Ruby trunk - 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:

```
$ ./ruby -e '"\u{1F9DE 200D 2640 FE0F}".each_grapheme_cluster.to_a.length.display'
2
```

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 trunk - Feature #15182: Update extended grapheme cluster implemen...	**Closed**

Associated revisions

Revision 0409290e - 11/26/2018 09:03 AM - duerst (Martin Dürst)

add tests for grapheme clusters using Unicode Emoji test data

Add file test/ruby/enc/test_emoji_breaks.rb to test String#each_grapheme_cluster
test data provided by Unicode (at https://www.unicode.org/Public/emoji/#{EMOJI_VERSION}/).

Lines containing emoji for genies, zombies, and wrestling are ignored
because there seems to be a bug (#15343) in the implementation.

git-svn-id: svn+ssh://ci.ruby-lang.org/ruby/trunk@65990 b2dd03c8-39d4-4d8f-98ff-823fe69b080e

Revision 65990 - 11/26/2018 09:03 AM - duerst (Martin Dürst)

add tests for grapheme clusters using Unicode Emoji test data

Add file test/ruby/enc/test_emoji_breaks.rb to test String#each_grapheme_cluster
test data provided by Unicode (at https://www.unicode.org/Public/emoji/#{EMOJI_VERSION}/).

Lines containing emoji for genies, zombies, and wrestling are ignored
because there seems to be a bug (#15343) in the implementation.

Revision 65990 - 11/26/2018 09:03 AM - duerst (Martin Dürst)

add tests for grapheme clusters using Unicode Emoji test data

Add file test/ruby/enc/test_emoji_breaks.rb to test String#each_grapheme_cluster
test data provided by Unicode (at https://www.unicode.org/Public/emoji/#{EMOJI_VERSION}/).

Lines containing emoji for genies, zombies, and wrestling are ignored
because there seems to be a bug (#15343) in the implementation.

Revision a96a594f - 12/02/2018 10:07 AM - duerst (Martin Dürst)

solve the genie/zombie/wrestlers bug

enc/unicode.c: - Add U+1F93C (WRESTLERS), U+1F9DE (GENIE), and U+1F9DF
to onigenc_unicode_GCB_ranges_E_Base.
- Add comments with character names.
test/ruby/enc/test_emoji_breaks.rb: Activate tests for genie/zombie/wrestlers.
This closes issue #15343.

git-svn-id: svn+ssh://ci.ruby-lang.org/ruby/trunk@66133 b2dd03c8-39d4-4d8f-98ff-823fe69b080e

Revision 66133 - 12/02/2018 10:07 AM - duerst (Martin Dürst)

solve the genie/zombie/wrestlers bug

enc/unicode.c: - Add U+1F93C (WRESTLERS), U+1F9DE (GENIE), and U+1F9DF
to onigenc_unicode_GCB_ranges_E_Base.
- Add comments with character names.
test/ruby/enc/test_emoji_breaks.rb: Activate tests for genie/zombie/wrestlers.
This closes issue #15343.

Revision 66133 - 12/02/2018 10:07 AM - duerst (Martin Dürst)

solve the genie/zombie/wrestlers bug

enc/unicode.c: - Add U+1F93C (WRESTLERS), U+1F9DE (GENIE), and U+1F9DF
to onigenc_unicode_GCB_ranges_E_Base.
- Add comments with character names.
test/ruby/enc/test_emoji_breaks.rb: Activate tests for genie/zombie/wrestlers.
This closes issue #15343.

Revision f43a2a5a - 12/02/2018 09:41 PM - duerst (Martin Dürst)

make sure all nodes are freed on error in node_extended_grapheme_cluster()

regparse.c: In function node_extended_grapheme_cluster(), introduce function-global
array node_array and use it for sequence and alternate construction. This is done
so that in case of error, all nodes that have already been constructed can be
correctly freed. (issue #15343)

git-svn-id: svn+ssh://ci.ruby-lang.org/ruby/trunk@66135 b2dd03c8-39d4-4d8f-98ff-823fe69b080e

Revision 66135 - 12/02/2018 09:41 PM - duerst (Martin Dürst)

make sure all nodes are freed on error in node_extended_grapheme_cluster()

regparse.c: In function node_extended_grapheme_cluster(), introduce function-global
array node_array and use it for sequence and alternate construction. This is done
so that in case of error, all nodes that have already been constructed can be
correctly freed. (issue #15343)

Revision 66135 - 12/02/2018 09:41 PM - duerst (Martin Dürst)

make sure all nodes are freed on error in node_extended_grapheme_cluster()

regparse.c: In function node_extended_grapheme_cluster(), introduce function-global
array node_array and use it for sequence and alternate construction. This is done
so that in case of error, all nodes that have already been constructed can be
correctly freed. (issue #15343)

Revision b56e266d - 12/02/2018 11:28 PM - duerst (Martin Dürst)

remove unnecessary settings with NULL_NODE in \X implementation

Remove unnecessary settings of node_array elements to NULL_NODE.
We can do this because we initialize the whole array to NULL_NODEs
and set everything again to NULL_NODEs when creating a sequence or
alternative node.

Also, fix an index error in the initialization of node_array.
(issue #15343)

git-svn-id: svn+ssh://ci.ruby-lang.org/ruby/trunk@66139 b2dd03c8-39d4-4d8f-98ff-823fe69b080e

Revision 66139 - 12/02/2018 11:28 PM - duerst (Martin Dürst)

remove unnecessary settings with NULL_NODE in \X implementation

Remove unnecessary settings of node_array elements to NULL_NODE.
We can do this because we initialize the whole array to NULL_NODEs
and set everything again to NULL_NODEs when creating a sequence or
alternative node.

Also, fix an index error in the initialization of node_array.
(issue #15343)

Revision 66139 - 12/02/2018 11:28 PM - duerst (Martin Dürst)

remove unnecessary settings with NULL_NODE in \X implementation

Remove unnecessary settings of node_array elements to NULL_NODE.
We can do this because we initialize the whole array to NULL_NODEs
and set everything again to NULL_NODEs when creating a sequence or
alternative node.

Also, fix an index error in the initialization of node_array.
(issue #15343)

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 regparse.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_elf.txt added

- File debug_X_genie.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)

- Backport changed from 2.4: UNKNOWN, 2.5: UNKNOWN to 2.4: UNKNOWN, 2.5: REQUIRED

- Assignee changed from naruse (Yui NARUSE) to duerst (Martin Dürst)

- Status changed from Open to Closed

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)