[C#爬蟲_HtmlAgilityPack使用]_如何透過C#爬蟲批量將當前網頁圖片全下載下來

coolmandiary.blogspot.com (https://coolmandiary.blogspot.com/2020/11/chtmlagilitypackc.html) · by Samuel

(https://1.bp.blogspot.com/-LTPubESPJnA/X6D3_6EFsoI/AAAAAAAAPzk/nOayeN17SRE_lNJxiTFHppVTga-9mnTzwCLcBGAsYHQ/s620/spider.jpg)

一個網頁上若要去捕抓下載所有png , jpg , gif的圖檔

此時就要透過網頁爬蟲(web crawler)

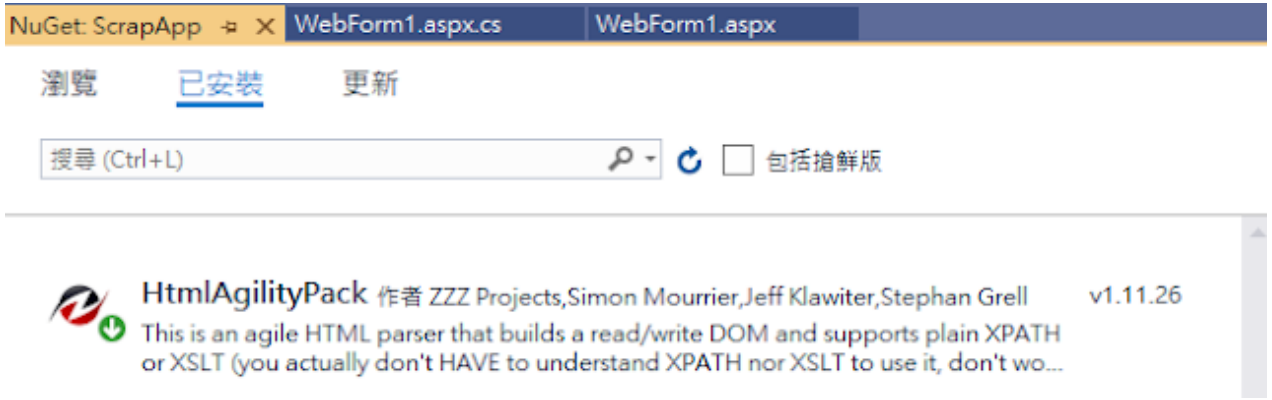
在此我們透過C#搭配vs2019 和HtmlAgilityPack這個套件進行開發

HtmlAgilityPack套件

https://html-agility-pack.net/

授權採用 MIT license

nuget上也可直接配置安裝

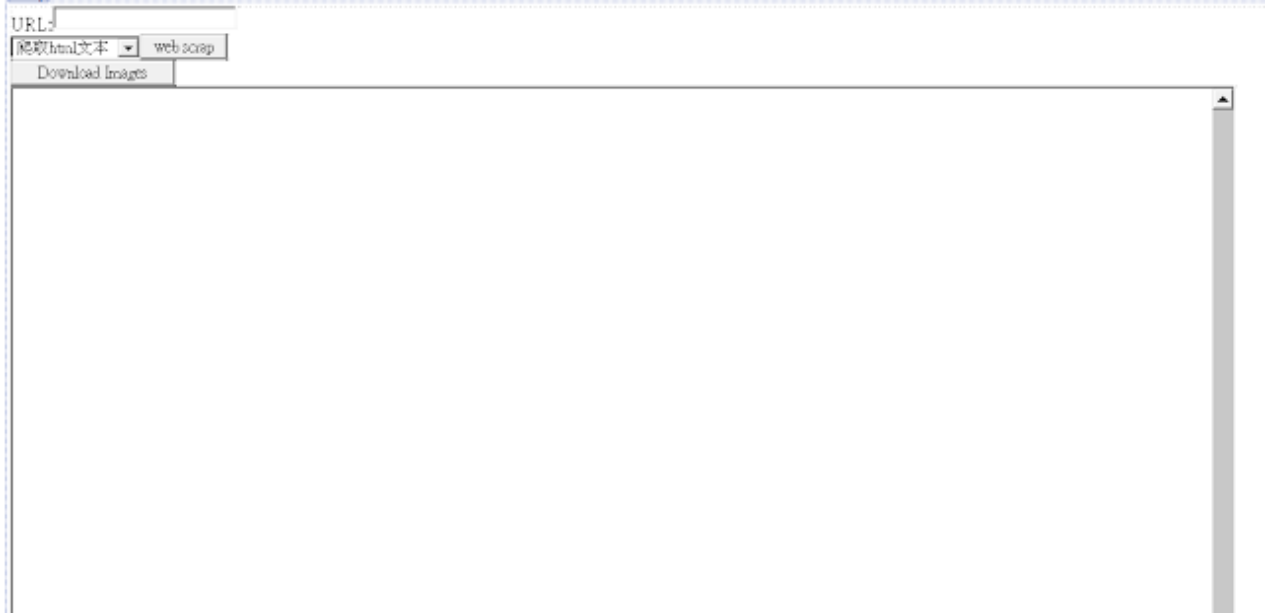


(https://1.bp.blogspot.com/--owe7I_Bx0E/X6D8pJl-joI/AAAAAAAAPz8/Nu6pdiz5vx4Xr7fjqdP0GnZ7HB7UGMJngCLcBGAsYHQ/s649/sx.png)

首先aspx網頁介面部分

WebForm1.aspx 程式碼

```
tWireup="true" CodeBehind="WebForm1.aspx.cs" Inherits="ScrapApp.WebForm1" %>



1999/xhtml">

 content="text/html; charset=utf-8"/>









ver">

l1" runat="server" Text="URL:"></asp:Label>
tURL" runat="server"></asp:TextBox>

D="ddl_action" runat="server">
alue="1">爬取html文本</asp:ListItem>
alue="2">爬取圖片連結</asp:ListItem>

Scrap" runat="server" Text="web scrap" OnClick="btnScrap_Click" />

BatchDownload" runat="server" Text="Download Images" OnClick="btnBatchDownload_Click"/>

rea1" runat="server"></textarea>
```



(https://1.bp.blogspot.com/-STCMzDgJ0YI/X6D-vY-7vAI/AAAAAAAAP0I/jncC7kXTGDc4_GUTGWz4rlHhQMcIhPsSQCLcBGAsYHQ/s1011/qq12.png)

WebForm1.aspx.cs 程式碼

```
42
43
```

```
using System;
using System.Collections.Generic;
using System.IO;
using System.Linq;
using System.Net;
using System.Text;
using System.Web;
using System.Web.UI;
using System.Web.UI.WebControls;
using HtmlAgilityPack;

namespace ScrapApp
{
    public partial class WebForm1 : System.Web.UI.Page
    {
        //https://stackoverflow.com/questions/307688/how-to-download-a-file-from-a-url-
        //https://stackoverflow.com/questions/2113924/how-can-i-use-html-agility-pack-t
        protected void Page_Load(object sender, EventArgs e)
        {

        }

        protected void btnScrap_Click(object sender, EventArgs e)
        {
            if (ddl_action.SelectedValue == "1")
            {
                string strHtmlDocText = WebCrawler.GetHtmlDocText(txtURL.Text);
                TextArea1.InnerText = strHtmlDocText;
            }
            else if (ddl_action.SelectedValue == "2")
            {
                string strLinksOfImage = WebCrawler.GetAllImageLinks(txtURL.Text);
                TextArea1.InnerText = strLinksOfImage;
            }
        }

        protected void btnBatchDownload_Click(object sender, EventArgs e)
        {
            WebCrawler.BatchDownloadImages(txtURL.Text, @"D:\ImgData");
        }
```

```
    }
}
```

兩個Class程式

WebUtility.cs

定義從URL獲取副檔名

```
 1
 2
 3
 4
 5
 6
 7
 8
 9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
```

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Web;

namespace ScrapApp
{
    public class WebUtility
    {




        /// <summary>
        /// 從URL中取得副檔名
        /// </summary>
        /// <param name="strURL"></param>
        /// <returns></returns>
        public static string GetFileExtensionFromUrl(string strURL)
        {
            strURL = strURL.Split('?')[0];
            strURL = strURL.Split('/').Last();
            return strURL.Contains('.') ? strURL.Substring(strURL.LastIndexOf('.')) : '
        }
    }
}
```

WebCrawler.cs

```
using HtmlAgilityPack;
using System;
using System.Collections.Generic;
using System.IO;
using System.Linq;
using System.Net;
using System.Text;
using System.Web;

namespace ScrapApp
{
    public class WebCrawler : WebUtility
    {

        //https://stackoverflow.com/questions/26189953/how-to-get-current-domain-name-i
        //https://docs.microsoft.com/zh-tw/dotnet/api/system.uripartial?view=netcore-3.
        //https://docs.microsoft.com/zh-tw/dotnet/api/system.uri.getleftpart?view=netcc

        public static string GetHtmlDocText(string strURL)
        {
            return GetHtmlDocObj(strURL).Text;
        }

        private static HtmlDocument GetHtmlDocObj(string strURL)
        {
            using (WebClient webClient = new WebClient())
            {
                using (MemoryStream memoryStream = new MemoryStream(webClient.DownloadD
                {
                    HtmlDocument doc = new HtmlDocument();
                    doc.Load(memoryStream, Encoding.UTF8);
                    return doc;
                }
            }
        }


        public static void BatchDownloadImages(string strURL, string saveDir, string fi
        {
            try
            {
                HtmlDocument doc = GetHtmlDocObj(strURL);
```

```
            Uri myUri = new Uri(strURL);
            string Uri = myUri.GetLeftPart(UriPartial.Authority);//獲取URI 的配置和
            var img_urls = doc.DocumentNode.Descendants("img")
                            .Select(ele => ele.GetAttributeValue("src", null))
                            .Where(s => !String.IsNullOrEmpty(s));
            List<string> lsImgUrl = img_urls.ToList();
            int idx = beginIdx;
            foreach (string item in lsImgUrl)
            {
                string imgURL = "";
                if (!item.StartsWith("http"))
                {
                    imgURL = item.Insert(0, myUri.GetLeftPart(UriPartial.Authority)
                }
                else
                {
                    imgURL = item;
                }
                string fileExt = GetFileExtensionFromUrl(imgURL);
                string SaveFilePath = Path.Combine(saveDir, fileName + String.Forma
                WebClient webClientImg = new WebClient();
                webClientImg.DownloadFile(imgURL, SaveFilePath);
                //webClientImg.DownloadFile(imgURL, String.Format(@"D:\ImgData\img_
                idx += interval;
            }
        }
        catch (Exception ex)
        {
            throw;
        }
    }


    public static string GetAllImageLinks(string strURL)
    {
        HtmlDocument doc = GetHtmlDocObj(strURL);
        Uri myUri = new Uri(strURL);
        string Uri = myUri.GetLeftPart(UriPartial.Authority);
        var img_urls = doc.DocumentNode.Descendants("img")
                        .Select(ele => ele.GetAttributeValue("src", null))
```

```
                        .Where(s => !String.IsNullOrEmpty(s));
            List<string> lsImgUrl = img_urls.ToList();
            StringBuilder sbResult = new StringBuilder();
            foreach (string item in lsImgUrl)
            {
                string imgURL = "";
                //https://www.taifex.com.tw
                if (!item.StartsWith("http"))
                {
                    imgURL = item.Insert(0, myUri.GetLeftPart(UriPartial.Authority));
                }
                else
                {
                    imgURL = item;
                }
                sbResult.AppendLine(imgURL);
            }
            return sbResult.ToString();
        }

    }
}
```

在此用這個網頁做測試

https://www.taifex.com.tw/cht/5/stockMargining

最終結果

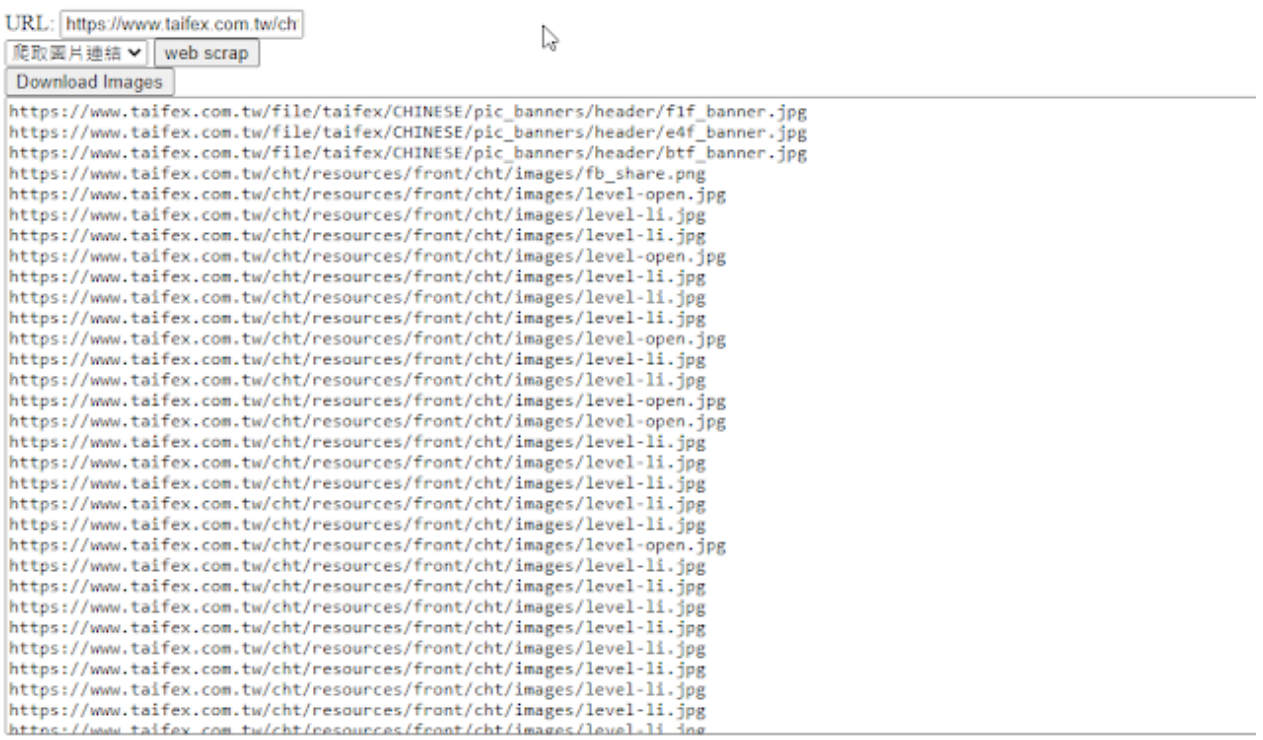
URL: https://www.taifex.com.tw/cht

爬取html文本 web scrap

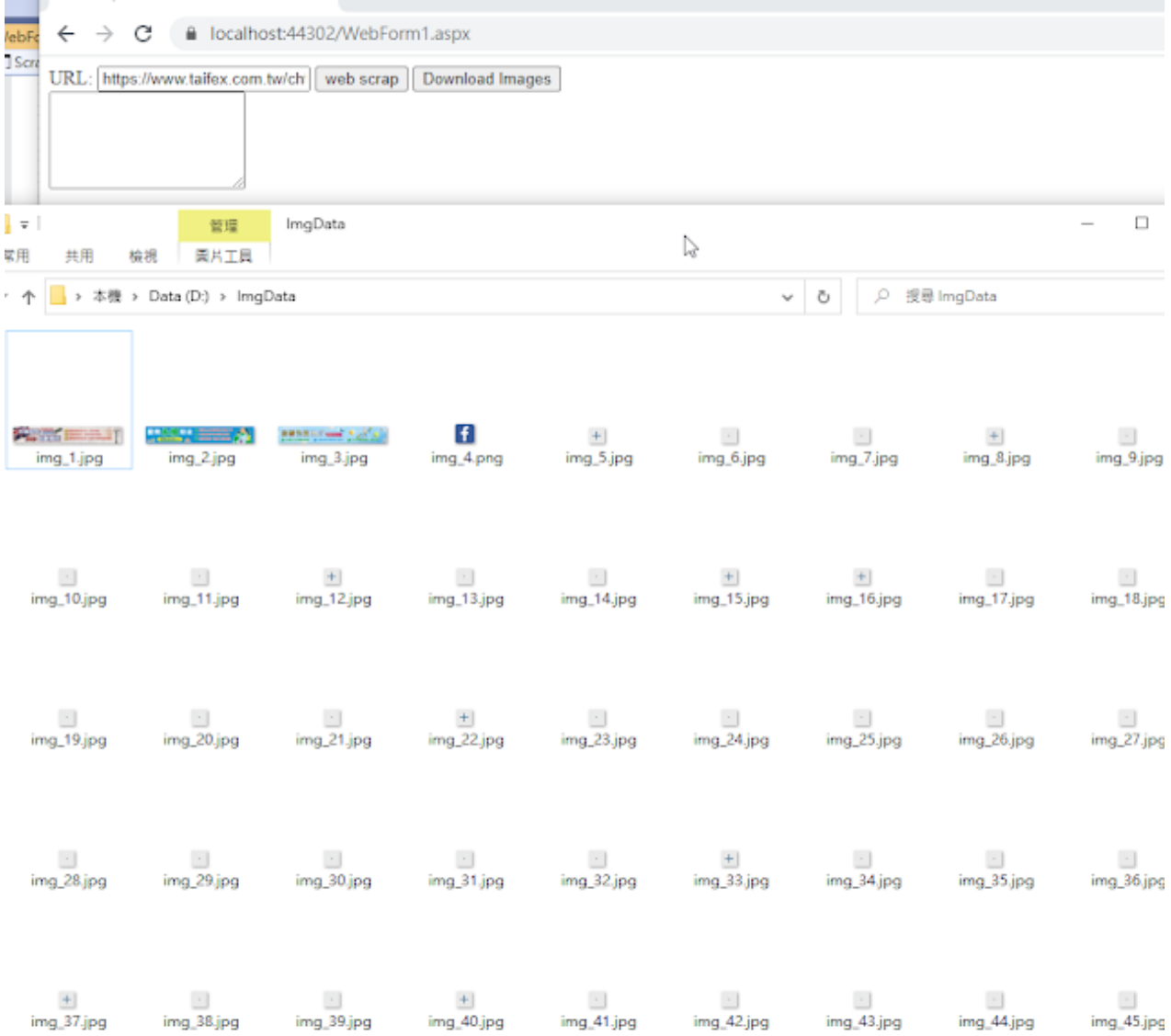
Download Images

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4/strict.dtd">
<html xmlns="http://www.w3.org/1999/xhtml" lang="zh-TW">

<head><meta http-equiv="X-UA-Compatible" content="IE=edge"></head><!doctype html>
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
  <link rel="shortcut icon" type="image/x-icon" href="/cht/resources/front/cht/images/favicon.ico" />
  <meta property="og:image" content="https://www.taifex.com.tw/chinese/images/fb_logo.jpg" />
  <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
  <meta http-equiv="X-UA-Compatible" content="IE=edge">
  <title>結算業務/保證金/保證金一覽表/股票類</title>


  <link rel="stylesheet" type="text/css" href="/cht/resources/front/cht/css/global.css" />
        <link rel="stylesheet" type="text/css" href="/cht/resources/front/cht/css/content.css" />
        <link rel="stylesheet" type="text/css" href="/cht/resources/front/cht/css/index.css" />
        <link rel="stylesheet" type="text/css" href="/cht/resources/front/cht/css/sitemap.css" />
        <link rel="stylesheet" type="text/css" href="/cht/resources/front/cht/css/jquery-ui.css" />
        <link rel="stylesheet" type="text/css" href="/cht/resources/front/js/slick/slick.css" />
        <link rel="stylesheet" type="text/css" href="/cht/resources/front/js/slick/slick-theme.css" />
  <script src="/cht/resources/front/js/jquery-3.1.1.min.js" type="text/javascript"></script>
  <script src="/cht/resources/front/js/jquery-ui-1.12.1.js" type="text/javascript" ></script>
  <script src="/cht/resources/front/js/jquery.ui.datepicker-zh-TW.js" type="text/javascript" ></script>
  <script src="/cht/resources/front/js/jquery.placeholder.js" type="text/javascript"></script>
  <script src="/cht/resources/front/js/jquery.ifixpng.js" type="text/javascript" ></script>
  <script src="/cht/resources/front/js/common.js" type="text/javascript"></script>
  <script src="/cht/resources/front/js/chart/Chart.min.js" type="text/javascript" ></script>
  <script src="/cht/resources/front/js/chart/chartjs-plugin-annotation.min.js" type="text/javascript" ></script>
```

(https://1.bp.blogspot.com/-aTAtw-h_0vU/X6EAa2BQImI/AAAAAAAAP0U/RoYSwnqo_7IvVKE9rCFAXobd9gix00NHgCLcBGAsYHQ/s1005/ww1.png)



(https://1.bp.blogspot.com/-Lzky8GE8Z5o/X6EAbTlfh1I/AAAAAAAAP0Y/UpZz734on-UWddnnv7hX4V2cIBZSPmwYgCLcBGAsYHQ/s916/ww2.png)



(https://1.bp.blogspot.com/-1OokeqrsYxY/X6D7LeaWvtI/AAAAAAAAPzw/NHFPtrrh400JpvK03xhRVq1G_g572RvZQCLcBGAsYHQ/s963/scx.png)

coolmandiary.blogspot.com (https://coolmandiary.blogspot.com/2020/11/chtmlagilitypackc.html) · by Samuel