正規表現の \z の話

あるちょっと遅そうな正規表現を
高速化した話

Kazuhiro NISHIYAMA

第67回 Ruby関西 勉強会
2015/06/13

自己紹介

- twitter や github では @znz

Ruby 関連

- Ruby (CRuby) のコミッター
- Ruby 関係でいろいろ
  - るびま (Rubyist Magazine) とか
    - http://magazine.rubyist.net/
  - るりま (Ruby リファレンスマニュアル) とか
    - http://docs.ruby-lang.org/ja/
  - Ruby 関連のイベントでの発表

とある正規表現

/(ん|ン)$/

- 遅そう
- 文字列末尾以外にもマッチする

文字クラスを使う

```
/[んン]$/
```

- 速くなった(ベンチマークは後で)

\z を使う

```
/[んン]\z/
```

- $: 行末にマッチ
- \z: 文字列末尾のみにマッチ
- もっと速くなった

ベンチマーク

```
#!/usr/bin/env ruby
# coding: utf-8
require 'benchmark'
s = 'ん' * 10000 + 'ン'
n = 1000
Benchmark.bmbm do |x|
  x.report('/(ん|ン)$/')  { n.times { /(ん|ン)$/  =~ s } }
  x.report('/[んン]$/')   { n.times { /[んン]$/   =~ s } }
  x.report('/(ん|ン)\z/') { n.times { /(ん|ン)\z/ =~ s } }
  x.report('/[んン]\z/')  { n.times { /[んン]\z/  =~ s } }
end
```

ベンチマーク結果 (1)

```
Rehearsal ------------------------------------------
/(ん|ン)$/    1.560000   0.000000   1.560000 (  1.561225)
/[んン]$/     0.850000   0.000000   0.850000 (  0.855042)
/(ん|ン)\z/   0.020000   0.000000   0.020000 (  0.015481)
/[んン]\z/    0.010000   0.000000   0.010000 (  0.015520)
--------------------------------- total: 2.440000sec

                user     system      total        real
/(ん|ン)$/    1.200000   0.000000   1.200000 (  1.205324)
/[んン]$/     0.900000   0.000000   0.900000 (  0.896039)
/(ん|ン)\z/   0.020000   0.000000   0.020000 (  0.015274)
/[んン]\z/    0.010000   0.000000   0.010000 (  0.015871)
```

ベンチマーク結果 (2)

```
Rehearsal ------------------------------------------
/(ん|ン)$/    1.300000   0.000000   1.300000 (  1.299798)
/[んン]$/     0.840000   0.000000   0.840000 (  0.834954)
/(ん|ン)\z/   0.010000   0.000000   0.010000 (  0.014847)
/[んン]\z/    0.020000   0.000000   0.020000 (  0.014750)
--------------------------------- total: 2.170000sec

                user     system      total        real
/(ん|ン)$/    1.110000   0.000000   1.110000 (  1.105155)
/[んン]$/     0.850000   0.000000   0.850000 (  0.853977)
/(ん|ン)\z/   0.010000   0.000000   0.010000 (  0.015622)
/[んン]\z/    0.020000   0.000000   0.020000 (  0.015683)
```

まとめ

- 1文字の場合は選択 (ん|ン) ではなく文字クラス [んン] を使おう
- 文字列末尾は $ ではなく \z