

Ruby - Bug #15343

String#each_grapheme_cluster wrongly splits some emoji (genie, zombie, wrestling)

11/26/2018 09:02 AM - duerst (Martin Dürst)

Status:	Closed	
Priority:	Normal	
Assignee:	duerst (Martin Dürst)	
Target version:	2.6	
ruby -v:	ruby 2.6.0dev (2018-11-26 trunk 65989) [x86_64-cygwin]	Backport: 2.4: UNKNOWN, 2.5: REQUIRED
Description All the codepoint combinations that turn up in the various emoji files provided by Unicode (currently we use those at https://www.unicode.org/Public/emoji/5.0/) are recognized as grapheme clusters by String#each_grapheme_cluster, except those relating to genies, zombies, and wrestling (THIS IS NOT A JOKE!). Taking an example from https://www.unicode.org/Public/emoji/5.0/emoji-zwj-sequences.txt, line 396: <pre>\$./ruby -e '"\u{1F9DE 200D 2640 FE0F}"'.each_grapheme_cluster.to_a.length.display'</pre> <pre>2</pre> The correct result is 1, not 2. The sequence of codepoints represents a woman genie. I will commit the file test/ruby/enc/test_emoji_breaks.rb, which excludes genie, zombie, and wrestling emoji to make sure the tests pass. I would like to make sure that this is correct for Unicode 10.0.0 before moving to Unicode 11.0.0. I will try to find out how to fix this by myself, but would definitely appreciate help.		
Related issues: Blocks Ruby - Feature #15182: Update extended grapheme cluster implementation... Closed		

History

#1 - 11/26/2018 09:03 AM - duerst (Martin Dürst)

- Blocks Feature #15182: Update extended grapheme cluster implementation for Unicode 11 added

#2 - 11/26/2018 08:12 PM - shevegen (Robert A. Heiler)

This issue is epic due to its title alone! (I don't quite know whether there are indeed genie and zombie emojis yet but it makes me curious.)

except those relating to genies, zombies, and wrestling (THIS IS NOT A JOKE!).

Awww :)

#3 - 11/29/2018 04:12 AM - duerst (Martin Dürst)

Some data points from a discussion between @naruse (Yui NARUSE) and myself:

- Up to elf (U+1F9DD) is Emoji_Modifier_Base, but genie (U+1F9DE) isn't.
- Emoji_Modifier only includes skin tones (U+1F3FB-1F3FF, light skin tone..dark skin tone)
- For experts, that seems to make sense, because there are apparently light and dark elves, but all the zombies have the same half-dead skin color.
- For 'wrestling' again, it doesn't allow skin colors.
- So the error seems to appear when an emoji takes male/female specifiers, but isn't allowed to take skin tones.
- As we are going to rewrite the underlying implementation (function node_extended_grapheme_cluster in reparsed.c), we may not care to fix this bug anymore. But if somebody finds a fix, they may want to apply it to older versions of Ruby (2.5 and 2.4).

#4 - 11/30/2018 05:15 AM - duerst (Martin Dürst)

- File debug_X_genie.txt added
- File debug_X_elf.txt added

I had my computer spend about 10h to compile Ruby with regexp debug flags activated. It took that long because while Ruby is building, it starts running Ruby scripts with lots of regexp debug output. (I probably should have deactivated document building and used 2>/dev/null for a bit of speedup.)

Then I was able to try out the above example, attached as debug_X_genie.txt (the exact command was: `./ruby --disable-gems -e "\u{1F9DE 200D 2640 FE0F}" =~ /X/ 2>debug_X_genie.txt`).

I also did the same for the 'elf' emoji: `./ruby --disable-gems -e "\u{1F9DD 200D 2640 FE0F}" =~ /X/ 2>debug_X_elf.txt`. File also attached.

The files only differ at the end, when the actual match happens.

#5 - 12/02/2018 10:27 AM - duerst (Martin Dürst)

- Status changed from Open to Closed
- Assignee changed from naruse (Yui NARUSE) to duerst (Martin Dürst)
- Backport changed from 2.4: UNKNOWN, 2.5: UNKNOWN to 2.4: UNKNOWN, 2.5: REQUIRED

Working through Unicode Standard Annex #29 (version 31, for Unicode 10.0.0), I'm not sure all of the code in `node_extended_grapheme_cluster()` (in `regparse.c`) is perfect. But this solves an obvious bug, and we'll leave it at that for Unicode 10.0.0.

Files

debug_X_genie.txt	30.2 KB	11/30/2018	duerst (Martin Dürst)
debug_X_elf.txt	29.9 KB	11/30/2018	duerst (Martin Dürst)