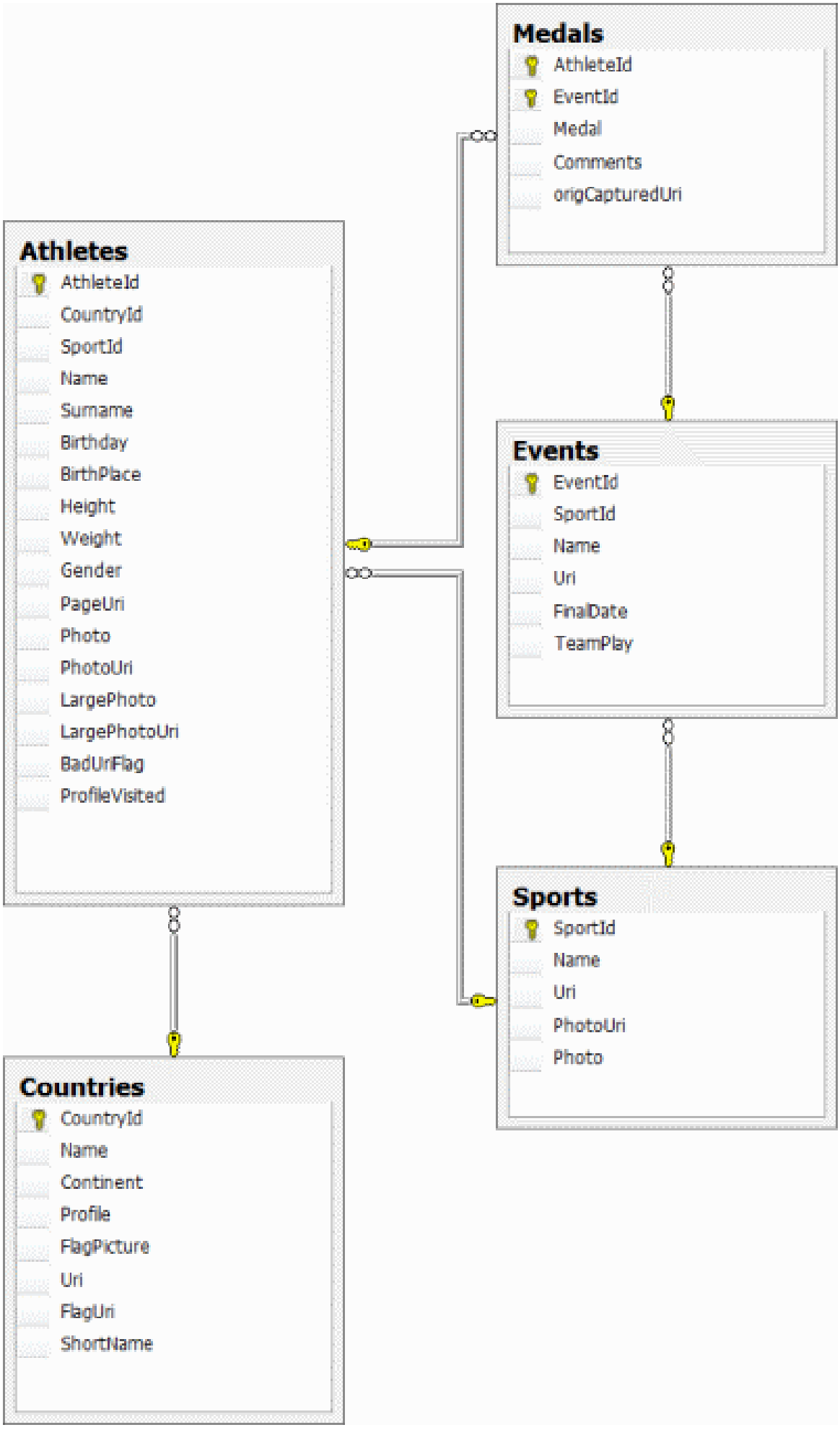
عنوان: **استخراج اطلاعات از صفحات وب با کمک HtmlAgilityPack**
نویسنده: ایمان عبیدی
تاریخ: ۱۳۹۱/۰۶/۰۳ ۲۳:۵۸
آدرس: www.dotnettips.info
گروه‌ها: XPATH, xml, Html Agility Pack, FireBug

چند مدت پیش موقعی که تب المپیک بود و جدول http://www.london2012.com/medals/medal-count/ رو زیاد نگاه می‌کردم به نظرم رسید که کاشکی به اطلاعاتی مثل اینکه چند نفر از مدال آورها خانم و یا آقا هستند و یا اینکه در روزهای مختلف تعداد مدال‌ها چطور توزیع می‌شند و بشه با یک jQuery UI Slider روزهای مختلف رو انتخاب کرد و جدول رو دید.
برای این کار اولین چیزی که لازم بود دریافت و ذخیره اطلاعات بود که من برای این کار از Entity framework 4.1 Database-first و کتابخانه HAP - htmlagilitypack استفاده کردم . طراحی دیتابیس نهایی به این صورت شد


Medals
AthleteId
EventId
Medal
Comments
origCapturedUri
Athletes
AthleteId
CountryId
SportId
Name
Surname
Birthday
BirthPlace
Height
Weight
Gender
PageUri
Photo
PhotoUri
LargePhoto
LargePhotoUri
BadUriFlag
ProfileVisited
Events
EventId
SportId
Name
Uri
FinalDate
TeamPlay
Sports
SportId
Name
Uri
PhotoUri
Photo
Countries
CountryId
Name
Continent
Profile
FlagPicture
Uri
FlagUri
ShortName

خوب در تلاش اول و مبتدیانه و بدون استفاده از این کتابخانه مفید چون اکثر صفحات وب XHTML نیستند و بالاخره چند تگ درست بسته نشده دارند و شما اگر بخواهید در آبجکت XmlDocument این htmlهای به ظاهر سالم رو لود کنید فورا با استثنای زیر مواجه می‌شوید

```
XmlException Was unhandeled
The 'img' start tag on line 1 position 1604 does not match the end tag of 'a'. Line 1, position 1766
```

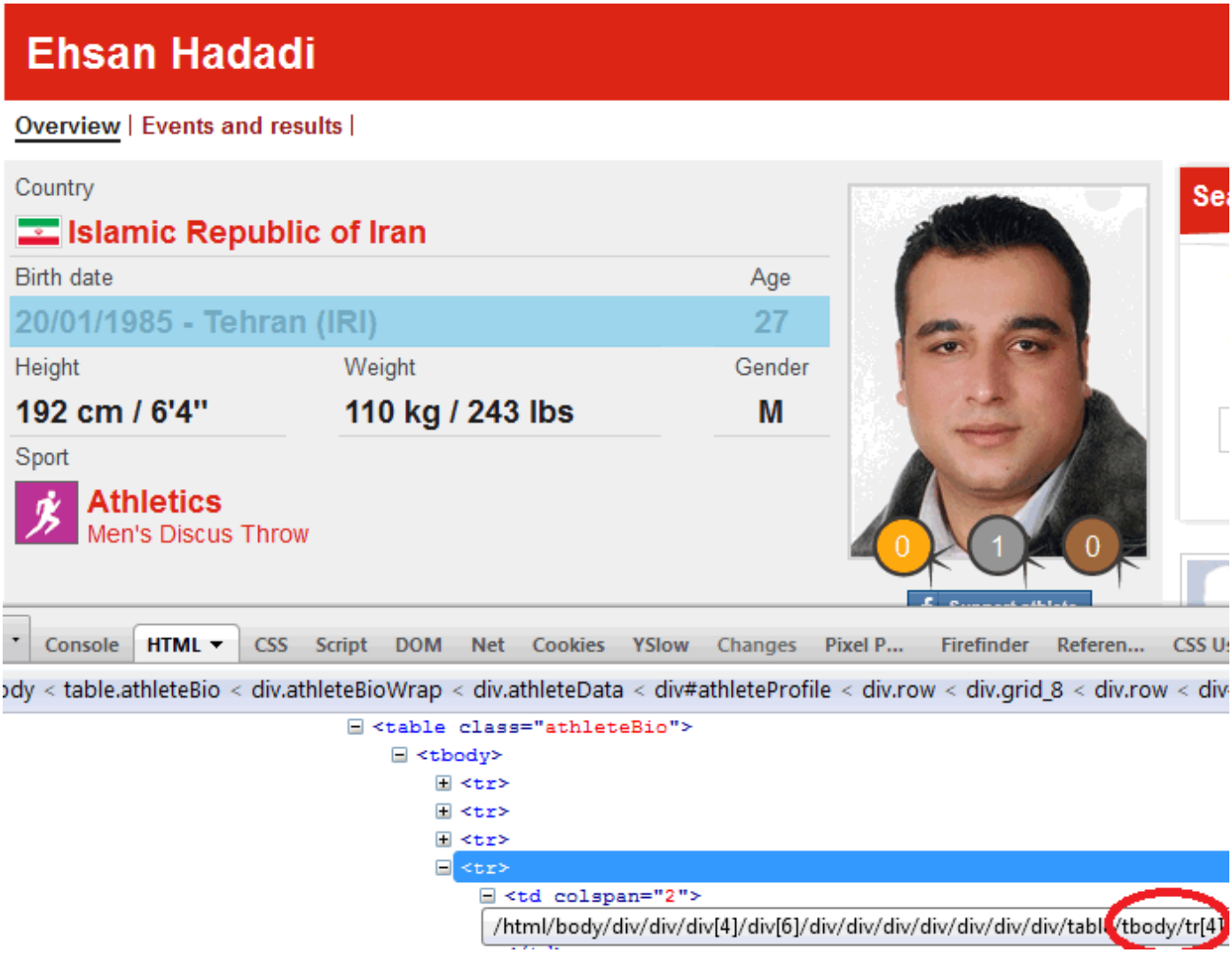
راه حل ساده اینه که این کتابخونه رو با کمک NuGet نصب کنید

```
PM> Install-Package HtmlAgilityPack
```

و از اینجا به بعد با کدی مثل این میتونید از کلاس HtmlDocument و مشابه XmlDocument ولی بدون ارور استفاده کنید.
مثلا با کد زیر میشه تاریخ تولد یک ورزشکار رو بدست آورد .توابع دیگه ای که خیلی جاها میتونه بدرد خورد GetAttributeValue و ChildNodes هست که یک نمونه نحوه استفادشو در ادامه میبینید

```
HtmlDocument xhtml = Crawler.GetXHtmlFromUri("http://www.london2012.com/athlete/hadadi-ehsan-
1077408/");
HtmlNode tempNode = xhtml.DocumentNode.SelectSingleNode("//table[@class='athleteBio']/tbody/tr[4]");



string temp = tempNode.FirstChild.FirstChild.InnerText.Replace(" ", "").Trim();
athlete.Birthday = DateTime.Parse(temp.Substring(0, 10), new CultureInfo("en-GB"));



tempNode = xhtml.DocumentNode.SelectSingleNode("//div[@class='athletePhotoMedals']/div/div/img");
athlete.LargePhotoUri = tempNode.GetAttributeValue("src", "");
```

البته تابع GetXHtmlFromUri رو جدا باید با کمک HttpWebRequest بنویسید و توی خوده HAP متاسفانه چنین تابعی توکار نشده نکته اصلی هم پیدا کردن محل دقیق اطلاعاته که با ابزاری مثل Firebug خیلی راحت‌تر میشه این کارو انجام داد. کافیه روی تاریخ تولد راست کلیک و inspect element by Firebug رو بزنید و حالا اگر تویه dom روی هر المنت html نگه دارید بهتون XPath کامل رو میده که میتونید تویه تابع DocumentNode.SelectSingleNode ازش استفاده کنید.




برای درک بهتر XPath هم این 2 تا صفحه xpath_syntax و xpath_examples خیلی میتونه کمکتون بکنه.

## نظرات خوانندگان

نویسنده: مهدی پایروند
تاریخ: ۱۳۹۱/۰۶/۰۴ ۱۲:۵۳

این روش برای استخراج مطالب همین سایت خیلی مفیده!

نویسنده: محسن کریمی
تاریخ: ۱۳۹۱/۰۶/۰۴ ۱۳:۵۷

با تشکر از مطلبتون
نکته: این Htmlagilitypack متاسفانه با صفحات فارسی و UTF8 مشکلات زیادی داره و واقعا نمیشه ازش استفاده کرد.

نویسنده: جمال
تاریخ: ۱۳۹۱/۰۶/۰۴ ۱۴:۱۹

Crawler.GetXHtmlFromUri
شی Crawler ازچه نوعیه؟موقع اجرا خطا میگیره؟

نویسنده: وحید نصیری
تاریخ: ۱۳۹۱/۰۶/۰۴ ۱۴:۲۸

شروع کار به این صورت هم می‌تواند باشد:

```
var doc = new HtmlDocument
                {
                    OptionCheckSyntax = true,
                    OptionFixNestedTags = true,
                    OptionAutoCloseOnEnd = true,
                    OptionDefaultStreamEncoding = Encoding.UTF8
                };
doc.LoadHtml(content);
```

OptionDefaultStreamEncoding رو به UTF8 تنظیم کنید.

نویسنده: محسن کریمی
تاریخ: ۱۳۹۱/۰۶/۰۴ ۱۵:۴۳

با تشکر

ولی طبق تجربه خود من کد بالا هم کمک نمی‌کنه و با تنظیم OptionDefaultStreamEncoding به UTF8 مشکل حل نمیشه ولی یه راه که قبلا من پیدا کرده بودم تو خوندن کد صفحات اینه که شما صفحات رو به صورت استریم دریافت کرده بعد توسط متد LoadHtml بخونید به این صورت مشکل برطرف میشه!(البته این تو سایت فارسی که من قصد خوندشو داشتم، بود با سایتهای دیگه تست نکردم)

```
var request = (HttpWebRequest)WebRequest.Create("آدرس سایت");
request.Method = "GET";
using (var response = (HttpWebResponse)request.GetResponse())
{
    using (var stream = response.GetResponseStream())
    {
        htmlDoc.Load(stream, Encoding.UTF8);
     }
}
```

نویسنده: وحید نصیری
تاریخ: ۱۳۹۱/۰۶/۰۴ ۱۶:۴۹

بستگی داره content نظر قبلی رو به چه فرمتی (چه Encoding ایی ) از وب دریافت کردید. مابقی آن توسط این کتابخانه بدون مشکل پردازش می‌شود.

```
using System.Net;
//...
var content = new WebClient { Encoding = Encoding.UTF8 }.DownloadString(url);
```

نویسنده: ایمان عبیدی
تاریخ: ۱۳۹۱/۰۶/۰۴ ۲۰:۴۶

Crawler همونطور که در متن هم نوشته شده دست سازه و مهم نیست و تابع GetXHtmlFromUri میتونه مثل نمونه زیر باشه و دقت کنید خالی نبودن UseAgent خیله مهمه وگرنه ارور The remote server returned an error: (409) Conflict رو میده.
من با همین تابع یک سایت فارسی رو چک کردم و اروری نداد و متن فارسی قابل کوئری گرفتن بود.
کامل‌تر و با ارور هندلینگ بهترش رو میتونید در برنامه مفید plrip آقای وحید نصیری ببینید

```
private static HtmlDocument GetXHtmlFromUri(string uri)   {
            HtmlDocument htmlDoc = new HtmlDocument()
            {
                OptionCheckSyntax = true,
                OptionFixNestedTags = true,
                OptionAutoCloseOnEnd = true,
                OptionDefaultStreamEncoding = Encoding.UTF8
            };

            var request = (HttpWebRequest)WebRequest.Create(uri);
            request.Method = "GET";



            //important
             request.UserAgent = "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; WOW64;
Trident/5.0)";
             request.Accept = "text/html";
            request.AutomaticDecompression = DecompressionMethods.GZip | DecompressionMethods.Deflate;

            using (var response = request.GetResponse() as HttpWebResponse)
            {
                using (var stream = response.GetResponseStream())
                {
                    htmlDoc.Load(stream, Encoding.UTF8);
                }
            }
            return htmlDoc;
        }
```

اینم روش دوم که بازم UserAgent باید اضافه بشه

```
private static HtmlDocument GetXHtmlFromUri2(string uri)        {
            WebClient client = new WebClient() { Encoding = Encoding.UTF8 };
            client.Headers.Add("user-agent", "Mozilla/5.0 (compatible; MSIE 9.0; Windows NT 6.1; WOW64;
Trident/5.0)");

            HtmlDocument htmlDoc = new HtmlDocument()
            {
                OptionCheckSyntax = true,
                OptionFixNestedTags = true,
                OptionAutoCloseOnEnd = true,
                OptionDefaultStreamEncoding = Encoding.UTF8
            };

            htmlDoc.LoadHtml(client.DownloadString(uri));

            return htmlDoc;
        }
```

نویسنده: افشار محبی
تاریخ: ۱۳۹۱/۰۶/۰۵ ۹:۳۷

من هم از این ابزار در کارهایم استفاده کرده‌ام. خیلی خوب جواب می‌دهد.

نویسنده: ایمان عبیدی
تاریخ: ۱۳۹۱/۰۶/۰۶ ۰:۴

برای مشاهده نتایج بدست آمده رده بندی المپیک 2012 لندن به همراه اطلاعات جنسیت مدال گیرها و همچنین وضعیت جدول در روز هایه مختلف میتونید به لینک هایه زیر مراجعه کنید.
توبه این صفحات از پلاگین tableSorter و یکم جاوا اسکریپت هم در لینک اول برای کش کردن اطلاعات json استفاده کردم .
http://olympics2012.iabidi.ir/Home/DailyMedals
http://olympics2012.iabidi.ir/Home/RankingsFull

نویسنده: پریسا
تاریخ: ۱۳۹۲/۰۲/۰۲ ۲۰:۵۸

چرا وقتی XPath از Firebug
استفاده می‌کنم با استفاده از SelectSingleNode
جواب دقیقی نمیده یا هیچی نمیاره

نویسنده: وحید نصیری
تاریخ: ۱۳۹۲/۰۲/۰۳ ۱۴:۱۴

علت این است که html ایی که در فایرباگ بررسی میشه عموما به دلیل یک سری از نرمال سازی‌ها توسط موتور فایرفاکس و همچنین خودش، با html اصلی یک سایت متفاوت است. به همین جهت XPath استخراجی از آن روی سایت اصلی کار نخواهد کرد. یک برنامه کمکی برای یافتن XPathها به همان نحوی که هستند.

نویسنده: mahsan
تاریخ: ۱۳۹۲/۰۳/۲۳ ۱۰:۳۸

این متد رو در چه صفحه ای باید بنویسم؟ آیا باید در همون صفحه ای که میخوام اطلاعات لود بشه بنویسم؟
uri مقدارش را باید داخل تابع بگیره؟ در page load فرمم چی بنویسم؟

نویسنده: وحید نصیری
تاریخ: ۱۳۹۲/۰۳/۲۳ ۱۱:۲۹

- تفاوتی نمی‌کنه کجا فراخوانی بشه ؛ در page load یا در یک روال رخداد گردان کلیک و یا در یک سرویس مستقل.
- بهتره نتیجه رو بعد از فراخوانی برای مدتی کش کنی، تا هربار اطلاعات از وب درخواست نشود.

نویسنده: mahsan
تاریخ: ۱۳۹۲/۰۳/۲۳ ۱۲:۱۰

ببخشین که دوباره مزاحمتون شدم: وقتی من کد بالا رو در page load کپی میکنم زیر بعضی کلمات مانند , Crawler
xhtml,HtmlNode,Parse, Birthday, LargePhotoUri, GetAttributeValue خط قرمز میکشه ! چرا؟ آیا HtmlDocument یک سند html که من باید ایجادش میکردم؟ فقط همین کدها رو بنویسم؟ جواب میگیرم با باید چیز دیگه ای هم اضافه بشه؟

نویسنده: وحید نصیری
تاریخ: ۱۳۹۲/۰۳/۲۳ ۱۲:۴۱

- شما باید از طریق نیوگت با دستور Install-Package HtmlAgilityPack این بسته رو نصب کنید. یا اینکه فایل DLL اون رو از سایتش دریافت و به ارجاعات پروژه اضافه کنید.
- کدهای کلاس Crawler چند کامنت بالاتر ارسال شدن. تابع GetXHtmlFromUri که ملاحظه می‌کنید .
- مواردی مانند Birthday, LargePhotoUri یک سری متغیر هستند که از طرف نویسنده مقاله تعریف شدن. مهم نیستند. حذفشون کنید.
- یک مثال دیگر در مورد استفاده از کتابخانه HtmlAgilityPack با کد قابل دریافت .

نویسنده: mahsan
تاریخ: ۱۳۹۲/۰۳/۲۵ ۹:۶

سلام ممنون که جوابم رو دادین
من کلاسی بنام Crawler ایجاد کردم ولی حالا زیر

```
OptionDefaultStreamEncoding
```

OptionAutoCloseOnEnd OptionFixNestedTags OptionCheckSyntax
و
LoadHtml

خط قرمز میکشه :( لطف میکنید بگید دلیلش چیه و باید چیکار کنم؟
و در فرمم هم زیر این کلمه GetXHtmlFromUri
مینویسه
generate metod stub for 'GetXHtmlFromUri in "crawel

وقتی من متد GetXHtmlFromUri را در کلاس crawel تعریف کردم جرا باید این پیغام رو بده ؟ آیا باید این گزینه رو انتخاب کنم؟

نویسنده: وحید نصیری
تاریخ: ۱۳۹۲/۰۳/۲۵ ۹:۴۳

پیشنهاد من این است که یک دوره سی شارپ مقدماتی رو بگذرونید. با مباحثی مانند نحوه تعریف فضای نام و روش فراخوانی یک متد استاتیک از کلاس متناظر با آن آشنا شوید.
یک دوره خوب مقدماتی سی‌شارپ

نویسنده: ایمان توکلی
تاریخ: ۱۳۹۲/۱۱/۲۶ ۱۷:۱۰

با سپاس
در صورتی که مقداری در querystring مربوط به صفحه اضافه شود و در هر درخواست این مقدار تغییر کند چطور می‌توان صفحه را خواند مثال: **http://website.com?log=1731004**
سایت برای هر بازدید یک لاگ جدید می‌نویسد یعنی هر درخواست جدید در صورتی که لاگ معتبر نباشد به صفحه دیگر ارسال می‌کند.حال اطلاعات این صفحه را چطور می‌توان خواند ؟

نویسنده: محسن خان
تاریخ: ۱۳۹۲/۱۱/۲۶ ۱۷:۲۱

این نوع مسایل رو نمی‌تونید با HtmlAgilityPack حل کنید. HtmlAgilityPack فقط کارش آنالیز اطلاعات ثابتی هست که بهش می‌دید و نهایتا خواندن نودهای HTML نهایی. موردی که عنوان کردید نیاز به آنالیز همزمان هدرهای دریافتی به همراه جاوا اسکریپت سایت مورد نظر داره.