STDIN.each_codepoint hangs indefinitely on certain codepoints

Description
Piping certain Unicode data into Ruby and calling each_codepoint on STDIN results in a hang. Control-c does not exit, but control-z will. I've tested this in 2.2.0, 2.2.2, and trunk.

Please see https://gist.github.com/cantino/48f977bee6a39e33f80f for reproduction. There is an example file with various Unicode Emoji characters in it (borrowed from http://apps.timwhitlock.info/emoji/tables/unicode).

I'd like to help fix this if anyone can point me in the right direction. I fiddled with the file until I found a state where it works if one character is removed from the front of the file, so hopefully that's useful information.

Related issues:
Related to Ruby master - Feature #1667: IO#codepoints, IO#each_codepoint, and...

Associated revisions
Revision cfa7550b - 08/15/2015 01:15 AM - nobu (Nobuyoshi Nakada)
io.c: read more data

- io.c (rb_io_each_codepoint): read more data when read partially. [ruby-core:70379] [Bug #11444]

git-svn-id: svn+ssh://ci.ruby-lang.org/ruby/trunk@51583 b2dd03c8-39d4-4d8f-98ff-823fe69b080e

Revision 51583 - 08/15/2015 01:15 AM - nobu (Nobuyoshi Nakada)
io.c: read more data

- io.c (rb_io_each_codepoint): read more data when read partially. [ruby-core:70379] [Bug #11444]

Revision 51583 - 08/15/2015 01:15 AM - nobu (Nobuyoshi Nakada)
io.c: read more data

- io.c (rb_io_each_codepoint): read more data when read partially. [ruby-core:70379] [Bug #11444]

Revision 51583 - 08/15/2015 01:15 AM - nobu (Nobuyoshi Nakada)
io.c: read more data

- io.c (rb_io_each_codepoint): read more data when read partially. [ruby-core:70379] [Bug #11444]

Revision 51583 - 08/15/2015 01:15 AM - nobu (Nobuyoshi Nakada)
io.c: read more data

- io.c (rb_io_each_codepoint): read more data when read partially. [ruby-core:70379] [Bug #11444]

Revision 51583 - 08/15/2015 01:15 AM - nobu (Nobuyoshi Nakada)
io.c: read more data

- io.c (rb_io_each_codepoint): read more data when read partially. [ruby-core:70379] [Bug #11444]

Revision 84bf320b - 08/17/2015 02:57 AM - nobu (Nobuyoshi Nakada)
io.c: raise at incomplete char

- io.c (rb_io_each_codepoint): raise an exception at incomplete character before EOF when conversion takes place. [Bug #11444]
Revision 51594 - 08/17/2015 02:57 AM - nobu (Nobuyoshi Nakada)
io.c: raise at incomplete char
  
  * io.c (rb_io_each_codepoint): raise an exception at incomplete character before EOF when conversion takes place. [Bug #11444]

Revision 51594 - 08/17/2015 02:57 AM - nobu (Nobuyoshi Nakada)
io.c: raise at incomplete char
  
  * io.c (rb_io_each_codepoint): raise an exception at incomplete character before EOF when conversion takes place. [Bug #11444]

Revision 51594 - 08/17/2015 02:57 AM - nobu (Nobuyoshi Nakada)
io.c: raise at incomplete char
  
  * io.c (rb_io_each_codepoint): raise an exception at incomplete character before EOF when conversion takes place. [Bug #11444]

Revision 51594 - 08/17/2015 02:57 AM - nobu (Nobuyoshi Nakada)
io.c: raise at incomplete char
  
  * io.c (rb_io_each_codepoint): raise an exception at incomplete character before EOF when conversion takes place. [Bug #11444]

Revision 51594 - 08/17/2015 02:57 AM - nobu (Nobuyoshi Nakada)
io.c: raise at incomplete char
  
  * io.c (rb_io_each_codepoint): raise an exception at incomplete character before EOF when conversion takes place. [Bug #11444]

Revision 51594 - 08/17/2015 02:57 AM - nobu (Nobuyoshi Nakada)
io.c: raise at incomplete char
  
  * io.c (rb_io_each_codepoint): raise an exception at incomplete character before EOF when conversion takes place. [Bug #11444]

Revision 0bf77265 - 08/19/2015 07:57 AM - nobu (Nobuyoshi Nakada)
io.c: initialize variable
  
  * io.c (rb_io_each_codepoint): fix use of uninitialized variable. [Bug #11444]

Revision 51638 - 08/19/2015 07:57 AM - nobu (Nobuyoshi Nakada)
io.c: initialize variable
  
  * io.c (rb_io_each_codepoint): fix use of uninitialized variable. [Bug #11444]

Revision 51638 - 08/19/2015 07:57 AM - nobu (Nobuyoshi Nakada)
io.c: initialize variable
  
  * io.c (rb_io_each_codepoint): fix use of uninitialized variable. [Bug #11444]

Revision 51638 - 08/19/2015 07:57 AM - nobu (Nobuyoshi Nakada)
io.c: initialize variable
  
  * io.c (rb_io_each_codepoint): fix use of uninitialized variable. [Bug #11444]

Revision 51638 - 08/19/2015 07:57 AM - nobu (Nobuyoshi Nakada)
io.c: initialize variable
  
  * io.c (rb_io_each_codepoint): fix use of uninitialized variable. [Bug #11444]

Revision 51638 - 08/19/2015 07:57 AM - nobu (Nobuyoshi Nakada)
io.c: initialize variable
  
  * io.c (rb_io_each_codepoint): fix use of uninitialized variable. [Bug #11444]

Revision 51638 - 08/19/2015 07:57 AM - nobu (Nobuyoshi Nakada)
io.c: initialize variable
  
  * io.c (rb_io_each_codepoint): fix use of uninitialized variable. [Bug #11444]

Revision 51638 - 08/19/2015 07:57 AM - nobu (Nobuyoshi Nakada)
io.c: initialize variable
  
  * io.c (rb_io_each_codepoint): fix use of uninitialized variable. [Bug #11444]

Revision 51638 - 08/19/2015 07:57 AM - nobu (Nobuyoshi Nakada)
io.c: initialize variable
  
  * io.c (rb_io_each_codepoint): fix use of uninitialized variable. [Bug #11444]

Revision 6d8ed07d - 11/29/2015 08:21 AM - nagachika (Tomoyuki Chikanaga)
merge revision(s) 51583,51638: [Backport #11444]
  
  * io.c (rb_io_each_codepoint): read more data when read partially.
Revision 52786 - 11/29/2015 08:21 AM - nagachika (Tomoyuki Chikanaga)
merge revision(s) 51583,51638: [Backport #11444]
* io.c (rb_io_each_codepoint): read more data when read partially. [Bug #11444]

Revision 4af525b1 - 11/30/2015 05:35 PM - nagachika (Tomoyuki Chikanaga)
merge revision(s) 51594: [Backport #11444]
* io.c (rb_io_each_codepoint): raise an exception at incomplete character before EOF when conversion takes place. [Bug #11444]

Revision 52804 - 11/30/2015 05:35 PM - nagachika (Tomoyuki Chikanaga)
merge revision(s) 51594: [Backport #11444]
* io.c (rb_io_each_codepoint): raise an exception at incomplete character before EOF when conversion takes place. [Bug #11444]

Revision f81b6694 - 12/01/2015 08:00 AM - usa (Usaku NAKAMURA)
merge revision(s) 51583,51594,51638: [Backport #11444]
* io.c (rb_io_each_codepoint): read more data when read partially. [ruby-core:70379] [Bug #11444]
* io.c (rb_io_each_codepoint): raise an exception at incomplete character before EOF when conversion takes place. [Bug #11444]

Revision 52831 - 12/01/2015 08:00 AM - usa (Usaku NAKAMURA)
merge revision(s) 51583,51594,51638: [Backport #11444]
* io.c (rb_io_each_codepoint): read more data when read partially. [ruby-core:70379] [Bug #11444]
* io.c (rb_io_each_codepoint): raise an exception at incomplete character before EOF when conversion takes place. [Bug #11444]

History

#1 - 08/15/2015 01:13 AM - nobu (Nobuyoshi Nakada)
- Description updated
- Backport changed from 2.0.0: UNKNOWN, 2.1: UNKNOWN, 2.2: UNKNOWN to 2.0.0: REQUIRED, 2.1: REQUIRED, 2.2: REQUIRED

#2 - 08/15/2015 01:15 AM - nobu (Nobuyoshi Nakada)
- Status changed from Open to Closed

Applied in changeset r51583.

io.c: read more data

* io.c (rb_io_each_codepoint): read more data when read partially. [ruby-core:70379] [Bug #11444]

#3 - 08/15/2015 03:30 AM - cantino (Andrew Cantino)
Thank you for fixing this, and in less than 24 hours!

#4 - 08/17/2015 01:37 AM - nobu (Nobuyoshi Nakada)
- Related to Feature #1667: IO#codepoints, IO#each_codepoint, and StringIO added

05/10/2020
Backported r51583 and r51638 into ruby_2_2 at r52786.
I didn't backport r51594 because it introduce an incompatibility.

Tomoyuki Chikanaga wrote:

But without r51549, the test never pass on mswin.
See http://ruby-mswin.cloudapp.net/vc10-x64/ruby-2.2/log/20151129T090947Z.log.html.gz#rubyspec

usa san: Thank you for your notice.
Hmm, I will backport r51594.

Backported r51594 into ruby_2_2 branch at r52804.

ruby_2_1 r52831 merged revision(s) 51583,51594,51638.