Jcodeを使おう(実用本位のJcode紹介)

Kiyoka Nishiyama
(http://www.netfort.gr.jp/~kiyoka/jcode_intro/index.html)

$Date: 2000/11/29 15:40:18 $

目次

1 はじめに

1.1 目的

- Jcode を即使えるようになる。(Perl の基礎知識は必要)

1.2 アプローチ (5W1H で見てみる)

1. what:Jcode とは何か
2. why:なぜ Jcode か？
3. when:どんな場面で使えるのか
4. where:どんな環境で使えるのか
5. who:だれが使うのか?(それは、もちろんあなたです。)
6. how:どうやって使うのか

2 Jcode 概説

2.1 what:何か

1. 日本語文字コード用汎用モジュールである
   - Perl Module 化されている (.pl 版もあるが、perl5 が普及している今日では、pm 版を使うべき)
2. 基本的な機能は、文字コード判別、変換
   - 本資料ではこの機能のみ紹介
3. サポートする文字コード
   - ascii,binary,euc,sjis,jis,ucs2,utf8
4. 開発者・開発チーム
   - オリジナルである jcode.pl は歌代和正氏が開発された。(現在もメンテナンスが行なわれている)

補足

この講演資料は、Jcode の全ての事柄を網羅しているわけではありません。
現時点で Jcode を使い始める上で知っておくべき情報をまとめたものです。

- Jcode.pm は jcode.pl のコードを元に、RingServer 上の OpenLab で開発されている

2.2 why:なぜ Jcodeか？

1. デバッグ済みコードを使おう
   - 車輪の再発明を避けよう
   - 要は、楽をしよう
2. なぜ Jcode が必要なのか?
   - perl はいろんな日本語文字コードを認識してくれない
     - perl5.6 を使えば utf-8 の文字コードは認識してくれるがそれ以外の文字コードは認識してくれない
     - perl5.6 でも未だ漢字コード変換はサポートされていない
   - Web 上の HTML ファイルは、sjis,jis,euc 等バラバラ。
     - HTML ファイルを処理するのに、あらゆるコードを処理するプログラムを書くの？->Jcode を使おう

2.3 when:どんな場面で使えるのか

1. HTML ファイルの文字コード変換・チェック (ヘッダの文字コード指定と、ファイルの文字コードがマッチしているかどうかなど)
2. perl プログラムから送信する Mail の文字コード変換
3. CGI へのフォーム入力データの文字コード変換/半角・全角変換 (詳しい方、補足お願いします。)

2.4 where:どんな環境で使えるのか

- jcode.pl なら perl4 以上で使用できる
- Jcode.pm なら perl5 以上で使用できる
  - 但し、perl スクリプト内で正規表現などを使って日本語を処理したい場合は、jperl4,jperl5 か、perl5.6 以上を使う必要があります。

2.5 who:誰が使うのか (それは、もちろんあなたです。)

- Perl で日本語を使う必要のある人は Jcode を使いましょう

2.6 how:どうやって使うのか

- 次章から次のような手順で見ていきます
  - 漢字コードについての基礎
  - getcode メソッドと convert メドッド
  - 簡単なデモの紹介 (perl5.6 のテスト・コード判別・nkf もどき)

3 漢字コードについての基礎 (深い話題は除く)

3.1 perlとマルチバイト処理の関係

- perl でマルチバイトを操作したい場合は、jperl か、perl5.6 以上を使う必要があります。
  - jperl：2 バイトの漢字を 1 文字として扱うことができる (jperl にはプラットフォームによって色々なバージョンがあるらしいので注意が必要)
  - perl5.6：utf-8(マルチバイト) を扱うことができる

    perl5.6 がリリースされた現在では、perl5.6 を使うのが正解でしょう。

    perl5.6 なら utf-8 を処理できるので、マルチバイトを処理する正規表現が直接記述できます。
- この後の話は、perl5.6 + Jcode.pm を前提に進めていきます

4 getcode メソッドと convert メソッド

- jcode は getcode と convert メソッドが基本です。

4.1 jcode.plレガシーインターフェースとオブジェクト指向インターフェース

- Jcode.pm を使用する場合 jcode.pl 風のレガシーインターフェースと、オブジェクト指向のインターフェースの両方が使えます。
- ここでは、jcode.pl 風のインターフェースでの使いかたをメインに紹介します。

4.2 getcode メソッド – エンコードを調べる

- `@result = getcode( \$line );`
  - $line のエンコード名を、”jis”, ”sjis”, ”euc”, ”ucs2”, ”utf8”, のいずれかで返します。$line 中にバイナリキャラクタを見つければ、”binary”を返します。
  - この関数をリストコンテキストで使うと、2つの要素を含むリストを返します。第1要素は、$line 中で、予想されているエンコードに一致したバイト数です。第2要素は、$line のエンコード名です。

4.3 convert メソッド – 任意のエンコードに変換する

- `convert(\$line, [$ocode, $icode, $opt])`
  - $line を、$ocode に与えられたエンコードに変換します。
  - 第2引数$ocode には、変換したいエンコードを”jis”, ”sjis”, ”euc”, ”ucs2”, ”utf8”のどれかで与えます。
  - $opt は、半角カナ等の処理に関するオプションです。(詳細は省略します。)

5 サンプルコードによるデモ

- Jcode の説明のために簡単なサンプルコードを作成しました
- 具体的にコードと動作を示します

5.1 perl5.6 のテスト

- perl5.6 の utf-8 対応がどんな状況かを見てみます
- ソースコード utf_test.pl[1]

5.2 文字コード判別 (getcode メソッド使用)

- 文字コードの判別の実験を行います。
- ソースコード jcodeis.pl[2]

[1] code/utf_test.pl
[2] code/jcodeis.pl

5.3 nkfもどき (convert メソッド使用)

- nkf のようなフィルターを作ってみます。
- ソースコード nkf.pl[3]

6 付録:その他のリソース

6.1 perlがインストールされている環境で見れる資料

- perldoc ドキュメント (コマンドラインで perldoc Jcode してみよう)
  - この資料では触れていませんが、ほかにも全角・半角変換や MIME データの作成など便利な機能があります。

6.2 インターネット上のリソース

- pkf

  歌代和正さん <utashiro@iij.ad.jp> 作の nkf の perl 版です。

  ftp://ftp.iij.ad.jp/pub/IIJ/dist/utashiro/perl/pkf-2.1[4]

- Jcode.pm

  DAN Kogaidan<kogai@dan.co.jp> さん作で、jcode.pl をモジュール化したものです。

  http://openlab.ring.gr.jp/Jcode/index-j.html[5]

  CPAN/authors/id/D/DA/DANKOGAI/[6]

- 「jcode.pl の私的な解説書」

  jcode.pl の日本語訳版 (非公式)[7]

- 「Manpage of UNICODE」

  UNICODE の manpage(日本語)[8]

[3] code/nkf.pl
[4] ftp://ftp.iij.ad.jp/pub/IIJ/dist/utashiro/perl/
[5] http://openlab.ring.gr.jp/Jcode/index-j.html
[6] ftp://ftp.dti.ad.jp/pub/lang/CPAN/authors/id/D/DA/DANKOGAI/
[7] http://www.mikeneko.ne.jp/~lab/kcode/jcode.html
[8] http://www.linux.or.jp/JM/html/LDP_man-pages/man7/unicode.7.html

6.3 書籍

- 「日本語情報処理」
  - 著者:Ken Lunde
  - ISBN 4-89052-708-7
  - 本体価格:4893 円
  - 概要:
    コンピュータによる日本語処理について、JIS 漢字コードや EUC、Unicode、各メーカー漢字コードなどの各コード体系、アプリケーションにおける日本語特有の問題についての対処法など詳解。
    ※原書が 3 年前に絶版なっているので、この翻訳本も入手できないかもしれません。
- 「CJKV Information Processing (Chinese, Japanese, Korean & Vietnamese Computing)」
  - By Ken Lunde
  - 1st Edition December 1998
  - 1-56592-224-7, Order Number: 2247
  - 1125 pages, $64.95
    O'Reilly の紹介ページへ[9]
- 本ドキュメントの PDF 版はこちら[10]

[9] http://www.oreilly.com/catalog/cjkvinfo/
[10] index.pdf