



NYCU Data Science - Homework 1

繳交期限: 2023/03/11 23:59

TA Hour: 週二 13:20~15:20

TA Email: yilun.ee08@nycu.edu.tw (<mailto:yilun.ee08@nycu.edu.tw>)

[90%] Part 1. Crawler

目標

爬PTT Beauty板2023一整年的文章，2023年的第一篇文章為 [正妹] 周子瑜

<https://www.ptt.cc/bbs/Beauty/M.1672503968.A.5B5.html>

(<https://www.ptt.cc/bbs/Beauty/M.1672503968.A.5B5.html>).。



需要繳交一份python腳本，接下來以 `{student_id}.py` 來代表這一腳本。其必須支援四種命令列介面功能：

- Crawl
- Push
- Popular
- Keyword

[24%] Crawl

後續所有功能（**Push**、**Popular**、**Keyword**）都是基於**Crawl**找到的文章在指定區間內找對應資訊，所以被**Crawl**忽略的文章不用被後續的功能考慮。

- 格式：

```
$ python {student_id}.py crawl
```

範例：

```
$ python 0850726.py crawl
```

- 功能：

- 爬2023年所有文章。
- 忽略分類為 [公告] 和 Fw:[公告] 的文章。
- 忽略沒有標題缺少對應網址的文章。

- 輸入：

沒有輸入。

- 輸出：

- 格式說明：
在當前資料夾輸出兩個檔案：

- articles.jsonl
包含所有文章。
- popular_articles.jsonl
包含所有推爆的文章。

兩個檔案的格式均為jsonlines，其中每一個json代表一篇文章的資訊，文章不需要按照日期排序，其json格式為：

```
{  
    "date": "{date}",  
    "title": "{標題}",  
    "url": "{文章網址}"  
}
```

- 範例 articles.jsonl 的前5行：

```
{"date": "0101", "title": "[正妹] 周子瑜", "url": "https://www.ptt.cc/bbs/Beauty/M.1561111111.A.0101.html"},  
 {"date": "0101", "title": "[帥哥] 星野結衣的老公 星野源", "url": "https://www.ptt.cc/bbs/Beauty/M.1561111111.A.0101.html"},  
 {"date": "0101", "title": "[正妹] 六兔興旺", "url": "https://www.ptt.cc/bbs/Beauty/M.1561111111.A.0101.html"},  
 {"date": "0101", "title": "[神人] 淘寶靴子model", "url": "https://www.ptt.cc/bbs/Beauty/M.1561111111.A.0101.html"},  
 {"date": "0101", "title": "[正妹] 架乃ゆら (架乃由羅)", "url": "https://www.ptt.cc/bbs/Beauty/M.1561111111.A.0101.html"},  
 ...
```

建議用append模式開檔，間歇寫入已經爬取的文章資訊，避免遇到Exception時完全沒有儲存結果。

為了更方便實作多執行緒爬蟲，文章順序不會影響評分。

測試時第一個指令一定會測試**Crawl**，所以後續所有功能都可以使用 `articles.jsonl` 和 `popular_articles.jsonl` 的結果。

- 日期、標題和URL：

批踢踢實業坊 > 看板 Beauty

推文數 - 嘘文數

日期

標題

序號	內容	日期
30	[正妹] aespa Karina wafie708	1/01 ...
[公告]	水桶 lude71 soulknight	1/01 ...
X6	[帥哥] 二兵 蔡育成 succhungzen52	1/01 ...
[公告]	水桶 yokann soulknight	1/02 ...
7	[正妹] Cosplay 048 日本 换装娃娃 Gentlemon	1/02 ...
9	[正妹] 張馨予 gogoe04	1/02 ...
19	[正妹] Ella Netzer 以色列女孩 Kjartan	1/02 ...
[正妹]	馬甲 hateonas	1/02 ...
4	[正妹] 李世榮 ashiol	1/02 ...
19	[正妹] Lia ReiKuromiya	1/02 ...
4	[正妹] 雜 spermjuice	1/02 ...
39	[正妹] 女友感 asxc530530	1/02 ...
28	[正妹] 太妍 wafie708	1/02 ...
6	[正妹] Lisa Matthews 美國模特兒 175cm Kjartan	1/02 ...

批踢踢實業坊 > 看板 Beauty

URL

看板 Beauty

https://www.ptt.cc/bbs/Beauty/M.1672503968.A.5B5.html

作者 ReiKuromiya (ReiKuromiya)

標題 [正妹] 周子瑜

時間 Sun Jan 1 00:26:06 2023

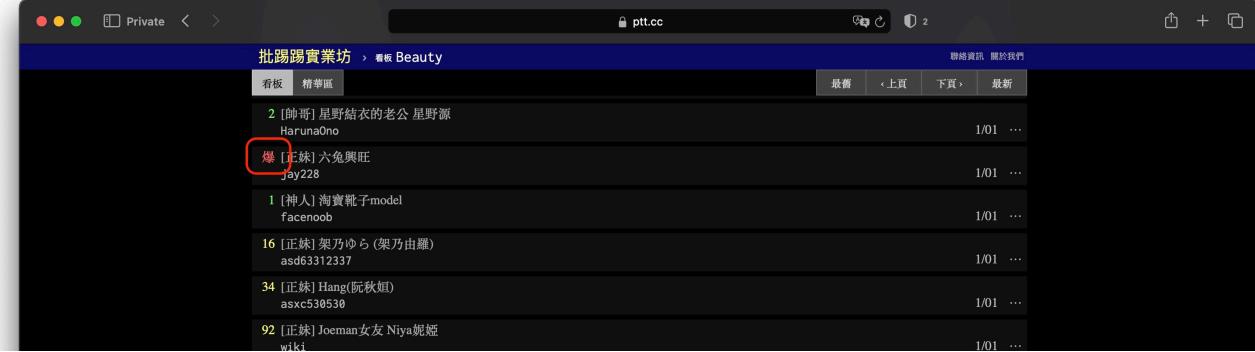
<https://i.imgur.com/BdmZ7Ps.jpg>

<https://i.imgur.com/bBiw4IS.jpg>

- 推爆定義

一篇文章被推文時，文章前會顯示被推文（低調推除外）的次數，但如果被噓文時，會抵銷掉推文的次數，而推文-噓文大於100時，就會變成「推爆」的狀況。但進入爆的狀況之後，系統就不會再計算推文數，而是持續以100來計數，因此即使推文數遠超100，但這時有人噓

文時，數字還是會恢復到99，這時要再有人推才會再變成爆的狀態。（或可稱為「再推爆一次」）



[21%] Push

請基於**Crawl**找到的文章在指定區間內找對應資訊，所有被**Crawl**忽略的文章同樣不用被考慮。

- 格式：

```
$ python {student_id}.py push {start_date} {end_date}
```

範例：

```
$ python 0850726.py push 0304 1231
```

- 功能：

找出在 `{start_date}` (含) 跟 `{end_date}` (含) 之間的以下資訊：

- 推文和噓文兩種行為各自的總數量。
- 進行推文行為最多次的前10名 `user_id`。
- 進行噓文行為最多次的前10名 `user_id`。

- 輸入：

- `{start_date}`、`{end_date}`

格式均為MMDD，例如3/4為 0304，12/31為 1231。

- 輸出：

- 在當前資料夾輸出一個json檔，檔名請按照以下格式：

`push_{start_date}_{end_date}.json`

範例：

`push_0304_1231.json`

- json格式：

```
{  
    "push": {  
        "total": {推文總數},  
        "top10": [  
            {"user_id": "{user id}", "count": {推文數}},  
            {"user_id": "{user id}", "count": {推文數}},  
            ...  
        ]  
    },  
    "boo": {  
        "total": {噓文總數},  
        "top10": [  
            {"user_id": "{user id}", "count": {噓文數}},  
            {"user_id": "{user id}", "count": {噓文數}},  
            ...  
        ]  
    }  
}
```

top10 是依照 count 排序由大到小排序，如果 count 相同，則 user_id 字典序 (Lexicographical Order) 較大者排序在前。如果不滿10人則只需列出僅有的人。

- 輸出範例：

```
{  
  "push": {  
    "total": 1040,  
    "top10": [  
      {"user_id": "maxxxxxx", "count": 6},  
      {"user_id": "Krishna", "count": 6},  
      {"user_id": "yggyygy", "count": 5},  
      {"user_id": "tyrande", "count": 5},  
      {"user_id": "monarch0301", "count": 5},  
      {"user_id": "johnwu", "count": 5},  
      {"user_id": "cityhunter04", "count": 5},  
      {"user_id": "adamlovedogc", "count": 5},  
      {"user_id": "abellea85209", "count": 5},  
      {"user_id": "Lailungsheng", "count": 5}  
    ]  
  },  
  "boo": {  
    "total": 247,  
    "top10": [  
      {"user_id": "QVQ9487", "count": 6},  
      {"user_id": "theclgy2001", "count": 4},  
      {"user_id": "cczoz", "count": 4},  
      {"user_id": "zss40401", "count": 3},  
      {"user_id": "cityhunter04", "count": 3},  
      {"user_id": "yushenglu", "count": 2},  
      {"user_id": "un94su3", "count": 2},  
      {"user_id": "srmember", "count": 2},  
      {"user_id": "sion1993", "count": 2},  
      {"user_id": "saw6904", "count": 2}  
    ]  
  }  
}
```

- 推文、噓文、中立留言說明



中立留言

推文

噓文

推文自動更新已關閉

本網站已依台灣網站內容分級規定處理。此區域為限制級，未滿十八歲者不得瀏覽。

返回看板

推文

中立留言

→ BBWAS: 我以為她要唱愛神

推 bettys111: 真的有點微妙

推 s610052003: 比之前辮子頭好太多了

→ dingading: https://youtu.be/PKEmgy6_qIE

P GEM (鄧詩穎)13歲唱歌比賽面試片

Watch on YouTube

Share

223.137.236.143 01/03 09:45:59

218.164.140.153 01/03 11:09

223.138.129.229 01/03 11:41

114.45.15.140 01/03 13:28

39.14.50.129 01/03 13:28

36.236.171.197 01/03 14:26

106.64.154.180 01/03 15:47

49.216.225.54 01/03 18:46

111.254.217.82 01/03 20:02

114.136.189.36 01/04 12:09

49.216.24.177 01/05 12:48

220.143.30.249 01/10 01:15

39.9.137.188 01/12 15:00

[21%] Popular

請基於**Crawl**找到的文章在指定區間內找對應資訊，所有被**Crawl**忽略的文章同樣不用被考慮。

- 格式：

```
$ python {student_id}.py popular {start_date} {end_date}
```

範例：

```
$ python 0850726.py popular 0304 1231
```

- 功能：

找出在 `{start_date}` (含) 跟 `{end_date}` (含) 之間的以下資訊：

- 推爆數量。
- 推爆文章中的所有圖片URL，包括內文和留言中的圖片URL。
- 圖片URL定義：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}
格式均為MMDD，例如3/4為0304，12/31為1231。

- 輸出：

- 在當前資料夾輸出一個json檔，檔名請按照以下格式：
popular_{start_date}_{end_date}.json
範例：
popular_0304_1231.json
- json格式：

```
{
    "number_of_popular_articles": {推爆數量},
    "image_urls": [
        "{url_1}",
        "{url_2}",
        ...
    ]
}
```

- 輸出範例：

```
{
    "number_of_popular_articles": 2,
    "image_urls": [
        "https://i.imgur.com/UDJQEyi.jpg",
        "https://i.imgur.com/jUrvWQM.jpg",
        "https://i.imgur.com/lU5JTIT.jpg",
        "http://i.imgur.com/spn4dNg.jpg",
        ...
    ]
}
```

評分時不比較順序。不用刪除重複的URL。

- 推文、噓文、中立留言說明

中立留言

→ BBWAS: 我以為她要唱愛神
推 bettys111: 真的有點微妙
推 s610052003: 比之前辮子頭好太多了
→ dingading: https://youtu.be/PKEmgy6_q1E

推文

推 chemistryxixi: 化妝技術越來越好
推 paris27xi: 再美也是美瓊（攤手）
推 qinggg: 推推很賣唱
推 linsred1006: 想帶他吃麥當勞
噓 un94su3: 美在哪？
噓 marchmaymay: 蛛
→ OGC218: 偶台育？
噓 pase139: 哪裡張鳳書了？
噓 jhihheng: 塑膠 充氣娃娃是不是

噓文

39.14.50.129 01/03 13:28
36.236.171.197 01/03 14:26
106.64.154.180 01/03 15:47
49.216.225.54 01/03 18:46
111.254.217.82 01/03 20:02
114.136.189.36 01/04 12:09
49.216.24.177 01/05 12:48
220.143.30.249 01/10 01:15
39.9.137.188 01/12 15:00

[24%] Keyword

請基於**Crawl**找到的文章在指定區間內找對應資訊，所有被**Crawl**忽略的文章同樣不用被考慮。

- 格式：

```
$ python {student_id}.py keyword {start_date} {end_date} {keyword}
```

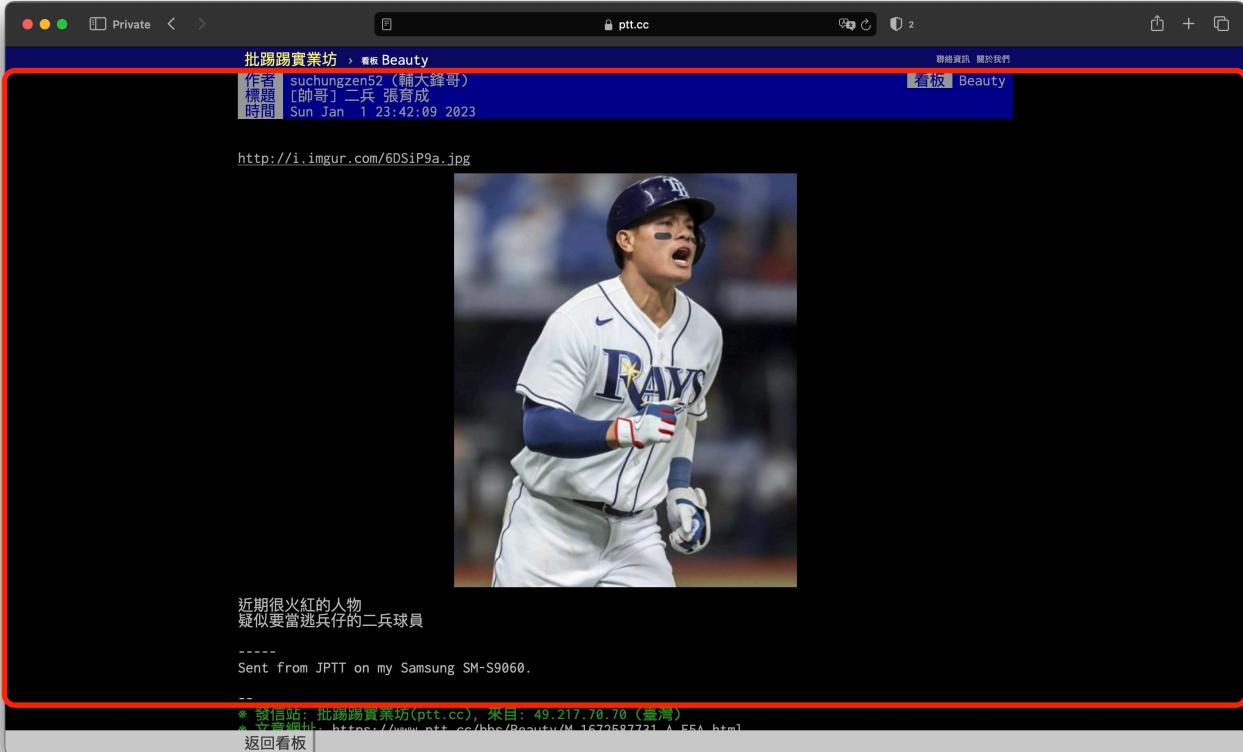
範例：

```
$ python 0850726.py keyword 0304 1231 正妹
```

- 功能：

找出在 `{start_date}` (含) 和 `{end_date}` (含) 之間的文章其內文必須包含 `{keyword}`，並統計這些文章的以下資訊：

- 文章中的所有圖片URL，包括內文和留言中的圖片URL。
- 圖片URL定義：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。
- 內文範圍說明：
 - 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}`。
 - 如果「※ 發信站」不存在，則忽略這篇文章。
 - 內文標題和文章列表顯示的標題可能不同，以內文為準。
 - 網址也在keyword匹配範圍內。



- 輸入:

- {keyword}
保證不包含空白字元 (space、tab ...) 。
- {start_date}、{end_date}
日期格式均為MMDD，例如3/4為 0304，12/31為 1231。

- 輸出:

- 在當前資料夾輸出一個json檔，檔名請按照以下格式：
keyword_{start_date}_{end_date}_{keyword}.json
範例：
keyword_0304_1231_正妹.json
- json格式：

```
{  
    "image_urls": [  
        "{url_1}",  
        "{url_2}",  
        ...  
    ]  
}
```

- 輸出範例：

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

評分時不比較順序。不用刪除重複的URL。

繳交內容

只能繳交一個 .py 檔，名稱為 {student_id}.py，請將 {student_id} 替換為你的學號。

如果繳交格式有任何錯誤，Part1最後成績 $\times 0.95$ 。

測試環境

- 作業系統: Ubuntu 22.04。
- Python 3.10.x。
- 請用 `os.cpu_count()` 取得邏輯處理器數量。
- 只能使用 Python 3.10.x 預設套件和以下套件:

requests	beautifulsoup4	lxml	scrapy
pyquery	click	tqdm	pandas
httpx	pydantic		

評分

- 評分方式
 - 助教對 {student_id}.py 進行測試。
 - 除了 **Crawl** 以外，每種功能測試多組參數。
- 配分

功能	配分	時限(單次執行)
Crawl	24%	15分鐘
Push	21%	15分鐘
Popular	21%	15分鐘
Keyword	24%	15分鐘

如果你的程式執行超過時限會被強制kill。

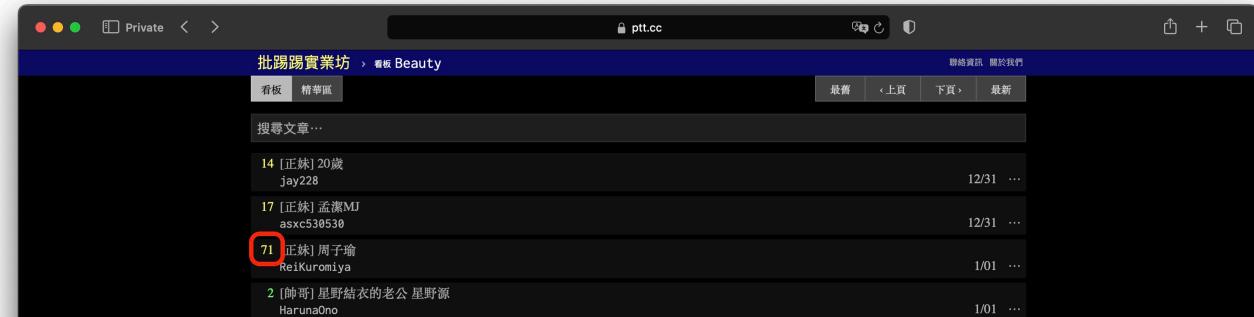
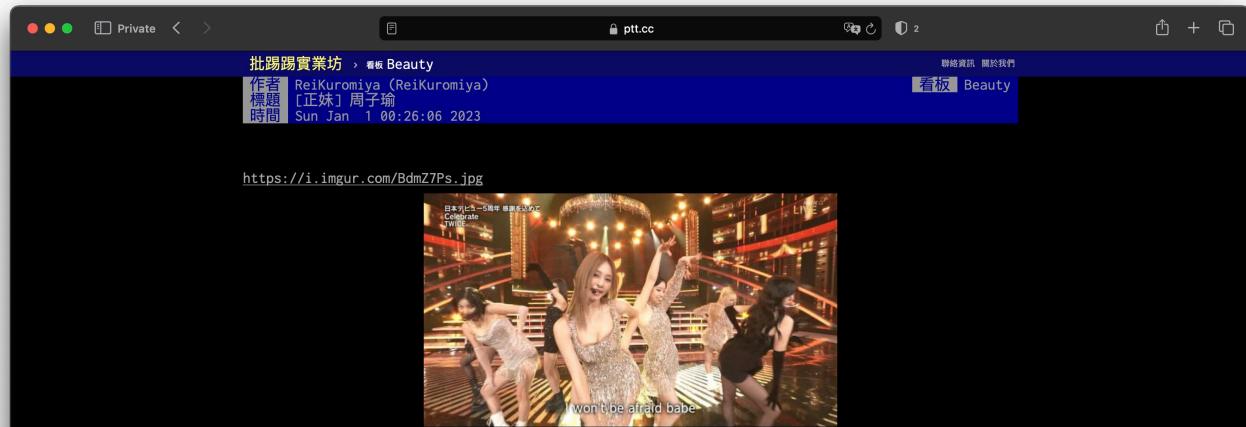
如果沒有輸出檔案，則該筆測試資料0分。

[15%] Part 2. Popular Photo Prediction

建構一個二元分類器來分辨一張圖是不是出自熱門貼文。

熱門貼文定義為文章列表中顯示(推文-噓文)>35的文章。

- 熱門貼文範例：



- 非熱門貼文範例：

批踢踢實業坊 > 看板 Beauty

作者 chirex (不含銅鈸鉛)
標題 [帥哥] 翁立友
時間 Sun Jan 1 20:04:22 2023

聯絡資訊 | 關於我們 | 看板 Beauty

47歲，恭喜走出尾牙的陰影。
結婚了。
<https://i.imgur.com/PEX361R.jpg>

[帥哥] 上班不要看的小士
chirex 1/01 ...

19 [帥哥] 翁立友
chirex 1/01 ...

爆 [正妹] 河北彩花
jay228 1/01 ...

21 [正妹] 日本小馮
asxc530530 1/01 ...

1 [正妹] 川村那月
ReiKuromiya 1/01 ...

本網站已依台灣網站內容分級規定處理。此區域為限制級，未滿十八歲者不得瀏覽。

說明

需要繳交一份壓縮檔，接下來以 `main.py` 代表執行腳本。其必須提供一個命令列介面功能。

- 格式：

```
$ python main.py {test_file}.json
```

範例：

```
$ python main.py images/test.json
```

- 輸入：

{test_file}.json 的格式為：

```
{  
    "image_paths": [  
        "{path_1}",  
        "{path_2}",  
        "{path_3}",  
        ...  
    ]  
}
```

- 圖片路徑為相對 main.py 的本地路徑，可以直接讀檔。
- 圖片副檔名可能為 jpg、jpeg、png 之一。

範例：

```
{  
    "image_paths": [  
        "images/image_1.jpg",  
        "images/image_1.png",  
        "images/image_1.jpeg",  
        ...  
    ]  
}
```

- 輸出：

在當前資料夾輸出一個json檔，檔名為 image_predictions.json 。

- 格式：

```
{  
    "image_predictions": [  
        {is_image_1_popular},  
        {is_image_2_popular},  
        {is_image_3_popular},  
        ...  
        {is_image_i_popular},  
        ...  
    ]  
}
```

- {is_image_i_popular} 表示第i張圖是不是來自熱門貼文，1 代表是，0 代表不是。
注意這個變數的型別是整數。

- 範例：

```
{  
    "image_predictions": [  
        0,  
        1,  
        1,  
        ...  
    ]  
}
```

繳交內容

繳交一個 .zip 檔，名稱為 {student_id}.zip，請將 {student_id} 替換為你的學號。上傳檔案大小限制為250MB。壓縮檔的根目錄中必須含有一個 main.py 為主要執行腳本，例如：

```
0850726.zip
├── main.py
└── model.py
└── subdirectory/ckpt.pth
```

請注意，在產生壓縮檔時如果是對整個資料夾進行壓縮，要避免根目錄是為一個資料夾而缺少 main.py，錯誤範例：

```
0850726.zip
├── 0850726/main.py
├── 0850726/model.py
└── 0850726/subdirectory/ckpt.pth
```

如果繳交格式有任何錯誤，Part2最後成績 $\times 0.7$ 。

測試環境

- 作業系統：Ubuntu 22.04。
- Python 3.10.x。
- 只有CPU，型號 Intel i7-9700K。
- 只能使用 Python 3.10.x 預設套件和以下套件：

tensorflow-cpu==2.15.0.post1	tqdm	pandas
torch==2.2.0	torchvision==0.17.0	click
scikit-learn==1.4.1.post1	opencv-python==4.9.0.80	

請確保在沒有GPU的環境下可以正常執行。

評分

- 評分方式
 - 對 main.py 進行測試。
 - 大約100張圖，均取自2023表特版內文或留言。
 - 時限15分鐘
- 配分
F1-score至少要大於0.51才能得分。

F1-score	配分
\geq Top 20%	15
\geq Top 40%	13
\geq Top 60%	11
> 0.51	10

如果你的程式執行超過時限會被強制kill。

如果沒有輸出檔案，則該筆測試資料0分。