigation links include '中大Portal', '校園活動', and '捐贈中大'. A search bar is located at the top right. The main banner features a scenic view of the university campus and a 'TOP 4 IN TAIWAN' ranking from US News. A black sidebar on the right displays a congratulatory message about NCU's achievement in the 2019 US News global ranking. Below the banner, there are three news thumbnails: one about the award ceremony for the 'Chen Shun-ying Outstanding Student Award', another about the Kunqu Opera Museum's 20th anniversary, and a third about the International Atmosphere Radar School.

| 中大新聞 [更多]

朱順一合勤獎學金頒獎典禮 以「教育」逆轉人生

朱順一合勤獎學金於11月27日在中央大學舉行頒獎典禮，獎勵12位學業優良、2位運動績優的學生。全體獲獎同學非常感謝朱順一董事長的獎助，期許自己未來有能力時，貢獻所學回饋社會，幫助更多優秀年輕人圓夢！

老戲迷・新創意 中大崑曲博物館慶賀週年

古老的崑曲藝術，如何在博物館展現新創意？中央大學崑曲博物館為慶賀成立一週年，結合新一期「樂府傳聲——崑劇傳字輩特展」，推出崑曲水袖多媒體互動體驗，除了欣賞珍貴文物，還可以學崑曲，體驗崑曲之美。

ISAR-NCU TAIWAN 12th-16th November 2018 國際大氣雷達學校

提昇第三世界研究水準 2018國際大氣雷達學校在中大

ISAR-NCU-2018 國際大氣雷達學校今年11月於本校太空及遙測研究中心舉行，主要目的是提昇第三世界的研究水準。本次共計19位國外學員參與，分別來自泰國、馬來西亞、菲律賓、越南、印尼及印度等六個國家。

urllib.request.urlopen() 取得網頁資料_2

- ◆ 此例子會擷取網頁的「網址」、「讀取狀態」、「網頁表頭」、「網頁資料」等。
- ◆ 擷取下來的網頁格式分別以Byte格式與字串格式呈現其差異，如下圖所示。

```
1 #下載中央大學網頁資訊
2 import urllib.request as ur
3 url='http://www.ncu.edu.tw/'
4 res=ur.urlopen(url)
5 print('1. 網址：',res.geturl())
6 print('2. 讀取狀態：',res.status)
7 print('3. 網頁表頭資訊：',res.getheaders())
8 content=res.read()
9 print('4. 網頁資料(Byte格式)：',content)
10 print()
11 print()
12 print()
13 print('5. 網頁資料(字串格式)：',content.decode())
```

1. 網址：<https://www.ncu.edu.tw/>
2. 讀取狀態：200
3. 網頁表頭：[('Date', 'Mon, 03 Dec 2018 01:27:59 GMT'), ('Server', 'Apache'), ('Vary', 'Accept-Encoding'), ('X-Frame-Options', 'SAMEORIGIN'), ('Connection', 'close'), ('Transfer-Encoding', 'chunked'), ('Content-Type', 'text/html; charset=UTF-8')]
4. 網頁資料(Byte方式)：`b'<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict//EN" "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd">\r\n<html xmlns="http://www.w3.org/1999/xhtml" class="no-js">\r\n<head>\r\n<link rel="SHORTCUT ICON" href="/assets/images/NCU.ico" />\r\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />\r\n<meta name="google-site-verification"`

urlopen()範例說明

1. 匯入urllib.request套件，並設定網址給變數url。(line2~3)
2. 使用urllib.parse.urlopen()函式取得變數url的網頁內容，將結果指定給uo物件並輸出(line4~5)
3. 輸出uo物件的status屬性，此處會印出「200」，表示資料的傳輸成功。(line6)
4. 使用getheaders()函式取得網頁的表頭資訊並輸出。(line 7)
5. 使用read()函式取得網頁的原始內容並輸出，預設格式為Byte。(line 8~9)
6. 使用decode()函式將Byte格式的網頁資料轉成字串後輸出。(line 10)

```
1 #下載中央大學網頁資訊
2 import urllib.request
3 url='http://www.ncu.edu.tw/'
4 uo=urllib.request.urlopen(url)
5 print('1. 網址:',uo.geturl())
6 print('2. 讀取狀態:',uo.status)
7 print('3. 網頁表頭:',uo.getheaders())
8 content=uo.read()
9 print('4. 網頁資料(Byte方式):',content)
10 print()
11 print()
12 print()
13 print('5. 網頁資料(字串方式):',content.decode())
```

requests

requests 套件簡介

- ◆ requests是一個模擬html request功能的第三方套件。
- ◆ 使用get()函式，可以讀取網頁的資料，get()函式會對伺服器(Server)提出取得網頁資料的請求(Request)，伺服器接到請求後，回應(Response)網頁的原始碼內容。
- ◆ requests套件的一些物件資料項目如下：

屬性/函式	說明
status_code	請求結果代碼，例如：200(請求成功)、404(找不到)、.....
headers	請求所回傳的標頭內容，其型態為dict
encoding	所回傳內容的編碼方式
text	所回傳的文字內容



HTTP 回應狀態一覽

- ◆ 200：此次HTTP交談傳輸成功了，沒有出錯
- ◆ 401(Unauthorized，用戶未經認證)：尚未通過認證就嘗試存取動作
- ◆ 400(Bad Request，錯誤的請求)：以不正確的方式存取網站伺服器
- ◆ 403(Forbidden，被禁止)：類似於401，但沒有登入的可能性
- ◆ 404(Not Found，找不到)：嘗試存取根本不存在的網頁
- ◆ 500 (Internal Server Error，伺服器內部錯誤)：代表其他所有種類的泛用型錯誤狀態碼

requests 套件相關函式

- ◆ requests 套件相關函式如下：

```
r=rq.request(method, url, **kwargs)
#取得header
r_header=rq.head(url, **kwargs)
#取得資料
r_get=rq.get(url, params=None, **kwargs)
#新增資料
r_post=rq.post(url, data=None, json=None, **kwargs)
#替換資料(新增或完整更新資料)
r_put=rq.put(url,data=None, **kwargs)
#部分更新資料
r_patch=rq.patch(url, data=None, **kwargs)
#刪除資料
r_delete=rq.delete(url, **kwargs)
#取得可用的http方法
r_options=rq.options(url, **kwargs)
```



requests 套件 - 網頁擷取範例.1

- ◆ requests 非常便利，輕鬆就能抓取網站資料，再放入程式中或是Python工作階段。
- ◆ 底下範例就從開放資料網站(<https://data.gov>)抓取資料。請注意要檢查status_code是否為 200，確認是否有抓取成功。
- ◆ 若是JSON格式的話，requests可以使用Python字典容器的介面存取HTTP回應資料，裡頭含有資料與資料項目串列。

```
import requests
response=requests.get("https://data.ntpc.gov.tw/od/data/api/E09B35A5-A738-48CC-B0F5-570B67AD9C78?format=json")
print(response.status_code)
data=response.json()
print(data[0])
print(data[1])
print(data[2])
```

```
200
{'ID': '010001', 'AVAILABLECAR': '41'}
{'ID': '010002', 'AVAILABLECAR': '-9'}
{'ID': '010003', 'AVAILABLECAR': '-9'}
```

requests 套件 - 網頁擷取範例.2

- ◆ 以get()函式讀取網頁的原始碼內容，以政府資料開放平台為例，其網址為 <https://data.gov.tw>。
- ◆ 在使用get()函式後，獲得的網頁原始碼部分內容如下。



```
1 #從網頁擷取原始碼內容
2 import requests
3 url='https://data.gov.tw/'
4 html_body=requests.get(url)
5 html_body.encoding='utf-8'
6 print(html_body.text)
```

```
<!DOCTYPE html>
<html lang="zh-hant" dir="ltr">
<head profile="http://www.w3.org/1999/xhtml/vocab">
<meta charset="utf-8">
<meta http-equiv="X-UA-Compatible" content="IE=edge">
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta name="MobileOptimized" content="width" />
<meta name="HandheldFriendly" content="true" />
<meta name="viewport" content="width = device-width, initial-scale = 1.0, minimum-scale = 1, maximum-scale = 1, user-scalable = no" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-60x60.png" rel="apple-touch-icon" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-76x76.png" rel="apple-touch-icon" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-120x120.png" rel="apple-touch-icon" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-152x152.png" rel="apple-touch-icon" />
<link rel="shortcut icon" href="https://data.gov.tw/sites/datagov/themes/thetheme/favicon.ico" type="image/vnd.microsoft.icon" />
<title>政府資料開放平臺 | 政府資料開放平臺</title>
```

requests 套件 - 網頁擷取範例.2

1. 汇入requests套件。(line 2)
2. 将政府資料開放平台之網指定給變數 url。(line 3)
3. 使用requests套件的get()函式取得網頁原始碼內容。(line 4)
4. 設定網頁編碼為「utf-8」。(line 5)
5. 印出該網頁的原始碼內容。(line 6)

```
1 #從網頁擷取原始碼內容
2 import requests
3 url='https://data.gov.tw/'
4 html_body=requests.get(url)
5 html_body.encoding='utf-8'
6 print(html_body.text)
```

```
<!DOCTYPE html>
<html lang="zh-hant" dir="ltr">
<head profile="http://www.w3.org/1999/xhtml/vocab">
  <meta charset="utf-8">
  <meta http-equiv="X-UA-Compatible" content="IE=edge">
  <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta name="MobileOptimized" content="width" />
<meta name="HandheldFriendly" content="true" />
<meta name="viewport" content="width = device-width, initial-scale = 1.0, minimum-scale = 1, maximum-scale = 1, user-scalable = no" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-60x60.png" rel="apple-touch-icon" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-76x76.png" rel="apple-touch-icon" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-120x120.png" rel="apple-touch-icon" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-152x152.png" rel="apple-touch-icon" />
```

requests 套件 - 網頁擷取範例.3

```
import requests

url = 'https://data.gov.tw/'

res = requests.get(url)                      # 產生回傳物件 res

if res.status_code == requests.codes.ok: # 回傳物件狀態
    print("取得網頁內容成功")
else:
    print("取得網頁內容失敗")

print(res.text)                             # 顯示網頁內容(原始碼)

print('網頁內容大小:', len(res.text))       # 顯示網頁大小
```

```
In [1]: runfile('C:/Users/Alex/Desktop/untitled1.py', wdir='C:/Users/Alex/Desktop')
取得網頁內容成功
<!DOCTYPE html>
<html lang="zh-hant" dir="ltr">
<head profile="http://www.w3.org/1999/xhtml/vocab">
    <meta charset="utf-8">
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
<meta name="MobileOptimized" content="width" />
<meta name="HandheldFriendly" content="true" />
<meta name="viewport" content="width = device-width, initial-scale = 1.0, minimum-scale = 1, maximum-scale = 1, user-scalable = no" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-60x60.png" rel="apple-touch-icon" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-76x76.png" rel="apple-touch-icon" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-120x120.png" rel="apple-touch-icon" />
<link href="/sites/datagov/themes/thetheme/images/apple-touch-icon-152x152.png" rel="apple-touch-icon" />
<link rel="shortcut icon" href="https://data.gov.tw/sites/datagov/themes/thetheme/favicon.ico" type="image/vnd.microsoft.icon" />
<title>政府資料開放平臺 | 政府資料開放平臺</title>
<link type="text/css" rel="stylesheet" href="https://data.gov.tw/sites/datagov/files/css/css_1QaZfjVpwP_oGNqdtWCSpJT1EMqXdMiU84ekLLxQnc4.css" media="all" />
<link type="text/css" rel="stylesheet" href="https://data.gov.tw/sites/datagov/files/css/css_aVIJ7_1N_en1st3uEjKpqa0c38dcjDoECmLS0RmmT-I.css" media="all" />
<link type="text/css" rel="stylesheet" href="https://data.gov.tw/sites/datagov/files/css(css_PNK0Up2mTwJ7V7xDtJ3lT5nyPpjW5ZSJ6uBQgg1IYQE.css" media="all" />
<link type="text/css" rel="stylesheet" href="https://data.gov.tw/sites/datagov/files/css/css_VGhXE00ZN6KL74X_YguXPQGR8m8wzJW22WYXUZeYt6M.css" media="all" />
<script src="https://data.gov.tw/sites/datagov/files/js/js_xvYJgU6LChHqbcSh4y1AvdXfD5QBIwT3GVGVUeuksbM.js"></script>
<script>document.createElement( "picture" );</script>
<script src="https://data.gov.tw/sites/datagov/files/js/js_RjHLIt2ubvq_rxC9A8eAGN368iBRymHeCN60H9r7Ywg.js"></script>
<script>(function(i,s,o,g,r,a,m){i["GoogleAnalyticsObject"] = r;i[r]=i[r]||function(){
    (i[r].q=i[r].q||[]).push(arguments),i[r].l=1*new Date();a=s.createElement(o),m=s.getElementsByTagName(o)[0];a.async=1;a.src=g;m.parentNode.insertBefore(a,m)})(window,document,"script","https://www.google-analytics.com/analytics.js","ga");ga("create", "UA-74374928-2", {"cookieDomain":"auto"});ga("set", "anonymizeIp", true);ga("send", "pageview");</script>
<script src="https://data.gov.tw/sites/datagov/files/js/js_zNhtv2vchr2qpcDnIVurDuqYwQnuL65pE0fpH_XmHf0.js"></script>
<script src="https://data.gov.tw/sites/datagov/files/js/js_Qvc34jnCdaHTAF-rd6i6U_u0BZzo0dpYarc7DJ2nIo0.js"></script>
```

requests 套件 - 搜尋指定字串

- ◆ 設計一個Python程式，讓使用者先輸入要搜尋的網頁網址，再輸入在該網頁要搜尋的字串。

請輸入您要搜尋的網址:<http://www.thu.edu.tw>

請輸入您要搜尋的字串:東
「東」字串在網頁中找到143筆!

requests 套件 - 搜尋指定字串

1. 匯入requests套件。(line 2)
2. 使用input()函式讀入使用者輸入網址並指定給變數 url。(line 3)
3. 使用requests套件的get()函式取得指定網頁的原始碼內容。(line 4)
4. 設定網頁編碼為「utf-8」。(line 5)
5. 使用splitlines()函式，去除換行符號後將每一列存成串列。(line 6)
6. 將字串個數的計算變數n設為「0」。(line 7)
7. 使用input()函式讀入使用者輸入的字串並指定變數keyword。(line 8)
8. 每列去尋找keyword字串，找到則將變數n的值加1 (line 9-11)
9. 印出該字串在網中出現的筆數。(line 12)

```
1 #從網頁擷取原始碼內容
2 import requests
3 url=input('請輸入您要搜尋的網址：')
4 html_body=requests.get(url)
5 html_body.encoding='utf-8'
6 htmllist=html_body.text.splitlines()
7 n=0
8 keyword=input('請輸入您要搜尋的字串：')
9 for row in htmllist:
10     if keyword in row:
11         n+=1
12 print('『%s』字串在網頁中找到%s筆！' %(keyword,n))
```

Q & A