

情資爬蟲介紹與實作

李炅倫

108年12月23日

Outline

- 利用requests 取得網站資訊
- 使用Beautifulsoup解析網站
- 情資爬蟲實作－PTT爬蟲
- 情資搜集系統－Elasticsearch建置與應用

Python 環境安裝-mac

- Python版本 3.7.3
- 先安裝 Homebrew
 - `/usr/bin/ruby -e "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/master/install)"`
 - `brew install python3==3.7.3`
- 利用 pip 安裝套件
 - `pip3 install requests`
 - `pip3 install jupyter`
 - `pip3 install BeautifulSoup4`

Python 環境安裝-ubuntu

- Python版本 3.7.3
- `sudo apt-get update`
- `sudo apt-get install python3.7`
- `sudo apt install python3-pip`

python套件安裝-pip3語法介紹

- pip3是python 的套件管理工具
- 可以用來管理當前安裝的python 套件
- pip3 list 列出所有安裝的項目
- pip3 install 套件名稱
- pip3 uninstall 套件名稱

Get vs. Post (2/2)

- GET : 發送requests , Server 回傳資料
 - URL 會隨著不同的網頁改變
 1. <http://www.taipeibo.com/yearly/2017>
 2. <http://www.taipeibo.com/yearly/2016>
- POST : 發送requests並附帶資料 , Server 回傳資料
 - 網址不會改變 , 但是網頁內容會隨著使用者不同的requests而改變

爬蟲原理

- 利用呼叫http 連上網站請求先取得網站原始碼
- 將原始碼根據所想要的資訊進行解析
- 解析完之後將資料轉成所想要的資料型態儲存
- 爬蟲腳本自動化
- 將爬蟲程式部署並與資料庫串接

利用request取得網站資訊

- 可以利用request 呼叫http method 與網站溝通，取得網站資料
- request 可以使用許多的http請求
- GET 請求範例

```
# 引入 requests 模組
import requests

# 使用 GET 方式下載普通網頁
r = requests.get('https://www.google.com.tw/')
```


BeautifulSoup + regular expression

- BeautifulSoup 可以幫你解析HTML
- regular expression 可以按照你所制定的規則回傳字串
- BeautifulSoup + regular expression 即可以利用所制定的規則取出目標HTML的標籤

爬蟲範例：ptt網站爬取

- 取得八卦版的今日發文
- 文章的標題
- 連結網址
- 作者
- 內文
- 並將所取得資訊變成json檔

ptt網站爬取- 設定cookie(1/2)

- 初次進入ptt web版時，會詢問是否滿18歲，這其實是一個設定cookie的動作

本網站已依網站內容分級規定處理

警告：您即將進入之看板內容需滿十八歲方可瀏覽。

若您尚未年滿十八歲，請點選離開。若您已滿十八歲，亦不可將本區之內容派發、傳閱、出售、出租、交給或借予年齡未滿18歲的人士瀏覽，或將本網站內容向該人士出示、播放或放映。

我同意，我已年滿十八歲
進入

未滿十八歲或不同意本條款
離開

ptt網站爬取- 設定cookie(2/2)

- Over18的value設定成1即表示已確認過

The screenshot shows the Chrome DevTools Application tab. The left sidebar is expanded to 'Cookies' for the URL 'https://www.ptt.cc'. The main pane displays a table of cookies:

Name	Value
__cfduid	df92a1dd91770e5eea27089184ae1a5821557290474
_ga	GA1.2.935473292.1557290475
_gat	1
_gid	GA1.2.1392886811.1567561322
over18	1

爬蟲範例：ptt網站爬取流程

- 從首頁進入
- 設定cookie
- 取得此頁所有文章的連結
- 對各文章原始碼進行處理

從首頁進入並設定cookie

```
In [1]: import requests
        from bs4 import BeautifulSoup
        from datetime import timedelta
        import json

        #從首頁進入並設定cookie

        ptt = "https://www.ptt.cc"
        url = 'https://www.ptt.cc/bbs/Gossiping/index.html'
        s = requests.Session()
        s.post(ptt + "/ask/over18", data={'yes': 'yes'})
        page = s.get(url).text
```

先取得文章標題位置

- 文章標題目標位於 <div class="title" 中>

```
In [3]: ##發現文章位於 div 標籤的 r-ent class先往下解析一層
soup = BeautifulSoup(page, features="html.parser")
divs = soup.find_all('div', class_='r-ent')
for div in divs:
    print(div)
```

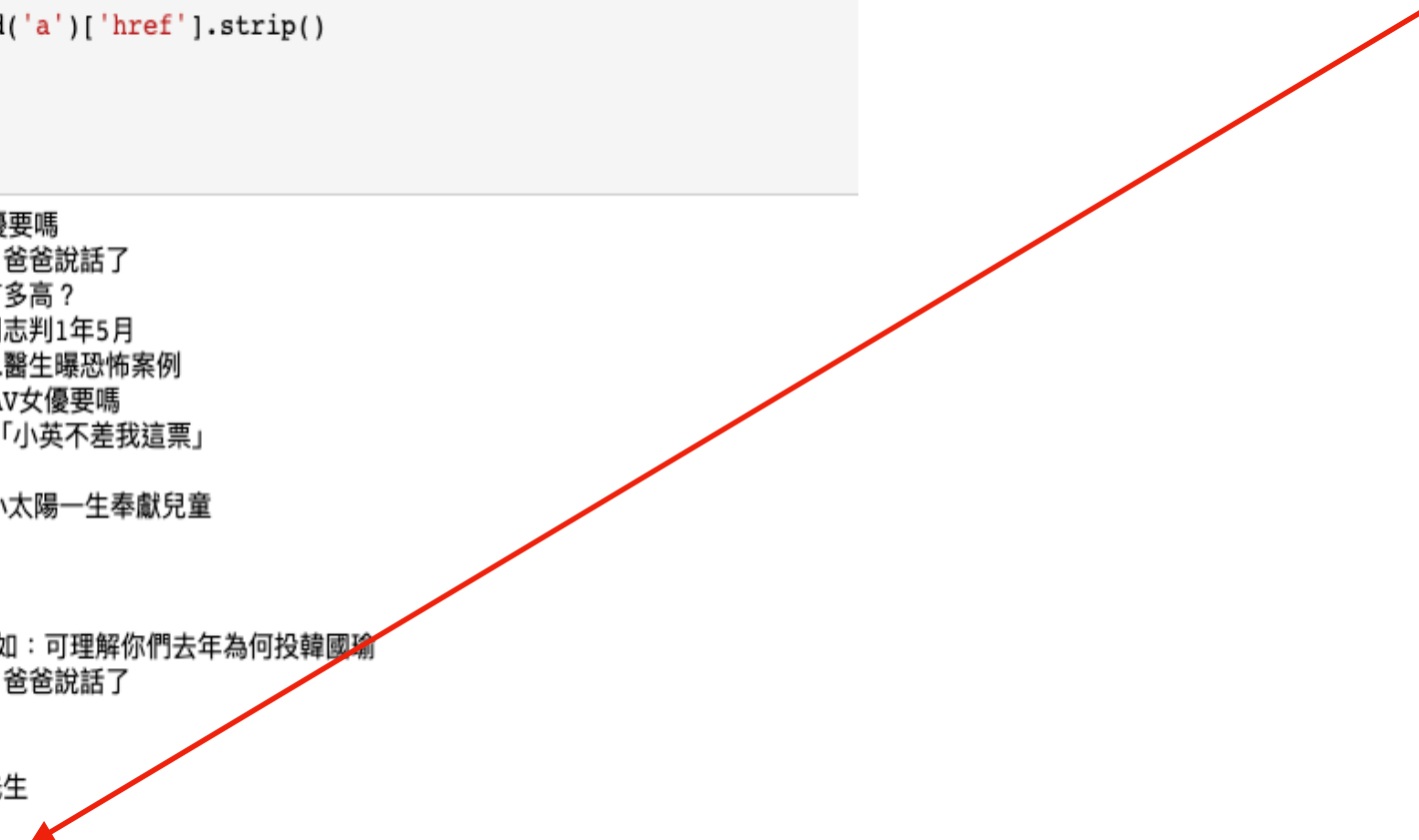
```
</div>
</div>
<div class="r-ent">
<div class="nrec"><span class="hl f2">9</span></div>
<div class="title">
<a href="/bbs/Gossiping/M.1577084594.A.75E.html">Re: [問卦]請問1996年的PTT素質有多高?</a>
</div>
<div class="meta">
<div class="author">dake</div>
<div class="article-menu">
<div class="trigger">...</div>
<div class="dropdown">
<div class="item"><a href="/bbs/Gossiping/search?q=thread%3A%5B%E5%95%8F%E5%8D%A6%5D%E8%AB%8B%E5%95%8F1996%E5%B9%B4%E
7%9A%84PTT%E7%B4%A0%E8%B3%AA%E6%9C%89%E5%A4%9A%E9%AB%98%EF%BC%9F">搜尋同標題文章</a></div>
<div class="item"><a href="/bbs/Gossiping/search?q=author%3Adake">搜尋看板內 dake 的文章</a></div>
</div>
</div>
<div class="date">12/23</div>
<div class="mark"></div>
</div>
```

往下取得標籤並將剛剛程式結合

```
In [4]: ##發現文章位於 div 標籤的 r-ent class先往下解析一層
soup = BeautifulSoup(page, features="html.parser")
divs = soup.find_all('div', class_='r-ent')
for div in divs:
    ptt = "https://www.ptt.cc"
    soup_title = BeautifulSoup(str(div), features="html.parser")
    ##解析標題的div結構
    title = soup_title.find_all('div', class_='title')
    news_title = title[0].text.strip()
    ##取得標題文章內文之連結
    try:
        link = title[0].find('a')['href'].strip()
    except:
        continue
    print(news_title)
    print(ptt + link)
```

- 還需要處理掉協尋與公告文章

[問卦] 給你三千萬跟三個正妹AV女優要嗎
Re: [新聞] 女嬰遭韓國瑜又抱又親 爸爸說話了
Re: [問卦] 請問1996年的PTT素質有多高?
Re: [新聞] 患愛滋卻與人性交 男同志判1年5月
Re: [新聞] 悚! 女嬰遭韓國瑜捧吻...醫生曝恐怖案例
Re: [問卦] 給你三千萬跟三個正妹AV女優要嗎
Re: [新聞] 北漂年輕人沒買到車票「小英不差我這票」
[問卦] 哆啦A夢哪集劇場版最好看?
Re: [新聞] 林良96歲辭世 永遠的小太陽一生奉獻兒童
[問卦] 爺爺追殺孫女的卦
[爆卦] 自稱嬰兒媽媽的友人道歉了
[問卦] 愛一個人需要理由嗎?
Re: [新聞] 大眾走高雄第2天 蔡壁如:可理解你們去年為何投韓國瑜
Re: [新聞] 女嬰遭韓國瑜又抱又親 爸爸說話了
[問卦] 明年1/11投票
[公告] 八卦板板規(2019.08.21)
[協尋] 協尋雲林縣崙背鄉前代表廖先生
[協尋] 妹妹與友人失蹤 代po
[協尋] 行車記錄器 高雄區/鳳山區
[公告] 一劍無悔,十二月置底閒聊文



全部完成之後將程式碼函數化

```
In [7]: #取得文章標題與連結
def get_links_from_index(page):
    soup = BeautifulSoup(page, features="html.parser")
    divs = soup.find_all('div', class_='r-ent')
    linkList = list()
    for div in divs:
        soup_title = BeautifulSoup(str(div), features="html.parser")
        title = soup_title.find_all('div', class_='title')
        news_title = title[0].text.strip()
        try:
            link = title[0].find('a')['href'].strip()
        except:
            continue
        if '[公告]' in news_title or '[協尋]' in news_title:
            continue
        linkList.append([news_title, link])
    # 因為每一個 index.html 中的文章，最新的那篇是在最底下，所以做個 reversed
    # 這樣最新的文章就會是在 linkList[0]
    linkList = list(reversed(linkList))
    return linkList
```

```
In [14]: ##將上述程式碼組合
linkList = get_links_from_index(page)
for title, link in linkList:
    print(title)
    print(ptt + link)
```

```
[問卦] 明年1/11投票
https://www.ptt.cc/bbs/Gossiping/M.1577085532.A.8C6.html
Re: [新聞] 女嬰遭韓國瑜又抱又親 爸爸說話了
https://www.ptt.cc/bbs/Gossiping/M.1577085532.A.D73.html
Re: [新聞] 大眾走高雄第2天 蔡壁如：可理解你們去年為何投韓國瑜
https://www.ptt.cc/bbs/Gossiping/M.1577085512.A.197.html
[問卦] 愛一個人需要理由嗎？
https://www.ptt.cc/bbs/Gossiping/M.1577085507.A.3D5.html
[爆卦] 自稱嬰兒媽媽的友人道歉了
https://www.ptt.cc/bbs/Gossiping/M.1577085335.A.3EB.html
[問卦] 爺爺追殺孫女的卦
https://www.ptt.cc/bbs/Gossiping/M.1577085324.A.9BF.html
Re: [新聞] 林良96歲辭世 永遠的小太陽一生奉獻兒童
https://www.ptt.cc/bbs/Gossiping/M.1577085316.A.07A.html
[問卦] 哆啦A夢哪集劇場版最好看？
https://www.ptt.cc/bbs/Gossiping/M.1577085282.A.EEE.html
Re: [新聞] 北漂年輕人沒買到車票「小英不差我這票」
https://www.ptt.cc/bbs/Gossiping/M.1577085121.A.DBB.html
Re: [問卦] 給你三千萬跟三個正妹AV女優要嗎
https://www.ptt.cc/bbs/Gossiping/M.1577085006.A.766.html
Re: [新聞] 悚！女嬰遭韓國瑜捧吻...醫生曝恐怖案例
https://www.ptt.cc/bbs/Gossiping/M.1577084900.A.9EC.html
Re: [新聞] 患愛滋卻與人性交 男同志判1年5月
https://www.ptt.cc/bbs/Gossiping/M.1577084753.A.4AF.html
Re: [問卦] 請問1996年的PTT素質有多高？
https://www.ptt.cc/bbs/Gossiping/M.1577084594.A.75E.html
Re: [新聞] 女嬰遭韓國瑜又抱又親 爸爸說話了
https://www.ptt.cc/bbs/Gossiping/M.1577084583.A.B79.html
[問卦] 給你三千萬跟三個正妹AV女優要嗎
https://www.ptt.cc/bbs/Gossiping/M.1577084510.A.56F.html
```

對文章內文原始碼處理

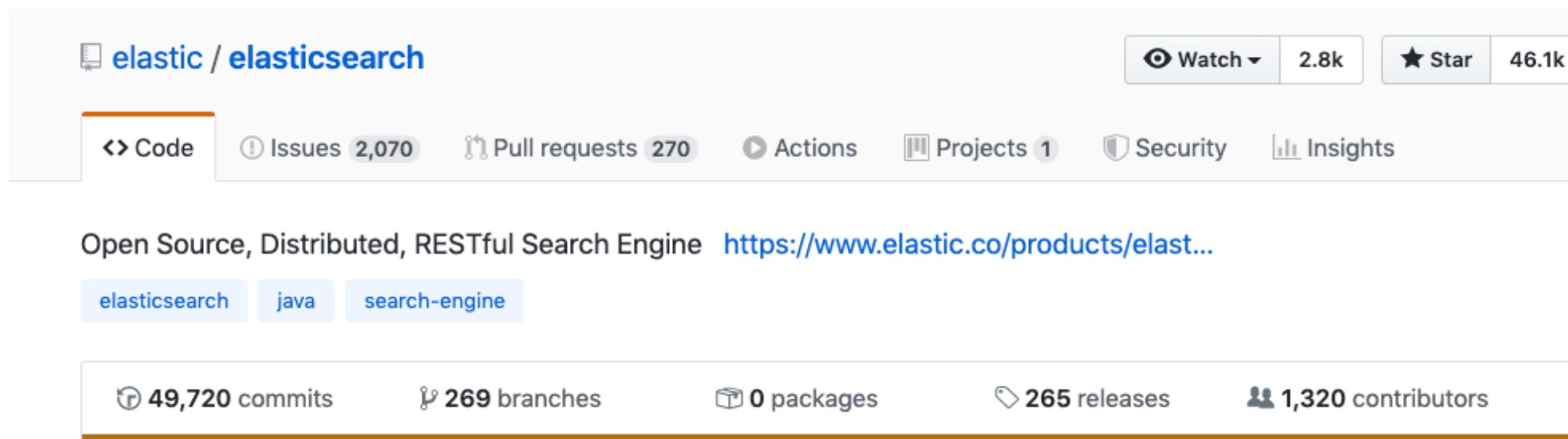
```
In [13]: ##將上述程式碼組合
linkList= get_links_from_index(page)
for title, link in linkList:
    print(title)
    print(ptt + link)
    html = s.get(ptt + link).text
    content = get_news_content(ptt + link,html)
```

```
[問卦] 明年1/11投票
https://www.ptt.cc/bbs/Gossiping/M.1577085532.A.8C6.html
{'Title': '[問卦] 明年1/11投票', 'URL': 'https://www.ptt.cc/bbs/Gossiping/M.1577085532.A.8C6.html', 'Author': 'lamigo (lamigo)', 'Description': '\n剛睡醒時時看到line出現明年1月的班表出現了\n\n當天被排到要上班,家住新莊上單地點在桃園\n\n往年上班日投票都是雙薪,然後兩個小時可以外出輪流投票\n\n可是新莊到桃園2小時騎車真的很趕,去年11/24看到新聞\n\n排隊大排長龍,就沒去排了\n\n請問明年的1/11投票日會像去年的11/24日,排隊投票\n\n像排東京迪士尼樂園一樣嗎\n\n--\n我話說完 誰贊成 誰反對\n\n--\n* 發信站: 批踢踢實業坊(ptt.cc), 來自: 61.228.157.183 (臺灣)\n'}
Re: [新聞] 女嬰遭韓國瑜又抱又親 爸爸說話了
https://www.ptt.cc/bbs/Gossiping/M.1577085532.A.D73.html
{'Title': 'Re: [新聞] 女嬰遭韓國瑜又抱又親 爸爸說話了', 'URL': 'https://www.ptt.cc/bbs/Gossiping/M.1577085532.A.D73.html'}
```

- 集合取出內文函數與title函數，即可撈取特定頁面資料
- 觀察PTT URL變化可寫成爬蟲排程撈取大量資料

什麼是Elasticsearch

- 是在Github上的一個開用專案
- 開放原始碼：免費
- 新一代的搜尋引擎
- 擁有龐大的活躍社群
- 善於處理巨量情資



環境安裝

- JDK8 update 20 或以上的版本並設定環境變數(windows)
- 下載Elasticsearch (<https://www.elastic.co/cn/products/elasticsearch>)
- 使用5.5.3版本的elasticsearch

特點

- 分散式資料庫
- JSON document storage
- Serve search requests in ms
- RESTful API operation
- Build index

Distributed 分散式架構

- Distributed 為新型態資料庫NoSQL的重要特徵
- 為了應付大量資料與效能瓶頸
- Elasticsearch 屬於Document oriented NoSQL
 - Using JSON as document
 - Schema-free
- Elasticsearch 的分散式架構具有
 - Horizontally scalable
 - Easy replication



RESTful API易於使用

- Restful 是一種架構風格，漸成Web service 的主流
- 基於HTTP協定，定義資源(URL)，並善用HTTP method
- Elasticsearch API = RESTful http requests + JSON body

- 搜尋語法為：



```
{
  "query": {
    "bool": {
      "must": [
        {
          "match": {
            "Title": "蔡英文"
          }
        }
      ],
      "must_not": [],
      "should": []
    }
  },
  "from": 0,
  "size": 10,
  "sort": [],
  "aggs": {}
}
```

```
"took": 255,
"timed_out": false,
"_shards": {
  "total": 60,
  "successful": 60,
  "skipped": 0,
  "failed": 0
},
"hits": {
  "total": 4510654,
  "max_score": 12.823797,
  "hits": [
    {
      "_index": "sec_ptt-201902",
      "_type": "ptt",
      "_id": "M.1550631257.A.EC0.html-2-wy9968",
      "_score": 12.823797,
      "_source": {
        "BBS": "八卦",
        "Board": "Gossiping",
        "Latitude": null,
        "Type": "req",
        "Floor": 2,
        "City": null,
        "Req": "推 wy9968: 88888 02/20 10:54 ",
        "Region": null,
        "Ip": "",
        "URL": "/bbs/Gossiping/M.1550631257.A.EC0.html",
        "Title": "[問卦] 蔡英文的英文能力",
        "JiebaKeyword": [
          "88888"
        ],
        "Date": "2019-02-20T10:54:00",
        "Author": "wy9968",
        "Longitude": null,
        "PublishTime": 1550631240000,
        "Push": "推",
        "Message": "88888"
      }
    }
  ]
}
```

Search Engine 搜尋引擎

- Elasticsearch 使用 Apache Lucene 為核心是一個成熟且開源的全文檢索系統
- 為NoSQL Database 的一個分類。是為了搜尋而生的資料庫，通常具有全文檢索，地理搜尋等高階功能



綜合各方優點

- 事實上，Elasticsearch沒有哪個部分是革命性的技術創新。全文檢索、分析系統、分散式資料庫早就被實作完了。
- 其創新在於將這些技術結合起來成為一個單一系統



啟動Elasticsearch — 1

- 解壓縮完後進入Elasticsearch 目錄
- 下指令 `./bin/elasticsearch` (ubuntu 或 mac OS)
- 解壓縮完後在 bin 目錄下，直接點擊elasticsearch.bat (windows)
- 接著去網頁上 輸入 <http://localhost:9200/> 確認

啟動Elasticsearch — 2

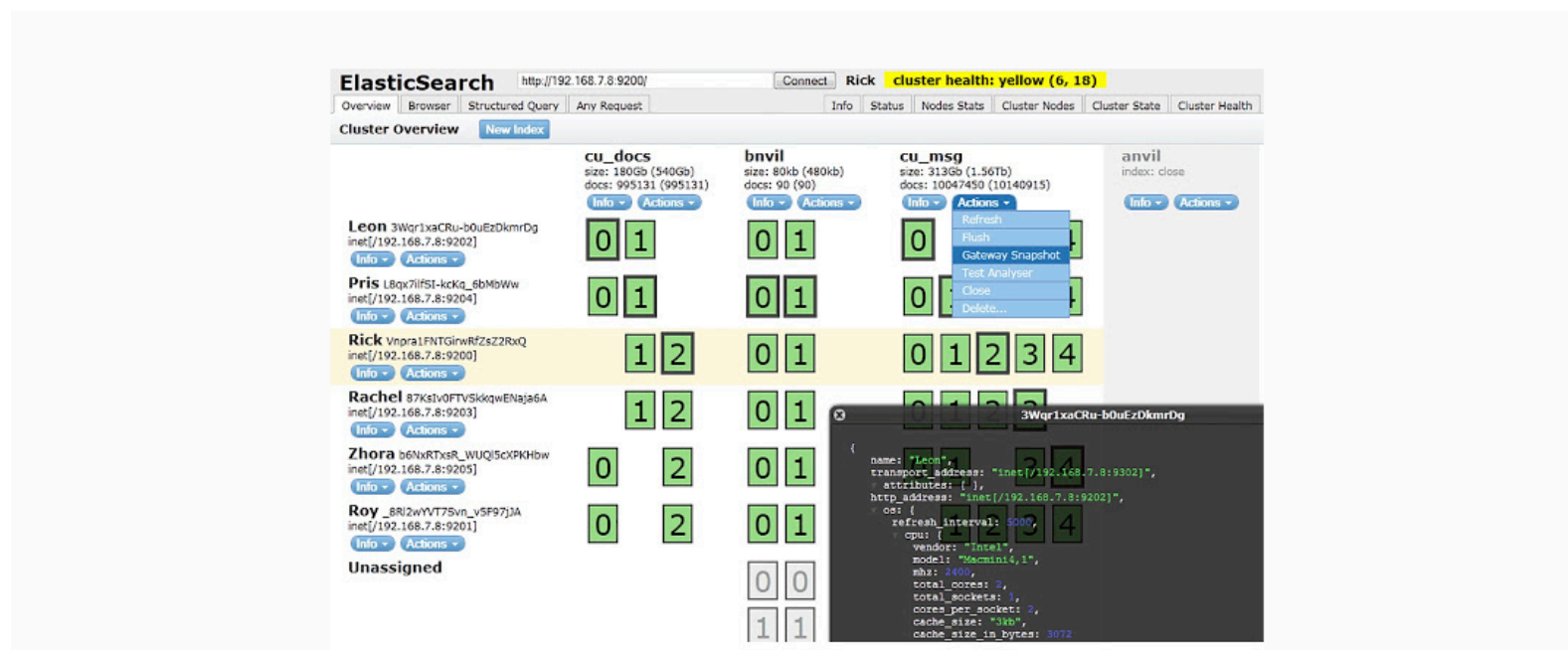
```
└─ ./bin/elasticsearch
[2019-12-23T10:05:34,924][INFO ][o.e.n.Node               ] [] initializing ...
[2019-12-23T10:05:35,051][INFO ][o.e.e.NodeEnvironment ] [17Xo2TU] using [1] data paths, mounts [[/ (/dev/disk1s1)]] , net usable_space [55.5gb], net total_space [465.6gb], spins? [unknown], types [apfs]
[2019-12-23T10:05:35,051][INFO ][o.e.e.NodeEnvironment ] [17Xo2TU] heap size [1.9gb], compressed ordinary object pointers [true]
[2019-12-23T10:05:35,387][INFO ][o.e.n.Node               ] node name [17Xo2TU] derived from node ID [17Xo2TU1S3Cm7iC3_AhYGw]; set [node.name] to override
[2019-12-23T10:05:35,387][INFO ][o.e.n.Node               ] version[5.6.3], pid[73173], build[1a2f265/2017-10-06T20:33:39.012Z], OS[Mac OS X/10.14.6/x86_64], JVM[Oracle Corporation/Java HotSpot(TM) 64-Bit Server VM/1.8.0_102/25.102-b14]
[2019-12-23T10:05:35,387][INFO ][o.e.n.Node               ] JVM arguments [-Xms2g, -Xmx2g, -XX:+UseConcMarkSweepGC, -XX:CMSInitiatingOccupancyFraction=75, -XX:+UseCMSInitiatingOccupancyOnly, -XX:+AlwaysPreTouch, -Xss1m, -Djava.awt.headless=true, -Dfile.encoding=UTF-8, -Djna.nosys=true, -Djdk.io.permissionsUseCanonicalPath=true, -Dio.netty.noUnsafe=true, -Dio.netty.noKeySetOptimization=true, -Dio.netty.recycler.maxCapacityPerThread=0, -Dlog4j.shutdownHookEnabled=false, -Dlog4j2.disable.jmx=true, -Dlog4j2.skipJansi=true, -XX:+HeapDumpOnOutOfMemoryError, -Des.path.home=/Users/liguilun/Desktop/技術安裝檔/ELK/elasticsearch-5.6.3]
[2019-12-23T10:05:36,286][INFO ][o.e.p.PluginsService     ] [17Xo2TU] loaded module [aggs-matrix-stats]
[2019-12-23T10:05:36,286][INFO ][o.e.p.PluginsService     ] [17Xo2TU] loaded module [ingest-common]
[2019-12-23T10:05:36,286][INFO ][o.e.p.PluginsService     ] [17Xo2TU] loaded module [lang-expression]
[2019-12-23T10:05:36,287][INFO ][o.e.p.PluginsService     ] [17Xo2TU] loaded module [lang-groovy]
[2019-12-23T10:05:36,287][INFO ][o.e.p.PluginsService     ] [17Xo2TU] loaded module [lang-mustache]
[2019-12-23T10:05:36,287][INFO ][o.e.p.PluginsService     ] [17Xo2TU] loaded module [lang-painless]
[2019-12-23T10:05:36,287][INFO ][o.e.p.PluginsService     ] [17Xo2TU] loaded module [parent-join]
[2019-12-23T10:05:36,287][INFO ][o.e.p.PluginsService     ] [17Xo2TU] loaded module [percolator]
[2019-12-23T10:05:36,287][INFO ][o.e.p.PluginsService     ] [17Xo2TU] loaded module [reindex]
[2019-12-23T10:05:36,287][INFO ][o.e.p.PluginsService     ] [17Xo2TU] loaded module [transport-netty3]
[2019-12-23T10:05:36,287][INFO ][o.e.p.PluginsService     ] [17Xo2TU] loaded module [transport-netty4]
[2019-12-23T10:05:36,288][INFO ][o.e.p.PluginsService     ] [17Xo2TU] no plugins loaded
[2019-12-23T10:05:38,055][INFO ][o.e.d.DiscoveryModule    ] [17Xo2TU] using discovery type [zen]
[2019-12-23T10:05:38,692][INFO ][o.e.n.Node               ] initialized
[2019-12-23T10:05:38,692][INFO ][o.e.n.Node               ] [17Xo2TU] starting ...
[2019-12-23T10:05:38,894][INFO ][o.e.t.TransportService   ] [17Xo2TU] publish_address {127.0.0.1:9300}, bound_addresses {[fe80::1]:9300}, {[::1]:9300}, {127.0.0.1:9300}
[2019-12-23T10:05:41,952][INFO ][o.e.c.s.ClusterService    ] [17Xo2TU] new_master {17Xo2TU}{17Xo2TU1S3Cm7iC3_AhYGw}{IrYZDkU6RIqS7ZgJRfMxlQ}{127.0.0.1}{127.0.0.1:9300}, reason: zen-disco-elected-as-master ([0] nodes joined)
[2019-12-23T10:05:41,974][INFO ][o.e.h.n.Netty4HttpServerTransport] [17Xo2TU] publish_address {127.0.0.1:9200}, bound_addresses {[fe80::1]:9200}, {[::1]:9200}, {127.0.0.1:9200}
[2019-12-23T10:05:41,974][INFO ][o.e.n.Node               ] [17Xo2TU] started
[2019-12-23T10:05:42,430][INFO ][o.e.g.GatewayService     ] [17Xo2TU] recovered [22] indices into cluster_state
[2019-12-23T10:05:44,262][INFO ][o.e.c.r.a.AllocationService] [17Xo2TU] Cluster health status changed from [RED] to [YELLOW] (reason: [shards started [[sec_cisco][3]] ...]).
```

啟動Elasticsearch — 3

```
{
  "name" : "17Xo2TU",
  "cluster_name" : "elasticsearch",
  "cluster_uuid" : "aV9-SPCnRoeOAhc7bfpjCg",
  "version" : {
    "number" : "5.6.3",
    "build_hash" : "1a2f265",
    "build_date" : "2017-10-06T20:33:39.012Z",
    "build_snapshot" : false,
    "lucene_version" : "6.6.1"
  },
  "tagline" : "You Know, for Search"
}
```

安裝外掛插件方便瀏覽資料

- Google chrome 提供視覺化插件瀏覽資料
- chrome 插件 搜尋 Elasticsearch Head



結構介紹

- Index(名詞)：等同於一般DB中的database
- Index(動詞)：儲存文件的動作
- Type：等同於DB中的table，但於ES6版本之後開始廢棄Type
- Field：等同於DB中的column
- Field Type：指定Field欄位所儲存的型態，如果沒有預先設定型態ES會自動偵測並給予資料型態

結構介紹

The screenshot displays a database management interface with five indices at the top: **sec_ptt-201903** (size: 1.78Gi, docs: 1,758,186), **sec_ptt-201902** (size: 2.00Gi, docs: 1,956,328), **sec_ptt-201901** (size: 1.59Gi, docs: 1,615,771), **sec_ptt-201812** (size: 2.49Gi, docs: 2,409,075), and **sec_ptt-201811** (size: 1.89Gi, docs: 1,894,111). Each index has buttons for '訊息' (Message) and '動作' (Action).

The main window shows the mapping for **sec_ptt-201902**. The JSON structure is as follows:

```
{
  "mappings": {
    "ptt": {
      "properties": {
        "Description": {
          "type": "text",
          "fields": {
            "keyword": {
              "ignore_above": 256,
              "type": "keyword"
            }
          }
        },
        "Message": {
          "type": "text",
          "fields": {
            "keyword": {
              "ignore_above": 256,
              "type": "keyword"
            }
          }
        },
        "Ip": {
          "type": "text",
          "fields": {
            "keyword": {
              "ignore_above": 256,
              "type": "keyword"
            }
          }
        },
        "Latitude": {
          "type": "float"
        },
        "Pushcount": {
          // ...
        }
      }
    }
  }
}
```

- Index(名詞)：等同於一般DB中的database

- Field：等同於DB中的column

- Field Type：表示Field中的型態

檢視文章方式

Elasticsearch **secbuzzer-es** 叢集健康值: green (3650 of 3650)

總覽 索引 資料瀏覽 基本查詢 [\[+\]](#) 複合查詢 [\[+\]](#)

搜尋 的文件，查詢條件:

must

ptt.Message

match

蔡英文

+

-

搜尋 返回格式:

Table

 顯示數量:

10

☐ 顯示查詢語句

查詢 5 個分片中用的 5 個. 116275 命中. 耗時 2.810 秒

_index	_type	_id	_score ▲	Push	Req	Message	Title
sec_ptt-201911	ptt	M.1573283535.A.89C.html-66-pkpk23456	16.73382	→	→ pkpk23456: 蔡英文? 42.77.222.85 11/09 15:44	蔡英文?	[爆卦] LIVE 實現居住正義 打破金權政治
sec_ptt-201911	ptt	M.1573302491.A.9C5.html-279-xianfen	16.73382	推	推 xianfen: 挺蔡英文180.204.212.195 11/09 22:58	挺蔡英文	[新聞] 蔡英文: 台灣守不住就是民主自由失敗
sec_ptt-201911	ptt	M.1573217573.A.6B1.html-13-vacuityhu	16.73382	噓	噓 vacuityhu: 蔡英文 211.76.55.190 11/08 20:53	蔡英文	[問卦] 姓蔡要取什麼名字比較特別?
sec_ptt-201911	ptt	M.1573212123.A.A6A.html-200-Deltoid	16.73382	→	→ Deltoid: 蔡英文了 223.138.86.66 11/08 20:32	蔡英文了	[新聞] 馬英九: 蔡英文是史上「最自我感覺良好」
sec_ptt-201911	ptt	M.1573479774.A.DD1.html-34-adamas0422	16.73382	→	→ adamas0422: 蔡英文了 175.97.17.186 11/12 18:49	蔡英文了	Re: [討論] 香港最終會面臨何種命運?
sec_ptt-201911	ptt	M.1573453059.A.58B.html-54-ThreekRoger	16.73382	→	→ ThreekRoger: 蔡英文: 「喔」 125.227.122.115 11/11 14:41	蔡英文: 「喔」	[新聞] 冷處理「國政配」? 蔡英文競辦: 沒有進
sec_ptt-201911	ptt	M.1573556593.A.200.html-4-Robben	16.73382	推	推 Robben: 蔡英文 49.216.102.214 11/12 19:04	蔡英文	[新聞] 快訊 / 宋楚瑜副手是誰? 親民黨證實: 1
sec_ptt-201911	ptt	M.1572775569.A.BED.html-21-royroy666	16.73382	→	→ royroy666: 爆蔡英文 61.224.51.151 11/03 18:08	爆蔡英文	[問卦] 所以PTT到底有沒有五毛
sec_ptt-201911	ptt	M.1573865990.A.512.html-23-fatout	16.73382	→	→ fatout: , 蔡英文 39.9.132.194 11/16 10:14	, 蔡英文	Re: [爆卦] 今日被民進黨封殺之法案
sec_ptt-201911	ptt	M.1574048126.A.7B2.html-12-grayoasis	16.73382	推	推 grayoasis: 蔡英文122.116.200.241 11/18 11:36	蔡英文	[問卦] 現在誰有資格當志玲姊姊的接班人?

下Query語法進行複雜的全文檢索

Elasticsearch

http://192.168.70.182:30200/

連接

secbuzzer-es

叢集健康值: green (3650 of 3650)

總覽

索引

資料瀏覽

基本查詢 [\[+\]](#)

複合查詢 [\[+\]](#)

▶ 歷史記錄

▼ 查詢

http://192.168.70.182:30200/sec_ptt-*/

_search

POST

```
{
  "query": {
    "bool": {
      "must": [
        {
          "match": {
            "Message": "蔡英文"
          }
        }
      ],
      "must_not": [],
      "should": []
    }
  },
  "from": 0,
  "size": 10,
  "sort": [],
  "aggs": {}
}
```

送出

驗證 JSON

☒ 易讀

▶ 結果轉換器

▶ 重複請求

▶ 顯示選項

```
{
  "took": 18225,
  "timed_out": false,
  "num_reduce_phases": 2,
  "_shards": {
    "total": 930,
    "successful": 930,
    "skipped": 0,
    "failed": 0
  },
  "hits": {
    "total": 1756873,
    "max_score": 25.29984,
    "hits": [
      {
        "_index": "sec_ptt-201807",
        "_type": "ptt",
        "_id": "M.1531757856.A.45E.html-42-kof2200",
        "_score": 25.29984,
        "_source": {
          "Message": "蔡英文?",
          "PublishTime": 1531792260000,
          "Title": "[新聞] 鴻海接班人條件 郭董：被我罵愈兇機會愈",
          "Latitude": null,
          "Region": null,
          "BBS": "科技",
          "Date": "2018-07-17T09:51:00",
          "Req": "推 kof2200: 蔡英文? 07/17 09:51 ",
          "Push": "推",
          "Ip": "",
          "Author": "kof2200",
          "Floor": 42,
          "URL": "/bbs/Tech_Job/M.1531757856.A.45E.html",
          "Board": "Tech_Job",
          "JiebaKeyword": [
            "蔡英文"
          ],
          "Longitude": null,
          "City": null,
          "Type": "req"
        }
      }
    ]
  }
}
```

更多複雜語法可上ES文件翻閱

- <https://www.elastic.co/guide/en/elasticsearch/reference/5.5/docs.html> 官網文件
- 常用Query
 - term query : 下關鍵字對所想撈取的欄位全文搜尋
 - range query : 若該欄位有設定時間或著數量可進行時間排序搜尋或著數量搜尋
 - from/size : 設定回傳資料返回數量
 - count : 可直接取出該搜尋條件數量