## 財報檔案處理

Change Name - os.rename

/網路爬蟲/資料處理/





資料來源: https://mops.twse.com.tw/mops/web/t203sb02

下載

下載

下載

下載

Q1

下載

下載

下載

下載

Q2

下載

下載

下載

下載

Q3

下載

下載

下載

下載

Q4

105

104

103

102

年度\季別

IFRSs前

## 檔 名 規 則

Taiwan IFRS: tifrs Ex: tifrs-財報別-市場別-行業別-報告別-公司代號-年度季別.xml

財報/財測別:

□ 首次使用: fr0

□ 一般:fr1

□ 追朔重大 /追朔: fr2

□ 財務預測:ff

□ 財測-預結:fc

市場別:

□ 上市 / 上櫃: m1

□ 興櫃 / 興櫃(已申請上市) / 公開發行 / 未公開發行: m2

行業別代號:

□ 金融業: BASI (basi)

□ 證券期貨業: BD (bd)

□ 一般工商業: CI (ci)

□ 金控業: FH (fh)

□ 保險業: INS (ins)

□ 異業合併: MIM (mim)

報告別代號:

□ 合併財報:cr

□ 個體財報:er

□ 個別財報:ir

2018



.html

.xml

引用: http://kdh74616.blogspot.com/2017/05/project-taiwan-xbrl-filter-python-script.html

```
import Get_Data as gd
                                   在 List 中放入
所要判斷的季別
    Seasons = [ '2015Q2' ] 
    LeNum = len(Seasons)
    for Season in Seasons:
        Path = "C:\\Senior Project\\HTML\\" + Season + "\\"
10
        for Num in range(LeNum):
11
12
                                          判斷 List 中的季別 防止
           if Season == Seasons[Num]:
13
                                           For Loop 重複問題
14
               print(Season)
15
               print("----")
16
17
               gd.Rename_XML(Path)
               gd.Download_CR_XML(Path, Season)
               gd.Download_IR_XML(Path, Season)
20
```



```
xename_XML(Path):
Stock_List_Get_StockNumber()
                                            方便檔名做使用
StockLen = len(Stock_List)
for root, dirs, files in walk(Path):
   for file in files:
       FullPath = join(root, file)
       FileName = join(file)
       CR = '-cr-'
       IR = '-ir-'
        for i in range(StockLen):
           Number, Name, Group = Stock List[i]
           Number List = '-' + Number + '-'
           if Number_List in FileName:
                                            的判斷避免後續有
                                                 空資料問題
               if CR in FileName:
                   folder = Path + "CR\\"
                   if not os.path.isdir(folder):
                      os.makedirs(folder)
                   os.rename(FullPath, Path + "CR\\" + Number + ".xml")
                       n FileName:
                   folder = Path + "IR\\"
                   if not os.path.isdir(folder):
                      os.makedirs(folder)
                   os.rename(FullPath, Path + "IR\\" + Number + ".xml")
```

```
import os
   from os import walk
  from os.path import join
  def Get_StockNumber():
     Results = []
         with open('Stock Symbol.csv', encoding="utf-8-sig") as File:
            F_Read = File.readlines()
            for Stock_List in F_Read:
               S = Stock_List.replace("\u3000", ",")
               S = S.split(',')
               Results.append([S[0].strip(), str(S[1].strip()), str(S[2].strip())])
         print('讀不到')
      return Results
```

## 重點程式碼解說

## 1 def Download\_CR\_XML(Path, Season): Stock\_List = Get\_StockNumber() StockLen = len(Stock\_List) for i in range(StockLen): Number, Name, Group = Stock\_List[i] 判斷 該分類(CR) StockName = Number + ".xml" 是否有檔案 if StockName in os.listdir("C:\\Senior Project\\HTML\\" + Season + "\\CR\\"): URL = "https://mops.twse.com.tw/server-java/t164sb01?t203sb01Form=t203sb01Form8step=18CO ID=" + Number + "8SYEAR=" + Season[0:4] + "8SSEASON=" + Season[5:6] + "8REPORT ID=C" headers = {"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/106.0.0.0 Safari/537.36"} folder = "C:\\Senior Project\\HTML\\" + Season + "\\" if not os.path.isdir(folder): os.makedirs(folder) 利用 List 與 random 隨機 delay 時間 避免 delay\_choices = [3, 5, 10, 6, 7, 11] 判定為機器人 delay = random.choice(delay\_choices) time.sleep(delay) 抓取過快 導致IP被鎖 Web\_URL = requests.get(URL, headers=headers) 利用 wget 套件 將 Web\_Content = wget.download(Web\_URL.url, folder + str(Number) + '.html') 整個網頁抓取下來 print(Number + "Fail")

