「使用」 cartoonmad(動漫狂) 爬蟲

- 專人維護板 (HSSLC)
  - https://github.com/HSSLC/cartoonmad-dlr
  - 下載快速、互動式介面
- 陽春版 (atrisk, 自製)
  - 下載慢、提示少
- 陽春版的優點是掌控度佳，未來要擴充會較易實現

航海王（海賊王）動漫狂首頁

WTGroup - Synology DiskSta | GitHub - HSSLC/cartoonmad

https://github.com/HSSLC/cartoonmad-dlr

HSSLC / cartoonmad-dlr

Code | Issues | Pull requests | Actions | Projects | Security | Insights

master | 1 branch | 0 tags | Go to file | Code

專人維護版的首頁

先下載 zip 檔，之後再解壓縮後即可跑

HSSLC 修正錯誤 701e783 on 24 Jun 7 commits

README.md	新增更多資訊	6 months ago
down.py	修正錯誤	6 months ago
main.py	新增資訊與改進語法	6 months ago

程式由兩個檔案組成

README.md

cartoonmad-dlr

自動多執行緒批量下載動漫狂的漫畫，可完全掛機運作

僅作為個人使用，勿作為商業用途，版權由包括但不限於原作者與受到原作者合

執行環境

此程式由Python製作，執行需要Python環境

至於如何安裝Python環境就自己去查囉，裝起來其實不難

Googleは君のよい友達

必要模組

- requests
- BeautifulSoup4

本程式需要用到的第三方模組

使用方法

1. 執行 `main.py`
2. 輸入動漫狂中作品的目錄頁面網址，例如 `https://www.cartoonmad.com/comic/xxxx.html`
3. 程式會提示編號與章節對應表，照表輸入要下載的章節，範例：

```
編號 對應名稱
0    第 001 話
1    第 002 話
2    第 003 話
```

C:\Windows\system32\cmd.exe - python main.py

```
D:\Slides_2020\wt_demo\2020_12\25\code>cd cartoonmad-dlr-master

D:\Slides_2020\wt_demo\2020_12\25\code\cartoonmad-dlr-master>dir
 磁碟區 D 中的磁碟沒有標籤。
 磁碟區序號:  6810-D4D0

 D:\Slides_2020\wt_demo\2020_12\25\code\cartoonmad-dlr-master 的目錄

2020/12/25  上午 08:48    <DIR>          .
2020/12/25  上午 08:48    <DIR>          ..
2020/12/25  上午 08:48             1,481 down.py
2020/12/25  上午 08:48             2,586 main.py
2020/12/25  上午 08:48             1,832 README.md
               3 個檔案           5,899 位元組
               2 個目錄  848,912,715,776 位元組可用

D:\Slides_2020\wt_demo\2020_12\25\code\cartoonmad-dlr-master>python main.py
僅作為個人使用，勿作為商業用途，版權由包括但不限於原作者與受到原作者合法授權者擁有
輸入作品目錄URL:https://www.cartoonmad.com/comic/1152.html
標題: 海賊王
編號 對應名稱
0    第 001 卷
1    第 002 卷
2    第 003 卷
```

1

2

```
D:\Slides_2020\wt_demo\2020_12\25\code\cartoonmad-dlr-master>dir
 磁碟區 D 中的磁碟沒有標籤。
 磁碟區序號:  6810-D4D0

 D:\Slides_2020\wt_demo\2020_12\25\code\cartoonmad-dlr-master 的目錄

2020/12/25  上午 08:49    <DIR>          .
2020/12/25  上午 08:49    <DIR>          ..
2020/12/25  上午 08:48             1,481 down.py
2020/12/25  上午 08:48             2,586 main.py
2020/12/25  上午 08:48             1,832 README.md
2020/12/25  上午 08:48    <DIR>          __pycache__
2020/12/25  上午 08:49    <DIR>          海賊王
               3 個檔案           5,899 位元組
               4 個目錄  848,908,349,440 位元組可用
```

跑完後的結果：多出一個海賊王子目錄

```
386  第 997 話
387  第 998 話
388  第 999 話
輸入上列編號選擇下載話數(ex:1-2 5-8 10 將會下載編號1, 2, 5, 6, 7, 8, 10的章節)
388
多執行緒正在暴走中
暴走結束
```

3

執行專人維護的程式

陽村版程式的執行過程

C:\Windows\system32\cmd.exe

```
D:\Slides_2020\wt_demo\2020_12\25\code>python cartonmad_demo_0.py
Got 1152/999/001.jpg
Got 1152/999/002.jpg
Got 1152/999/003.jpg
Got 1152/999/004.jpg
Got 1152/999/005.jpg
Got 1152/999/006.jpg
Got 1152/999/007.jpg
Got 1152/999/008.jpg
Got 1152/999/009.jpg
Got 1152/999/010.jpg
Got 1152/999/011.jpg
Got 1152/999/012.jpg
Got 1152/999/013.jpg
Got 1152/999/014.jpg
Got 1152/999/015.jpg

D:\Slides_2020\wt_demo\2020_12\25\code>_
```

cartonmad_demo_0.cfg - 記事本

檔案(F) 編輯(E) 格式(O) 檢視(V) 說明

```
# the 1st line is not used
1152
-1


# line #2: comic identify
# line #3: start index
# line #4: end index(not included)
# line #5: step
```

控制下載漫畫和範圍

第 2 行：漫畫代號

第 3 行：起始章節

第 4 行：結束章節

第 5 行：間隔章節

注意： cfg 檔案的 3,4,5 行參考 python 的 slice 語法

陽春版程式的原始碼

cartonmad_demo_0.py - D:\Slides_2020\wt_demo\2020_12\25\code\cartonmad_demo_0.py (3.8.5)

File Edit Format Run Options Window Help

```python
import requests, bs4
import os

with open('cartonmad_demo_0.cfg', mode='rt', encoding='ascii') as f:
    selections = f.readlines()

ALBUM = selections[1].strip() #ALBUM = '1152'
choose = []
for selection in selections[2:5]:
    try:
        int_value = int(selection.strip())
    except ValueError:
        int_value = None
    choose.append(int_value)

URL_HOME = 'https://www.cartoonmad.com/comic/{}.html'.format(ALBUM)
DOMAIN = 'https://www.cartoonmad.com'

link_res = requests.get(URL_HOME)
link_res.raise_for_status()
link_soup = bs4.BeautifulSoup(link_res.content, 'html.parser')
chapters = link_soup.select('fieldset#info a[target="_blank"]')

headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like
    'Referer':'https://www.cartoonmad.com/comic/' # hint by cartoonmad-dlr
}

for chapter in chapters[choose[0]:choose[1]:choose[2]]:

    CHAPTER = ''.join(ch for ch in chapter.text if ch.isdecimal())
    SUB_DIR = '{}/{}/'.format(ALBUM, CHAPTER)
    os.makedirs(SUB_DIR, exist_ok=True)
    # html of first image
    old_url = DOMAIN + chapter['href']
    chapter_res = requests.get(old_url, headers=headers) # add headers
    chapter_res.raise_for_status()

    while True:
        page_soup = bs4.BeautifulSoup(chapter_res.content, 'html.parser')
        pages = page_soup.select('a.pages')
        if not pages:
            break

        image = page_soup.select('img[oncontextmenu]')
        new_url = image[0]['src'] # raw url instead of old need-to-be-merged one

        image_res = requests.get(new_url, headers = headers)
        image_res.raise_for_status()
        image_res2 = requests.get(image_res.url)
        image_res2.raise_for_status()

        image_filename= SUB_DIR + os.path.basename(image_res.url)

        with open(image_filename, mode='wb') as image_fp_w:
            for chunk in image_res.iter_content(100000):
                image_fp_w.write(chunk)

        print('Got {}'.format(image_filename))
        # html of non-first image
        old_url = DOMAIN + '/comic/'+pages[-1]['href']
        chapter_res = requests.get(old_url, headers=headers) # add headers
        chapter_res.raise_for_status()
```

記得安裝 requests 和 beautifulsoup4 這兩個第三方模組才能跑程式