

高級打字員的技術雲

還有技術債...

2017-10-27

[C#] 使用 CSS Selector方式來解析爬蟲Html網頁並修改文件 with AngleSharp

 3823  0  .Net Framework  2019-11-25

Parse Html jQuery like CSS/HTML selector in C#

前言

在2010年左右，.net技術如果想parse解析Html網頁，只有[Html Agility Pack](#)這個選擇

HtmlAgilityPack由於是以XML角度看待Html，抓取網頁標籤資料使用XPath+Linq的寫法

對於習慣寫前端jQuery的人來說相當不好上手

如今事隔多年，.net解析Html網頁的第三方套件百家爭鳴，在[nuget官網](#)上看得我眼花撩亂XD

陸續出現採用CSS Selector的寫法來解析網頁的套件也不少，終於可以在後端使用類似jQuery CSS Selector方式來抓取網頁標籤資料

今天要介紹的一款就是號稱解析效能很好和HtmlAgilityPack有得拼的[AngleSharp](#)

實作

從Nuget即可安裝，第一個就是

請用「空白」區分關鍵字



本頁段落

■ [](#)

線上工具

[Convert C# to VB.NET](#)

[Json Parser Online](#)

[json2csharp - generate c# classes from json string](#)

[繁簡轉換工具](#)



論壇討論區

[MSDN フォーラム \(日本\)](#)

[MSDN 論壇 \(台灣\)](#)

[MSDN 论坛 \(中国\)](#)

其他部落客

anglesharp   ☐ 包括搶鮮版



AngleSharp 依 AngleSharp, 1.2M 項下載

✓ v0.9.9



AngleSharp is the ultimate angle brackets parser library. It parses HTML5, CSS3, and XML to construct a DOM based on the official W3C specification.



AngleSharp.Scripting.JavaScript 依 AngleSharp, 38.9K 項下載

v0.5.1

Integrates a JavaScript engine (Jint) to AngleSharp.



TextDiscovery.AngleSharp 依 David West, 473 項下載

v1.0.3

TextDiscovery AngleSharp implementations of IDomInterpreter, IDomNodeFactory, and IHtmlConverter. Enables the following capabilities: mark search hits in the DOM, create HTML excerpts at a given word...



AngleSharp.io 依 AngleSharp, 3.01K 項下載

v0.3.2

Providers additional requesters and IO helpers for AngleSharp.

專案環境必須是 .net Framework 4 以上，3.5 以下的話，從 Nuget 會安裝失敗

以抓取奇摩電影海報圖為例

The Will Will Web (Will 保哥)

推薦部落格

博客頻道



雷神索爾3：諸神黃昏

上映日期：10月24日



愛的徒勞

上映日期：10月25日



翻滾吧！男人

上映日期：10月27日



叛逆的麥田捕手

服務中心 | 隱私權 | 建議 |

Elements Console Sources Network Performance Memory Application Security Audits

```

▼<div class="l_box_inner_content">
  ▼<div class="l_box_inner" data-module="0" style="display: block; z-index: 1; opacity: 1;">
    ▼<ul class="movie_ind_list slick-initialized slick-slider">
      ▼<div class="slick-list draggable">
        ▼<div class="slick-track" style="opacity: 1; width: 80000px; transform: translate3d(0px, 0px, 0px);">
          ::before
          ▼<li class="slick-slide slick-current slick-active" data-slick-index="0" aria-hidden="false" tabindex="0">
            ▼<div class="_slickcontent">
              ▼<div class="movie_foto"> == $0
                
                <div class="movie_ovrbox">...</div>
              </div>
            </div>
            <div class="movielist_info_inner">...</div>
          </div>
        </div>
      </ul>
    </div>
  </div>

```

```

using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
/*引用命名空間*/
using AngleSharp;
using AngleSharp.Dom;

```

```

namespace ConsoleApp2Test
{
    class Program

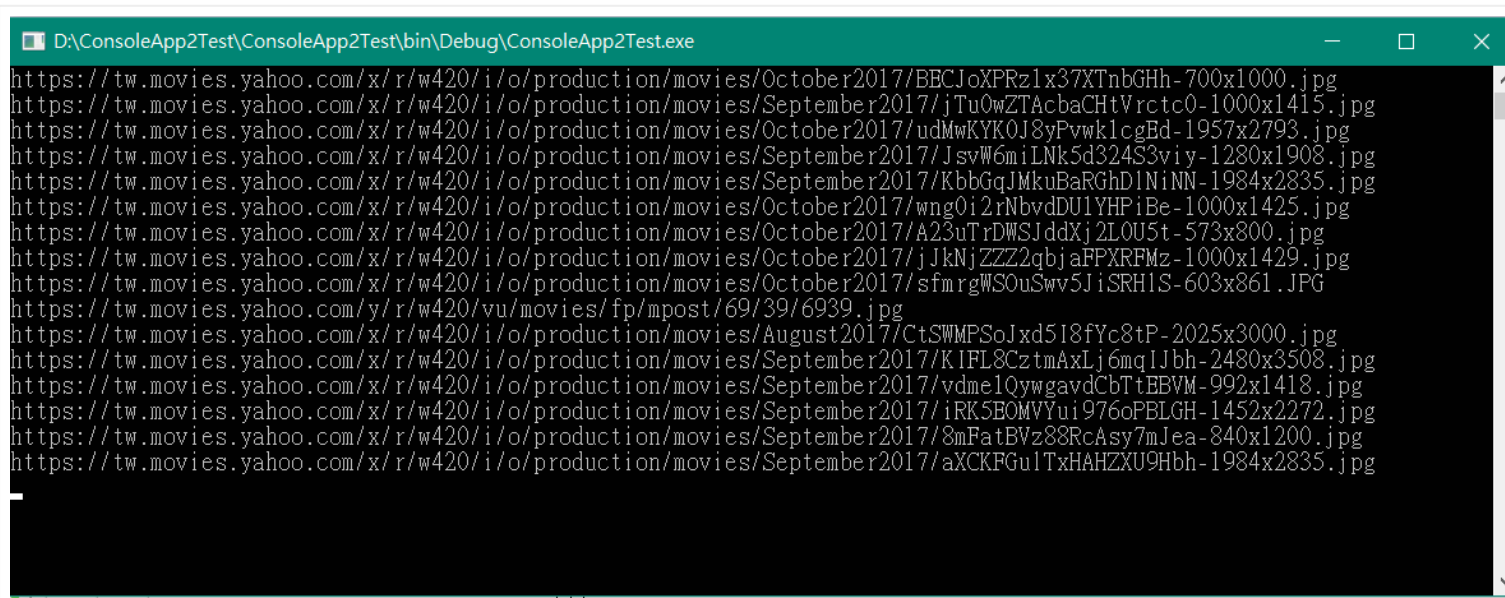
```

C#

```
static void Main(string[] args)
{
    IConfiguration config = Configuration.Default.WithDefaultLoader();
    string url = "https://tw.movies.yahoo.com";
    IDocument doc = BrowsingContext.New(config).OpenAsync(url).Result;

    /*CSS Selector寫法*/
    IHtmlCollection<IElement> imgs = doc.QuerySelectorAll("div.movie_foto img");
    foreach (IElement img in imgs)
    {
        Console.WriteLine(img.GetAttribute("src"));
    }
    Console.ReadKey();
}
}
```

執行結果：



```
D:\ConsoleApp2Test\ConsoleApp2Test\bin\Debug\ConsoleApp2Test.exe
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/October2017/BECJoXPRz1x37XTnbGHh-700x1000.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/September2017/jTu0wZTAcbACHtVrctc0-1000x1415.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/October2017/udMwKYK0J8yPwklcgBd-1957x2793.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/September2017/JsvW6miLNk5d324S3viy-1280x1908.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/September2017/KbbGqJMkuBaRGhDINiNN-1984x2835.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/October2017/wng0i2rNbvdDU1YHPiBe-1000x1425.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/October2017/A23uTrDWSJddXj2LOU5t-573x800.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/October2017/jJkNjZZZ2qbjafPXRFMz-1000x1429.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/October2017/sfmrgWSOuSwv5JiSRHIS-603x861.JPG
https://tw.movies.yahoo.com/y/r/w420/vu/movies/fp/mpost/69/39/6939.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/August2017/CtSWMPSoJxd5I8fYc8tP-2025x3000.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/September2017/KIFL8CztmAxLj6mqIJbh-2480x3508.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/September2017/vdme1QywgavdCbTtBBVM-992x1418.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/September2017/iRK5BOMVYui976oPBLGH-1452x2272.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/September2017/8mFatBVz88RcAsy7mJea-840x1200.jpg
https://tw.movies.yahoo.com/x/r/w420/i/o/production/movies/September2017/aXCKFGulTxHAHZXU9Hbh-1984x2835.jpg
```

再看看其他官方[AngleSharp examples](#)，也支援JavaScript engine

整體而言，真是不錯的套件，後續看好它的發展

2019-10-20補充

如果需要把讀取出來的HTML代碼做修改可以參考以下

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
/*引用命名空間*/
using AngleSharp;
using AngleSharp.Dom;
namespace ConsoleApp1_Selector
{
    class Program
    {
        static void Main(string[] args)
        {
            //html代碼
            var source = @"
                <!DOCTYPE html>
                <html>
                    <meta charset=utf-8>
                    <meta name=viewport content=""initial-scale=1, width=c
                    <title>Test Page</title>
                    <style>
                        *{margin:0;padding:0}html,code{font:15px/22px arial,
                </style>
```

```
<!-- 第一張圖alt沒給值也沒有等號 -->
<img src=""Content/1.jpg" />
<br/>
<!-- 第二張圖alt沒給值也沒有等號 -->
<img alt src=""Content/2.jpg" />
<br/>
<img alt="" src=""Content/3.jpg" />
<br/>
<img alt='' src='Content/4.jpg' />
<br/>
<img alt= src=Content/5.jpg />
<br/>
<img alt=test src=Content/6.jpg >
</div>";
```

```
IDocument document = BrowsingContext.New(Configuration.Default.WithDefaultOptions)
    .OpenAsync(req => req.Content(source)).Result;
```

```
IEnumerable<IElement> imgs= document.QuerySelectorAll("img");//取得所有img
```

```
int i = 1;
```

```
foreach (IElement img in imgs)
```

```
{
```

```
    img.SetAttribute("alt", "alt_" + i);//設定img的alt屬性
```

```
    i++;
```

```
}
```

```
//將修改後的html代碼輸出
```

```
Console.WriteLine(document.ToHtml());
```

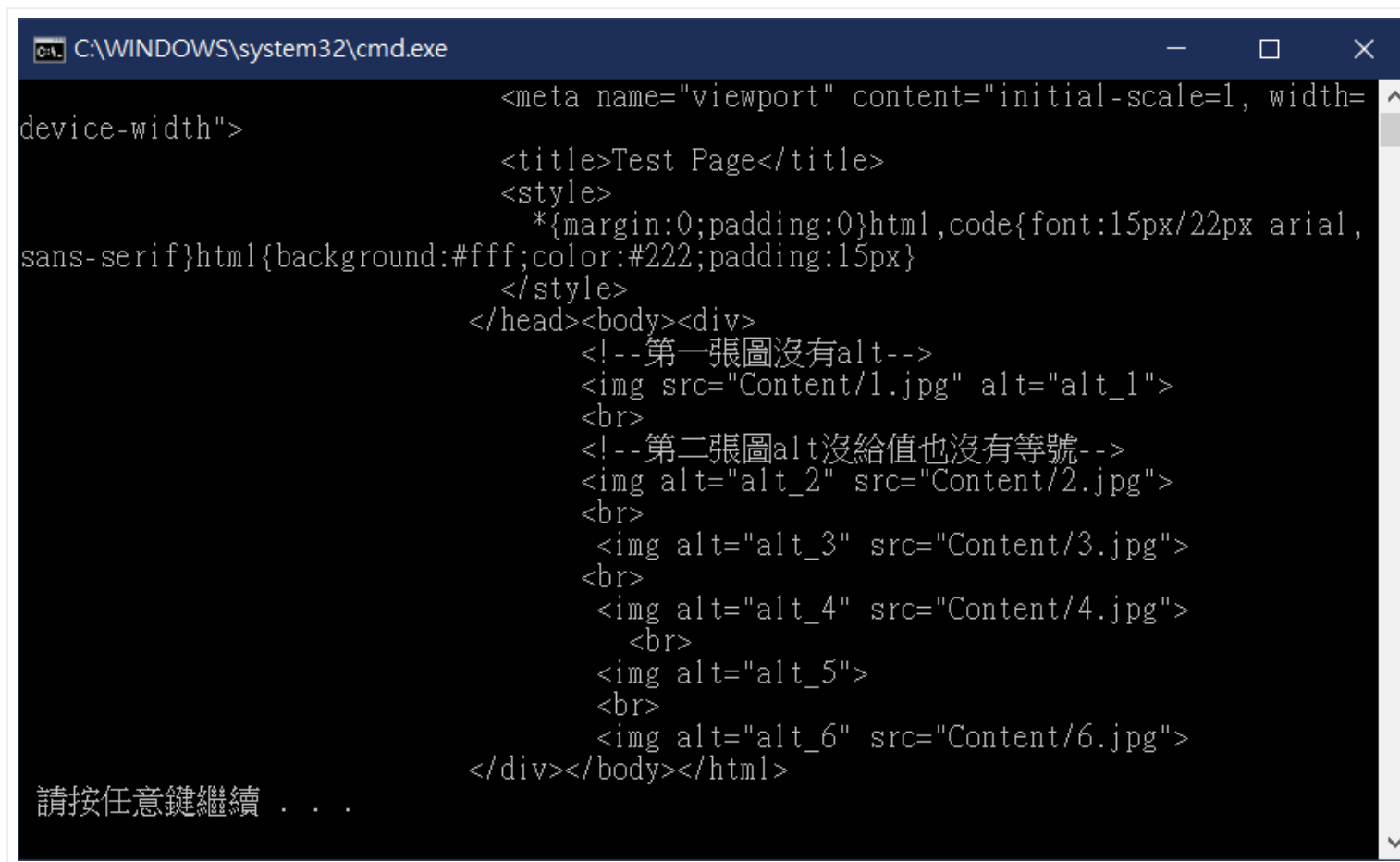
```
Console.ReadKey();
```

```
}
```

```
}
```

```
}
```

執行結果



```
C:\WINDOWS\system32\cmd.exe

<meta name="viewport" content="initial-scale=1, width=device-width">
<title>Test Page</title>
<style>
    *{margin:0;padding:0}html,code{font:15px/22px arial,sans-serif}html{background:#fff;color:#222;padding:15px}
</style>
</head><body><div>
    <!--第一張圖沒有alt-->
    
    <br>
    <!--第二張圖alt沒給值也沒有等號-->
    
    <br>
    
    <br>
    
    <br>
    <img alt="alt_5">
    <br>
    
</div></body></html>

請按任意鍵繼續 . . .
```

經過我的實測，以下HTML Element解析出來會失敗

1. ``：角括號後接續空白
2. ``：AngleSharp看不懂 `alt= src`

總之HTML結構確保愈完整愈好

官方[Example Code - AngleSharp](#)

參考文章

[How to parse HTML in .NET – Pavel Nasovich's Blog](#)

[一个犀利的 HTML 解析器 —— Less.Html](#)

[回首頁](#)

系列文章

[\[ASP.net\] 利用SQL Server的活動監視器\(圖形化介面\)瞭解Connection Pool運作](#)

[\[C#/VB.net\] 一次只執行一個WinForm視窗程式](#)

[\[.net C#\] 列舉型別的轉型](#)

[\[C#\] 網頁Html轉PDF檔\(一行程式碼解決\)](#)

[\[C#/VB.net\] 集合的亂數排序：Java有shuffle方法一行Code馬上實現 · C#/VB.net 就用.....](#)

[\[C#/ASP.net WebForm\] 把動態網頁轉成靜態網頁](#)

[\[C#\] 用反射\(映射\)移除if…else陳述式](#)

[\[C#\] 計算地圖上兩點座標距離的演算法](#)

[\[C#/ASP.net\] 經緯度轉行政區 \(利用Google反地理編碼\)](#)

[C#] 利用ASP.net WebService和Windows Service實作Android手機的訊息推播(舊版C2DM)

[C#] 利用ASP.net和Windows Service實作Android手機的訊息推播(2012/6月底GCM版)

[C#] 利用ASP.net和Console專案實作iOS的訊息推播

[C#] EPPlus 讀寫(read/write) Excel檔案 懶人包範例程式碼

[C#.net] 產生JSON字串的幾種方式整理

[C#] 利用Dictionary集合和委派完整移除if else的分支判斷

[C#/Facebook API] 利用Facebook API 發文，適合前後端平台(使用者無需輸入帳密方式)

[C#/Facebook API] 抓取粉絲專頁的塗鴉牆訊息

[C#/NHibernate] 10分鐘快速體驗NHibernate

[C#] 等比例縮圖的程式碼

[ASP.net MVC 4][小技巧] 如何得知用戶有沒有開啟該封Email

[C#/WinForm] 無名網站備份相簿(下載實體照片)的懶人程式分享

[ASP.net MVC] 將HTML轉成PDF檔案，使用iTextSharp套件的XMLWorkerHelper (附上解決顯示中文問題)

[ASP.net MVC] ASP.net MVC整合FormsAuthentication表單驗證登入 - 簡易範例程式碼

[C#] 判斷集合陣列裡的日期(或數字)是否連續n次

[C#] 抓取Java .JRE檔和.Net Console .EXE檔 主控台應用程式輸出的文字

[C#/XML/ASP.net MVC] 控制字元導致輸出XML失敗「', hexadecimal value 0x , is an invalid character」解法

分享自己存取資料庫使用的SqlHelper類別，ADO.net技術

[C#] 使用NetOffice、PptxTemplater第三方元件 讀寫PowerPoint

[\[C#/\].net\] 使用HttpWebRequest來Post資料](#)

[\[C#\] EntityFramework交易寫法 Sample Code](#)

[\[C#.Net\] 壓縮圖片，指定JPG的壓縮品質](#)

在 THE BLOG OF TYPEWRITER職人 上還有

[\[.Net Core\] 在.Net Core ...](#)

3年前 • 1 則留言

Read appsettings.json in
.Net Core Console
Application

[IT 技術人的部落格平台](#)

7年前 • 1 則留言

快來開設專屬部落格！分享
心得、散播你的熱情 :D

[\[安裝\] Visual Studio
2012的Windows ...](#)

7年前 • 1 則留言

[安裝] Visual Studio 2012的
Windows Service服務安裝方
式

[\[ASP.net MVC\] ckedito
網頁編輯器 ...](#)

7年前 • 1 則留言

[ASP.net MVC] ckeditor網頁
編輯器 字型、圖片上傳功能
的快速安裝筆記

[♥ Favorite](#)[🐦 Tweet](#)[f 分享](#)[按最新排序 ▾](#)

通過以下方式登錄

或注冊一個 **DISQUS** 帳號 [?](#)

來做第一個留言的人吧！

[訂閱](#)[將Disqus添加到您的網站](#)[添加 Disqus](#)[不要出售我的數據](#)