Best way to fetch a varying HTML tag

Asked 16 years, 3 months ago Modified 10 years ago Viewed 547 times

5

I'm trying to fetch some HTML from various blogs and have noticed that different providers use the same tag in different ways.

For example, here are two major providers that use the meta name generator tag differently:

- Blogger: `<meta content='blogger' name='generator'/>` (content first, name later and, yes, single quotes!)
- WordPress: `<meta name="generator" content="WordPress.com" />` (name first, content later)

Is there a way to extract the value of content for all cases (single/double quotes, first/last in the row)?

P.S. Although I'm using Java, the answer would probably help more people if it where for regular expressions generally.

html regex language-agnostic

Share

Improve this question

Follow

edited Dec 13, 2014 at 8:17

J0e3gan

8,928 ● 10 ● 56 ● 81

asked Aug 28, 2008 at 2:23

pek

18k ● 28 ● 88 ● 99

8 Answers

Sorted by: Highest score (default)

The answer is: **don't use regular expressions**.

Seriously. Use a SGML parser, or an XML parser if you happen to know it's valid XML (probably almost never true). You will absolutely screw up and waste tons of time trying to get it right. Just use what's already available.

Share Improve this answer

Follow

answered Aug 28, 2008 at 2:31

Brad Wilson

70.4k ● 9 ● 77 ● 85

Actually, you should probably use some sort of HTML parser where you can inspect each node (and therefore node attributes) in the DOM of the page. I've not used any of these for a while so I don't know the pros and cons but here's a list http://java-source.net/open-source/html-parsers

Share Improve this answer

Follow

answered Aug 28, 2008 at 2:30

martinatime
2,487 ● 1 ● 17 ● 23

2

Those differences are not really important according to the XHTML standard.

In other words, they are exactly the same thing.

Also, if you replace double quotes with single quotes would be the same.

The typical way of 'normalizing' an xml document is to pare it using some API that treats the document as its Infoset representation. Both DOM and SAX style APIs work that way.

If you want to parse them by hand (or with a RegEx) you have to replicate all those things in your code and, in my opinion, that's not practical.

Share Improve this answer Follow

edited Aug 28, 2008 at 2:34

answered Aug 28, 2008 at 2:28

Sergio Acosta
11.4k ● 12 ● 63 ● 91

2

Note: single quotes (even no quotes, if the value doesn't contain a space) is valid according to the W3C HTML spec. Quote:

> By default, SGML requires that all attribute values be delimited using either double quotation marks (ASCII decimal 34) or single quotation marks (ASCII decimal 39)... In certain cases, authors may specify the value of an attribute without any quotation marks.

Also, don't forget that the order of attributes can be reversed and that other attributes can appear in the tag.

Share Improve this answer

Follow

answered Aug 28, 2008 at 2:56

Grey Panther

13.1k ● 7 ● 49 ● 67

1

You may want to give Java's HTMLEditorKit a shot. It is easy to experiment with to see if the parsing provides what you are looking for.

Share Improve this answer

Follow

answered Aug 28, 2008 at 3:24

Preston

11 ● 2

0

Ok, since you are looking for language-agnostic then you can try a REGEX like `/<meta\s.*content=.*>/` and take the result from that and parse out the specific values that you are looking for. I'm by no means a REGEX expert so there is probably a better way but in using the tool at

http://www.codehouse.com/webmaster_tools/regex/ I matched both of the strings you provided.

Share Improve this answer

Follow

answered Aug 28, 2008 at 3:20

martinatime

2,487 ● 1 ● 17 ● 23

0

If you must use regex, here is a regex to get just the content part:

```
content\s*=\s*['"].*?['"]
```

returns

```
content = "blogger"
```

and

```
content='Worpress.com'
```

respectively. I'm no regex expert, but it gets those when given your examples in regexpal.

Once you get that you can get everything between the quotes however you choose, be it another regex (which is just immoral at that point) or just looping over the characters.

Share Improve this answer

answered Aug 28, 2008 at 3:38

Follow

dwestbrook
5,268 ● 3 ● 26 ● 20

0

If your using java you may want to look at tagsoup, which is a SAX-compliant parser for "[parsing] HTML as it is found in the wild".

Share Improve this answer

Follow

answered Aug 28, 2008 at 12:53

Peter Stuifzand
5,084 ● 1 ● 25 ● 28