عنوان: **حذف تگ‌های زاید دریافتی از متون MS-Word**
نویسنده: وحید نصیری
تاریخ: ۱۳۸۷/۰۸/۱۱ ۱۷:۳۷:۲۰
آدرس: www.dotnettips.info
برچسب‌ها: Regular expressions

یکی از مشکلاتی که من همیشه با کاربران عادی دارم بحث انتقال مطالب از Word مایکروسافت به ادیتورهای WYSWING تحت وب است. برای مثال شما سایت پویایی را درست کرده‌اید که کاربران می‌توانند مطالب آنرا ویرایش یا کم و زیاد کنند.
اگر مطلب از ابتدا در این نوع ادیتورها تایپ و آماده شود هیچ مشکلی وجود نخواهد داشت چون خروجی اکثر آنها استاندارد است، اما متاسفانه خروجی وب word بسیار مشکل‌زا است (copy/paste معمولی مطالب آن در یک ادیتور تحت وب) و خصوصا برای نمایش تایپ فارسی در وب اصلا مناسب نیست. یعنی هیچ الزامی وجود ندارد که اندازه فونت‌ها در متن نهایی نمایش داده شده در وب یکسان باشند یا خطوط در هم فرو نروند و یا عدم تناسب اندازه قلم متن صفحه با قلم استفاده شده در CSS سایت (که شکل ناهماهنگ و غیرحرفه‌ای را حاصل خواهد کرد) و امثال آن. اینجاست که کار شما زیر سؤال می‌رود! "این برنامه درست کار نمیکنه! متن من به‌هم ریخته شده و امثال این"
این کاربر عادی عموما یک تایپیست است یا یک منشی که به او گفته شده است شما از امروز موظفید مطالبی را در این سایت قرار دهید. بنابراین این کاربر حتما از word استفاده خواهد کرد (برای پیش نویس مطالب). همچنین عموما هم مرورگر "سازمانی" مورد استفاده، هنوز که هنوز است همان IE6 است (در اکثر شرکت‌ها و خصوصا ادارات) و مهم نیست که الان آخرین نگارش IE یا فایرفاکس و تمام هیاهوهای مربوطه به کجا ختم شده‌اند. حتما باید سایت با IE6 هم سازگار باشد. بنابراین از برنامه IE tester غافل نشوید.
و دست آخر شما هم نمی‌توانید به کاربر عادی ثابت کنید که این خروجی وب word اصلا استاندارد نیست (حتما کار شما است که مشکل دارد نه شرکت معظم مایکروسافت!). یا اینکه به آنها بگوئید اصلا مجاز نیستید در وب همانند یک فایل word از چندین نوع قلم مختلف فارسی غیراستاندارد استفاده کنید چون ممکن است کاربری این نوع قلم مورد استفاده شما را نداشته باشد و نمایش نهایی به هم ریخته‌تر از آنی خواهد بود که شما فکرش را می‌کنید! یا اینکه با استفاده از این روش حجم نهایی صفحه حداقل 50 کیلو بایت بیشتر خواهد شد (بدلیل حجم بالای تگ‌های زاید word) و نباید کاربران دایال آپ را فراموش کرد.
مدتی در اینباره جستجو کردم و نتیجه حاصل این بود که تمامی روش‌ها به یک مورد ختم می‌شود: حذف تگ‌های غیراستاندارد word هنگام دریافت مطلب و پیش از ذخیره سازی آن در دیتابیس
یک سری از ادیتورهای متنی تحت وب مانند FCK editor این قابلیت را به صورت خودکار اضافه کرده‌اند و حتی اگر کاربر متنی را از word در آنها Paste کند پیغامی را در همین رابطه دریافت خواهد کرد (شکل زیر) و البته کاربر می‌تواند گزینه لغو یا خیر را نیز انتخاب کند و دوباره همان وضعیت قبل تکرار خواهد شد. (یا حتی دکمه مخصوص کپی از word را هم به نوار ابزار خود اضافه کرده‌اند)

برای این منظور تابع زیر تهیه شده‌است که من همواره از آن استفاده می‌کنم و تا به امروز مشکل پاسخ پس دادن به کاربران عادی را به این صورت حل کرده‌ام!
این تابع تمامی تگ‌های اضافی و غیراستاندارد word متن دریافتی از یک ادیتور WYSWING را حذف می‌کند و به این صورت متن نهایی نمایش داده شده در سایت، تابع CSS مورد استفاده در سایت خواهد شد و نه حجم بالایی از تگ‌های غیراستاندارد word. (ممکن است کاربر در ابتدا کمی جا بخورد ولی مهم نیست! سایت باید استاندارد نمایشی خودش را از CSS آن دریافت کند و نه از تگ‌های word)

```
using System.Text.RegularExpressions;
 /// <summary>
 /// Removes all FONT and SPAN tags, and all Class and Style attributes.
 /// Designed to get rid of non-standard Microsoft Word HTML tags.
 /// </summary>
 public static string CleanMSWordHtml(string html)
 {
     try
     {
         // start by completely removing all unwanted tags
```

```
        html = Regex.Replace(html, @"<[/]?(font|span|xml|del|ins|[ovwxp]:\w )[^>]*?>", "",
RegexOptions.IgnoreCase);
        // then run another pass over the html (twice), removing unwanted attributes
        html = Regex.Replace(html, @"<([^>]*)(?:class|lang|style|size|face|[ovwxp]:\w
)=(?:'[^']*'|""[^""]*""|[^\s>] )([^>]*)>", "<$1$2>", RegexOptions.IgnoreCase);
        html = Regex.Replace(html, @"<([^>]*)(?:class|lang|style|size|face|[ovwxp]:\w
)=(?:'[^']*'|""[^""]*""|[^\s>] )([^>]*)>", "<$1$2>", RegexOptions.IgnoreCase);
        return RemoveHTMLComments(html);
    }
    catch
    {
        return html;
    }
}

public static string RemoveHTMLComments(string html)
{
    try
    {
        Regex _Regex = new Regex("((<!-- )((?!<!-- ).)*( -->))(\r\n)*", RegexOptions.Singleline);
        return _Regex.Replace(html, string.Empty);
    }
    catch
    {
        return html;
    }
}
```

متد RemoveHTMLComments را عمدا جدا قرار دادم تا مشخص‌تر باشد. پس از تمیزکاری اولیه، ممکن است دسته‌گل‌های تیم مایکروسافت به صورت کامنت باقی بمانند که باید آنها را هم تمیز کرد! :)

نظرات خوانندگان

نویسنده: Shaho
تاریخ: ۰۲:۲۶:۰۰ ۱۳۸۷/۰۸/۱۷

اقا خیلی مرسی!

نویسنده: سیدمحمدرضا فخری
تاریخ: ۱۱:۲۰:۲۰ ۱۳۸۸/۰۲/۳۱

سلام. خیلی ممنون از این کد بسیار مفید. اما یک مطلب و آن اینکه عبارت class=MsoNormal با این تابع حذف نمیشه. ممنون میشم بفرمائید چه تغییری در کد بدهیم.

نویسنده: وحید نصیری
تاریخ: ۱۴:۲۴:۱۹ ۱۳۸۸/۰۲/۳۱

class ها و lang|style|size|face|[ovwxp باید حذف بشه. ولی اگر روش فوق راضی کننده نبود از روش مقاله زیر هم می‌توان استفاده کرد:

http://www.codinghorror.com/blog/archives/000485.html

یک روش دیگر هم این است که کلا هرچی تگ html است را یکجا حذف کرد. روش کار به صورت زیر است:

http://gibbons.co.za/archive/2005/01/28/249.aspx

نویسنده: سیدمحمدرضا فخری
تاریخ: ۱۱:۴۵:۰۹ ۱۳۸۸/۰۳/۰۲

سلام. با تشکر از شما، برنامه مربوط به کدینگ هارور رو قبلا تست کردم، مشکل داشت. کدی شکه شما زحمت کشیدید هم غیر از مسئله ذکر شده درکامنت اول، خوب کار میکرد فقط مشکل اینست که اختصاصی به فرمت های ورد ندارد و همه استایل ها را پاک میکند که برای ما بعنوان یک بلاگ سرویس قابل استفاده نیست، چون کاربران پس از کپی پیست، استایل های خودشان را هم اضافه میکنند و این کد همه را پاک میکند. اگر برنامه بود که فقط بتواند اضافات ورد را پاک کند خوب بود. و اینکه بصورت جاوا اسکریپت باشد تا بتوان درکلاینت هم از آن استفاده کرد(من متاسفانه رگولار اکسپرشن را عمیق کار نکرده ام، همین رگولاررا میتوان در جاوا هم بکار برد؟).

نویسنده: وحید نصیری
تاریخ: ۱۲:۴۶:۳۷ ۱۳۸۸/۰۳/۰۲

سلام،
بله در حالت جاوا اسکریپتی توسط FCK-Editor هم کار شده که می‌شود از آن ایده گرفت:

http://dev.fckeditor.net/browser/FCKeditor/trunk/editor/dialog/fck_paste.html

به تابع CleanWord آن در صفحه فوق مراجعه نمائید.

نویسنده: سیدمحمدرضا فخری
تاریخ: ۱۴:۱۹:۴۸ ۱۳۸۸/۰۳/۰۲

ممنون