String#succ

田中 哲
akr@fsij.org
産業技術総合研究所 / FSIJ
2008-06-22

String#succ

- "a".succ #=> "b"
- "z".succ #=> "aa"
- "ko1".succ #=> "ko2"
- "ko9".succ #=> "kp0"
- s = "ko1"; 381.times { s.succ! }
  p s #=> "ma2"
- s = "matz"; 166760647.times { s.succ! }
  p s #=> "nakada"

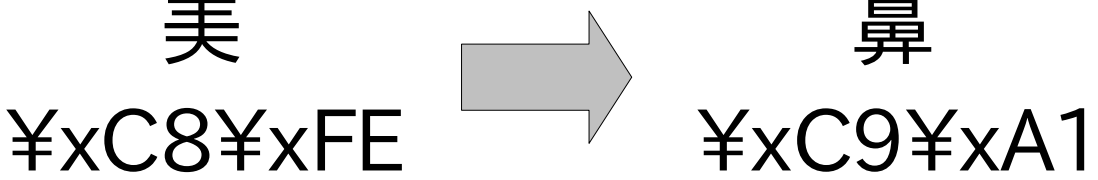
M17N

- "あ".succ #=> "い" # EUC-JP
- "元".succ #=> "原" # EUC-JP
- "元".succ #=> "兄" # UTF-8
- s = "ささだ"; 1674553600.times { s.succ! }
  p s #=> "なかだ" # EUC-JP

- "美".succ #=> "鼻" # EUC-JP
  "美" == "¥xC8¥xFE"
  "鼻" == "¥xC9¥xA1"

CSI

- succ は EUC-JP の知識を持っていない
- エンコーディング API に「次の文字」を求めるものはない
- 「美」から「鼻」をどう求めるかが問題
- succ は入力にない文字を意図的に生成する唯一のメソッド



encoding validation

- あるバイト列が正しいエンコーディングか調べられる

- "¥xC8¥xFE" は EUC-JP として正しい
- "¥xC8¥xFF" は EUC-JP として正しくない
- "¥xC9¥x00" は EUC-JP として正しくない
- 中略
- "¥xC9¥xA0" は EUC-JP として正しくない
- "¥xC9¥xA1" は EUC-JP として正しい→次の文字

問題: 遅いことがある

- UTF-8
- "¥xF0¥x9F¥xBF¥xBF" の次は
  "¥xF0¥xA0¥x80¥x80"
  49345回試さないと見つからない

encoding validation (2)

- 正しいエンコーディングの途中であるか判断できる
- バイトを追加しても正しくならないことを判断できる
- "¥xF0¥x9F¥xBF¥xC0" ..
  "¥xF0¥x9F¥xBF¥xFF"　　64回
- "¥xF0¥x9F¥xC0" ..
  "¥xF0¥x9F¥xFF"　　64回
- "¥xF0¥xA0¥x00" ..
  "¥xF0¥xA0¥x7F"　　128回
- "¥xF0¥xA0¥x80¥x00" ..
  "¥xF0¥xA0¥x80¥x7F"　　128回
- "¥xF0¥xA0¥x80¥x80"　　計385回で済む

まとめ

- succ には高度な技術が投入されている
- CSI には高度な技術が必要