Backport193 - Backport #8516

IO#readchar returns wrong codepoints when converting encoding

06/12/2013 04:15 AM - bbxiao1 (Xiao Ba)

Status:	Closed	
Priority:	Normal	
Assignee:	usa (Usaku NAKAMURA)	

Description

I am trying to parse plain text files with various encodings that will ultimately be converted to UTF-8 strings. Non-ascii characters work fine with a file encoded as UTF-8, but problems come up with non-UTF-8 files.

$ file -i utf_8.txt
utf_8.txt: text/plain; charset=utf-8

$ file -i iso_8859_1.txt
iso_8859_1.txt: text/plain; charset=iso-8859-1

Code:

```
utf_8_file = "utf_8.txt"
iso_file = "iso_8859_1.txt"

puts "Processing #{utf_8_file}"
File.open(utf_8_file) do |io|
line, char = "", nil

until io.eof? || char == ?\n || char == ?\r
char = io.readchar
puts "Character #{char} has #{char.each_codepoint.count} codepoints"
puts "Character #{char} codepoints: #{char.each_codepoint.to_a.join}"
puts "SLICE FAIL" unless char == char.slice(0,1)
line << char
end

line
end
puts "\n"
puts "Processing #{iso_file}"
File.open(iso_file) do |io|
io.set_encoding("#{Encoding::ISO_8859_1}:#{Encoding::UTF_8}")
line, char = "", nil

until io.eof? || char == ?\n || char == ?\r
char = io.readchar
puts "Character #{char} has #{char.each_codepoint.count} codepoints"
puts "Character #{char} codepoints: #{char.each_codepoint.to_a.join(', ')}"
puts "SLICE FAIL" unless char == char.slice(0,1)
line << char
end

line
end
```

Output:

```
Processing utf_8.txt
Character á has 1 codepoints
Character á codepoints: 225
Character Á has 1 codepoints
Character Á codepoints: 193
Character ð has 1 codepoints
Character ð codepoints: 240
Character
has 1 codepoints
Character
```

codepoints: 10

Processing iso_8859_1.txt
Character á has 2 codepoints
Character á codepoints: 195, 161
SLICE FAIL
Character Á has 2 codepoints
Character Á codepoints: 195, 129
SLICE FAIL
Character ð has 2 codepoints
Character ð codepoints: 195, 176
SLICE FAIL
Character
has 1 codepoints
Character
codepoints: 10

With the ISO-8859-1 encoded file, readchar is returning the character bytes when I would expect UTF-8 codepoints.

Associated revisions

Revision ab64f237 - 06/12/2013 03:44 AM - nobu (Nobuyoshi Nakada)

io.c: fix 7bit coderange condition

- io.c (io_getc): fix 7bit coderange condition, check if ascii read data instead of read length. [ruby-core:55444] [Bug #8516]

git-svn-id: svn+ssh://ci.ruby-lang.org/ruby/trunk@41250 b2dd03c8-39d4-4d8f-98ff-823fe69b080e

Revision 41250 - 06/12/2013 03:44 AM - nobu (Nobuyoshi Nakada)

io.c: fix 7bit coderange condition

- io.c (io_getc): fix 7bit coderange condition, check if ascii read data instead of read length. [ruby-core:55444] [Bug #8516]

Revision 41250 - 06/12/2013 03:44 AM - nobu (Nobuyoshi Nakada)

io.c: fix 7bit coderange condition

- io.c (io_getc): fix 7bit coderange condition, check if ascii read data instead of read length. [ruby-core:55444] [Bug #8516]

Revision 41250 - 06/12/2013 03:44 AM - nobu (Nobuyoshi Nakada)

io.c: fix 7bit coderange condition

- io.c (io_getc): fix 7bit coderange condition, check if ascii read data instead of read length. [ruby-core:55444] [Bug #8516]

Revision 41250 - 06/12/2013 03:44 AM - nobu (Nobuyoshi Nakada)

io.c: fix 7bit coderange condition

- io.c (io_getc): fix 7bit coderange condition, check if ascii read data instead of read length. [ruby-core:55444] [Bug #8516]

Revision 41250 - 06/12/2013 03:44 AM - nobu (Nobuyoshi Nakada)

io.c: fix 7bit coderange condition

- io.c (io_getc): fix 7bit coderange condition, check if ascii read data instead of read length. [ruby-core:55444] [Bug #8516]

Revision 41250 - 06/12/2013 03:44 AM - nobu (Nobuyoshi Nakada)

io.c: fix 7bit coderange condition

- io.c (io_getc): fix 7bit coderange condition, check if ascii read data instead of read length. [ruby-core:55444] [Bug #8516]

Revision 63bc35ff - 06/12/2013 02:23 PM - nagachika (Tomoyuki Chikanaga)

merge revision(s) 41250: [Backport #8516]

```
    * io.c (io_getc): fix 7bit coderange condition, check if ascii read
      data instead of read length. [ruby-core:55444] [Bug #8516]
```

git-svn-id: svn+ssh://ci.ruby-lang.org/ruby/branches/ruby_2_0_0@41260 b2dd03c8-39d4-4d8f-98ff-823fe69b080e

Revision fbc47099 - 06/26/2013 07:13 AM - usa (Usaku NAKAMURA)

merge revision(s) 41250: [Backport #8516]

```
    * io.c (io_getc): fix 7bit coderange condition, check if ascii read
      data instead of read length. [ruby-core:55444] [Bug #8516]
```

git-svn-id: svn+ssh://ci.ruby-lang.org/ruby/branches/ruby_1_9_3@41644 b2dd03c8-39d4-4d8f-98ff-823fe69b080e

Revision 41644 - 06/26/2013 07:13 AM - usa (Usaku NAKAMURA)

merge revision(s) 41250: [Backport #8516]

```
* io.c (io_getc): fix 7bit coderange condition, check if ascii read
  data instead of read length. [ruby-core:55444] [Bug #8516]
```

History

#1 - 06/12/2013 12:44 PM - nobu (Nobuyoshi Nakada)

- Status changed from Open to Closed

- % Done changed from 0 to 100

This issue was solved with changeset r41250.
Xiao, thank you for reporting this issue.
Your contribution to Ruby is greatly appreciated.
May Ruby be with you.

---

io.c: fix 7bit coderange condition

- io.c (io_getc): fix 7bit coderange condition, check if ascii read data instead of read length. [ruby-core:55444] [Bug #8516]

#2 - 06/12/2013 12:45 PM - nobu (Nobuyoshi Nakada)

- Backport changed from 1.9.3: UNKNOWN, 2.0.0: UNKNOWN to 1.9.3: REQUIRED, 2.0.0: REQUIRED

#3 - 06/12/2013 11:04 PM - nagachika (Tomoyuki Chikanaga)

- Tracker changed from Bug to Backport

- Project changed from Ruby trunk to Backport200

- Status changed from Closed to Assigned

- Assignee set to nagachika (Tomoyuki Chikanaga)

#4 - 06/12/2013 11:23 PM - nagachika (Tomoyuki Chikanaga)

- Status changed from Assigned to Closed

This issue was solved with changeset r41260.
Xiao, thank you for reporting this issue.
Your contribution to Ruby is greatly appreciated.
May Ruby be with you.

---

merge revision(s) 41250: [Backport #8516]

```
* io.c (io_getc): fix 7bit coderange condition, check if ascii read
  data instead of read length. [ruby-core:55444] [Bug #8516]
```

#5 - 06/12/2013 11:24 PM - nagachika (Tomoyuki Chikanaga)

- Project changed from Backport200 to Backport193

- Status changed from Closed to Assigned

- Assignee changed from nagachika (Tomoyuki Chikanaga) to usa (Usaku NAKAMURA)

#6 - 06/26/2013 04:13 PM - usa (Usaku NAKAMURA)

- Status changed from Assigned to Closed

This issue was solved with changeset r41644.
Xiao, thank you for reporting this issue.

Your contribution to Ruby is greatly appreciated.
May Ruby be with you.

---

merge revision(s) 41250: [Backport #8516]

```
* io.c (io_getc): fix 7bit coderange condition, check if ascii read
  data instead of read length. [ruby-core:55444] [Bug #8516]
```

Files

utf_8.txt	7 Bytes	06/12/2013	bbxiao1 (Xiao Ba)
iso_8859_1.txt	4 Bytes	06/12/2013	bbxiao1 (Xiao Ba)