

If you read this file `_as_is_`, just ignore the funny characters you see. It is written in the POD format (see `perlpod manpage`) which is specially designed to be readable as is.

The following documentation is written in EUC-CN encoding.

如果你用一般的文字编辑器浏览这份文件, 请忽略文中奇特的注记字符. 这份文件是以 POD (简明文件格式) 写成; 这种格式是为了能让人直接阅读, 而特别设计的. 关于此格式的进一步信息, 请参考 `perlpod` 线上文件.

1 NAME

`perlcn` - 简体中文 Perl 指南

2 DESCRIPTION

欢迎来到 Perl 的天地!

从 5.8.0 版开始, Perl 具备了完善的 Unicode (统一码) 支援, 也连带支援了许多拉丁语系以外的编码方式; CJK (中日韩) 便是其中的一部份. Unicode 是国际性的标准, 试图涵盖世界上所有的字符: 西方世界, 东方世界, 以及两者间的一切 (希腊文, 叙利亚文, 亚拉伯文, 希伯来文, 印度文, 印地安文, 等等). 它也容纳了多种作业系统与平台 (如 PC 及麦金塔).

Perl 本身以 Unicode 进行操作. 这表示 Perl 内部的字符串数据可用 Unicode 表示; Perl 的函式与算符 (例如正规表示式比对) 也能对 Unicode 进行操作. 在输入及输出时, 为了处理以 Unicode 之前的编码方式存放的数据, Perl 提供了 `Encode` 这个模块, 可以让你轻易地读取及写入旧有的编码数据.

`Encode` 延伸模块支援下列简体中文的编码方式 ('gb2312' 表示 'euc-cn'):

```
euc-cn    Unix 延伸字符集, 也就是俗称的国标码
gb2312-raw 未经处理的 (低比特) GB2312 字符表
gb12345    未经处理的中国用繁体中文编码
iso-ir-165 GB2312 + GB6345 + GB8565 + 新增字符
cp936     字码页 936, 也可以用 'GBK' (扩充国标码) 指明
hz        7 比特逸出式 GB2312 编码
```

举例来说, 将 EUC-CN 编码的档案转成 Unicode, 祇需键入下列指令:

```
perl -Mencoding=euc-cn,STDOUT,utf8 -pe1 < file.euc-cn > file.utf8
```

Perl 也内附了 "piconv", 一支完全以 Perl 写成的字符转换工具程序, 用法如下:

```
piconv -f euc-cn -t utf8 < file.euc-cn > file.utf8
piconv -f utf8 -t euc-cn < file.utf8 > file.euc-cn
```

另外, 利用 `encoding` 模块, 你可以轻易写出以字符为单位的程序码, 如下所示:

```
#!/usr/bin/env perl
# 启动 euc-cn 字符串解析; 标准输出及标准错误都设为 euc-cn 编码
use encoding 'euc-cn', STDIN => 'euc-cn', STDOUT => 'euc-cn';
print length("骆驼");      # 2 (双引号表示字符)
print length('骆驼');      # 4 (单引号表示字节)
print index("谆谆教诲", "蛔蛔"); # -1 (不包含此子字符串)
print index('谆谆教诲', '蛔蛔'); # 1 (从第二个字节开始)
```

在最后一列例子里, "谆" 的第二个字节与 "谆" 的第一个字节结合成 EUC-CN 码的 "蛔"; "谆" 的第二个字节则与 "教" 的第一个字节结合成 "蛔". 这解决了以前 EUC-CN 码比对处理上常见的问题.

额外的中文编码

如果需要更多的中文编码, 可以从 CPAN (<http://www.cpan.org/>) 下载 Encode::HanExtra 模块. 它目前提供下列编码方式:

gb18030 扩充过的国标码, 包含繁体中文

另外, Encode::HanConvert 模块则提供了简繁转换用的两种编码:

big5-simp Big5 繁体中文与 Unicode 简体中文互转
gbk-trad GBK 简体中文与 Unicode 繁体中文互转

若想在 GBK 与 Big5 之间互转, 请参考该模块内附的 b2g.pl 与 g2b.pl 两支程序, 或在程序内使用下列写法:

```
use Encode::HanConvert;  
$euc_cn = big5_to_gb($big5); # 从 Big5 转为 GBK  
$big5 = gb_to_big5($euc_cn); # 从 GBK 转为 Big5
```

进一步的信息

请参考 Perl 内附的大量说明文件 (不幸全是用英文写的), 来学习更多关于 Perl 的知识, 以及 Unicode 的使用方式. 不过, 外部的资源相当丰富:

提供 Perl 资源的网址

<http://www.perl.com/>

Perl 的首页 (由欧莱礼公司维护)

<http://www.cpan.org/>

Perl 综合典藏网 (Comprehensive Perl Archive Network)

<http://lists.perl.org/>

Perl 邮递论坛一览

学习 Perl 的网址

<http://www.oreilly.com.cn/html/perl.html>

简体中文版的欧莱礼 Perl 书籍

Perl 使用者集会

<http://www.pm.org/groups/asia.shtml#China>

中国 Perl 推广组一览

Unicode 相关网址

<http://www.unicode.org/>

Unicode 学术学会 (Unicode 标准的制定者)

<http://www.cl.cam.ac.uk/%7Emgk25/unicode.html>

Unix/Linux 上的 UTF-8 及 Unicode 问答问

3 SEE ALSO

Encode, Encode::CN, encoding, perluniintro, perlunicode

4 AUTHORS

Jarkko Hietaniemi <jhi@iki.fi>

Autrijus Tang (唐宗汉) <autrijus@autrijus.org>