

Forum:[Hack Forums](#)

Thread: [Need help with regex code](#)

dark.team - June 23, 2015, 4:51 am

Quote:>fireup;</td><td class="nw">45fcea35db75fab8ce78e2fccd6b4563 </td><td class="nw">

i have the above code in richtext box i would to get the hash 45fcea35db75fab8ce78e2fccd6b4563 by using regex how can i do that.Thanks

Only3 - June 23, 2015, 5:02 am

```
MsgBox(System.Text.RegularExpressions.Regex.Match(">fireup;</td><td class=""nw"">45fcea35db75fab8ce78e2fccd6b4563&nbsp;</td><td class=""nw"">", """)>(.*)&").Groups(1).Value)
```

dark.team - June 23, 2015, 5:17 am

(Jun 23, 2015 05:02 AM)Only3 Wrote: [\[CITE\]](#)
MsgBox(System.Text.RegularExpressions.Regex.Match(">fireup;</td><td class=""nw"">45fcea35db75fab8ce78e2fccd6b4563 </td><td class=""nw"">", """)>(.*)&").Groups(1).Value)

Your code works fine but i changed the code like this but its not working.
Quote:TextBox1.Text = System.Text.RegularExpressions.Regex.Match(RichTextBox2.Text, """)>(.*)&").Groups(1).Value

RichTextBox2.Text contains some website source....Thanks for the help.

Only3 - June 23, 2015, 5:35 am

```
System.Text.RegularExpressions.Regex.Match(RichTextBox2.Text, "class=""nw"">(.*)&").Groups(1).Value
```

VmX5 - June 23, 2015, 9:47 am

Well If the string will ALWAYS start and ALWAYS end with the same you could go the easy way and do this:

```
Code Dim i as String = ">fireup;</td><td class=""nw"">45fcea35db75fab8ce78e2fccd6b4563&nbsp;</td><td class=""nw"">"
i = i.Replace(">fireup;</td><td class=""nw"">", "")
i = i.Replace("&nbsp;</td><td class=""nw"">", "")
Msgbox("Your Final String Will Be : " & i)
```

LUOIdRentA - June 23, 2015, 11:32 am

```
Code Dim hash As String = Regex.Match(RichTextBox2.Text, ">fireup;<Vtd><td\sclass=""nw"">(.*);<Vtd><td\sclass=""nw"">").Groups(1).Value
```

Is it possible to provide us with the full source code? That way we can alter the regex pattern to fit the source code better.

+Haze - June 23, 2015, 11:37 am

You shouldn't really be using Regex to parse a HTML document. Regex is a terribly inefficient way

to parse HTML. If computers had the capability to cry then your computer would cry each time you attempted to parse HTML with Regex.

This is Regex abuse. Search up on Finite State Automaton, you'll see why you should never try to parse HTML with Regex and how inefficient it can be. Rather, you should use a HTML Parser such as HTMLAgilityPack.

Regex was designed to match Regular Expressions and not context free languages like HTML.

Only3 - June 23, 2015, 11:45 am

(Jun 23, 2015 11:37 AM)+Haze Wrote: [\[CITE\]](#) You shouldn't really be using Regex to parse a HTML document. Regex is a terribly inefficient way to parse HTML. If computers were able to cry then your computer would cry each time you attempted to parse HTML with Regex.

This is Regex abuse. Search up on Finite State Automaton, you'll see why you should never try to parse HTML with Regex. Rather, you should use a HTML Parser such as HTMLAgilityPack.

Regex was designed to match Regular Expressions and not context free languages like HTML.
who say ? some dumb guy from stackoverflow
Dont [\(Click to View\)](#)wb ?

+Haze - June 23, 2015, 11:49 am

(Jun 23, 2015 11:45 AM)Only3 Wrote: [\[CITE\]](#) who say ? some dumb guy from stackoverflow
Dont [\(Click to View\)](#)wb ?

Anybody that knows what they're doing would recommend you not to parse HTML with Regex. While there was a guy I worked with that got roasted by two of my senior co-workers simply because he suggested using to scrape data using a regex exp on a html page.
I mean sure it works, but it's not recommended to use.

Using Regex to parse HTML is just as bad as using `Me.CheckForIllegalCrossThreadCalls = False`

Read up on the theory of finite state automaton, and you'll see why Regex shouldn't be parsed with them.

LUOIdRentA - June 23, 2015, 11:55 am

(Jun 23, 2015 11:49 AM)+Haze Wrote: [\[CITE\]](#) Anybody that knows what they're doing would recommend you not to parse HTML with Regex. While there was a guy I worked with that got roasted by two of my senior co-workers simply because he suggested using to scrape data using a regex exp on a html page.
I mean sure it works, but it's not recommended to use.

Using Regex to parse HTML is just as bad as using `Me.CheckForIllegalCrossThreadCalls = False`

Read up on the theory of finite state automaton, and you'll see why Regex shouldn't be parsed with them.

Doesn't it depend how much you're scraping? If you just want one string in an html source, wouldn't using regex be acceptable? I thought like the HTML Agility Pack was only for scraping large amounts. If I'm wrong, please correct me. I'm still learning about all this stuff so I wouldn't be surprised if I was wrong.

+Haze - June 23, 2015, 12:01 pm

(Jun 23, 2015 11:55 AM)thevegita730 Wrote: [\[CITE\]](#) Doesn't it depend how much you're scraping? If you just want one string in an html source, wouldn't using regex be acceptable? I thought like the

HTML Agility Pack was only for scraping large amounts. If I'm wrong, please correct me. I'm still learning about all this stuff so I wouldn't be surprised if I was wrong.

Well that's the thing, Regex is a quick fix for parsing HTML. Ideally you should be using HTMLAgilityPack regardless of how much you're intending to scrape. It's good to put good practice into use as much as possible. That's like saying, "I should only use Delegates and Invoking when multi threading when I'm doing big tasks, but when I'm doing a quick task I should use "Me.CheckForCrossThreadCalls = False".

See how that mentality is bad?

Also, Parsing a HTML file with HTMLAgilityPack is so sweet, it's so much nicer and easier to use, and at the end of the day it's easier to write maintainable code with it. Any dev can hop into your code straight from where you left with ease.

LUOIdRentA - June 23, 2015, 1:55 pm

(Jun 23, 2015 12:01 PM)+Haze Wrote: [\[CITE\]](#) Well that's the thing, Regex is a quick fix for parsing HTML. Ideally you should be using HTMLAgilityPack regardless of how much you're intending to scrape. It's good to put good practice into use as much as possible. That's like saying, "I should only use Delegates and Invoking when multi threading when I'm doing big tasks, but when I'm doing a quick task I should use "Me.CheckForCrossThreadCalls = False".

See how that mentality is bad?

Also, Parsing a HTML file with HTMLAgilityPack is so sweet, it's so much nicer and easier to use, and at the end of the day it's easier to write maintainable code with it. Any dev can hop into your code straight from where you left with ease.

I see, thanks for that comment, I learned a lot from that. I shall definitely look into using the HTMLAgilityPack more. I also understand and agree with you completely on the example that you gave about multithreading.

Only3 - June 23, 2015, 2:36 pm

? <https://code.msdn.microsoft.com/VBASPNETStripHtmlCode-f6ace456>

+Haze - June 23, 2015, 2:43 pm

(Jun 23, 2015 02:36 PM)Only3 Wrote: [\[CITE\]](#) ?

<https://code.msdn.microsoft.com/VBASPNETStripHtmlCode-f6ace456>

What's that suppose to prove?

Not too long ago, I applied for a job at a big software company. To verify my account the company sent me a link and both my username and password I had just registered with in plaintext. This obviously means my password is not safely stored (not hashed, though it could be encrypted).

But, does that also mean that just because a huge cooperation does something that we should follow the same principles?

Look at Adobe, before they used to store their passwords in plain text, does that mean we should also store ours in plain text just because adobe is doing so?

CodeAssassin - June 23, 2015, 2:51 pm

(Jun 23, 2015 02:43 PM)+Haze Wrote: [\[CITE\]](#) What's that suppose to prove?

Not too long ago, I applied for a job at a big software company. To verify my account the company sent me a link and both my username and password I had just registered with in plaintext. This

obviously means my password is not safely stored (not hashed, though it could be encrypted).

But, does that also mean that just because a huge cooperation does something that we should follow the same principles?

Look at Adobe, before they used to store their passwords in plain text, does that mean we should also store ours in plain text just because adobe is doing so?

To be fair adobe are not MS and MS invented the .net library so if regex is fine with them its fine with me :)

Edit: Also I used to work for a large company as vb.net dev(I am now freelance as a c# developer) , The boss said he dose not care what methods are used as long as it dose the job and works, Its good code..

As many other business's are run time is the key factor and keeping to the SLA agreements are the only true thing to worry about.

+Haze - June 23, 2015, 3:08 pm

(Jun 23, 2015 02:51 PM)sweetysucksclock Wrote: [\[CITE\]](#) To be fair adobe are not MS and MS invented the .net library so if regex is fine with them its fine with me :)

Edit: Also I used to work for a large company as vb.net dev(I am now freelance as a c# developer) , The boss said he dose not care what methods are used as long as it dose the job and works, Its good code..

As many other business's are run time is the key factor and keeping to the SLA agreements are the only true thing to worry about.

Microsoft inventing the .NET library? What does that link in with your argument?

No offense but I think your boss is either very old or is also giving bad advice. So if what your boss says is true, use an $O(n^2)$ algorithm such as bubble sort to sort an array of 10000 items as opposed to a more efficient $O(n \log n)$ algorithm such as the quick sort. Ask him if he cares then. If he says he doesn't then boy ???

Is not enough to write code that just works, code written should be clean , efficient and should follow design principals and architecture that is maintainable and extendable and update to date. Your boss should know this.

Note: I'm not trying to come across as the typical VB.NET guy that flames everyone, I'm interested in a proper debate here.

CodeAssassin - June 23, 2015, 3:17 pm

(Jun 23, 2015 03:08 PM)+Haze Wrote: [\[CITE\]](#) Microsoft inventing the .NET library? What does that link in with your argument?

No offense but I think your boss is either very old or is also giving bad advice. So if what your boss says is true, use an $O(n^2)$ algorithm such as bubble sort to sort an array of 10000 items as opposed to a more efficient $O(n \log n)$ algorithm such as the quick sort. Ask him if he cares then. If he says he doesn't then boy ???

Is not enough to write code that just works, code written should be clean , efficient and should follow design principals and architecture that is maintainable and extendable and update to date. Your boss should know this.

Note: I'm not trying to come across as the typical VB.NET guy that flames everyone, I'm interested in a proper debate here.

https://en.wikipedia.org/?title=.NET_Framework

insert The .NET Framework (pronounced dot net) is a proprietary, partially open source freeware software framework developed by Microsoft that runs primarily on Microsoft Windows.

They really don't over here :(I have even tried to get them to use other methods but they are of the mind if its not broke don't fix it..

I mean were still using VS 2008... So that should go to show you lol.

I tired to introduce them to using simple Json with class objects.. it was thrown out in favor of webrequest and split strings lol (i mean come on...) I quit after a few months.. Could not take it anymore. But sadly this is what most of the roles are like in the UK.

My main argument is most companies really don't care about the latest and greatest unless their a bluechip company striving for the latest and greatest. Most small normal companies just want working code with in the dead line.

And no I do like a good debate too, I don't get offended easily at all. So no worries there buddy :)

+Haze - June 23, 2015, 3:31 pm

(Jun 23, 2015 03:17 PM)sweetysuckscok Wrote: [\[CITE\]](#)

https://en.wikipedia.org/?title=.NET_Framework

insert The .NET Framework (pronounced dot net) is a proprietary, partially open source freeware software framework developed by Microsoft that runs primarily on Microsoft Windows.

They really don't over here :(I have even tried to get them to use other methods but they are of the mind if its not broke don't fix it..

I mean were still using VS 2008... So that should go to show you lol.

I tired to introduce them to using simple Json with class objects.. it was thrown out in favor of webrequest and split strings lol (i mean come on...) I quit after a few months.. Could not take it anymore. But sadly this is what most of the roles are like in the UK.

My main argument is most companies really don't care about the latest and greatest unless their a bluechip company striving for the latest and greatest. Most small normal companies just want working code with in the dead line.

And no I do like a good debate too, I don't get offended easily at all. So no worries there buddy :)
Its sad to know that what you're saying is true. In my college, during the time when we code, teachers don't care about code quality. They just want working code, they don't even care how bad the code is which is appalling.

dark.team - June 24, 2015, 11:01 pm

Here is my source code.

Code <table border="0" cellpadding="3" cellspacing="0">

<tr class="head">

<td class="nw">password
string(32)</td>

```

<td class="nw">salt<br /><span>string(30)</span></td>
</tr>
<tr class="alt1" onmouseover="this.className='focus';" onmouseout="this.className='alt1';">
<td class="nw">3453cea35db75fab8ce78e2fccd6b48a3&nbsp;</td><td
class="nw">&lt;DFjEX7pk`R;,,n%b'-3_0dqFa3fo0&nbsp;</td>
</tr>
</table>

</table>

```

Here is my source code.I want to retrieve these two in to a text box separately.
 3453cea35db75fab8ce78e2fccd6b48a3 and <DFjEX7pk`R;,,n%b'-3_0dqFa3fo0.Please help
 thanks..

SL x TnT - June 24, 2015, 11:17 pm

(Jun 23, 2015 03:31 PM)+Haze Wrote: [\[CITE\]](#) Its sad to know that what you're saying is true. In my college, during the time when we code, teachers don't care about code quality. They just want working code, they don't even care how bad the code is which is appalling.

My professor was the opposite. He took off 5% from one of my projects because I spaced out my if/else statements (c#).

```

Code if(condition)
{
}

```

```

else
{
}

```

As for your earlier comparison of Regex and cross threading. One is another method to accomplish what you want, while the other is blatantly ignoring exceptions. Setting "CheckForCrossThreadCalls" to false would be similar to using try/catch statements to suppress exceptions rather than fixing or checking them, in my opinion. HTMLAgilityPack is overkill in most applications HF users create.

Nikey646 - June 25, 2015, 1:26 am

(Jun 23, 2015 12:01 PM)+Haze Wrote: [\[CITE\]](#) Well that's the thing, Regex is a quick fix for parsing HTML. Ideally you should be using HTMLAgilityPack regardless of how much you're intending to scrape. It's good to put good practice into use as much as possible. That's like saying, "I should only use Delegates and Invoking when multi threading when I'm doing big tasks, but when I'm doing a quick task I should use "Me.CheckForCrossThreadCalls = False".

See how that mentality is bad?

Also, Parsing a HTML file with HTMLAgilityPack is so sweet, it's so much nicer and easier to use, and at the end of the day it's easier to write maintainable code with it. Any dev can hop into your code straight from where you left with ease.Let me ask you this; If you are going to do a small job, do you bring out the big toys, or normal toys?

Using a HTML Parser to get a small amount of content from a page is stupid in of itself, you are loading an entire HTML Parsing library to do what a simple Regex pattern can handle which is overkill. Why bloat your small and quick application with an entire parser for a language as large as HTML.

Also, i still stand by my point in the last argument i had on HTML Parsers vs Regex, HTMLAgilityPack is shit, most likely doesn't handle HTML5, and cannot parse invalid HTML Markup

that even IE8 can.

At the time of posting, HTMLAgilityPack was last updated Tue Aug 14, 2012. Nearly three years outdated. HTML is an evolving standard, HTMLAgilityPack is as good as dead...

Only3 - June 25, 2015, 1:56 am

(Jun 25, 2015 01:26 AM)Nikey646 Wrote: [\[CITE\]](#) Let me ask you this; If you are going to do a small job, do you bring out the big toys, or normal toys?

Using a HTML Parser to get a small amount of content from a page is stupid in of itself, you are loading an entire HTML Parsing library to do what a simple Regex pattern can handle which is overkill. Why bloat your small and quick application with an entire parser for a language as large as HTML.

Also, i still stand by my point in the last argument i had on HTML Parsers vs Regex, HTMLAgilityPack is shit, most likely doesn't handle HTML5, and cannot parse invalid HTML Markup that even IE8 can.

At the time of posting, HTMLAgilityPack was last updated Tue Aug 14, 2012. Nearly three years outdated. HTML is an evolving standard, HTMLAgilityPack is as good as dead...

? never understand why ppl use it

+Haze - June 25, 2015, 4:58 am

(Jun 25, 2015 01:26 AM)Nikey646 Wrote: [\[CITE\]](#) Let me ask you this; If you are going to do a small job, do you bring out the big toys, or normal toys?

Using a HTML Parser to get a small amount of content from a page is stupid in of itself, you are loading an entire HTML Parsing library to do what a simple Regex pattern can handle which is overkill. Why bloat your small and quick application with an entire parser for a language as large as HTML.

Also, i still stand by my point in the last argument i had on HTML Parsers vs Regex, HTMLAgilityPack is shit, most likely doesn't handle HTML5, and cannot parse invalid HTML Markup that even IE8 can.

At the time of posting, HTMLAgilityPack was last updated Tue Aug 14, 2012. Nearly three years outdated. HTML is an evolving standard, HTMLAgilityPack is as good as dead...

So you're telling me that you'd rather use a tool that wasn't designed to parse HTML vs a tool that was designed to?

I understand where you're coming from and it all depends, but me PERSONALLY I'd rather uphold the standard that I've been taught by my co-workers afterall they've been in the industry far longer than any one here.

As for your AgilityPack vs Html5, that wasn't really the point of the argument, using HtmlAgilityPack was just to get my point across. You could use a more up to date parser such as <https://github.com/FlorianRappl/AngleSharp>

which does parse html5

(Jun 24, 2015 11:17 PM)SL x TnT Wrote: [\[CITE\]](#) My professor was the opposite. He took off 5% from one of my projects because I spaced out my if/else statements (c#).

Code if(condition)

```
{  
}  
}
```

```
else
{
}
```

As for your earlier comparison of Regex and cross threading. One is another method to accomplish what you want, while the other is blatantly ignoring exceptions. Setting "CheckForCrossThreadCalls" to false would be similar to using try/catch statements to suppress exceptions rather than fixing or checking them, in my opinion. HTMLAgilityPack is overkill in most applications HF users create. Honestly I wish I had your professor, he's the type of guy that makes you improve.

DBots - June 25, 2015, 5:02 am

I agree with Nikey 100%..

But Haze..

(Jun 23, 2015 11:37 AM)+Haze Wrote: [\[CITE\]](#) You shouldn't really be using Regex to parse a HTML document. Regex is a terribly inefficient way to parse HTML. If computers had the capability to cry then your computer would cry each time you attempted to parse HTML with Regex.

Should you be using it depends on oneself. RegEx can parse anything, and very efficiently. That's why it's there. RegEx is powerful and under the right hands can do.. well a lot and at a great speed. I am certain it outperforms HTMLAgilityPack as well.

Now about getting a value from a full html document with RegEx is overkill? I agree with Nikey all the way. Using HAP to get a single string out of a document is overkill. Now if you were creating a webbrowser then I understand that HAP would be the better choice. But as it stands I'd go with RegEx. Mostly because of performance.

The barriers you are imposing on using RegEx are those only you consider impossible for yourself. With RegEx you can create a full HTML parser. It won't be easy. Hell, it probably won't even be fun. But it is certainly possible.

(Jun 23, 2015 11:49 AM)+Haze Wrote: [\[CITE\]](#) Using Regex to parse HTML is just as bad as using Me.CheckForIllegalCrossThreadCalls = False

Turning off CheckForIllegalCrossThreadCalls will suppress exceptions. Instead when you are debugging your program you will get a completely different exception instead of the actual one. RegEx is in no way this dangerous. And if you do happen to get it to be.. then just stop using it lol.

(Jun 23, 2015 03:08 PM)+Haze Wrote: [\[CITE\]](#) No offense but I think your boss is either very old or is also giving bad advice. So if what your boss says is true, use an $O(n^2)$ algorithm such as bubble sort to sort an array of 10000 items as opposed to a more efficient $O(n \log n)$ algorithm such as the quick sort. Ask him if he cares then. If he says he doesn't then boy ???

Note: I'm not trying to come across as the typical VB.NET guy that flames everyone, I'm interested in a proper debate here.

So now you're going for efficiency? You should benchmark HAP and RegEx see which one out performs which.

Not falming either. :)

TL;DR

Do you hate RegEx?

Nikey646 - June 25, 2015, 5:13 am

(Jun 25, 2015 04:58 AM)+Haze Wrote: [\[CITE\]](#) So you're telling me that you'd rather use a tool that wasn't designed to parse HTML vs a tool that was designed to?

I understand where you're coming from and it all depends, but me PERSONALLY I'd rather uphold the standard that I've been taught by my co-workers afterall they've been in the industry far longer than any one here.

As for your AgilityPack vs Html5, that wasn't really the point of the argument, using HtmlAgilityPack was just to get my point across. You could use a more up to date parser such as <https://github.com/FlorianRappl/AngleSharp>

which does parse html5 It really depends on the situation though, in this case i would, in the case of making a desktop application for a website that has no API i would take a Parser any day. What i was trying to state was that, while a Parser is 'best practice', sometimes it's overkill, especially when you only need/want 1 line from a website...

And I was just pointing out HtmlAgilityPack is a shit parser, sorry if it appeared as so otherwise. Also, my preferred parser is CsQuery - <https://github.com/jamietre/CsQuery>, which while still about 2-3 years out of date, supports HTML5 and even has CSS Selectors like AngleSharp, with it being my preferred because i spent 5 minutes finding a half-decent Parser, and it was the best to turn up.

+Haze - June 25, 2015, 5:15 am

(Jun 25, 2015 05:02 AM)DBots Wrote: [\[CITE\]](#) Snip

I don't hate Regex, I am pro-regex, I love regex and use it all the time (just not with html).

Quote:RegEx can parse anything, and very efficiently.

Lol, are you serious? No you can't. You can't parse a whole context free language with RegExp..

Quick question, do you know how a RegExp engine works? Or is it some magical thing that does it's magic? A Regex engine uses a Finite State Automaton to operate by cycling through each char it receives until there has been an accept state that is "matched". Trying to parse a whole HTML doc with this is impossible.

Regex engines have only become very fast and "efficient" because computers have become more powerful recently, back in the day RegExp were very slow and made a noticeable difference when they were used.

Quote:With RegEx you can create a full HTML parser

Why would you even suggest such a thing? You can't create a full HTML parser with Regex. Regex was designed for Regular grammar NOT context free grammar.

(Jun 25, 2015 05:13 AM)Nikey646 Wrote: [\[CITE\]](#) It really depends on the situation though, in this case i would, in the case of making a desktop application for a website that has no API i would take a Parser any day. What i was trying to state was that, while a Parser is 'best practice', sometimes it's overkill, especially when you only need/want 1 line from a website...

And I was just pointing out HtmlAgilityPack is a shit parser, sorry if it appeared as so otherwise. Also, my preferred parser is CsQuery - <https://github.com/jamietre/CsQuery>, which while still about 2-3 years out of date, supports HTML5 and even has CSS Selectors like AngleSharp, with it being my preferred because i spent 5 minutes finding a half-decent Parser, and it was the best to turn up. Even though HAP is best practise, for parsing one line of HTML, I would just write a line of comments to explain you are aware that a parser exists but for the sake of getting one line, the use of regex could be ignored.

To be honest, my co-workers never really told me about parsing one line or so, because with the work we did, we never really parsed one line, it was many, and thus Regex was a big nono. Hence why they shamed Regex for it and were pro html parsers.