



Data Science Coding Assignment #1 Crawler

Submission Deadline:

2019/10/01 23:55

Submit to E3

Hard deadline, No extensions

Goal

Crawl PTT Beauty板

- 爬2018一整年的文章
- 統計日期內推文跟噓文的數量
- 找出日期內最會推跟最會噓的人各前10名
- 統計日期內爆文的數量
- 抓取日期內爆文的所有圖片URL
- 關鍵字查詢

Goal

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

批踢踢實業坊 > 看板 Beauty

聯絡資訊 關於我們

看板 楊華區

S [正妹] 12歲羅刹網紅「童顏巨乳」 網友驚：發育
3/04 GhostFather

爆 [正妹] 同事
3/05 dani0921

10 [正妹] 小隻馬
3/05 HaChoo

4 [女神] 金髮正妹
3/05 archalkyrie

X4 M [正妹] 一點點努力
3/05 centergym

14 [正妹] 17歲正妹Duda Reis
3/05 edisonjuly

6 [女神] 寶特瓶廣告混血長腿學生妹
3/05 dontz

71 [公告] 不願上表格 & 優文推薦 & 梅學建議專區
10/04 ffwind

! [公告] 表特板板規 (2015.2.12)
2/12 GeminiMan

爆 M [公告] 對於謾罵，希望大家將心比心
4/27 ffwind

爆 M [公告] 版規修訂 - 意淫文字
10/05 GeminiMan

11 M [公告] 偷拍相關板規修訂
10/10 GeminiMan

◎ 發信站: 批踢踢實業坊(ptt.cc), 來自: 36.225.58.160
◎ 編輯網: dani0921 (36.225.58.160), 03/05/2018 00:19:11

推 goanchor: 推第一個超正
03/05 00:22

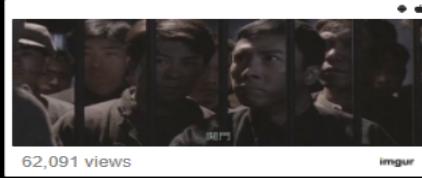
推 bruce666: 已羨慕
03/05 00:22

推 blessRM: 我選第二個
03/05 00:23

推 jou0129: 空姐?
03/05 00:23

推 tpwin7: 李剴!!!
03/05 00:31

→ tpwin7: 開門 <https://i.imgur.com/rXr1JR4.jpg>
03/05 00:31



62,091 views imgur

批踢踢實業坊 > 看板 Beauty

聯絡資訊 關於我們

看板 楊華區

作者 dani0921 (想睡覺)
標題 [正妹] 同事
時間 Mon Mar 5 00:18:20 2018

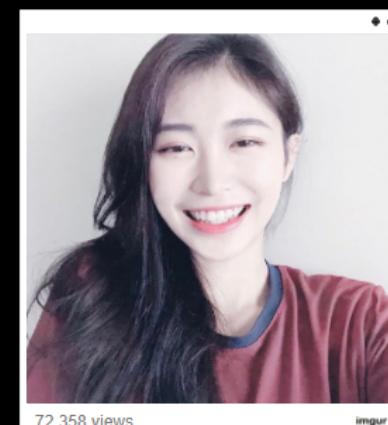
兩個同事笑起來都好美

同事1
<https://i.imgur.com/iMRYkF7.jpg>



71,963 views imgur

https://i.imgur.com/7tRPEkV.jpg



推 sexen: 交出來
推 mhtvpz: 2號大心!!!!!!
推 whatupjk: 啊問公司？？我去應徵掃地就好我要求不多！
推 unitehung: 我當顧客就好
→ tomwu1993: 歐三笑啊？為什麼你同事那麼正
→ tomwu1993: 樓下幫推餅乾，門問問
推 leo91240: 樓上啥意思？
→ tomwu1993: 就是又驚訝又感嘆怎麼別人同事這麼正，還一次兩個，剛
→ tomwu1993: 好推CD的意思
推 hank85202: 原po也可愛
推 lp3388: 該到哪裡報到 我是來應徵工作的

03/05 00:34
03/05 00:35
03/05 00:36
03/05 00:39
03/05 00:42
03/05 00:42
03/05 00:43
03/05 00:45
03/05 00:45
03/05 00:56
03/05 01:07

2018第一篇文章

[正妹] 有村架純

<https://www.ptt.cc/bbs/Beauty/M.1514740613.A.FF1.html>

批踢踢實業坊 › 看板 Beauty

聯絡資訊 關於我們

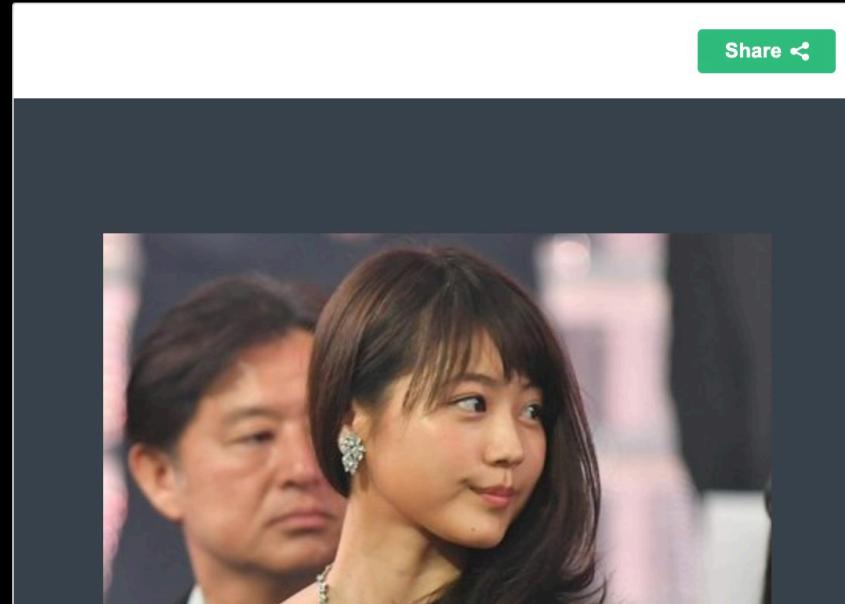
作者 as314 (幽默販賣機)

看板 Beauty

標題 [正妹] 有村架純

時間 Mon Jan 1 01:16:50 2018

<https://i.imgur.com/RIo8fVu.jpg>



Requirements

- Implement with python **3.6**
- Strictly follow input/output formats
- Do not copy/paste others' codes
 - You can refer to the codes on GitHub or anywhere else
 - But you need to write your own code

Functions to implement

四種功能：crawl, push, popular, keyword

- **python {studentID}.py crawl**
 - 爬2018年一整年文章
- **python {studentID}.py push start_date end_date**
 - 數推文噓文和找出前10名最會推跟噓的人
- **python {studentID}.py popular start_date end_date**
 - 找爆文和圖片URL
- **python {studentID}.py keyword {keyword} start_date end_date**
 - 找內文中含有{keyword}的文章中的所有圖片

爆文的定義

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

批踢踢實業坊 › 看板 Beauty

看板 精華區

32	[正妹] 腿不錯鄰家女孩	3/04 maxxxxxx
爆	[正妹] 時間小偷企鵝	3/04 onwaytothend
X4	(K:X4 7 days) <soundbox>	3/04 -
	[正妹] 玩cosplay的妹仔	3/04 manu1119
3	[正妹] 三張	3/04 joker00507
22	[正妹] 非常豔Anastasia Martzipanov	3/04 edisonjuly

詳細的功能與input/output

1. Crawl

Python {studentID}.py crawl

沒網址的不用爬，
可忽略。

- Input: N/A
- 程式內容：
 - 爬2018年所有文章
 - 忽略分類為”公告”的文章
- Output:
 - 存成兩個檔案：
 - 1. all_articles.txt(所有文章)
 - 2. all_popular.txt (所有爆文)
 - 檔案內容格式：
 - 日期,標題,URL (逗號後無空格)

範例：

304,[正妹] 12歲羅莉網紅「童顏巨乳」 網友驚：發育,https://www.ptt.cc/bbs/Beauty/M.1520178886.A.658.html
305,[正妹] 同事,https://www.ptt.cc/bbs/Beauty/M.1520180302.A.086.html
305,[正妹] 小隻馬,https://www.ptt.cc/bbs/Beauty/M.1520182931.A.B07.html
305,[神人] 金髮正妹,https://www.ptt.cc/bbs/Beauty/M.1520183422.A.929.html
305,[正妹] 一點點兒,https://www.ptt.cc/bbs/Beauty/M.1520196795.A.9F5.html

批踢踢實業坊 > 看板 Beauty

看板 精華區

S [正妹] 12歲羅莉網紅「童顏巨乳」 網友驚：發育
3/04 GhostFather

爆 [正妹] 同事
3/05 dani0921

10 [正妹] 小隻馬
3/05 HaChooo

4 [神人] 金髮正妹
3/05 archvalkyrie

X4 M [正妹] 一點點兒
3/05 centergym

看板

精華區

日期

最舊

< 上頁

下頁 >

最新

S [正妹] 12歲羅莉網紅「童顏巨乳」 網友驚：發育
3/04 GhostFather

爆 [正妹] 同事
3/05 dani0921

10 [正妹] 小隻馬
3/05 HaChooo

4 [神人] 金髮正妹
3/05 archvalkyrie

X4 M [正妹] 一點點兒
3/05 centergym

14 [正妹] 17歲正妹Duda Reis
3/05 edisonjuly

推文數

標題

6 [神人] 寶味味廣告混血長腿學生妹
3/05 dontz

1 [正妹] 水野朝陽
3/05 ClownT

71 [公告] 不願上表特 & 優文推薦 & 檢舉建議專區
10/04 ffwind

忽略公告

[公告] 表特板板規 (2015.2.12)
2/12 GeminiMan

爆 M [公告] 對於謾罵，希望大家將心比心
4/27 ffwind

爆 M [公告] 板規修訂 - 意淫文字
10/05 GeminiMan

11 M [公告] 偷拍相關板規修訂
10/10 GeminiMan

批踢踢實業坊 > 看板 Beauty

聯絡資訊 關於我們

作者 dani0921 (想睡覺)
標題 [正妹] 同事
時間 Mon Mar 5 00:18:20 2018

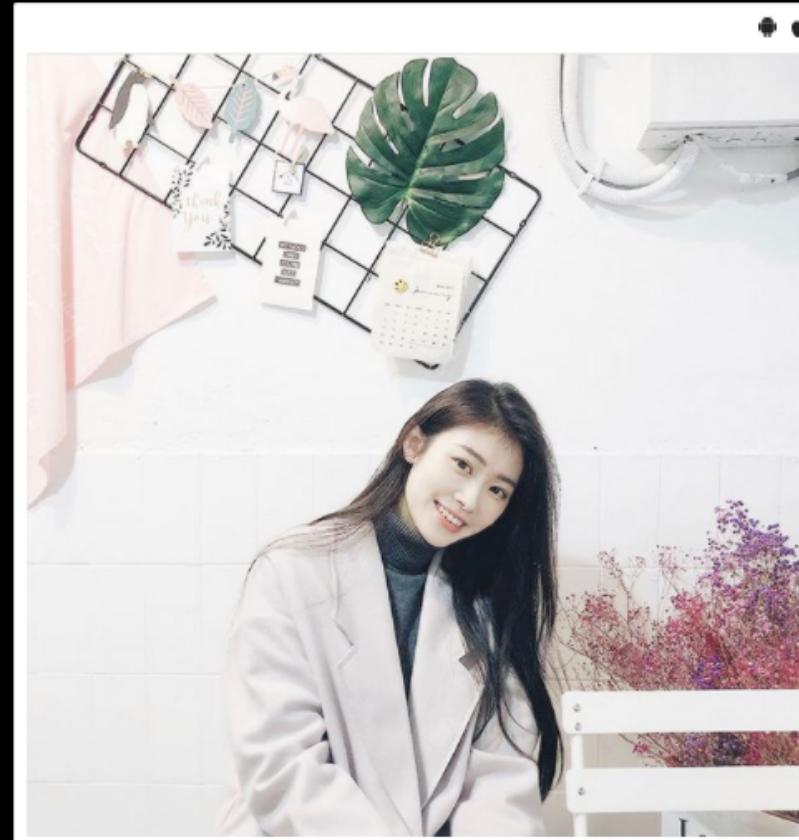
看板 Beauty

兩個同事笑起來都好美

同事1

<https://i.imgur.com/iMRYkF7.jpg>

URL



71,963 views

imgur

2. Push -1

- Python {studentID}.py push start_date end_date
- Input:
 - start_date
 - end_date

因為爬下來的資料是2018一整年，所以只要指定幾月幾日就好。
如果我們要找3月4號到10月20號，那

1. start_date就會是304
2. end_date就會是1020

這兩天的資料也必須包含在內

2. Push -2

**Python {studentID}.py push start_date
end_date**

- 程式內容：
 - 找出在 start_date(含)跟end_date (含)之間的：
 - 推文跟噓文的數量
 - 推最多文前10名的user id
 - 噓最多文前10名的user id
- Output：
 - 將結果輸出至push[start_date-end_date].txt
e.g., push[117-1230].txt
 - 檔案的格式請見下下頁

逐個character比較
length小的priority
比較高
數字的priority比字母高
大寫比小寫高
一樣就比下一個

順序範例:

11

123

13

aa

aaa

abc

b1

b12

ba

白石醫生 我喜歡妳!!!!

--
 ※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 36.237.57.209
 ※ 文章網址: https://www.ptt.cc/bbs/Beauty/M_1519658712.A.3B4.html
 嘘 img14999: <http://i.imgur.com/BwpzRez.jpg>

02/26 23:26

虚文

推文



※ 編輯: gj94ek (36.237.57.209), 02/26/2018 23:32:12

推 Boboinlyz: \白石\八緋山\八返島\

推 bearhaha: 有結衣我就推

推 Wi11iam0703: 是我最愛的日劇!!!!大推

推 Geoffrey314: 結衣

推 potato31627: 結衣結衣得第一~

推 kkkk666: 幹新木優子才第一啦

推 goodzoro: Gakki我的<3，話說一樓好氣喔，都不用桶一樓這種人嗎

→ yoyoruru: 我要新的實習醫生跟護士 XD

推 als8855: 推

User id

02/26 23:37

02/26 23:44

02/27 00:07

02/27 00:37

02/27 00:46

02/27 00:47

02/27 01:09

02/27 01:31

02/27 01:59

2. Push -3

Output Format:

推文數量

格式 : all like: 推文數量

噓文數量

格式 : all boo: 噓文數量

推文前10名(rank是1到10)

格式 : like #rank: userid 推文數

噓文前10名(rank是1到10)

格式 : boo #rank: userid 噓文數

```
all like: 201923
all boo: 32380
like #1: sulin209 2708
like #2: kai6366 1991
like #3: lovegogi 1123
like #4: htc10 890
like #5: Krishna 887
like #6: lck0 765
like #7: starryice 735
like #8: bbflisky 656
like #9: William0703 591
like #10: yoyonigo 506
boo #1: OCG5566 702
boo #2: htc10 639
boo #3: NeGe56 287
boo #4: BradleyBeal 223
boo #5: bh0925 216
boo #6: zxnstu3104 149
boo #7: yoyonigo 147
boo #8: yuigood 143
boo #9: jay3u7218 130
boo #10: dontkissme 126
```

3. Popular -1

Python {studentID}.py popular start_date end_date

- Input:
 - start_date
 - end_date
- 程式內容：
 - 找出在 start_date(含)跟end_date (含)之間的：
 - 輸出爆文數量
 - 輸出爆文內的所有圖片的URL，包括在推文的圖片
 - 圖片URL請抓文字的網址，並且要以jpg, jpeg, png, gif為結尾
(不限大小寫)
- Output：
 - 將結果輸出至popular [start_date-end_date].txt
 - e.g., popular[505-1101].txt
 - (格式請見下頁)

圖片非此結尾，可忽略。

3. Popular -2

- Output Format:
 - 第一行輸出 “number of popular articles: n”
 - n就是爆文數量
 - 接下來就一行一個URL，列出爆文中的圖片URL(含推文中的)
 - 圖片URL請抓文字的網址，並且要以jpg, jpeg, png, gif為結尾(不限大小寫)

```
number of popular articles: 2
https://i.imgur.com/zmA38zn.jpg
https://i.imgur.com/rvRNDHB.jpg
https://i.imgur.com/E8245cB.jpg
https://i.imgur.com/yzFekwn.jpg
https://i.imgur.com/mfjcIFm.jpg
```

作者 gj94ek (我帥的太沉重)

標題 [正妹] 空中急診英雄 Code Blue

時間 Mon Feb 26 23:24:55 2018

看板 Beauty

Code Blue電影版今年7月27號要在日本上映了

演員：新垣結衣 戸田惠梨香 比嘉愛未 新木優子 馬場富美加

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

圖片URL



56,873 views

imgur

白石醫生 我喜歡妳!!!!

圖片URL

※ 發信站: 批踢踢實業坊(ptt.cc), 來自: 36.237.57.209

※ 文章網址: <https://www.ptt.cc/bbs/Beauty/M.1510658712.A.3B4.html>

噓 mg14999: <http://i.imgur.com/BwpzRez.jpg>

02/26 23:26



31,309 views

imgur

※ 編輯: gj94ek (36.237.57.209), 02/26/2018 23:32:12

推 Boboinlyz: \白石\八紘山\八重島\

02/26 23:37

推 bearhaha: 有結衣我就推

02/26 23:44

推 William0703: 是我最愛的日劇！！！大推

02/27 00:07

推 Geoffrey314: 結衣

02/27 00:37

推 potato31627: 結衣結衣得第一~

02/27 00:46

推 kkkk666: 幹新木優子才第一啦

02/27 00:47

推 goodzoro: Gakki我的<3，話說一樓好氣喔，都不用桶一樓這種人嗎

02/27 01:09

→ yoyoruru: 我要新的實習醫生跟護士 XD

02/27 01:31

推 als8855: 推

02/27 01:59

4. Keyword -1

Python {studentID}.py keyword {keyword} start_date end_date

- Input:
 - {keyword} (欲尋找的關鍵字)
 - start_date
 - end_date
- 程式內容：
 - 找出在 start_date(含) 跟 end_date (含) 之間的且包含 {keyword} 的文章中所有圖片的 URL，包括在推文的圖片
 - 請注意哦，這邊只有要圖片 URL，文章內容不用
- Output：
 - 將結果輸出至 keyword({keyword})[start_date-end_date].txt
e.g., keyword(正妹)[505-1101].txt
 - (格式請見下頁)

4. Keyword -2

- Output Format:
 - 一行一個URL，列出圖片URL(含推文中的)
 - 圖片URL請抓文字的網址，並且要以jpg, jpeg, png, gif為結尾(不限大小寫)

```
https://i.imgur.com/zmA38zn.jpg
https://i.imgur.com/rvRNDHB.jpg
https://i.imgur.com/E8245cB.jpg
https://i.imgur.com/yzFekwn.jpg
https://i.imgur.com/mfjcIFm.jpg
```

4. Keyword -3 相關文章

從「作者」（含）開始到綠色的「發信站上一行的『--』(不含)」間，只要有出現keyword就算是包含keyword

- 所以推文不算
- 圖中綠色的字不算
- Beauty算（右上角）
- 「18:26:55 2018」算

批踢踢實業坊 > 看板 Beauty

作者 SHu410502152 (SuHuZen)
標題 一張驚為天人(?)
時間 Mon Mar 5 18:26:55 2018

<https://i.imgur.com/Uf06WdK.jpg>



59 views imgur

※ 發信站：批踢踢實業坊(ptt.cc),來自：118.163.131.88
※ 文章網址：<https://www.ptt.cc/bbs/Beauty/M.1520245617.A.545.html>
→ tpwin7：沒驚到
※ 編輯：SHu410502152 (118.163.131.88), 03/05/2018 18:28:18
→ yuigood：這批很純
噓 GGrundela：嗯 真是嚇死惹
→ yoyonigo：？
→ loneleonlmc：算漂亮但還不到驚為天人
→ hawaii987：沒很差啊 我可以<3

聯絡資訊 關於我們

看板 Beauty

03/05 18:27

03/05 18:28

03/05 18:28

03/05 18:35

03/05 18:36

評分方式與配分

- **評分方式**
 - 助教執行上傳的程式(**同一支程式！**)
 - 助教會測試多種input parameters
 - 助教會保證日期區間內一定會有足夠(≥ 10)的推噓文id
 - 輸出檔案請和{studentID}.py在同一個資料夾
- **配分 (total 103%)**
 - crawl: 25%
 - push: 24% (3個input cases，每個8%)
 - popular: 24% (3個input cases，每個8%)
 - Keyword: 30% (3個input cases，每個10%)
 - 沒輸出檔案就沒有分

Crawling時間限制

- 目前各個功能的時間限制如下

Crawl: 30分鐘

Push: 150分鐘

Popular: 30分鐘

Keyword: 150分鐘

也就是你的程式運行上列的功能時最久不能超過這個時間，如果超過的話會直接卡掉，該 testcase 直接算 0 分

Implementation Hints

- 第一個會跑的指令是crawl
- 其餘指令必須運用crawl的output來繼續動作 (別每次一直重抓)
 - 1. all_articles.txt
 - 2. all_popular.txt

(像是用all_articles.txt裡面的URL去抓推噓文)

執行環境

- OS: Ubuntu 18.04.3 LTS
- Python version: python3.6
- CPU: Intel(R) Core(TM) i7-8700K CPU @ 3.70GHz 6 Cores 12 Threads
 - Multiprocess\thread 請參考這個核心數
- RAM: 15G
 - 如果單次執行超過這個上限會直接被kill

繳交內容 - 1

- 請繳交一個**{studentID}.zip**壓縮檔到new e3，並且保證使用**右鍵->解壓縮至此**或是使用指令**unzip {studentID}.zip**解壓縮後有一個名稱為**{studentID}**的資料夾產生，**{studentID}**資料夾中至少包含以下兩個檔案(檔案名稱請完全相同)：
 - {studentID}.py
 - requirements.txt

繳交內容 - 2

- `{studentID}.py`
 - 就是python code
- `requirements.txt`
 - 在你的執行環境中使用以下指令產生
`pip3 freeze > requirements.txt`
 - 如果使用anaconda，在你的active environment 中使用以下指令產生
`conda list --export > requirements.txt`

繳交內容 - 3

- requirements.txt 範例

```
asn1crypto==0.24.0
attrs==17.4.0
Automat==0.6.0
cffi==1.11.5
constantly==15.1.0
cryptography==2.2.1
cssselect==1.0.3
hyperlink==18.0.0
idna==2.6
incremental==17.5.0
lxml==4.2.1
parsel==1.4.0
pyasn1==0.4.2
pyasn1-modules==0.2.1
pycparser==2.18
PyDispatcher==2.0.5
pyOpenSSL==17.5.0
```

Q and A

- 請問顯示本文已被刪除的文章需要抓入嗎？
- A: 沒有URL的就不用爬
- 如果文章內沒有出現※ 發信站？
- 那就是不符合格式，請把此篇文章忽略
也就是在做crawl的時候就用自己的方式把這篇忽略掉

- If you have any question about HW#1, please email to 吳易倫 or post on Facebook group.
 - w86763777@gmail.com