

Forum: [Hack Forums](#)

Thread: [Get text off website using webclient?](#)

Proioxix - December 31, 2012, 3:11 pm

I am just learning httpwebrequest. I never had much of a use for it but I want to learn it now. I looked at some tutorials online and such and figured what I wanted to do was make a proxy grabber. I figured it wouldn't be so difficult all I would have to do is grab the text off the website.

Here is what I have so far

```
Code Dim webStream As Stream
```

```
Dim webResponse As String
```

```
Dim req As HttpWebRequest
```

```
Dim res As HttpWebResponse
```

```
req = WebRequest.Create("http://www.hidemyass.com/proxy-list/")
```

```
req.UserAgent = ("Mozilla/6.0 (Windows NT 6.2; WOW64; rv:16.0.1) Gecko/20121011  
Firefox/16.0.1")
```

```
req.Method = "GET"
```

```
res = req.GetResponse()
```

```
webStream = res.GetResponseStream()
```

```
Dim webStreamReader As New StreamReader(webStream)
```

```
While webStreamReader.Peek >= 0
```

```
webResponse = webStreamReader.ReadToEnd()
```

```
End While
```

```
frmWb.Show()
```

```
frmWb.WebBrowser1.DocumentText = webResponse
```

It opens another form i made real quick that has a webbrowser in it to display the response. It pretty much brings up the page in all text, no css or pictures. But I'm wondering how I can get the proxies off there? I'm not asking to be spoonfed I'm just wondering where I can find a decent reference on how to get the text I'm looking for off the page

Cody8295 - December 31, 2012, 4:32 pm

If you're just doing a GET request, perhaps the WebClient is a better option.

<http://msdn.microsoft.com/en-us/library/system.net.webclient.aspx>

Proioxix - December 31, 2012, 5:06 pm

(Dec 31, 2012 05:32 PM)Cody8295 Wrote: [\[CITE\]](#) If you're just doing a GET request, perhaps the WebClient is a better option.

<http://msdn.microsoft.com/en-us/library/system.net.webclient.aspx>

thanks for showing me this. I feel this will be a lot easier with a webclient but I've run into a snag. Not sure if you can help but the html code for the website is... confusing to say the least

this html code here

Spoiler [\(Click to View\)](#)

```
Code </style><span class="DjBk">61</span><span class="55">.</span><span  
class="EV5C">22</span><div style="display:none">22</div><span class="Wl4p">49</span><span  
style="display:none">100</span><span style="display:none">122</span><div
```

```

style="display:none">122</div><span style="display:none">124</span><span
class="EV5C">124</span><span></span><span style="display:none">207</span><span
class="129">.</span><span style="display:none">10</span><span></span><span
class="Wl4p">40</span><span class="zBUq">99</span><div style="display:none">99</div><span
style="display:none">109</span><span class="V7CG">109</span><span></span><span
style="display:none">217</span><span style="display:none">219</span><span
class="EV5C">219</span><span></span><span
style="display:none">232</span><span></span><span class="zBUq">234</span><div
style="display:none">234</div><span class="7">.</span><span style="display:
inline">28</span></span></td>
<td>
8080</td>

```

seems to just output this

(either that or I'm way to dumb and just missing the obvious)

this is my new code here in replacement of the httpwebrequest

```
Code Dim wc As New WebClient
```

```
    wc.Headers.Add("user-agent", "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.2; WOW64;
Trident/6.0; .NET4.0E; .NET4.0C; .NET CLR 3.5.30729; .NET CLR 2.0.50727; .NET CLR
3.0.30729)")
```

```
    Dim html As Byte() = wc.DownloadData("http://www.hidemyass.com/proxy-list/")
```

```
    Dim utf As New UTF8Encoding()
```

```
    frmWb.Show()
```

```
    frmWb.WebBrowser1.DocumentText = utf.GetString(html)
```

```
    RichTextBox1.Text = utf.GetString(html)
```

Cody8295 - December 31, 2012, 6:29 pm

(Dec 31, 2012 06:06 PM)Proioxix Wrote: [\[CITE\]](#) this is my new code here in replacement of the httpwebrequest

```
Code Dim wc As New WebClient
```

```
    wc.Headers.Add("user-agent", "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 6.2; WOW64;
Trident/6.0; .NET4.0E; .NET4.0C; .NET CLR 3.5.30729; .NET CLR 2.0.50727; .NET CLR
3.0.30729)")
```

```
    Dim html As Byte() = wc.DownloadData("http://www.hidemyass.com/proxy-list/")
```

```
    Dim utf As New UTF8Encoding()
```

```
    frmWb.Show()
```

```
    frmWb.WebBrowser1.DocumentText = utf.GetString(html)
```

```
    RichTextBox1.Text = utf.GetString(html)
```

You don't even need to do that much, you can minimize all that into

```
Code wc.DownloadString("url")
```

Proioxix - December 31, 2012, 6:47 pm

(Dec 31, 2012 07:29 PM)Cody8295 Wrote: [\[CITE\]](#) You don't even need to do that much, you can minimize all that into

```
Code wc.DownloadString("url")
```

Doing so results in the source still showing up...

However I peeked at some source for a program that does the same thing mine is trying to do (on the same website) and there was this

Code Dim r As New

```
System.Text.RegularExpressions.Regex("(\\d{1,4}\\.\\d{1,4}\\.\\d{1,4}\\.\\d{1,4})</span></td>\\D+(\\d{1,5})</td>\\D+<img src=\"\"\\S+\"\"\\D+\"\"flag\"\" />
(\\w+)\\D+\\S+\\D+(\\d+%)\\D+\\S+\\D+(\\d+%)\"\"\\D+<td>(\\w+)\\D+\"\">(\\w+)\"")
```

except that seems to have changed because that program doesn't work now. I'd say this project is a bust I'll just try something else because I can't find any sites with a free frequently updating proxy list

Tidas - December 31, 2012, 6:51 pm

(Dec 31, 2012 07:47 PM)Proioxis Wrote: [\[CITE\]](#) Doing so results in the source still showing up...

However I peeked at some source for a program that does the same thing mine is trying to do (on the same website) and there was this

Code Dim r As New

```
System.Text.RegularExpressions.Regex("(\\d{1,4}\\.\\d{1,4}\\.\\d{1,4}\\.\\d{1,4})</span></td>\\D+(\\d{1,5})</td>\\D+<img src=\"\"\\S+\"\"\\D+\"\"flag\"\" />
(\\w+)\\D+\\S+\\D+(\\d+%)\\D+\\S+\\D+(\\d+%)\"\"\\D+<td>(\\w+)\\D+\"\">(\\w+)\"")
```

except that seems to have changed because that program doesn't work now. I'd say this project is a bust I'll just try something else because I can't find any sites with a free frequently updating proxy list Because it downloads the entire string of the website. It doesn't magically know what you want to get from it.

Use RegEx to get the info.

Proioxis - December 31, 2012, 8:12 pm

(Dec 31, 2012 07:51 PM)Tidas Wrote: [\[CITE\]](#) Because it downloads the entire string of the website. It doesn't magically know what you want to get from it.

Use RegEx to get the info.

after spending a bit of time researching how to use regex i understand what you mean. Thanks for telling me that

??? Matto ??? - January 1, 2013, 8:17 am

If you struggle with RegEx a quick and dirty solution is to use the public GetBetweenAll and GetBetween functions.

Cody8295 - January 1, 2013, 8:47 am

(Jan 1, 2013 09:17 AM)??? Matto ??? Wrote: [\[CITE\]](#) If you struggle with RegEx a quick and dirty solution is to use the public GetBetweenAll and GetBetween functions.

I believe GetBetween uses Split(), which is, if anything, less dirty than Regex, especially for HTML.

??? Matto ??? - January 1, 2013, 8:49 am

You'll need to set up loops and cleanup the results, with a good RegEX string you can easily get all the matches in one search.

TrackPro+ - January 1, 2013, 8:55 am

(Jan 1, 2013 09:49 AM)??? Matto ??? Wrote: [\[CITE\]](#) You'll need to set up loops and cleanup the

results, with a good RegEX string you can easily get all the matches in one search.
That's what most are suggesting, but instead of suggesting, help?

??? Matto ??? - January 1, 2013, 9:01 am

I don't know how to write good RegEX strings? Instead of criticising my posts and getting off-topic, help?

OP: You can find loads of Regular Expressions to match IP addresses on Google, here are some (untested), obviously these will only get the IP not the port, this is a starting point for you if you want to use RegEX

```
\b\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3}\b
\b(?:\d{1,3}\.){3}\d{1,3}\b
```

Cody8295 - January 1, 2013, 9:28 am

(Jan 1, 2013 10:01 AM)??? Matto ??? Wrote: [\[CITE\]](#) I don't know how to write good RegEX strings? Instead of criticising my posts and getting off-topic, help?

OP: You can find loads of Regular Expressions to match IP addresses on Google, here are some (untested), obviously these will only get the IP not the port, this is a starting point for you if you want to use RegEX

```
\b\d{1,3}\.\d{1,3}\.\d{1,3}\.\d{1,3}\b
\b(?:\d{1,3}\.){3}\d{1,3}\b
```

RegEx isn't even appropriate for HTML anyways...

??? Matto ??? - January 1, 2013, 9:30 am

I thought you can use RegEX for any kind of text?

Cody8295 - January 1, 2013, 9:35 am

(Jan 1, 2013 10:30 AM)??? Matto ??? Wrote: [\[CITE\]](#) I thought you can use RegEX for any kind of text?

You can, but HTML tags aren't regular expressions, therefor RegEx isn't the best choice.

??? Matto ??? - January 1, 2013, 9:39 am

Not sure I follow, a regular expression is the search pattern e.g. `\b(?:\d{1,3}\.){3}\d{1,3}\b`

So what do you mean by HTML tags aren't regular expressions? You can use regex with strings containing HTML tags?

Cody8295 - January 1, 2013, 10:50 am

(Jan 1, 2013 10:39 AM)??? Matto ??? Wrote: [\[CITE\]](#) Not sure I follow, a regular expression is the search pattern e.g. `\b(?:\d{1,3}\.){3}\d{1,3}\b`

So what do you mean by HTML tags aren't regular expressions? You can use regex with strings containing HTML tags?

HTML = Hyper Text Markup Language

Any markup language is not a regular expressions. Regular expressions include string of text, not code.

??? Matto ??? - January 1, 2013, 11:16 am

Ah ok, thanks for explaining it. Never knew you couldn't parse markup lanugages with RegEX, I

assume its the same with XML?

Cody8295 - January 1, 2013, 11:30 am

(Jan 1, 2013 12:16 PM)??? Matto ??? Wrote: [\[CITE\]](#) Ah ok, thanks for explaining it. Never knew you couldn't parse markup lanugages with RegEX, I assume its the same with XML?

You can parse whatever you want, it will work fine, it's just not appropriate to do so.

ident - January 1, 2013, 2:32 pm

(Jan 1, 2013 10:28 AM)Cody8295 Wrote: [\[CITE\]](#) RegEx isn't even appropriate for HTML anyways... Care to explain how? HTML is a string. And dont post that silly ONE post. Explain from your own understanding.

Proioxix - January 1, 2013, 2:49 pm

its to my understanding, only after 30 or so minutes of researching regex yesterday that it can be used to find regular expressions in html code to get a specific text value within the html so you can output a string from the html code into the program without the tags

ident - January 1, 2013, 2:54 pm

Cody constantly refers to this post

<http://stackoverflow.com/questions/1732348/regex-match-open-tags-except-xhtml-self-contained-tag>
[s](#) But actually can't explain in his own words. The comment is not 100% wrong but holds little water. Regex can very much parse HTML. Strings are strings.

Cody8295 - January 1, 2013, 4:39 pm

(Jan 1, 2013 03:54 PM)ident Wrote: [\[CITE\]](#) Cody constantly refers to this post

<http://stackoverflow.com/questions/1732348/regex-match-open-tags-except-xhtml-self-contained-tag>
[s](#) But actually can't explain in his own words. The comment is not 100% wrong but holds little water. Regex can very much parse HTML. Strings are strings.

I've been explaining it for the past 2 pages, reread.

carrot223 - January 2, 2013, 2:58 am

That one SO question has spread a lot of confusion. RegEx is ideal for extracting specific information from specific html, like to OP requires.

Take note of the 2nd answer to that question.