備份 pchome 相簿和其照片文字敘述

pchome_album_ReadFile.py - D:\Python_projects\pchome_album\pchome_album_ReadFile.py (3.8.5)

File Edit Format Run Options Window Help

```python
import requests, bs4
import os
import re
import html5lib

with open('pchome_album.txt', mode='rt', encoding='utf-8') as f:
    selections = f.readlines()
    user_url = selections[0].strip().upper()
# check if url of the 1st line is valid or not
if not user_url.startswith('HTTP'):
    print('請以http開始')
    quit
if not 'PHOTO.PCHOME.COM.TW' in user_url:
    print('可能不是pchome相簿')
    quit

URL_HOME = user_url + '*{}'

options = []
for selection in selections[1:5]:
    try:
        album_range = int(selection.strip())
    except ValueError:
        album_range = None
    options.append(album_range)

headers = {'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit\
/537.36(KHTML, like Gecko) Chrome/50.0.2661.102 Safari/537.36'}

DOMAIN ='http://photo.pchome.com.tw'

user_page = 1

# collect links even if spanning multitple pages
links = []
while True:
    link_res = requests.get(URL_HOME.format(user_page), headers=headers)
    link_res.raise_for_status()
    link_soup = bs4.BeautifulSoup(link_res.content, 'html5lib')
    links_tmp = link_soup.select('div#alb > a') (1)
    if not links_tmp:
        break
    links.extend(links_tmp)
    user_page = user_page + 1

pattern = re.compile("[【】？！一-龥，。：“”、‘’；\w\s.,_-]")

album_index = options[3]
for link in links[options[0]:options[1]:options[2]]:
    album_url = DOMAIN + link['href'][:-1] + '*{}' # discard extra '/'
    # first page
    link2_res = requests.get(album_url.format(1), headers=headers)
    link2_res.raise_for_status()
    link2_soup = bs4.BeautifulSoup(link2_res.content, 'html5lib')
    title = link2_soup.select_one('div.tit a').text.strip()

    valid_name = "".join(ch for ch in title if pattern.match(ch)) (2)
    title = valid_name
    title = title.strip()

    abstract = link2_soup.select_one('div.ab').text.strip() (3)

    pics = link2_soup.select('div#pic > a') # up to 25 pics

    # non-first pages if any (4)
    page = 2
    while True:

        link2b_res = requests.get(album_url.format(page), headers=headers)
        link2b_res.raise_for_status()
        link2b_soup = bs4.BeautifulSoup(link2b_res.content, 'html5lib')
        pics_next = link2b_soup.select('div#pic > a')
        if not pics_next:
            break
        pics.extend(pics_next)
        page += 1

    folder_album = '{:03}_'.format(album_index) + title #new 2 to 3
    os.makedirs(folder_album, exist_ok=True)

    album_index += 1

    filename_descriptor = folder_album + '/' + title + '.txt'
    with open(filename_descriptor, mode='wt', encoding='utf-8') as text_fp_w:
        # save album name & description in the specific file
        text_fp_w.write(title)
        delimit_title = '\n' + '。'*len(title) + '\n'
        text_fp_w.write(delimit_title)
        text_fp_w.write(abstract)
        text_fp_w.write('\n\n')

        pic_index = 1
        for pic in pics:
            link3_res = requests.get(DOMAIN + pic['href'], headers=headers)
            link3_res.raise_for_status()
            #link_soup = bs4.BeautifulSoup(link_res.content, 'html.parser')#bad
            link3_soup = bs4.BeautifulSoup(link3_res.content, 'html5lib')

            subtitle = link3_soup.find('b', {'id':'PTSection'}).text (5)

            valid_name2 = "".join(c for c in subtitle if pattern.match(c))
            subtitle = valid_name2

            content = link3_soup.find('div', {'id':'PDSection'}).text (6)
            pic_big = link3_soup.find('img', {'id':'PhotoAreaEns'}) (7)
            # save photo name & description in the specific file
            text_fp_w.write(subtitle)
            delimit_subtitle = '\n' + '—'*len(subtitle) + '\n'
            text_fp_w.write(delimit_subtitle)
            text_fp_w.write(content)
            text_fp_w.write('\n\n')

            image_res = requests.get(DOMAIN + pic_big['src'], headers=headers)
            #image_res.raise_for_status()
            # some jpg files seem to get lost
            if image_res.ok == False:
                print('failed:', image_res.url)
                continue
            image_filename = folder_album + '/{:03}_' + subtitle + '.jpg'
            with open(image_filename.format(pic_index), mode='wb') as img_fp_w:
                for chunk in image_res.iter_content(100000):
                    img_fp_w.write(chunk)
            pic_index += 1
    print('Got', folder_album)
```

先選稍後要用到的模組

粗略檢查網址是否合理，若不則離開程式

對使用者指定的選項做些處理

收集該網站的所有相簿的位址連結

一次只處理一本相簿

過濾相簿命名和敘述

收集該本相簿內的照片網址連結 (一頁一頁收集)

建立該本相簿的目錄

準備檔案並寫入該本相簿的大略文字敘述

根據前面收集的連結，開始請求下載照片了

搜尋單一照片的文字敘述並寫入檔案

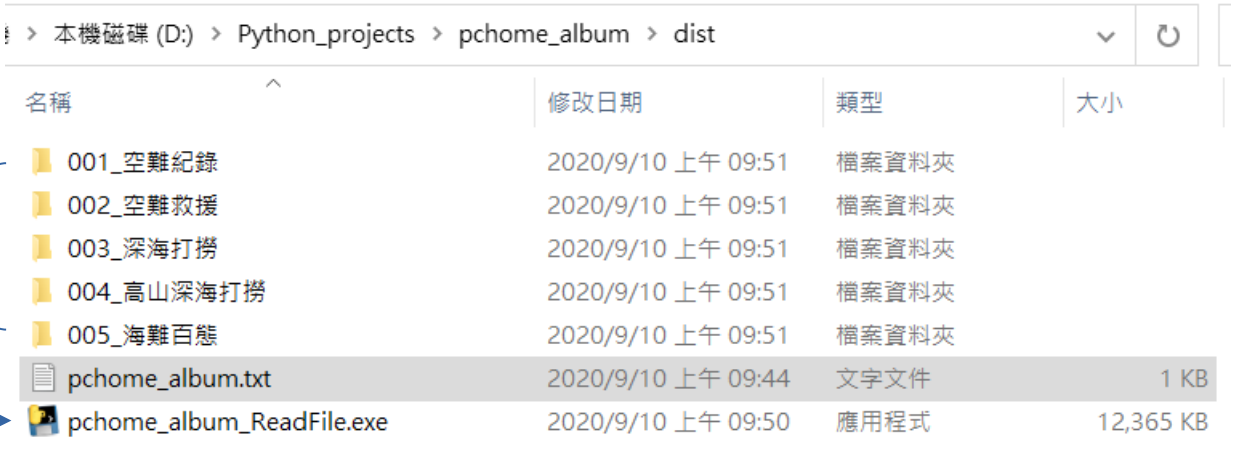
若成功抓到照片，則將其存成檔案。若失敗，則處理下一張

列個訊息說該本相簿處理結束

備份 pchome 相簿程式的執行結果

產生 5 本相簿

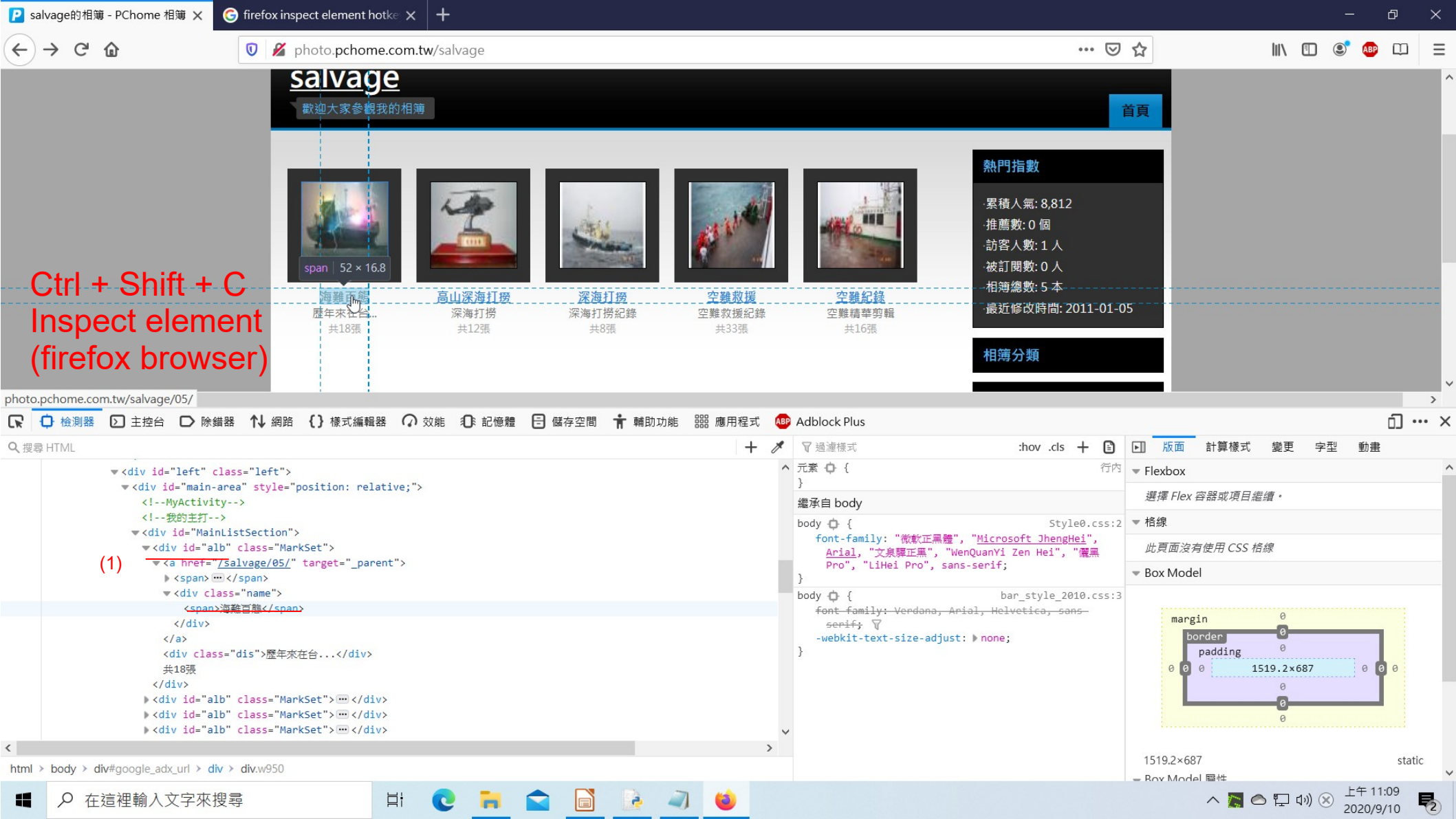
按這個檔案執行



原則上右面設定除了第一行網址改成你的外，其餘不須變動

魔鬼藏在細節裡

- 利用瀏覽器的開發者模切將網頁切換成人容易看的樹狀圖模式，並使用強大的 beautifulsoup4 模組將一團純文字的網頁轉換成有組織的物件
  - (1) links_tmp = link_soup.select('div#alb > a')
  - (2) title = link2_soup.select_one('div.tit a').text.strip()
  - (3) abstract = link2_soup.select_one('div.ab').text.strip()
  - (4) pics = link2_soup.select('div#pic > a')
  - (5) subtitle = link3_soup.find('b', {'id':'PTSection'}).text
  - (6) content = link3_soup.find('div', {'id':'PDSection'}).text
  - (7) pic_big = link3_soup.find('img', {'id':'PhotoAreaEns'})


salvage的相簿 - PChome 相簿
firefox inspect element hotke
photo.pchome.com.tw/salvage
salvage
歡迎大家參觀我的相簿
首頁
span | 52 × 16.8
海難百態
歷年來在台...
共18張
高山深海打撈
深海打撈
共12張
深海打撈
深海打撈紀錄
共8張
空難救援
空難救援紀錄
共33張
空難紀錄
空難精華剪輯
共16張
熱門指數
累積人氣: 8,812
推薦數: 0 個
訪客人數: 1 人
被訂閱數: 0 人
相簿總數: 5 本
最近修改時間: 2011-01-05
相簿分類
Ctrl + Shift + C
Inspect element
(firefox browser)
photo.pchome.com.tw/salvage/05/
檢測器
主控台
除錯器
網路
樣式編輯器
效能
記憶體
儲存空間
輔助功能
應用程式
Adblock Plus
搜尋 HTML

<!--MyActivity-->
<!--我的主打-->

(1)

…

海難百態

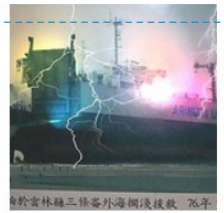
歷年來在台...
共18張

…
…
…
過濾樣式
:hov .cls
元素 {
}
行內
繼承自 body
body {
font-family: "微軟正黑體", "Microsoft JhengHei", Arial, "文泉驛正黑", "WenQuanYi Zen Hei", "儷黑 Pro", "LiHei Pro", sans-serif;
}
Style0.css:2
body {
font-family: Verdana, Arial, Helvetica, sans-serif;
-webkit-text-size-adjust: none;
}
bar_style_2010.css:3
版面
計算樣式
變更
字型
動畫
Flexbox
選擇 Flex 容器或項目繼續。
格線
此頁面沒有使用 CSS 格線
Box Model
margin
border
padding
1519.2×687
1519.2×687
static
html › body › div#google_adx_url › div › div.w950
在這裡輸入文字來搜尋
上午 11:09
2020/9/10

海難百態 - salvage的相簿 - PC

firefox inspect element hotke

photo.pchome.com.tw/salvage/05/



海難百態 共18張

a | 64 × 21.6

點閱數：981

歷年來在台灣各附近海域擱淺及沉沒之船舶海難照片

相簿標籤：

全站分類：拍賣市集| 個人分類：未分類

相片回應

訂閱的相簿 更多>>

我的收藏 更多>>

倒插沉沒

擱淺

擱淺

外傘頂洲擱淺

西子灣擱淺

photo.pchome.com.tw/salvage/05/

檢測器 主控台 除錯器 網路 樣式編輯器 效能 記憶體 儲存空間 輔助功能 應用程式 Adblock Plus

搜尋 HTML

```
<!--相簿資料 開始-->
<div id="ainfo">
  <div class="cover">…</div>
  <div class="dis">
    <div class="tit">
      <a href="/salvage/05/">海難百態</a>
      <span class="am">共18張</span>
    </div>
    <div class="t">…</div>
    <div class="ab" style="height:57px;overflow:hidden">…</div>
    <!--隱藏標籤-->
    <div id="hidden_tags" style="position:absolute; width:320px; z-index:999; background:#fef…px; line-heig
    0 50px; visibility: hidden;" onmouseout="HiddenTags()" onmouseover="ShowAllTags()">…</div> event
    <!--隱藏標籤-->
    <div class="cb" style="display:block;">…</div>
    <div class="cb" style="display:none;">…</div>
    <div class="cb">…</div>
  </div>
</div>
```

(2)

html › body

過濾樣式 :hov .cls

```
元素 {
}                                                    行內

html, body {                                         pstylist.css:1
  margin: ▸ 0;
  padding: ▸ 0;
}

body {                                               Style0.css:2
  font-family: "微軟正黑體", "Microsoft JhengHei",
  Arial, "文泉驛正黑", "WenQuanYi Zen Hei", "儷黑
  Pro", "LiHei Pro", sans-serif;
}

BODY {                                               Style0.css:1
  FONT-SIZE: ▸ 12px;
  COLOR: ▸ #000;
  FONT-FAMILY: ▸ arial;
  background: ▸ #a1a1a1;
  margin: ▸ 0px;
}

body {                                               topstyle.css:3
  margin: ▸ 0;
  FONT-FAMILY: ▸ Arial;
}
```

版面 計算樣式 變更 字型 動畫

Flexbox

選擇 Flex 容器或項目繼續。

格線

此頁面沒有使用 CSS 格線

Box Model

margin 0 border 0 padding 0 1519.2×1150

1519.2×1150 static

在這裡輸入文字來搜尋

上午 11:23 2020/9/10


海難百態 - salvage的相簿 - PC
photo.pchome.com.tw/salvage/05/
PChome 相簿
PChome | 說明 | (登入)
ktsalvage
歡迎大家參觀我的相簿
首頁
海難百態 共18張
相簿建立時間： 點閱數：
div.ab 400 × 57
歷年來在台灣各附近海域擱淺及沉沒之船舶海難照片
相簿分類
相片回應
訂閱的相簿
更多>>
檢測器
主控台
除錯器
網路
樣式編輯器
效能
記憶體
儲存空間
輔助功能
應用程式
Adblock Plus
搜尋 HTML
海難百態
共18張

…

(3)
歷年來在台灣各附近海域擱淺及沉沒之船舶海難照片

<!--隱藏標籤-->
<div id="hidden_tags" style="position:absolute; width:320px; z-index:999; background:#fef…px; line-heig
0 50px; visibility: hidden;" onmouseout="HiddenTags()" onmouseover="ShowAllTags()">… event
<!--隱藏標籤-->
…
…

全站分類：
拍賣市集
| 個人分類：
未分類
過濾樣式
:hov .cls
元素 {
height: 57px;
overflow: hidden;
}
行內
#ainfo .ab {
height: 90px;
font-size: 12px;
color: #777;
}
Style0.css:180
繼承自 div
#ainfo .dis {
font-size: 13px;
line-height: 20px;
color: #2b2b2b;
}
Style0.css:177
繼承自 div
.w950 {
text-align: left;
}
Style0.css:10
繼承自 body
版面
計算樣式
變更
字型
動畫
Flexbox
選擇 Flex 容器或項目繼續。
格線
此頁面沒有使用 CSS 格線
Box Model
margin
border
padding
400×57
400×57
static
html › body › div › div.w950 › table#cont › tbody › tr › td › div#cont › div.left › div#main-area › div#ainfo › div.dis › div.ab
在這裡輸入文字來搜尋
下午 12:43
2020/9/10


海難百態 - salvage的相簿 - PC
photo.pchome.com.tw/salvage/05/
海難百態 共18張
相簿建立時間： 點閱數：981
歷年來在台灣各附近海域擱淺及沉沒之船舶海難照片
相簿標籤：
全站分類：拍賣市集| 個人分類：未分類
相片回應
訂閱的相簿 更多>>
我的收藏 更多>>
img | 94.8 × 94.8
倒插沉沒
擱淺
擱淺
外傘頂洲擱淺
西子灣擱淺
photo.pchome.com.tw/salvage/106
檢測器 主控台 除錯器 網路 樣式編輯器 效能 記憶體 儲存空間 輔助功能 應用程式 Adblock Plus
(4)
<!--相簿資料 結束-->

<span id="Img[/imgdata
/czo2NDoiV0ZNVXp3QmppT0xqS3VMSmpBZjh2bTNxeTRiVldKRzFLWnhPVW9oQ1FOUnFzQzVJN1ZnaWFJemJCa01ZUXF4dCI7.jpg]"
class="PostponeImg Cover" style="width:90px; height:90px;">
<img src="/imgdata
/czo2NDoiV0ZNVXp3QmppT0xqS3VMSmpBZjh2bTNxeTRiVldKRzFLWnhPVW9oQ1FOUnFzQzVJN1ZnaWFJemJCa01ZUXF4dCI7.jpg"
onerror="this.src = '/img/spacei130.gif';" width="90" height="90" border="0">

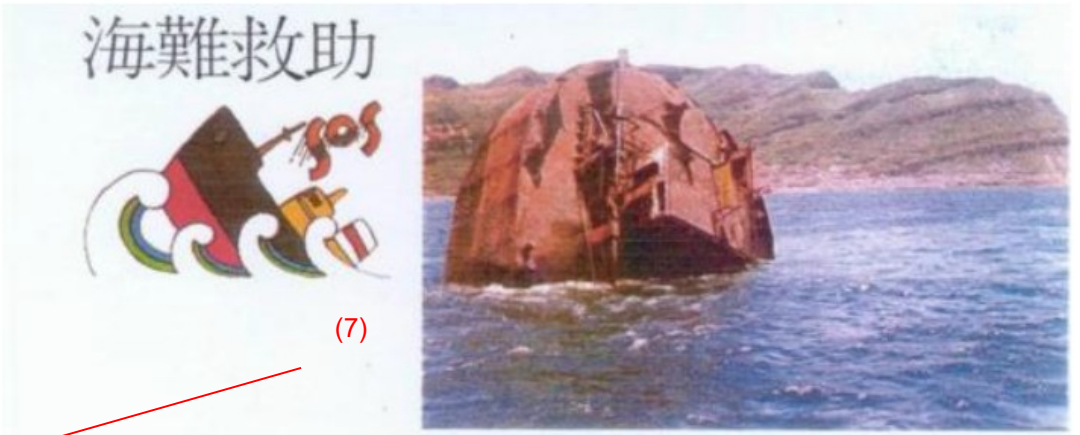
倒插沉沒

html body div div.w950 table#cont tbody tr td div#cont div.left div#main-area div#ainfo div.dis div.ab
#ainfo .ab { Style0.css:180
height: 90px;
font-size: 12px;
color: #777;
#ainfo .dis { Style0.css:177
font-size: 13px;
line-height: 20px;
color: #2b2b2b;
.w950 { Style0.css:10
text-align: left;
margin
border
padding
400×57
static
在這裡輸入文字來搜尋
下午 12:51
2020/9/10

海難百態 › 倒插沉沒 (1/18)　(5)

1 2 3 4 5 ›

```
<!--照片資料 start-->
<div class="picinfo">
  <div class="jumppic">…</div>
  <h1 style="margin-top:10px">
    <a href="/salvage/05/" title="海難百態">海難百態</a>
    <span class="ar">></span>
    <b id="PTSection" style="font-size:15px;" title="倒插沉沒">倒插沉沒</b>
    <span>(1/18)</span>
  </h1>
```



(7)

```
<div id="PhotoArea">
  <img id="PhotoAreaEns" src="/imgdata/czo0NDoiV0ZNVXp3QmppT0xqS3VMSmpBZjh2a1J6cmJGMERnOHNhM1Z4RUxVYnZvOD0iOw==.jpg"
  style="width:564px; height:239px; cursor:pointer;">
```

(6)

日本工作船倒插沉沒

```
<!--照片資料 start-->
<div class="picinfo">
  <div class="dis">
    <div id="PDSection">日本工作船倒插沉沒</div>
  </div>
</div>
<!--照片資料 end-->
```