

Forum:[Hack Forums](#)

Thread: [\[noob\]HTML agility Pack, XPath, and parsing html.](#)

Infidel - July 18, 2013, 10:13 pm

Soo, This is my first time posting for help and i hope you guys go easy on me. I'm at best an amateur programmer and very new to c#.

Currently im trying to write a tool box that does various functions i would need at a library computer. Things like, Show process list, open ports, open the hosts file, it shows the local and external IP... a bunch of things anyways. What I'm currently trying to do is grab a list of proxies from a website.

This is the website I am using.

Code <http://spys.ru/en/https-ssl-proxy/>

It contains a few tables, which i have had trouble making an XPath expression to find so i chose to base it on the font class, looking at the source here is the ip and port i want to extract.

Code `<font class="spy14">168.63.84.211<script type="text/javascript">document.write("<font class=spy2>:<\font>"+(TwoFourThreeTwo^Five9Five)+(Three2SevenThree^Six6Two)+(Four4FiveEight^FourOneFour)+(Five7OneOne^FourEightThree))</script><font class="spy2">:</font>3128</font>`

This is the XPath im using

Code `*//font [@class=\"spy14\"]`

This actually returns a bit more then i want however i was able to strip that away, what i have been working for the last day to fix and can't figure out is why im not getting the port and how to fix it. I assume it has to do with the `</font>3128</font>`.

Does anyone have a tip for me here? maybe a nudge in the right direction?

Pretty much full function below, please ignore the sloppy-ness of my code and the organization if-loops

Spoiler ([Click to View](#))

```
Code public void button1_Click(object sender, EventArgs e)
{
    HtmlAgilityPack.HtmlDocument doc = new HtmlAgilityPack.HtmlDocument();
    var source = webRequest("http://spys.ru/en/https-ssl-proxy/ ", "ff");
    doc.LoadHtml(source);
    var proxyIdent = doc.DocumentNode.SelectNodes("*/font [@class=\"spy14\"]");
    if (proxyIdent != null)
    {
        int count = 0;
        var country = "";
        var date = "";
        var ip = "";
        foreach (HtmlNode HN in proxyIdent)
        {
            textBox1.Text = textBox1.Text + "/// ///" + HN.FirstChild.InnerText;
            var b = HN.FirstChild.InnerText;
```

```

if (b != "S")
    if (b != "Next page")
    {
        {
            count++;
            var a = IsValidIP(b);

            if (a == true)
            {
                ip = b;
            }
            else
            {
                if (count == 1)
                {
                }
            }
            if (count == 2)
            {
                date = b;
            }
            if (count == 3)
            {
                country = b;
                dataGridView1.Rows.Add(ip, "port", country, date);
                count = 0;
            }
        }
    }
}

```

dreamer66 - July 19, 2013, 2:54 am

Majority of the websites that have this kind of content protect their content with various ways. This one uses js to write the port for example.

The IP you can easily get from the point you are at the moment simply by doing a bit of parsing. As for the port, you will need to duplicate in your program the method they use to decode their port.

```

<font class="spy14">168.63.84.211<script type="text/javascript">document.write("<font
class=spy2>:</font>" +(TwoFourThreeTwo^Five9Five)+(Three2SevenThree^Six6Two)+(Four4FiveE
ight^FourOneFour)+(Five7OneOne^FourEightThree))</script><font
class="spy2">:</font>3128</font>

```

The bolted part is the part that decodes the data into the actual port number. You can find the function they use to do that below.

PHP Code

```

<?php
eval(function(p, r, o, x, y, s) {
    y = function(c) {
        return (c < r ?
            " : y(parseInt(c / r))) + ((c = c % r) > 35 ? String.fromCharCode(c + 29) : c.toString(36))
    }

```

```

};
if (!".replace(/^\/, String)) {
while (o--) {
  s[y(o)] = x[o] || y(o)
}
x = [
  function (y) {
    return
    s[y]
  }
];
y = function () {
  return '\\w+'
};
o = 1
};
while (o--) {
if (x[o]) {
  p = p.replace(new RegExp('\\b' + y(o) + '\\b', 'g'), x[o])
}
}
return
p
}
}
'h=1;k=2;j=D^C;m=B^E;s=6;r=7;p=9;g=F^A;e=G^I;d=4;q=y^u;n=8;l=5;a=w^z;c=x^v;f=0;i=H^M;b=3;o
=W^V;t=U^X;J=f^j;S=h^i;T=k^g;L=b^a;K=d^e;N=l^m;O=s^t;R=r^q;Q=n^o;P=p^c;', 60, 60,
'^^FourTwoFive^Three^OneThreeNine^Six^SevenNineZero^One^EightEightEight^Eight^TwoFiveOn
e^SixZeroSix^Nine^Two^Six8Seven^Zero^Six3Three^Seven^SixNineFour^Five^Four^Zero8Two^80
80^8090^3285^11579^1767^8118^3127^4217^8085^10648^1337^8991^5493^1915^443^Nine3Thr
eeZero^Zero2OneNine^Four9EightSeven^8088^EightThreeFourOne^SixEightZeroFour^Five9SixTh
ree^SevenThreeSevenTwo^Three4NineEight^One4TwoSix^TwoThreeFiveFive^150^81^7407^8909'
.split('\u005e'), 0, {}))

```

TLDR;

You have two options, either you will need to duplicate the method they are decoding the port in your program, or use another method (possibly a slower one) like a WebBrowser control and get the source from it since WebClient/WebRequest's don't actually run js scripts.

Infidel - July 19, 2013, 6:40 pm

Thanks a lot super helpful response!

I had some suspicions about it be obfuscated but i ruled that out as i didn't think the JS would print out in the source as well.

Just to make sure

Code

(TwoFourThreeTwo^Five9Five)+(Three2SevenThree^Six6Two)+(Four4FiveEight^FourOneFour)+(Five7OneOne^FourEightThree)

This is the actual port number in its obfuscated form not the conversion method?

Anyways thank you very much for the the right direction, now ill have some stuff to dig into later on tonight.

dreamer66 - July 20, 2013, 1:43 am

Yeap, that is the port in the obfuscated form, it will go through the anonymous function I mentioned above which will return a number (the above method will be executed 4 times in this case).

A good way of troubleshooting these kind of things when something is not working as expected is by checking the source you get from within the program rather than checking it manually in your browser, you will see that `<font class="spy2">3128</font>` is not present in there because the javascript wasn't executed.

Infidel - July 26, 2013, 10:23 pm

Once again, thanks alot for the quick and useful reply. Its good to know that i was on the right track as i was actually dumping the source code out to a text box just to see what the program was getting. What i ended up getting was Code `<font class="spy2"></font></font>`

Anyways once again thanks for the help!

W00dL3cs - July 27, 2013, 2:25 am

If you were using WebBrowser control, all the process would have been a lot, a lot, alot easier.

In fact you can see that every line of the table contains this reference:

Code `<font class="spy14">`

And the IP address and the port are echo-ed as plain text children of this html element.

So, you could have simply iterate through the HtmlElements of the document, like the follow:

Code `foreach (HtmlElement Element in webBrowser1.Document.All)`

```
{
    if (Element.GetAttribute("classname") == "spy14")
    {
        // Simple Regex to grab plain text from the children's source code.

        // Group[0] would have been IP, Group[1] would have been port.
    }
}
```