TYPO3 Core - Bug #28567

Epic # 65814 (Closed): Make Indexed search extbase plugin shine

Ugly replacement character when removing whitespaces

2011-07-29 13:51 - Dimitri Koenig

Status:	Closed	**Start date:**	2011-07-29
Priority:	Should have	**Due date:**	
Assignee:	Tymoteusz Motylewski	**% Done:**	0%
Category:	Indexed Search	**Estimated time:**	0.00 hour
Target version:	7.5		
TYPO3 Version:	4.5	**Complexity:**	
PHP Version:		**Is Regression:**	No
Tags:		**Sprint Focus:**	

Description

tx_indexedsearch.php->markupSWpartsOfString($str) removes unnecessary whitespaces at the beginning:

```
$str = preg_replace('/\s\s+/',' ',$str);
```

But sometimes this produces ugly replacement characters (U+FFFD/65533).

Any solution?

History

#1 - 2012-04-13 12:06 - Tim Schenk

- Assignee set to Dimitri Koenig

Dimitri Koenig wrote:

> tx_indexedsearch.php->markupSWpartsOfString($str) removes unnecessary whitespaces at the beginning:
> [...]
>
> But sometimes this produces ugly replacement characters (U+FFFD/65533).
>
> Any solution?

i changed this to UTF-8 within a XClass-Extension

$str = str_replace(' ',' ',t3lib_parsehtml::bidir_htmlspecialchars($str,-1));
$str = preg_replace('/[\s][\s]+/u',' ',$str);

and ...

$parts = preg_split('/'.$regExString.'/iu', ' '.$str.' ', 20000, PREG_SPLIT_DELIM_CAPTURE);

what works for me. Might be a good idea to change this charset specific...

But there where a couple other issues with indexing or output of UTF-8 especially in chinese and ' ' in general, e.g. cropping of chinese search result outputs strange (question mark) characters. the solution was allowing cropping only at whitespaces, not within words...

I changed in t3lib_cs::entities_to_utf8

some lines to UTF-8:

$param2 = (is_numeric(ENT_HTML5))?ENT_HTML5:ENT_QUOTES;
$entities = get_html_translation_table(HTML_ENTITIES,$param2,'UTF-8');

and in function t3lib_cs::crop:

```
@if ($i === FALSE) { // $len outside actual string length
```

```
            return $string;
        } else {
            if ($len > 0) {
                if (strlen($string{$i})) {
                    $string = substr($string, 0, $i);
                    $lastWhiteSpace = strrpos($string, " ");
                    //t3lib_utility_Debug::debugInPopUpWindow(array("last"=>$lastWhiteSpace,"length"=>strlen($
string),"string"=>$string));
                    if($lastWhiteSpace){
                        $string = substr($string, 0, $lastWhiteSpace);
                    }
                    return $string . $crop;;

}
            } else {
                if (strlen($string{$i - 1})) {
                    $string =  substr($string, $i);
                    $firstWhiteSpace = strpos($string, " ");
                    //t3lib_utility_Debug::debugInPopUpWindow(array("first"=>$firstWhiteSpace,"length"=>strlen
($string),"string"=>$string));
                    if($firstWhiteSpace){
                        $string = substr($string, $firstWhiteSpace);
                    }
                    return $crop . $string;
                }
            }@
```

#2 - 2013-01-10 22:59 - Marc Véron

I can confirm problems with the line
$str = preg_replace('/\s\s+/',' ',$str);

Searching for text that contains the letter à between white space displays nasty results as follows:

demandé **?** être représentée directement
instead of
demandé à être représentée directement

It only happens with white space aroud (or after) à. If the letter à is embedded in a word, it always displays correctly.

I found out that the problem occurs in line 2022 of class.tx_indexedsearch.php
$str = preg_replace('/\s\s+/',' ',$str);

There is a hint on http://stackoverflow.com/questions/2050723 that PHP's regular expressions are not Unicode-aware.

For a quick hack, I added two lines to separate à from white space:
$str = str_replace('à','|à|',$str); //Hack MV
$str = preg_replace('/\s\s+/',' ',$str);
$str = str_replace('|à|','à',$str); //Hack MV

Results display now as expected, but it would be nice to have it fixed without this hack.

#3 - 2013-03-27 14:16 - Oliver Hader

- Target version set to 2222

#4 - 2013-03-27 14:17 - Oliver Hader

- Project changed from 1382 to TYPO3 Core

#5 - 2013-03-27 14:19 - Oliver Hader

- Category set to Indexed Search

#6 - 2013-03-27 14:19 - Oliver Hader

- Target version deleted (2222)

#7 - 2015-01-23 20:54 - Mathias Schreiber

- Target version set to 7.5

- TYPO3 Version set to 4.5

- Is Regression set to No

#8 - 2015-04-18 20:53 - Tymoteusz Motylewski

- Parent task set to #65814

#9 - 2015-09-15 13:31 - Tymoteusz Motylewski

- Assignee changed from Dimitri Koenig to Tymoteusz Motylewski

#10 - 2015-09-15 20:14 - Tymoteusz Motylewski

- Status changed from New to Closed

I have teste the issue with all kinds of utf characters (from http://www.cl.cam.ac.uk/~mgk25/ucs/examples/UTF-8-demo.txt as well as the one mentioned in the ticket) and couldn't reproduce.

It might be the case, that this bug was fixed with PHP upgrade to 5.4. In 5.4 they changed the default PHP charset to utf-8.
Also functions like htmlentities, html_entity_decode, htmlspecialchars etc are now utf-8 aware by default.

5.4 has just reached end of life, so we can safely close this ticket.

Please let me know if the issue is not solved for you with modern PHP version.