108502504 資工3A 林芸萱

WEB CRAWLER AND OPEN SOURCE IR SYSTEM

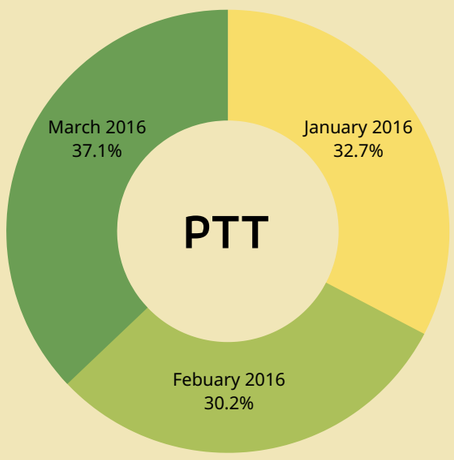
本次作業實作一個PTT C_Chat的web crawler，並將資訊存成JSON檔案，最後匯入Elasticsearch完成Open Source IR System

INTRODUCTION

PTT 為台灣的大型論壇之一，每天都有數千篇文章在上面發表。而八卦板、**C_Chat**、股票板更是熱門看板，即時在線人數都有數千人至萬人。而**PTT**只會保留一年的文章，因此有精華區的功能，將舊文保留下來。但精華區使用網頁板觀看時，沒有關鍵字搜尋系統，有諸多不便。因此本次藉由**web crawler c**和**elasticsearch system**，實作一個小型精華區關鍵字索引，方便使用者查找精華區文章。

DATA STATISTIC

本次爬蟲的資料主要為**PTT**的**C_Chat**精華區的備份區，由於備份區收錄了**2007-2020**年全部的文章，數量過於龐大，**C_Chat**為討論**ACG**的一個專板，而動畫討論多以**3**個月為斷點。因此只挑了**2016**的一個季度進行實作，本次挑選的為**2016**年**1**月~**3**月的精華區文章。



1月份有**6106**篇，**2**月份有**5652**篇，**3**月份有**6934**篇，共**18692**篇。

RELATED WORK

PTT crawler在網路上已有許多人實作，包括以**Beautiful soup**爬取單頁文章內容如**titile**、**author**、**article**等方法，也有爬取整頁文章列表標題的實作。

Beautifulsoup

快速解析網頁 **HTML** 碼，並可使用其中的**find()**函式，搜尋各種標籤，從中提取出使用者有興趣的資料。

Json

一種輕量級資料交換格式。其內容由屬性和值所組成，因此也有易於閱讀和處理的優勢。

Elasticsearch

Elasticsearch是一個基於**Lucene**庫的搜尋引擎。它提供了一個分散式、支援多租戶的全文搜尋引擎，具有**HTTP Web**介面和**JSON**文件。

METHOD

Figure1:C_Chat精華區架構

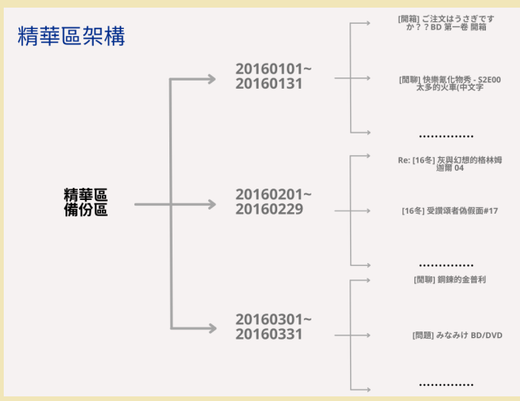


Figure2:DB Schema 架構

```
chat_mapping = {
    "properties": {
        "catalogueTitle": {
            "type": "keyword"
        },
        "title": {
            "type": "text",
            "analyzer": "ik_max_word",
            "search_analyzer": "ik_max_word"

        },
        "author": {
            "type": "keyword"
        },
        "board": {
            "type": "keyword"
        },
        "date": {
            "type": "keyword"
        },
        "article": {
            "type": "text",
            "analyzer": "ik_smart",
            "search_analyzer": "ik_smart"
        }
    }
}
```

01 PTT Crawler

先爬取並記錄精華區目錄標題，並以**regular expression**鎖定爬取年分**(2016)**，再爬取前三個月的文章列表，與每一篇文章的資訊：**author**、**date**、**board**、**date**、**article**等。之後，將其轉為**json**格式，寫入檔案。由於需要爬取的文章數較多，因此開了**5**個**thread**提高效率。

02 DB Schema

儲存的資訊有**catalogueTitle(**目錄索引**)**、**title(**文章標題**)**、**author**、**board**、**date**、**article**，**title**和**article**由於資訊較多，也希望可以做更精細的搜索，因此設成**text**，而**titlle**採用**ik_max_word**分詞，**artile**則因為文本資訊較大，也不希望分詞太瑣碎而出現相關度低的文章，因此使用**ik_smart**分詞。

EXPERIMENT

關鍵字

手遊 PS4 推薦 JUMP 排球少年

想要在article和title裡搜尋關鍵字，因此使用multi_match，因為會對關鍵字進行分詞，因此，為避免因為分詞，而造成搜尋出其他不相關的結果，所以選用"phrase"類型。如"排球少年"，可分為"排球" 、"少年"若是使用預設的，文本只要符合其一分詞即可。而若是要查詢作者，因為作者為keyword，因此不會替關鍵字進行分詞，結果也須完全符合輸入的關鍵字才能搜索的到。最後結果會依照score的分數進行排序。而在其中加入"highlight"屬性，可以標示出關鍵字的位置，方便查找。

```
搜尋作者請按a ，搜尋關鍵字請按/ ，結束搜尋請按0: /
請輸入搜尋關鍵字: 推薦
score: 10.721813
catalogue:◆2016/02/01 ~ 2016/02/29 -[5652]
author:nick861211 (我的妹妹)
title:◇[推薦] 求推薦治癒番
article:
小魯最近壓力很大

所以看完了輕鬆百合，黃金拼圖，點兔

想要繼續看這一類型的動畫來撫平我受傷的心靈

有其他推薦嗎？
==========
score: 10.721813
catalogue:◆2016/03/01 ~ 2016/03/31 -[6934]
author:Sasamumu (Samu)
title:◇[推薦] 求推薦純日常作品
article:
求推薦類似的作品＞＜！！

喵pass~~~

---
==========
score: 10.450887
catalogue:◆2016/02/01 ~ 2016/02/29 -[5652]
author:s8229926 (jameschang)
title:◇[推薦] 求推薦類似漫畫~
==========
```

Result 1:"推薦"之搜尋結果

```
搜尋作者請按a ，搜尋關鍵字請按/ ，結束搜尋請按0: /
請輸入搜尋關鍵字: 排球少年
score: 14.882776
catalogue:◆2016/02/01 ~ 2016/02/29 -[5652]
author:lisa6329 (瑪瑪)
title:◇[閒聊] 排球少年!! #195
==========
score: 13.656913
catalogue:◆2016/03/01 ~ 2016/03/31 -[6934]
author:lisa6329 (瑪瑪)
title:◇[閒聊]  排球少年!! #199 動搖
==========
score: 13.492786
catalogue:◆2016/03/01 ~ 2016/03/31 -[6934]
author:qhaabk (沒有人哪有那個國)
title:◇[問題] 運動型動畫
article:
現在再追的動畫有

飆速宅男1+2 排球少年1+2 鑽石王牌1+2

只是這禮拜就完了


請問最近還有運動型動畫嗎 謝謝
==========
score: 13.116716
catalogue:◆2016/03/01 ~ 2016/03/31 -[6934]
author:lisa6329 (瑪瑪)
title:◇[閒聊] 排球少年!! #196 背水一戰
==========
score: 13.116716
catalogue:◆2016/03/01 ~ 2016/03/31 -[6934]
author:lisa6329 (瑪瑪)
title:◇[閒聊] 排球少年!! #197 蛇vs貓
==========
```

Result 2:"排球少年"之搜尋結果

```
請輸入搜尋關鍵字: 手遊
D:\anaconda\lib\site-packages\elasticsearch\connection\base.py:200:
  warnings.warn(message, category=ElasticsearchWarning)
score: 9.380417
catalogue:◆2016/01/01 ~ 2016/01/31 -[6106]
author:pan46 (pan)
title:◇[手遊] 國產手遊 雷光少女
==========
score: 7.77117
catalogue:◆2016/01/01 ~ 2016/01/31 -[6106]
author:itohchen (伊藤)
title:◇[閒聊] PVP的手遊
article:
玩一個叫Gundam Conquest(GCQ)的日本手遊玩了兩年多
剛剛在跟香港盟友聊天的時候忽然發現

日本的手遊好像很少pvp元素

有玩過的像是CC PAD 勇者前線 之類的
大多都是收集腳色突破關卡為主
==========
score: 7.77117
catalogue:◆2016/02/01 ~ 2016/02/29 -[5652]
author:loveangel718 (エト我老婆)
title:◇[問題] 推薦手遊
article:
剛吃完年夜飯

不幸的是 筆電在前天螢幕壞掉了 還沒修理

FGO最近沒活動 這個卡池又沒興趣

體力花光了就放置了

漫畫動畫整天下來也看膩了(不管表裏)

求各位版眾推薦可以消磨時間的手遊

最好是那種無腦低能類型的
==========
```

Result 3:"手遊"之搜尋結果

```
請輸入搜尋關鍵字: Jump
score: 9.668464
catalogue:◆2016/03/01 ~ 2016/03/31 -[6934]
author:wl760713 (willy)
title:◇[閒聊] 獵人連載再開日期 4/18
article:
http://imgur.com/aYJFlmd.jpg

這期Jump只是情報
20號Jump 4/18才連載再開
還有一個月
==========
score: 9.497774
catalogue:◆2016/02/01 ~ 2016/02/29 -[5652]
author:godivan (加藤家的惠是我的!)
title:◇[心得]第八號學生食堂少女 01
article:
滿滿Jump風的王道系風格作品



這是我看完的感想


最弱的食堂--滿天V.S.最強的食堂--格蘭廚房



只不過女主角那個妹妹我只覺得.....根本沒有必要馬後炮說一切都是她的劇本和計畫.
==========
score: 9.464357
catalogue:◆2016/01/01 ~ 2016/01/31 -[6106]
author:bebopfan (bebopfan)
title:◇Fw: [少年] 崛越老師你內心堅強點好嗎?...
article:
[0;42mJUMP有因此逼你一週要畫兩話嗎? 沒有嘛!

說自己不適合JUMP體系? 拜託你嘛幫幫忙!
你到任何一家出版社畫週刊也是一週一話,
```

Result 4:"Jump"之搜尋結果

```
搜尋作者請按a ，搜尋關鍵字請按/ ，結束搜尋請按0: /
請輸入搜尋關鍵字: PS4
score: 17.815332
catalogue:◆2016/03/01 ~ 2016/03/31 -[6934]
author:JTOM (PEANUTS)
title:◇[PS4/V]Fate/EXTELLA
==========
score: 17.081627
catalogue:◆2016/03/01 ~ 2016/03/31 -[6934]
author:a031516462 (推坑的QB)
title:◇Re: [PS4/V]Fate/EXTELLA
article:
http://i.imgur.com/FQv3UuO.jpg
http://i.imgur.com/loFzHvo.jpg
■Fate/EXTELLA(フェイト/エクステラ)
ジャンル：アクション
機種：PS4
==========
score: 16.73293
catalogue:◆2016/03/01 ~ 2016/03/31 -[6934]
author:joug (好東西不容嗎)
title:◇[新聞] PS4 3.5版「武藏」發表 串流電腦玩PS4成
article:
[0;42mPS4 3.5版「武藏」發表 串流電腦玩PS4成真



SCEJA於前（3/1）日發表PS4的最新3.5版本的作業系統，代號MUSASHI（武藏），未來玩家將可以在PC、MAC上面直接串流PS4的畫面來遊玩
==========
score: 15.781723
catalogue:◆2016/01/01 ~ 2016/01/31 -[6106]
author:johnrolante (羅蘭特)
title:◇[實況] ps4 人中之龍0
==========
score: 15.781723
catalogue:◆2016/01/01 ~ 2016/01/31 -[6106]
author:johnrolante (羅蘭特)
title:◇[實況] ps4 人中之龍0
==========
```

Result 5:"PS4"之搜尋結果

CONCLUSION

在Elasticsearch中有各種搜索的方式，如multi_match、match、term等，根據不同的搜索，就會有分詞與否、是否可在不同的區段搜索等問題。可以根據自己的需求挑選搜索與匹配方式。而Elasticsearch的搜尋結果也幾乎都能與實際想要的結果相符合。

RESULT

可印出前15筆搜尋結果，與其相關資訊，本次礙於篇幅，titilec和article只印出有關鍵字相關的部分，如:在title有關鍵字就印出title，article若沒有關鍵字就沒有印出了。

而highlight關鍵字部分則使用ANSI碼，即可在cmd 印出相關指定的顏色。

REFERENCE

PTT_Crawler

Elasticserach