

Data Science HW1

Crawler (HW1-1) & Popularity Predictor (HW1-2)

Submission Deadline:

2023/3/6 23:59

Submit to E3

Hard deadline, No extensions

HW1-1 目標

爬 PTT Beauty 板

- 2022—整年的文章
- 計算日期內推文和噓文的數量
- 找出日期內最會推跟最會噓的人前10名
- 計算日期內爆文的數量
- 抓取日期內爆文的所有圖片URL
- 關鍵字過濾日期內的文章

Beauty 板

- <https://www.ptt.cc/bbs/Beauty/index.html>

A screenshot of the PTT Beauty board index page. The page has a dark blue header with the text "批踢踢實業坊" and "看板 Beauty". Below the header is a search bar and a navigation bar with buttons for "最舊", "上頁", "下頁", and "最新". The main content area displays a list of 12 posts, each with a title, author, and timestamp. The posts are as follows:

- 14 [神人] 中國主播-2022第一雪乳 Wiserwilly 1/01 ...
- 7 [正妹] 宋晶均 HarunaOno 1/01 ...
- 20 [神人] 這位女優是誰 ? justin21138 1/01 ...
- 1 [正妹] 大尺碼 | 肉特(801) ckpot 1/01 ...
- 爆 [正妹] 2022年，即將三十而立的女孩兒們 cjrmt 1/01 ...
- 6 [正妹] 香港日本混血模特兒 柳斐 Winnie Liu jor3twtw 1/01 ...
- 10 [正妹] 喜歡在床邊伸懶腰 asxc530530 1/01 ...
- 9 [正妹] 蘇慧倫 jerryuan 1/01 ...
- 56 [正妹] MOMO真的超會扭 playerunknow 1/01 ...
- [正妹] 大尺碼 | 肉特(A133) ckpot 1/01 ...

A screenshot of a post on the PTT Beauty board. The post includes the following information:

- 推 iloveq827: 五月
- 推 wafiea708: 朱喜悅要推一個
- wafiea708: <https://i.imgur.com/IHXSTkf.jpg>

The post also features a large image of a woman with long dark hair wearing a black sleeveless dress.

Three screenshots of posts on the PTT Beauty board, each featuring a different woman:

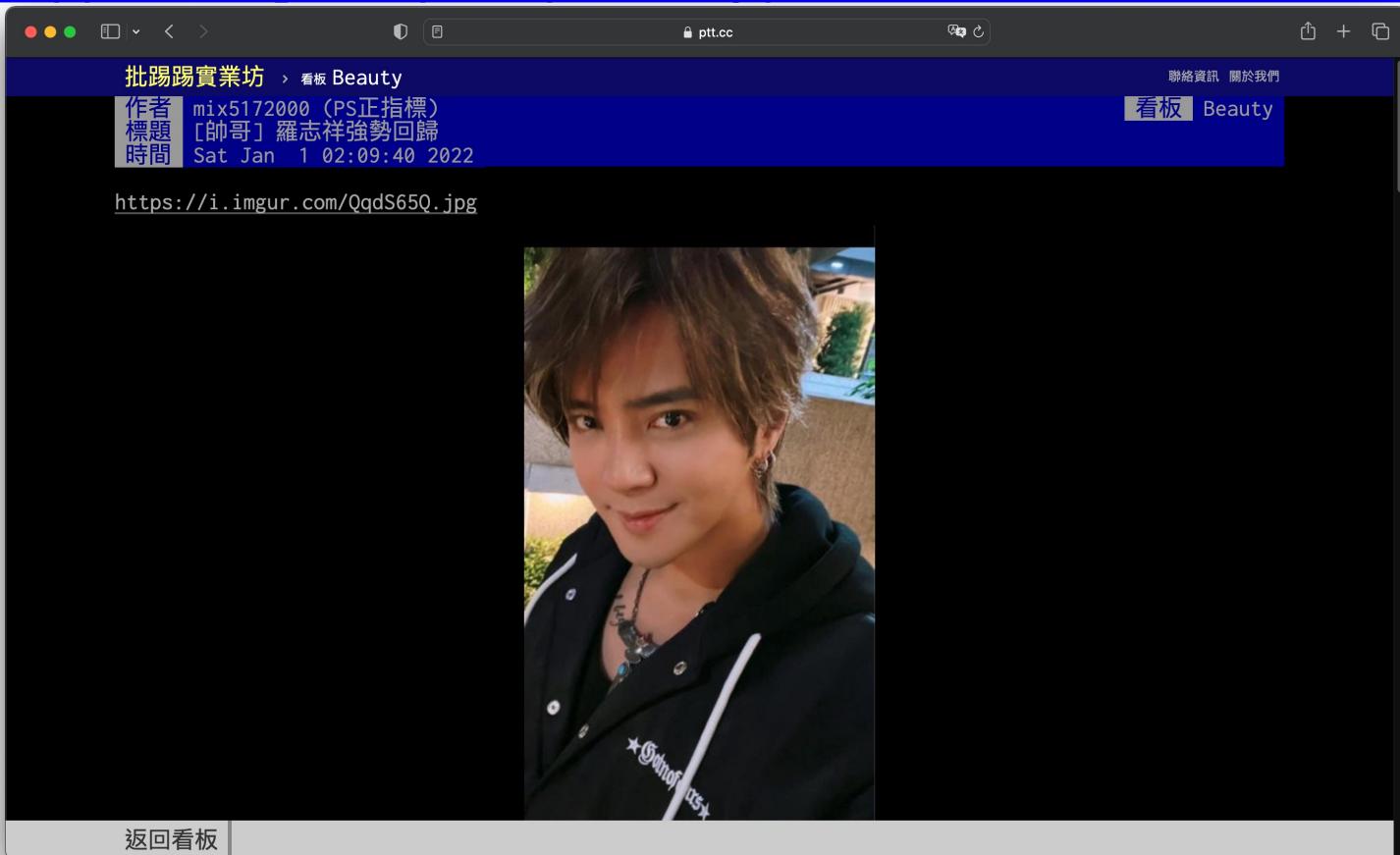
- The top screenshot shows a woman with short dark hair wearing a white lace-trimmed blouse over a red plaid skirt, standing outdoors in a park-like setting.
- The middle screenshot shows a woman with long dark hair holding a small white dog, standing against a pink background.
- The bottom screenshot shows a woman with long dark hair wearing a floral off-the-shoulder top, standing against a pink background.

Each screenshot includes a URL below the image: <https://i.imgur.com/LNFIIMk9.jpg>, <https://i.imgur.com/KpWP9Dm.jpg>, and <https://i.imgur.com/A2LqXBB.jpg>.

2022第一篇文章

[帥哥] 羅志祥強勢回歸

<https://www.ptt.cc/bbs/Beauty/M.1640974182.A.7DB.html>



實作四種 Command Line 功能

- `python {student_id}.py crawl`
爬 2022 年一整年文章
- `python {student_id}.py push {start_date} {end_date}`
計算推文/噓文和找出前 10 名最會推跟噓的人
- `python {student_id}.py popular {start_date} {end_date}`
找爆文和圖片 URL
- `python {student_id}.py keyword {keyword} {start_date} {end_date}`
找內文中含有 {keyword} 的文章中的所有圖片

詳細功能與輸入輸出格式

1. Crawl

python {student_id}.py crawl

例： python 0850726.py crawl

- 實作內容：
 - 爬2022年所有文章
 - 忽略分類為”[公告]”和”Fw: [公告]”的文章
 - 標題沒網址的可忽略
 - 不用按照日期順序
- 輸入：無
- 輸出：
 - 兩個檔案：
 - all_article.jsonl 包含所有文章
 - all_popular.jsonl 包含所有爆文
 - 兩個檔案的格式都是jsonlines，每個json的格式為：
{"date": "{date}", "title": "{標題}", "url": "{文章網址}"}
日期格式為MMDD，例如3/4為0304，12/31為1231

1. Crawl 輸出檔案範例

14 [神人] 中國主播-2022第一雪乳 WiserWilly	1/01	...
7 [正妹] 宋是昀 HarunaOno	1/01	...
20 [神人] 這位女優是誰 ? justin21138	1/01	...
1 [正妹] 大尺碼 肉特(801) ckpot	1/01	...
爆 [正妹] 2022年，即將三十而立的女孩兒們 cjrmt	1/01	...
6 [正妹] 香港日本混血模特兒 柳斐Winnie Liu jor03twtw	1/01	...
10 [正妹] 喜歡在床邊伸懶腰 asxc530530	1/01	...
9 [正妹] 蘇慧倫 jerryyuan	1/01	...
56 [正妹] MOMO真的超會扭 playerunknow	1/01	...

```
{"date": "0101", "title": "[神人] 中國主播-2022第一雪乳", "url": "https://www.ptt.cc/bbs/Beauty/M.1640975120.A.1C6.html"}  
{"date": "0101", "title": "[正妹] 宋是昀", "url": "https://www.ptt.cc/bbs/Beauty/M.1641000613.A.A6D.html"}  
{"date": "0101", "title": "[神人] 這位女優是誰 ?", "url": "https://www.ptt.cc/bbs/Beauty/M.1641008099.A.11D.html"}  
{"date": "0101", "title": "[正妹] 大尺碼 | 肉特(801) ", "url": "https://www.ptt.cc/bbs/Beauty/M.1641009158.A.39B.html"}  
{"date": "0101", "title": "[正妹] 2022年 即將三十而立的女孩兒們", "url": "https://www.ptt.cc/bbs/Beauty/M.1641026293.A.96D.html"}  
{"date": "0101", "title": "[正妹] 香港日本混血模特兒 柳斐Winnie Liu", "url": "https://www.ptt.cc/bbs/Beauty/M.1641031118.A.D1B.html"}  
{"date": "0101", "title": "[正妹] 喜歡在床邊伸懶腰", "url": "https://www.ptt.cc/bbs/Beauty/M.1641038162.A.8F6.html"}  
{"date": "0101", "title": "[正妹] 蘇慧倫", "url": "https://www.ptt.cc/bbs/Beauty/M.1641038666.A.3B9.html"}  
{"date": "0101", "title": "[正妹] MOMO真的超會扭", "url": "https://www.ptt.cc/bbs/Beauty/M.1641047155.A.D78.html"}  
...
```

爆文定義

- 當推文數 ≥ 100 的時候，那篇文章就是爆文，在標題旁邊會有一個紅色的爆

14 [神人] 中國主播-2022第一雪乳 WiserWilly	1/01	...
7 [正妹] 宋昱昀 HarunaOno	1/01	...
20 [神人] 這位女優是誰 ? justin21138	1/01	...
1 [正妹] 大尺碼 肉特(801) ckpot	1/01	...
爆 [正妹] 2022年，即將三十而立的女孩兒們 cjrmrt	1/01	...

批踢踢實業坊 > 看板 Beauty

聯絡資訊 | 關於我們

看板 精華區 ReiKuromiya

日期 最舊 < 上頁 下頁 > 最新 1/04 ...

爆 [正妹] 台灣很會做甜點的女孩 AmedRosario

32 [正妹] 戈偉如 YenFuOne

8 [正妹] 雪子大大 杉咲花 scott880514

13 [正妹] 俄羅斯蘿莉EvaR 粉色芭蕾服 playerunknown

[正妹] 大尺碼 | 肉特(805) ckpot

26 [正妹] Colleen Cole 美國模特兒 172cm Aotearoa

31 [正妹] 月尾島樂園員工 ReiKuromiya

6 [正妹] 142 ReiKuromiya

16 [正妹] 被逗笑的女警 teramars

6 [公告] 水桶 r66

[正妹] 大尺碼 | 肉特(806) ckpot

推文數

標題

忽略公告

The diagram illustrates how various search filters or sorting criteria are applied to the list of posts on the PTT board. The red annotations highlight the date, post count, and title filters, while the yellow annotation highlights the ignore announcement filter.

URL

https://www.ptt.cc/bbs/Beauty/M.1641000613.A.A6D.html

批踢踢實業坊 > 看板 Beauty

作者 HarunaOno (Scandal最高)
標題 [正妹] 宋是昀
時間 Sat Jan 1 09:30:08 2022

聯絡資訊 關於我們

看板 Beauty

<https://i.imgur.com/LNFIk9.jpg>



返回看板

2. Push

python {student_id}.py push {start_date} {end_date}
例：python 0850726.py push 0304 1231

- 實作內容：
找出在 start_date(含) 跟 end_date(含) 之間的：
 - 推文和噓文的總數量
 - 推文數最多的前10名
 - 噓文數最多的前10名
- 輸入：
 - {start_date}
 - {end_date}日期格式為MMDD，例如3/4為0304，12/31為1231

2. Push

- 輸出：
 - 將結果輸出至檔案

push_{start_date}_{end_date}.json

例: push_0304_1231.json

- json格式

{

```
"all_like": {總推文數量},  
"all_boo": {總噓文數量},  
"like 1": {"user_id": "{user id}", "count": {推文數}},  
"like 2": {"user_id": "{user id}", "count": {推文數}},  
...  
"boo 1": {"user_id": "{user id}", "count": {推文數}},  
"boo 2": {"user_id": "{user id}", "count": {推文數}},  
...  
}
```

2. Push 輸出範例

```
{  
  "all_like": 1210,  
  "all_boo": 217,  
  "like 1": {"user_id": "maxxxxxx", "count": 10},  
  "like 2": {"user_id": "popptt", "count": 7},  
  "like 3": {"user_id": "issemn", "count": 7},  
  "like 4": {"user_id": "highka1003", "count": 7},  
  "like 5": {"user_id": "asdfg5435", "count": 7},  
  "like 6": {"user_id": "HarunaOno", "count": 7},  
  "like 7": {"user_id": "playerunknow", "count": 6},  
  "like 8": {"user_id": "cms6384", "count": 6},  
  "like 9": {"user_id": "YummyMcGee", "count": 6},  
  "like 10": {"user_id": "s5689", "count": 5},  
  "boo 1": {"user_id": "gtoamdk7", "count": 5},  
  "boo 2": {"user_id": "supahotfire", "count": 3},  
  "boo 3": {"user_id": "ss810163", "count": 3},  
  "boo 4": {"user_id": "angeltear15", "count": 3},  
  "boo 5": {"user_id": "witness0828", "count": 2},  
  "boo 6": {"user_id": "straypoet", "count": 2},  
  "boo 7": {"user_id": "s0920142", "count": 2},  
  "boo 8": {"user_id": "ptt821105", "count": 2},  
  "boo 9": {"user_id": "playerunknow", "count": 2},  
  "boo 10": {"user_id": "peter28222", "count": 2}  
}
```

如果count相同，
則user id字典序
較小者rank較大。
也就是排序較後。
保證時間區間內
推文和噓文人數
都至少10人

批踢踢實業坊 > 看板 Beauty

推 c0961128831g: <https://i.imgur.com/YizXWWh.jpg> 223.141.11.184 01/03 20:42 聯絡資訊 關於我們

推 Ging: <https://i.imgur.com/rJqps9r.jpg> 111.243.43.2 01/03 20:44

推文

user id

虛文

推 freener: 哭哦
推 yhji: 不能只有我看到
虛 sses40713: 爸媽
虛 harrishu: 幹
推 eddysamsam: 不能只有我看到

42.72.201.118 01/03 20:48
106.1.175.180 01/03 20:50
58.114.171.88 01/03 21:03
101.10.7.121 01/03 21:10
42.73.188.197 01/03 21:18

返回看板

3. Popular

python {student_id}.py popular {start_date} {end_date}
例：python 0850726.py popular 0304 1231

- 實作內容：
找出在 start_date(含) 跟 end_date(含) 之間的：
 - 爆文數量
 - 爆文內的所有圖片的URL，包括推文和噓文的URL，
開頭必須是http://或https://，並且要以 jpg, jpeg, png, gif 為副檔名結尾，副檔名不限大小寫。
- 輸入：
 - {start_date}
 - {end_date}

日期格式為MMDD，例如3/4為0304，12/31為1231

3. Popular

- 輸出:
 - 將結果輸出至檔案
popular_{start_date}_{end_date}.json
例: popular_0304_1231.json
 - json格式

```
{  
    "number_of_popular_articles": {爆文數量},  
    "image_urls": [  
        "{圖片的URL_1}",  
        "{圖片的URL_2}",  
        ...  
    ]  
}
```

3. Popular 輸出範例

```
{  
  "number_of_popular_articles": 2,  
  "image_urls": [  
    "https://i.imgur.com/v7A79jh.jpg",  
    "https://i.imgur.com/7CjiiTK.jpg",  
    "http://i.imgur.com/uJP6lho.jpg",  
    ...  
  ]  
}
```

批踢踢實業坊 › 看板 Beauty

作者 HarunaOno (Scandal最高)
標題 [正妹] 宋是昀
時間 Sat Jan 1 09:30:08 2022

聯絡資訊 關於我們 看板 | Beauty 圖片 URL

<https://i.imgur.com/LNFIIMk9.jpg>

返回看板

圖片URL

批踢踢實業坊 > 看板 Beauty

推 wafiea708: 朱喜悅要推一個
→ wafiea708 <https://i.imgur.com/IHXSTkf.jpg>

123.193.199.130 01/01 15:42
123.193.199.130 01/01 15:43



A screenshot of a forum post from the "Beauty" board on PTT. The post contains a link to an image of a woman in a dark green dress. A red arrow points from the text "https://i.imgur.com/IHXSTkf.jpg" to the word "圖片URL" (Image URL) in a red box at the top right. The image itself shows a woman with long dark hair, wearing a dark green, form-fitting dress, standing against a dark background. The URL in the post is highlighted with a red oval.

返回看板

4. Keyword

python {student_id}.py keyword {keyword} {start_date} {end_date}
例: python 0850726.py keyword 正妹 0304 1231

- 實作內容：
 - 找出在 start_date(含) 和 end_date (含) 之間且包含 {keyword} 的文章中所有圖片的 URL
 - URL包含在推文和噓文的圖片，格式同 popular
 - {keyword} 只需考慮文章不需考慮推文和噓文
- 輸入：
 - {keyword}
 - {start_date}
 - {end_date}

{keyword} 不含空白字元

日期格式為MMDD，例如3/4為0304，12/31為1231

4. Keyword

- 輸出：
 - 將結果輸出至檔案
keyword_{keyword}_{start_date}_{end_date}.json
例: keyword_正妹_0304_1231.json
 - json格式：

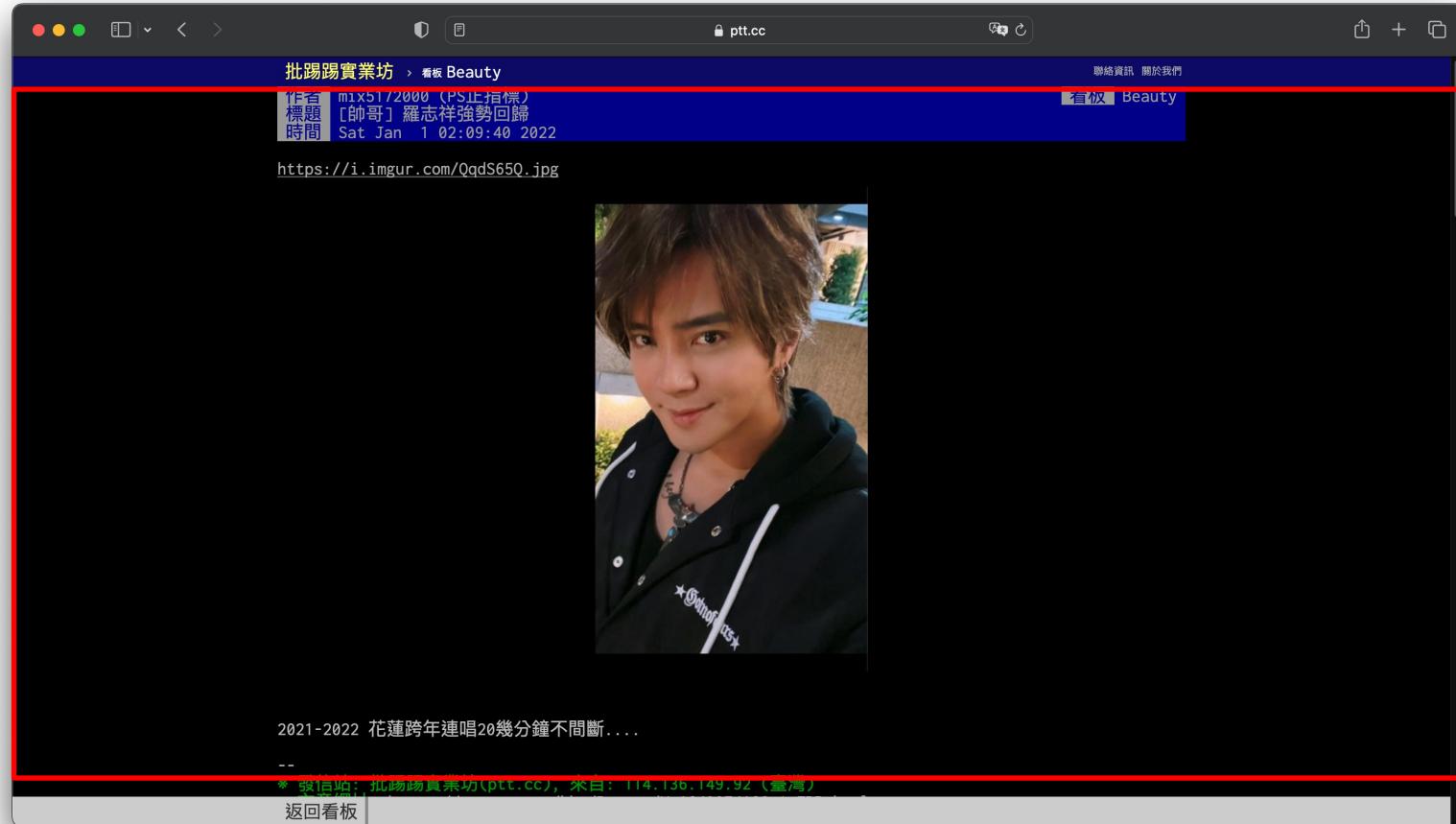
```
{  
    "image_urls": [  
        "{圖片的URL}",  
        "{圖片的URL}",  
        ...  
    ]  
}
```

4. Keyword 輸出範例

```
{  
  "image_urls": [  
    "https://i.imgur.com/LNFIMk9.jpg",  
    "https://i.imgur.com/KpWP9Dm.jpg",  
    "http://i.imgur.com/A2LqXBB.jpg",  
    ...  
  ]  
}
```

開頭是http://或https://，並且要以 jpg, jpeg, png, gif 為結尾，不限大小寫。

4. Keyword 判斷範圍



- 從「作者」（含）開始到綠色的「※ 發信站」（不含）之間，只要有出現 {keyword} 就算這篇文章包含 {keyword} 。
- 內文標題和文章列表標題可能不同，以內文為準。

綜合規則

- 如果文章列表的標題沒有URL可以連結到內文，則以上四種功能都不需考慮這篇文章。例如：某些被刪除的文章會沒有URL
- 以上四種功能都不需考慮類別為 “[公告]” 和 “Fw: [公告]” 的文章
- 如果文章內沒有出現綠色的「※ 發信站」，則 keyword 功能不需考慮此類文章，但 crawl、popular 和 push 需要考慮此類文章，這樣判斷邏輯比較簡單，可以參考投影片後面的提示
- 對於List類型的輸出，評分時不比較順序。不用刪除重複的URL
- 所有寫檔位置都是當前目錄

評分方式

- **評分方式**
 - 作業繳交截止後測試 `{student_id}.py`
 - 由助教測試多種輸入參數
- **配分 (total 90%)**
 - crawl: 24%
 - push: 21% (3個input cases，每個7%)
 - popular: 21% (3個input cases，每個7%)
 - Keyword: 24% (3個input cases，每個8%)
 - 沒輸出檔案就沒有分

時間限制

- 各個功能的時間限制如下
 - Crawl: 30分鐘
 - Push: 30分鐘
 - Popular: 30分鐘
 - Keyword: 30分鐘
- 如果你的程式運行一次的時間超過對應的時限，
則process會直接被kill

提示

- 第一個指令一定是 crawl
- 其餘任務請用 crawl 的 output 來繼續動作節省時間，不要每個任務都重新 crawl
- 像是用 all_article.jsonl 裡面的 URL 去抓時間區間內的推噓文

可用套件

- 所有python預設套件都可使用
- 需要另外安裝的套件必須包含在以下名單內
- 如果有名單外的套件需要使用可以和助教反應
- 更新此名單時會在e3公告讓所有人知道
- 請使用 Python 3.10.x

requests	beautifulsoup4	lxml
scrapy	pyquery	click
tqdm	pandas	

測試環境

如果嚴格遵守可用套件規範，不用擔心在批改時無法執行。

Docker：

- Base Image: ubuntu:22.04
- Python Version: Python 3.10.x
- RAM: 8G

如果單次執行超過這個上限可能會被系統自動kill

執行環境

如果想要使用和助教完全相同的環境，可以自行安裝 docker並pull助教提供的容器，會內建python 3.10.x 和所有名單內的套件

- Pull容器

```
docker pull yi1un/ds2023_hw1_1
```

- 在{student_id}.py所在的資料夾中執行docker，以push為例，windows必須使用PowerShell：

```
docker run --rm -v ${PWD}:/crawler \
yi1un/ds2023_hw1_1 \
python3 {student_id}.py push 0304 1231
```

繳交內容

- 請繳交一個 `{student_id}.py` 到 e3。`{student_id}` 請替換成你的學號
- Code 不會太長，所以請將全部東西寫在一個檔案

HW1-2 (15%)

- A small project from the crawled data!
- Goal: We want to build a binary classifier for classifying whether the image are in a popular article or not.
- Popular is 1 and unpopular is 0.
- [DEF] Popular is the article with more than 35 pushes.

Given the image

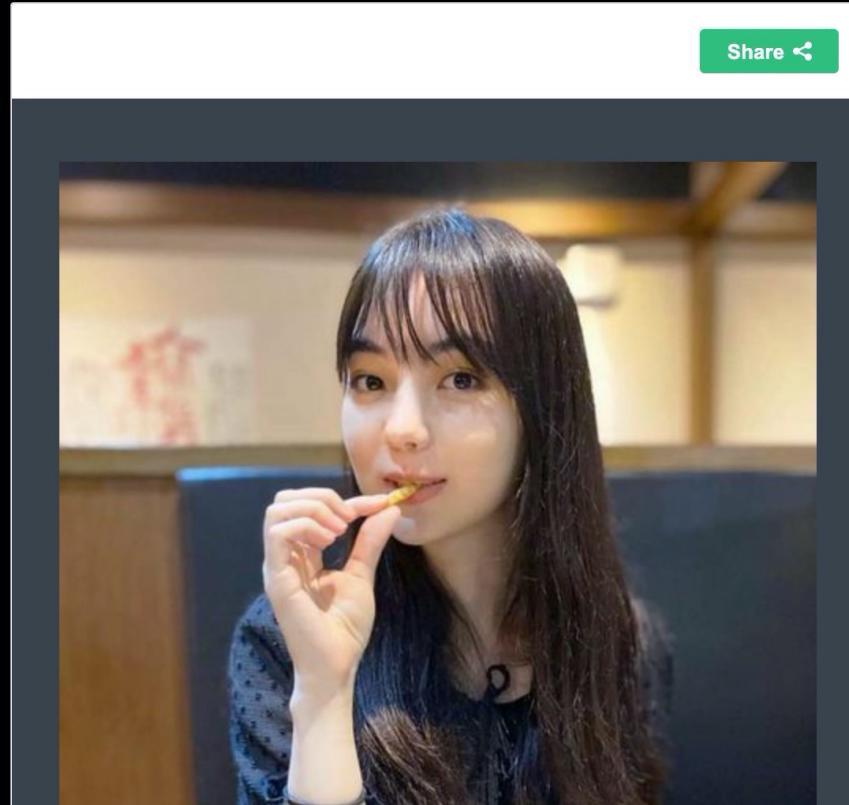
批踢踢實業坊 › 看板 Beauty

聯絡資訊 關於我們

作者 snh48spush (后里趙子龍)
標題 [正妹] 黑嘉嘉
時間 Sun Sep 13 08:22:18 2020

看板 Beauty

<https://i.imgur.com/TCFIMpU.jpg>



We want to get the popularity!

59 [正妹] 高中老師
teramars

9/13 ...

爆 [正妹] 黑嘉嘉
snh48spush

9/13 ...

51 [正妹] 三個小惡魔要如何收服
IS5F5566

9/13 ...

Difficult Example

批踢踢實業坊 › 看板 Beauty

聯絡資訊 關於我們

作者 ClownT (克朗)
標題 [帥哥] 金城武生氣
時間 Sat Apr 27 14:05:37 2019

看板 Beauty

表特一團亂
金城武生氣了

<https://imgur.com/tdjt7UQ.jpg>



推	MadAdipose:	抱歉那天遊戲卡關	04/27 14:07
推	kuroboy009:	一定是玩鬼武者一直死生77	04/27 14:13
推	linear:	怒	04/27 14:21
推	hdotistyle:	抱歉 之前隻狼一直卡關 讓大家困擾了	04/27 14:22
→	fakejoker:	我沒生氣啦 只是在沉思	04/27 14:34
推	j21802:	我沒笑的時候就這樣不好意思	04/27 15:27
推	saturn22k:	好久沒有表特板我很生氣	04/27 17:09
推	blue0310:	不好意思，因為我對這板主不是很滿意	04/27 17:34
推	loloman:	抱歉！見醜了	04/27 18:06
推	ncu1319:	(第一張) 武哥：「你到底是怎麼管板的？」	04/27 18:06
→	depe5175:	各位對不起 我森77	04/27 21:51
推	hogu134:	我沒有生氣！！！	04/28 01:06
→	steven56138:	抱歉我不願上表特	04/28 02:03
推	jason138:	對不起讓你看到我生氣的樣子	04/28 02:09

輸入輸出說明

- `python {student_id}_pred.py {image_path_list}.txt`
例：`python 0850726_pred.py test.txt`
- 輸入：
 - `{image_path_list}.txt` 的格式為一行一個絕對路徑，可直接讀取不需處理相對路徑：
`/mnt/data/01.jpg`
`/mnt/data/02.png`
...
- 輸出：
 - 輸出至 `{student_id}.txt`（當前資料夾下）
 - 格式為一行01字串，代表分類結果：
`10111000.....01`

HW1-2評分方式

- **評分方式**
 - 於作業繳交截止後測試`{student_id}_pred.py`
 - `{image_path_list}.txt` 中會有約100個圖片路徑
- **配分 (total 15%)**
 - Top-20%: 15
 - Top-40%: 12
 - Top-60%: 9
 - Top-80%: 6
 - F1-score is greater than 0.51: 3

可用套件

- 所有python預設套件都可使用
- 需要另外安裝的套件必須包含在以下名單內
- 如果有名單外的套件需要使用可以和助教反應
- 更新此名單時會在e3公告讓所有人知道
- 請使用 Python 3.10.x

<code>torch==1.13.1</code>	<code>torchvision==0.14.1</code>
<code>scikit-learn==1.2.1</code>	<code>opencv-python==4.7.0.68</code>
<code>tensorflow==2.11.0</code>	<code>timm==0.6.2</code>
<code>tqdm</code>	<code>click</code>
<code>joblib==1.2.0</code>	<code>pandas</code>

執行環境

Docker

- Image: ubuntu:22.04
- Python Version: Python 3.10
- RAM: 8G
 - 如果單次執行超過這個上限可能會被系統自動kill
- 測試環境不提供GPU

HW1-2繳交內容

- 請繳交一個`{student_id}.zip`壓縮檔到e3。其中含有以下檔案：
 - `{student_id}_pred.py`
 - 你的模型參數
- 注意，`{student_id}_pred.py`不能在任何資料夾底下，其必須在最外層。也就是說用unzip指令解壓縮後會直接看到這個檔案。

在Windows上的類似操作就是對壓縮檔右鍵選擇解壓縮至此，會在當前目錄直接看到
`{student_id}_pred.py`

執行環境 (Optional)

如果想要使用和助教完全相同的環境，可以自行安裝 docker並pull助教提供的容器，會內建python 3.10.x 和所有名單內的套件

- Pull容器

```
docker pull yi1un/ds2023_hw1_2
```

- 在{student_id}.py所在的資料夾中執行docker，以crawl為例，windows必須使用PowerShell：

```
docker run --rm -v ${PWD}:/crawler \
yi1un/ds2023_hw1_2 \
python3 {student_id}_pred.py
```

Potentially useful materials

- Face rating:
 - https://github.com/HuyTu7/face_rating
- Facial Beauty Prediction
 - <https://towardsdatascience.com/how-attractive-are-you-in-the-eyes-of-deep-neural-network-3d71c0755ccc>

其他規範

- 嚴格遵守所有功能的 command line arguments(CLI) 規範、輸入/輸出規範、執行時間、記憶體用量、壓縮檔格式等
- 如果是 CLI 、輸入/輸出、壓縮檔格式不合規定導致無法順利批改作業，較嚴重者成績會打9折
- 套件相容性問題有較大彈性空間，記得提前告知TA 以公告讓所有人知道
- 禁止抄襲來自任何地方的資源。可以參考 Github 或是任何技術文章，請吸收後根據自己的理解寫 code ，抓到抄襲就0分
- 如果有任何問題請寄信給助教
吳易倫 yilun.ee08@nycu.edu.tw