

Re: [PHP] generating an html intro text ...

Source: <http://coding.derkeiler.com/Archive/PHP/php.general/2007-06/msg01332.html>

- *From:* jochem@xxxxxxxxxxxxxx (Jochem Maas)
 - *Date:* Mon, 18 Jun 2007 14:18:15 +0200
-

tedd wrote:

At 11:39 AM +0200 6/14/07, Jochem Maas wrote:

original string:

....

The problem as I see it is covering all the possibilities that may occur even if the text is well formed. Like what if someone introduces a span that sets a color for a paragraph, such as:

```
<span color:"yellow"; >Dolore magna aliquam erat volutpat ut wisi enim  
ad minim veniam quis nostrud. Consectetur adipiscing elit sed diam  
nonummy nibh euismod tincidunt ut laoreet exerci tation ullamcorper  
suscipit lobortis! <b>Decima eodem modo </b>typi qui nunc nobis videntur  
parum clari fiant sollemnes in.</span>
```

And the `` tag as well as the `` tag is outside the 256 limit?

You would have to search out and pull in all closing tags.

So, I guess an algorithm could be:

roughly speaking yes this is what is would do, except:

First, grab 256 characters -- The string. If The string is shorter, then quit.

the algo should only be counting 'content characters', i.e. anything that is html markup should not go towards the string length count, additionally html entities such as '&';' should be considered as a single character.

Second, determine what tags are not closed.

Third, create closing tags and add them to the end of The string (in proper order).

Fourth, then remove the same number of non-html characters from the end of The string.

what the code should do (mmore or less) is quite clear – writing something flexible & robust to actually do it (and do it fast) is quite another matter.

I have been looking at Edward Vermillon's code but I suspect that what he sent me is not quite what I'm looking for for a number of reasons:

1. it deals primarily with custom bbcode like markup
2. I have a couple of doubts about the handling of html entities
3. performance

that said I still have to look at it in depth before making any real conclusions as to it's viability (and or the possiblity to rework the code to fit my needs).

I'm also looking at an alternative where by I go through the string and truncate it at the character (or characters that represent an html entity) that reresents the Nth 'content character' and then feeding the truncated string to the Tidy extension and let it figure out the html cleaning part ... does anyone have experience using tidy to clean (make valid) html snippets using Tidy, that they would like to share?

OR, just strip out the html tags (strip_tags) and go with straight text -- a lot easier.

that's not an option for me.

Cheers,

tedd

.