عنوان: **استخراج تمام XPathهای یک محتوای HTMLایی به کمک کتابخانه HtmlAgilityPack**
نویسنده: وحید نصیری
تاریخ: ۱۳۹۲/۰۳/۲۷ ۸:۰
آدرس: www.dotnettips.info
برچسب‌ها: Html Agility Pack

اولین قدم کار کردن با کتابخانه قدرتمند HtmlAgilityPack ، داشتن XPath معتبر و متناظر با یک گره خاص می‌باشد. هرچند به ظاهر تعدادی از مرورگرها با کمک افزونه‌های خود امکان استخراج این XPathها را فراهم کرده‌اند اما ... عموما این مقادیر ارائه شده، نادرست هستند و بر روی محتوای HTML اصلی یک سایت قابل اجرا نیستند؛ علت هم به نرمال سازی‌های انجام شده بر روی محتوای یک سایت، توسط موتور مرورگر بر می‌گردد.
خود کتابخانه HtmlAgilityPack به ازای هر HtmlNode ایی که ارائه می‌دهد، خاصیت XPath معتبری را نیز به همراه دارد. در ادامه قصد داریم از این امکان توکار استفاده کرده و کلیه XPathهای یک محتوای HTML ایی را استخراج کنیم.

پردازش تگ‌های تو در توی یک HTML به کمک کتابخانه HtmlAgilityPack

```
using System;
using System.Linq;
using System.Net;
using System.Text;
using HtmlAgilityPack;

namespace HapTests
{
    public class HtmlReader
    {
        public Action<string> ParseError { set; get; }

        public Func<HtmlNode, bool> ParserHtmlNode { set; get; }

        public void StartParsingHtml(Uri url)
        {
            using (var 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)");
                StartParsingHtml(client.DownloadString(url));
            }
        }

        public void StartParsingHtml(string htmlContent)
        {
            if (string.IsNullOrWhiteSpace(htmlContent))
                throw new ArgumentNullException("content");

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

            if (doc.ParseErrors != null && doc.ParseErrors.Any())
            {
                foreach (var error in doc.ParseErrors)
                {
                    if (ParseError != null)
                        ParseError(error.Code + " - " + error.Reason);
                }
            }

            if (!doc.DocumentNode.HasChildNodes)
                return;

            handleChildren(doc.DocumentNode.ChildNodes);
        }

        private void handleChildren(HtmlNodeCollection nodes)
        {
            foreach (var itm in nodes)
```

```
            {
                if (itm.Name.ToLower().Equals("html"))
                {
                    if (itm.Element("body") != null)
                        handleChildren(itm.Element("body").ChildNodes);
                }
                else
                    handleHtmlNode(itm);
            }
        }

        private void parserChildNodes(HtmlNode content)
        {
            foreach (var item in content.ChildNodes)
            {
                handleHtmlNode(item);
            }
        }

        private void handleHtmlNode(HtmlNode htmNode)
        {
            switch (htmNode.Name.ToLower())
            {
                case "html":
                case "body":
                    handleChildren(htmNode.ChildNodes);
                    break;

                default:
                    if (ParserHtmlNode == null)
                        throw new ArgumentNullException("ParserHtmlNode");

                    if (ParserHtmlNode(htmNode))
                        parserChildNodes(htmNode);

                    break;
            }
        }
    }
}
```

در اینجا کدهایی را ملاحظه می‌کنید که علاوه بر ارائه تنظیمات اولیه HtmlAgilityPack (خصوصا با درنظر گرفتن مباحث ورودی یونیکد)، به صورت بازگشتی (با توجه به اینکه الزاما مسیر یا Node خاصی مدنظر نیست)، کلیه گره‌های یک HTML را بررسی و ارائه می‌دهند.

این کد برای نوشتن مبدل‌های HTML به XYZ بسیار مناسب است. برای مثال اگر بخواهید یک مبدل HTML به PDF را تهیه کنید، کدهای ابتدایی آن همین موارد است:

```
new HtmlReader
{
    ParseError = error => Console.WriteLine(error),
    ParserHtmlNode = htmlNode =>
    {
        //switch(htmlNode.Name) { }
        return true; //it's a nested node.
    }
}.StartParsingHtml(html);
```

نمونه‌ای از نحوه استفاده از کدهای کلاس HtmlReader را ملاحظه می‌کنید.

در اینجا html، محتوای HTMLایی در حال بررسی است. ParserHtmlNode یک callback است. هر زمانیکه به یک گره HTML برخورد، آن‌را در اختیار شما قرار می‌دهد. در ادامه فرصت خواهید داشت تا برای نمونه یک swicth را تهیه کرده و مثلا به ازای تگ hr یک خط رسم کنید، به ازای تگ br یک سطر جدید را درنظر بگیرید و الی آخر. اگر خروجی این Func را true درنظر بگیرید، فرض بر این خواهد بود که گره جاری تو در تو است (حالت دنیای واقعی)؛ در غیراینصورت، یک سطح این گره، بیشتر بررسی نخواهد شد.

در این کلاس، ParseError نیز یک callback است و اگر کتابخانه HtmlAgilityPack، در حین آنالیز کدهای HTML دریافتی به خطایی برخورد، آن‌را گزارش خواهد داد.

در کلاس فوق، دو حالت برای متد StartParsingHtml در نظر گرفته شده است. در حالت اول، یک Uri یا آدرس اینترنتی دریافت

و سپس آنالیز می‌گردد. در حالت دوم، فرض بر این است که محتوای کدهای HTML مدنظر به هر نحوی پیشتر تهیه شده و به صورت string موجود است.

استخراج کلیه XPathها از یک فایل HTML به کمک کتابخانه HtmlAgilityPack

اکنون که یک HTML Parser عمومی را تهیه کرده‌ایم، استخراج XPathها توسط آن کار ساده‌ای خواهد بود. یک مثال کامل را در این زمینه در ادامه ملاحظه می‌کنید:

```
using System;
using System.Diagnostics;
using System.IO;
using System.Text;
using HtmlAgilityPack;

namespace HapTests
{
    class Program
    {
        static void Main(string[] args)
        {
            var html =
                @"<table width='750' border='0' style='font-size: 10pt; width: 736px' class='boxcar2
gerd'>
            <tbody><tr>
            <td height='70' colspan='4' class='boxcart1 gerd'>
            <iframe width='718' scrolling='no'>
            </iframe></td>
            </tr>
            <tr>
            <td height='70' colspan='4' class='boxcart1 gerd'>
    </td>
            </tr>
            <tr>
            <td width='193' height='36' class='boxcart2 gerd'>
            <a target='_self' href='Curr.cbi.2.php'>نرخ ارز مبادله ای بانک مرکزی</a></td>
            <td width='181' height='36' class='boxcart2 gerd'>
            <a target='_self' href='Curr.cbi.php'>نرخ ارز مرجع بانک مرکزی</a></td>
            <td width='149' height='36' class='boxcart2 gerd'>
            <a target='_self' href='curv.htm'>نمودار قیمت طلا</a></td>
            <td width='199' height='36' class='boxcart2 gerd'>
            <a target='_self' href='index.php'>قیمت طلا و سکه در بازار ایران</a></td>
            </tr>
            <tr>
            <td height='48' colspan='4' class='boxcart1 gerd'>
            <p dir='rtl'><span style='font-size: 13pt;'>تابلو آنلاین قیمت جهانی طلا و نقره ( دلار
            )</span></p></td>
            </tr>
            <tr>
            <td height='57' colspan='2' class='boxcart1 gerd'>قیمت لحظه ای هر انس
            نقره در بازارهای جهانی<br>
            <span style='font-size: 9pt;'>
            </span></td>
            <td height='57' colspan='2' class='boxcart1 gerd'>قیمت لحظه ای هر انس
            طلا در بازارهای جهانی<br>
            <span style='font-size: 9pt;'>
            </span></td>
            </tr>
            <tr>
            <td height='48' colspan='4' class='boxcart1 gerd'>
            <p dir='rtl'><span style='font-size: 13pt'>تابلو آنلاین قیمت طلا ، سکه
            ریال ( در بازار ایران و نقره )</span></p>
            </td>
            </tr>
            <tr>
            <td style='direction: rtl; font-size: 8pt' colspan='4'><div align='center'>
                                <table id='gold_tbl'><tbody><tr><th>قیمت طلا</th><th>قیمت
زنده</th><th>تغییر</th>
                                <th>کمترین</th><th>بیشترین</th><th>زمان</th></tr><tr><td>انس طلا
<sup>دلار</sup></td>
                                <td class='s0_1'>1,375.90</td><td class='c0_1 neg'>(-0.34%) -4.70</td>
                                <td class='l0_1'>1,374.90</td><td class='h0_1'>1,380.80</td><td class='z0_1
fa'>17:53</td>
                                </tr><tr><td>مثقال طلا</td><td class='s3_2'>5,290,000</td>
                                <td class='c3_2 pos'>(1.63%) 85,000</td><td class='l3_2'>5,200,000</td><td
class='h3_2'>5,320,000</td><td class='z3_2 fa'>17:50</td></tr><tr><td>18 گرم طلای</td>
                                <td class='s3_3'>1,221,200</td><td class='c3_3 pos'>(1.63%) 19,600</td><td
```

```
class='l3_3'>1,200,400</td><td class='h3_3'>1,228,100</td><td class='z3_3 fa'>17:50</td>
                            </tr><tr><td>انس نقره <sup>دلار</sup></td><td class='s0_5'>21.83</td><td
class='c0_5'>(0.00%) 0.00</td><td class='l0_5'>21.67</td><td class='h0_5'>21.96</td>
                            <td class='z0_5 fa'>17:53</td></tr></tbody></table><br><table
id='coin_tbl'><tbody><tr><th>سکه</th><th>قیمت زنده</th><th>تغییر</th><th>کمترین</th>
                            <th>بیشترین</th><th>زمان</th><th>ارزش طلا</th></tr><tr><td>بهار
آزادی</td><td class='s3_10'>12,650,000</td><td class='c3_10 pos'>(2.68%) 330,000</td>
                            <td class='l3_10'>12,320,000</td><td class='h3_10'>12,650,000</td><td
class='z4_10'>11,918,400</td><td class='z3_10 fa'>16:07</td></tr><tr><td>امامی</td>
                            <td class='s3_11'>12,960,000</td><td class='c3_11 pos'>(2.61%)
330,000</td><td class='l3_11'>12,630,000</td><td class='h3_11'>13,050,000</td><td
class='z4_11'>11,918,400</td>
                            <td class='z3_11 fa'>17:43</td></tr><tr><td>نیم</td><td
class='s3_12'>6,880,000</td><td class='c3_12 pos'>(2.69%) 180,000</td><td class='l3_12'>6,700,000</td>
                            <td class='h3_12'>6,900,000</td><td class='z4_12'>5,959,200</td><td
class='z3_12 fa'>16:08</td></tr><tr><td>ربع</td><td class='s3_13'>4,250,000</td><td class='c3_13
pos'>(2.41%) 100,000</td>
                            <td class='l3_13'>4,150,000</td><td class='h3_13'>4,300,000</td><td
class='z4_13'>2,978,100</td><td class='z3_13 fa'>17:42</td></tr><tr><td>گرمی</td><td
class='s3_14'>2,940,000</td>
                            <td class='c3_14 pos'>(3.16%) 90,000</td><td
class='l3_14'>2,850,000</td><td class='h3_14'>2,940,000</td><td class='z4_14'>1,465,400</td><td
class='z3_14 fa'>17:40</td></tr></tbody></table></div></td>
            </tr>
            </tbody></table>
                ";

            extractXPath(html);
            test(html);
        }

        /// <summary>
        /// Converts /#comment[1] to /comment()[1]
        /// or /#text[1] to /text()[1]
        /// </summary>
        private static string GetValidXPath(string xpath)
        {
            var index = xpath.LastIndexOf("/");
            var lastPath = xpath.Substring(index);

            if (lastPath.Contains("#"))
            {
                xpath = xpath.Substring(0, index);
                lastPath = lastPath.Replace("#", "");
                lastPath = lastPath.Replace("[", "()[");
                xpath = xpath + lastPath;
            }

            return xpath;
        }

        private static void extractXPath(string html)
        {
            var sb = new StringBuilder();
            new HtmlReader
            {
                ParseError = error => Console.WriteLine(error),
                ParserHtmlNode = htmlNode =>
                {
                    if (htmlNode is HtmlTextNode)
                    {
                        sb.AppendLine("Text NodeName: " + htmlNode.Name.Trim());
                        sb.AppendLine("InnerText: " + htmlNode.InnerText.Trim());
                    }
                    else
                    {
                        sb.AppendLine("NodeName: " + htmlNode.Name.Trim());
                        var nodeText = new StringBuilder();
                        for (int i = 0; (i < htmlNode.OuterHtml.Length && htmlNode.OuterHtml[i] !=
'>'); i++)
                            nodeText.Append(htmlNode.OuterHtml[i]);

                        nodeText.Append(">");

                        sb.AppendLine("Node Start: " + nodeText.ToString());
                    }

                    sb.AppendLine("XPath: " + GetValidXPath(htmlNode.XPath.Trim()));
                    sb.AppendLine(Environment.NewLine);

                    return true; //it's a nested node.
```

```
                }
            }.StartParsingHtml(html);

            File.WriteAllText("xpath.txt", sb.ToString());
            Process.Start("xpath.txt");
        }

        private static void test(string html)
        {
            var doc = new HtmlDocument
            {
                OptionCheckSyntax = true,
                OptionFixNestedTags = true,
                OptionAutoCloseOnEnd = true,
                OptionDefaultStreamEncoding = Encoding.UTF8
            };
            doc.LoadHtml(html);
            var node =
doc.DocumentNode.SelectSingleNode("/table[1]/tbody[1]/tr[7]/td[1]/div[1]/table[2]/tbody[1]/tr[6]/td[7]/
text()[1]");
            Console.WriteLine(node.InnerText);
        }
    }
}
```

در این مثال html مقداری است که از یک سایت عمومی دریافت شده است.
سپس نمونه‌ای دیگر از نحوه استفاده از کلاس HtmlReader قسمت قبل را در ادامه، در متد extractXPath ملاحظه می‌کنید. در اینجا کلاس HtmlReader در یک عملیات بازگشتی، کلیه گره‌های تو در توی HTML مورد نظر را آنالیز کرده و توسط callback ایی به نام ParserHtmlNode در اختیار ما قرار می‌دهد. اکنون که این htmlNode را داریم، خاصیت XPath آن دقیقا مقداری است که به دنبالش هستیم.
در اینجا چند نکته حائز اهمیت هستند:
- با بررسی HtmlTextNode، به نودهایی خواهیم رسید که دارای مقدار متنی هستند. در غیراینصورت این گره، خود ابتدای یک سری گره تو در توی دیگر است.
- XPath بازگشتی توسط کتابخانه HtmlAgilityPack نیاز به کمی تمیز سازی دارد. اینکار در متد GetValidXPath انجام شده است.
- در متد test انتهایی، نمونه‌ای از نحوه استفاده از XPathهای استخراجی را ملاحظه می‌کنید.

```
Text NodeName: #text
InnerText: 17:40
XPath: /table[1]/tbody[1]/tr[7]/td[1]/div[1]/table[2]/tbody[1]/tr[6]/td[7]/text()[1]
```

برای نمونه سه سطر فوق، یکی از مداخل فایل نهایی تولیدی مثال جاری است. اکنون که XPath را داریم، استفاده از آن جهت استخراج مقدار InnerText مدنظر، ساده خواهد بود.

نظرات خوانندگان

نویسنده: مهدی پایروند
تاریخ: ۱۳۹۲/۰۳/۲۷ ۸:۸

ممنون بابت مطلب مفیدتون، برای پردازش محتوای جاوا اسکریپت هم میشه از این کتابخانه استفاده کرد؟

نویسنده: وحید نصیری
تاریخ: ۱۳۹۲/۰۳/۲۷ ۸:۴۴

- برای استخراج هر نوع تگ قرار گرفته شده داخل HTML نهایی، میشه از این کتابخانه استفاده کرده. فقط کافی است در switch مطلب فوق، htmlNode.Nameها را تحت نظر قرار داد.
- برای اجرای کدهای جاوا اسکریپت در دات نت، یک سری موتور ویژه برای اینکار هست .

نویسنده: علی
تاریخ: ۱۳۹۲/۰۳/۲۷ ۱۰:۳۹

سلام جناب نصیری
آیا سرعت این کتابخانه از کتابخانه LINQ To HTML بیشتر هست؟ اگر مقایسه اس هم انجام بدهید عالی می‌شود

نویسنده: وحید نصیری
تاریخ: ۱۳۹۲/۰۳/۲۷ ۱۱:۳۵

برای انتخاب یک کتابخانه صرفا به سرعت آن نباید توجه کرد. این موارد برای انتخاب کتابخانه‌های ثالث، مهم هستند:
- آیا این کتابخانه محلی برای بحث و رفع اشکال دارد؟
- آیا سورس باز است؟ (غیر الزامی؛ اما یک امتیاز مثبت)
- آیا به همراه مثال‌های کاربردی است؟
- آیا مستندات قابل قبولی دارد؟
- آیا در جستجویی که انجام شده، کسی از آن در پروژه‌های خودش استفاده می‌کند؟
- آیا هر از چندگاهی به روز می‌شود؟ آخرین باری که به روز شده چه زمانی بوده؟
- آیا استفاده از آن در انواع و اقسام پروژه‌ها مجاز است؟ مجوز استفاده از آن به چه نحوی است؟

نویسنده: افشین
تاریخ: ۱۳۹۲/۰۳/۲۸ ۲:۹

بسیار مفید بود جناب نصیری.. ممنون.

یک مشکلی که هست، وقتی متن (این) صفحه رو با این روش پردازش می‌کنم، کاراکترهای نامفهومی نمایش داده می‌شه..
Encoding رو چطور تنظیم کنم، یا مشکل از جای دیگه ای هست؟

برای مثال InnerText این XPath:

```
/html[1]/body[1]/table[1]/tr[1]/td[1]/table[1]/tr[1]/td[2]/table[1]/tr[1]/td[1]/font[1]/td[2]/font[1]/td[1]/map[1]/tr[3]/td[1]/table[1]/tr[1]/td[1]/table[1]/tr[3]/td[1]/table[1]/tr[1]/td[1]/table[1]/tr[1]/td[1]/table[1]/tr[1]/td[2]/a[1]/td[1]/a[1]/table[3]/tr[2]/td[2]/table[1]/tr[1]/td[1]/div[1]/span[1]/span[1]/html[1]/head[1]/title[1]/text()[1]
```

نویسنده: وحید نصیری
تاریخ: ۱۳۹۲/۰۳/۲۸ ۹:۳۱

این صفحه 1256 است.

```
<meta http-equiv="Content-Type" content="text/html; charset=windows-1256">
```

در کدهای فوق به این شکل باید تنظیم شود:

```
using (var client = new WebClient { Encoding = Encoding.GetEncoding("windows-1256") })
```

نویسنده: صابر فتح الهی
تاریخ: ۱۳۹۲/۰۵/۱۶ ۱۸:۳۸

سلام؛ من از این کتابخانه استفاده کردم اما وقتی سایت دانلود شد که روش پردازش انجام بدم درخواست‌های نامتقارنی که بعدا لود میشن و محتوی صفحه تغییر میدن وجود ندارن چطور به اون محتویات دسترسی داشته باشم؟

نویسنده: وحید نصیری
تاریخ: ۱۳۹۲/۰۵/۱۶ ۱۸:۵۴

- این کتابخانه پردازشگر جاوا اسکریپتی نداره (همزمان و یا حالت‌های دیگری مانند Ajax ایی). فرضش بر این است که محتویات کامل رو در اختیارش قرار دادید.
- یک راه این است که از Web Control دات نت (موجود در WinForms و همچنین WPF) که در پشت صحنه از موتور کامل IE استفاده می‌کند، کمک بگیرید و زمانیکه Document آن کاملا load شد ، نتیجه آن‌را به این کتابخانه ارسال کنید.

نویسنده: صابر فتح الهی
تاریخ: ۱۳۹۲/۰۵/۱۶ ۱۹:۲۹

استفاده کردم حق با شما بود سایت کامل لود می‌کنه اما در خواست‌های Ajax پردازش نمی‌شه، یعنی این کامل شدن لود ربطی به درخواست‌های ای جکس نداره، خروجی ای جکسی وجود داره توی صفحه ولی در سورس html وجود ندارد