

Ruby - Bug #19455

Ruby 3.2: wrong Regexp encoding with non-ASCII comments

02/22/2023 10:18 AM - janosch-x (Janosch Müller)

Status:	Closed	
Priority:	Normal	
Assignee:		
Target version:		
ruby -v:	ruby 3.2.0 (2022-12-25 revision a528908271)	Backport: 3.0: DONTNEED, 3.1: DONTNEED, 3.2: DONE
Description		
comments and comment groups don't trigger the correct Regexp#encoding on Ruby 3.2 anymore:		
<pre># ruby 3.1 /#a/x.encoding # => #<Encoding:US-ASCII> # OK /(?#a)/.encoding # => #<Encoding:US-ASCII> # OK /#ü/x.encoding # => #<Encoding:UTF-8> # OK /(?#ü)/.encoding # => #<Encoding:UTF-8> # OK # ruby 3.2 /#a/x.encoding # => #<Encoding:US-ASCII> # OK /(?#a)/.encoding # => #<Encoding:US-ASCII> # OK /#ü/x.encoding # => #<Encoding:US-ASCII> # BUG /(?#ü)/.encoding # => #<Encoding:US-ASCII> # BUG /#ü/x.inspect # => "#\\xC3\\xBC/x" /(?#ü)/.inspect # => "/(?#\\xC3\\xBC)/" # bug is hidden if there are non-ascii chars outside comments /ü#ü/x.encoding # => #<Encoding:UTF-8> /ü(?#ü)/.encoding # => #<Encoding:UTF-8></pre>		
i think these changes might be the cause: https://github.com/ruby/ruby/commit/ec3542229b29ec93062e9d90e877ea29d3c19472#diff-c3675fa319803b2f5a775defa40694edb9a761baa3a54fa78e1fdef8f918cc7cR2837-R2890 @jeremyevans0 (Jeremy Evans) JFYI		

Associated revisions

Revision a8ba1ddd78544b4bda749051d44f7b2a8a0ec5ff - 04/24/2023 02:27 AM - jeremyevans (Jeremy Evans)

Use UTF-8 encoding for literal extended regexps with UTF-8 charactrs in comments

Fixes [Bug #19455]

Revision a8ba1ddd78544b4bda749051d44f7b2a8a0ec5ff - 04/24/2023 02:27 AM - jeremyevans (Jeremy Evans)

Use UTF-8 encoding for literal extended regexps with UTF-8 characters in comments

Fixes [Bug #19455]

Revision a8ba1ddd - 04/24/2023 02:27 AM - jeremyevans (Jeremy Evans)

Use UTF-8 encoding for literal extended regexps with UTF-8 characters in comments

Fixes [Bug #19455]

Revision be09d77b966c7bcc77957927f16cefe66b365495 - 07/17/2023 08:23 AM - nagachika (Tomoyuki Chikanaga)

merge revision(s) a8ba1ddd78544b4bda749051d44f7b2a8a0ec5ff: [Backport #19455]

Use UTF-8 encoding for literal extended regexps with UTF-8 characters in comments

```
Fixes [Bug #19455]
---
re.c | 9 ++++++--
test/ruby/test_regexp.rb | 7 ++++++
2 files changed, 15 insertions(+), 1 deletion(-)
```

Revision be09d77b966c7bcc77957927f16cfe66b365495 - 07/17/2023 08:23 AM - nagachika (Tomoyuki Chikanaga)

merge revision(s) a8ba1ddd78544b4bda749051d44f7b2a8a0ec5ff: [Backport #19455]

```
Use UTF-8 encoding for literal extended regexps with UTF-8 characters
in comments
```

```
Fixes [Bug #19455]
---
re.c | 9 ++++++--
test/ruby/test_regexp.rb | 7 ++++++
2 files changed, 15 insertions(+), 1 deletion(-)
```

Revision be09d77b - 07/17/2023 08:23 AM - nagachika (Tomoyuki Chikanaga)

merge revision(s) a8ba1ddd78544b4bda749051d44f7b2a8a0ec5ff: [Backport #19455]

```
Use UTF-8 encoding for literal extended regexps with UTF-8 characters
in comments
```

```
Fixes [Bug #19455]
---
re.c | 9 ++++++--
test/ruby/test_regexp.rb | 7 ++++++
2 files changed, 15 insertions(+), 1 deletion(-)
```

History

#1 - 02/22/2023 04:09 PM - jeremyevans0 (Jeremy Evans)

I'm not sure that this is a bug. If all non-comment characters considered in the regexp are in the US-ASCII range, it seems reasonable for US-ASCII to be used as the regexp encoding. I'll add this ticket to the next developer meeting and see what other committers think.

#2 - 02/23/2023 04:18 AM - mame (Yusuke Endoh)

@janosch-x Do you have any specific problem with this change? For example, a string that used to match no longer matches, or vice versa.

#3 - 02/23/2023 12:24 PM - janosch-x (Janosch Müller)

i don't have a problem with this myself and the matching behavior is not affected as far as i can tell.

notable behavioral differences are:

- `/#ü/x.source == 'ü'` used to be true but is now false
 - this might break some tests or metaprogramming (not very likely IMO)
- `/#{/#ü/x.source}/` now raises `ArgumentError` (invalid multibyte character)

#4 - 03/09/2023 11:07 AM - mame (Yusuke Endoh)

Discussed at the dev meeting. @matz (Yukihiro Matsumoto) said he would prefer 3.1 behavior if possible (but not high priority). @nobu (Nobuyoshi Nakada) said he would take a look.

#5 - 03/24/2023 08:08 PM - jeremyevans0 (Jeremy Evans)

I submitted a pull request to fix this: <https://github.com/ruby/ruby/pull/7592>

#6 - 04/24/2023 04:28 AM - jeremyevans (Jeremy Evans)

- Status changed from Open to Closed

Applied in changeset [git|a8ba1ddd78544b4bda749051d44f7b2a8a0ec5ff](https://github.com/ruby/ruby/pull/7592).

Use UTF-8 encoding for literal extended regexps with UTF-8 characters in comments

Fixes [Bug [#19455](#)]

#7 - 04/24/2023 10:37 AM - nagachika (Tomoyuki Chikanaga)

- Backport changed from 2.7: UNKNOWN, 3.0: UNKNOWN, 3.1: UNKNOWN, 3.2: UNKNOWN to 3.0: DONTNEED, 3.1: DONTNEED, 3.2: REQUIRED

#8 - 07/17/2023 08:26 AM - nagachika (Tomoyuki Chikanaga)

- Backport changed from 3.0: DONTNEED, 3.1: DONTNEED, 3.2: REQUIRED to 3.0: DONTNEED, 3.1: DONTNEED, 3.2: DONE

ruby_3_2 be09d77b966c7bcc77957927f16cefe66b365495 merged revision(s) a8ba1ddd78544b4bda749051d44f7b2a8a0ec5ff.