4-bytestring (most Emoji) Causes Wiki Cutoffs | Tiki Wiki CMS Groupware :: Development
4-bytestring (most Emoji) Causes Wiki Cutoffs

Status
Closed

Subject
4-bytestring (most Emoji) Causes Wiki Cutoffs

Version
17.x

Category
- Error

Feature
Wiki Syntax (text area, parser, external wiki, etc)

Resolution status
New

Submitted by
drsassafras

Lastmod by
drsassafras, Marc Laporte

Rating
(0)

Related-to
- Changing (modernizing) Tiki smileys (we should support Emoji)
- Replace Emoticons with Emoji
- Emojis crash Tiki (error 500)
- change the table definitions to support the new utf8mb4 character

Description

I have found that most emoji cause page cutoffs on wiki pages. It seems to be that all 4 bytestring unicode causes this most unfortunate error (including the smiley face)

3 bytestring unicode seems to be fine, including: ☺ ☹

I created a pastebin http://pastebin.com/DVf7S1AK with a few common emoji that will promptly kill your wiki page while saving an edit. These are easily usable by just about every mobile phone made since 2010 when unicode 6.0 came out. Desktop usage is also becoming more popular.

My first thought was that it was a HTML Purifier issue. I visited the HTML purifier website and there demo handles 4 bytestring unicode just fine. Then I thought it might be that some software of ours was using the Legacy CJK encoding, but 3 bytestring would not work if that was the case, so we are probably fine there.

That is as far as i have gotten. It wold be really nice if tiki could handle this now popular form of communication.

Emoji Unicode Implementation Info: https://gist.github.com/mranney/1707371
Unicode Lookup: http://unicode.scarfboy.com/

Solution
implemented in tiki 19

Priority
25

Demonstrate Bug (Tiki 19+)
Please demonstrate your bug on show2.tiki.org

Version: trunk ▼

Demonstrate Bug (older Tiki versions)

Ticket ID
6189

Created
Thursday 01 December, 2016 00:32:17 GMT-0000
by drsassafras

LastModif
Wednesday 17 October, 2018 01:14:45 GMT-0000

Comments

drsassafras 01 Dec 16 01:06 GMT-0000

I just noticed that the email notice of a page edit correctly displays the 4-bytestring unicode, so the issue lies somewhere between sending the email and displaying the page.

Jonny Bradley 01 Dec 16 17:25 GMT-0000

Hmm, confirmed - this really rings a bell with a mysql problem ages ago where wiki pages would be truncated at certain characters, i think some Greek ones afaicr, and i think there wasn't much we could do about it as it was a database issue... but i can't find any reference to it now, wonder if anyone else remembers?

drsassafras 02 Dec 16 05:08 GMT-0000

Good Catch. Thats it. Ive confirmed it. MySQL uf8 only supports 3 bytestring encoding.

The mysql uf8mb4 needs to be used to have full uf8 support. Now that every cell phone has a "kill tiki page keyboard" handy we should certanly address this.

uf8mb4 support was added in mysql 5.5.3. Right now we only specify that a mysql 5 database is required.

Care might also be taken in converting of fixed length char types. If they are specified as uf8mb4 then 4 bytes for each character need to be reserved. Varchar is of course not effected by this. I see though that we are not terribly good at choosing the most optimal type of encoding for data storage in tiki now. there is a lot of fixed length char values that would be just fine with ascii encoding (reserves 1 byte per

character) but is specified as uf8 (reserving 3 byes per character) ucs2 is also a good unicode format if only a couple thousand characters are needed. It of course reserves 2 characters... ok I think I'm off on a tangent here.

Marc Laporte 24 Apr 17 03:23 GMT-0000

Started utf8mb4

Attachments

filename	created	hits	comment	version	filetype

No attachments for this item

The original document is available at
https://dev.tiki.org/item6189-4-bytestring-most-Emoji-Causes-Wiki-Cutoffs