字符编码应用

版本：v0.7.0

Crifan Li

摘要

本文主要介绍了字符编码的各种应用，包括常见的字符编码有哪些类型，常见的各种编码规范和编程语言中对于各种字符编码的支持，常见的各种软件和工具对于字符编码的支持，常见的不同操作系统中对于不同编码的支持。

本文提供多种格式供：

在线阅读	HTML [1]	HTMLs [2]	PDF [3]	CHM [4]	TXT [5]	RTF [6]	WEBHELP [7]
下载（7zip压缩包）	HTML [8]	HTMLs [9]	PDF [10]	CHM [11]	TXT [12]	RTF [13]	WEBHELP [14]

HTML版本的在线地址为：

http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/html/char_encoding_usage.html

有任何意见，建议，提交bug等，都欢迎去讨论组发帖讨论：

http://www.crifan.com/bbs/categories/char_encoding_usage/

修订历史		
修订 0.7.0	2015-05-24	crl
1. 介绍字符编码的应用 2. 添加字符编码简明教程的链接		

[1] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/html/char_encoding_usage.html
[2] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/htmls/index.html
[3] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/pdf/char_encoding_usage.pdf
[4] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/chm/char_encoding_usage.chm
[5] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/txt/char_encoding_usage.txt
[6] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/rtf/char_encoding_usage.rtf
[7] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/webhelp/index.html
[8] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/html/char_encoding_usage.html.7z
[9] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/htmls/index.html.7z
[10] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/pdf/char_encoding_usage.pdf.7z
[11] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/chm/char_encoding_usage.chm.7z
[12] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/txt/char_encoding_usage.txt.7z
[13] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/rtf/char_encoding_usage.rtf.7z
[14] http://www.crifan.com/files/doc/docbook/char_encoding_usage/release/webhelp/char_encoding_usage.webhelp.7z

字符编码应用:

Crifan Li

版本：v0.7.0

出版日期 2015-05-24

[15] http://www.crifan.com/files/doc/docbook/soft_dev_basic/release/html/soft_dev_basic.html#cc_by_nc

目录

正文之前

1. 目的

本文旨在讲清楚字符编码的在各种不同语言，环境，工具，操作系统等方面的应用。

2. 声明

任何问题、意见、建议等，都欢迎发邮件一起探讨：admin (at) crifan.com。

第 1 章 常见字符编码类型

关于常见的字符编码的

- 详尽解释，可以参考：字符编码详解[1]
- 精简介绍，可以参考：字符编码简明教程[2]

接着再去看下面的介绍，则就很容易理解了：

目前一些常见的字符编码类型有：

1. 最常见编码类型：ascii utf8 gbk
2. 其次：gb2312,iso8859-1
3. 再次是其他的编码

[1] http://www.crifan.com/files/doc/docbook/char_encoding/release/html/char_encoding.html
[2] http://www.crifan.com/character_encoding_charset_simpile_tutorial/

第 2 章 常见的规范和编程语言中对于字符编码的支持

也可以叫做：字符编码在常见规范和语言中的应用

2.1. XML的编码声明

xml文件中，第一行的内容，就包含了字符编码声明

比如：

```
<?xml version='1.0' encoding="utf-8"?>
```

其中指定了编码类型encoding为UTF-8类型。

2.2. JSON中的字符编码

参考：序列化Python对象 - 深入Python 3[1]

第三，字符编码的问题是长期存在的。JSON 用纯文本编码数据，但是你知道，“不存在纯文本这种东西。” JSON必须以Unicode 编码(UTF-32, UTF-16, 或者默认的, UTF-8)方式存储, RFC 4627的第3节定义了如何区分使用的是哪种编码。

看起来好像说的是：JSON编码必须是Unicode编码？

有空再去确认。

2.3. HTML的charset

HTML中是通过charset来声明当前网页所使用的字符编码是何种类型的

最常见的比如：

```
<meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
```

其中指定了编码类型charset为UTF-8类型。

2.4. Python中的字符编码

Python 2.x中的str和unicode

Python 3.x中的bytes和str

[1] http://sebug.net/paper/books/dive-into-python3/serializing.html

系统默认编码：sys.defencoding？？？是ANSI

-->>导致很多人在使用Python时出现编码错误后，结果去想办法修改sys.defencoding的值

-->>实际上不应该这么做，而应该是搞清楚自己要处理的python文件的编码是什么，然后加上对应的编码声明

详见：

【整理】Python中用encoding声明的文件编码和文件的实际编码之间的关系 [2]

【整理】关于Python脚本开头两行的：#!/usr/bin/python和# -*- coding: utf-8 -*-的作用 – 指定文件编码类型 [3]

【总结】Python 2.x中常见字符编码和解码方面的错误及其解决办法 [4]

更多内容可参考另外一个教程：

Python专题教程：字符串和字符编码 [5]

Python中有个库，专门用来检测字符串是什么编码的。

详见：【已解决】windows下，安装python的chardet | 在路上[6]

2.5. C#中的字符编码

在用C#进行网页处理时，网络爬虫抓取得到网页的原始字符串后，需要解码才能得到unicode的字符串

详见：【整理】关于HTML网页源码的字符编码（charset）格式（GB2312，GBK，UTF-8，ISO8859-1等）的解释 | 在路上[7]中的"从原始html中解码为对应的unicode字符串"

2.6. Java中的字符编码

Java中第三方库函数可以用来动态地检测字符的编码类型：

- juniversalchardet [8]
- jchardet [9]
- cpdetector [10]
- ICU4J [11]

详见：java - What is the most accurate encoding detector? - Stack Overflow[12]

[2] http://www.crifan.com/python_string_encoding_declare_encoding_vs_file_real_encoding/
[3] www.crifan.com/python_head_meaning_for_usr_bin_python_coding_utf-8/
[4] http://www.crifan.com/summary_python_2_x_common_string_encode_decode_error_reason_and_solution/
[5] http://www.crifan.com/files/doc/docbook/python_topic_str_encoding/release/html/python_topic_str_encoding.html
[6] http://www.crifan.com/resolved_windows_install_the_python_chardet/
[7] http://www.crifan.com/summary_explain_what_is_html_charset_and_common_value_of_gb2312_gbk_utf_8_iso8859_1/
[8] http://code.google.com/p/juniversalchardet/
[9] http://jchardet.sourceforge.net/
[10] http://cpdetector.sourceforge.net/
[11] http://userguide.icu-project.org/conversion/detection
[12] http://stackoverflow.com/questions/3759356/what-is-the-most-accurate-encoding-detector

第 3 章 常见工具中所涉及的字符编码

字符编码在常见软件中的应用

3.1. Notepad中的字符编码

notepad 支持不多但也有最基本的。这就是别人拿来演示神奇的保存一个词 联通 打开后显示其他字符的奥秘所在

TODO：抽空找到网上流传很广的那个，写一个联还是通，换个编码保持，结果再打开显示出另外一个字的效果

3.2. Notepad2中的字符编码

3.3. Notepad++中的字符编码

详见：

【crifan推荐】轻量级文本编辑器，Notepad最佳替代品：Notepad++[1]中的：Notepad++的多种编码支持[2]

3.4. Eclipse中的字符编码

Eclipse中的默认的（文件的字符）编码

3.5. Sublime中的字符编码

sublime 也可以保存成不同编码，虽然支持的不够多

但是对于及时显示当前文件的编码，支持的不是足够好：

【基本解决】Sublime中竟然不能方便地找到哪里可以显示当前文件的编码类型 | 在路上 [3]

3.6. UltraEdit中的字符编码

[1] http://www.crifan.com/files/doc/docbook/rec_soft_npp/release/html/rec_soft_npp.html
[2] http://www.crifan.com/files/doc/docbook/rec_soft_npp/release/html/rec_soft_npp.html#npp_func_multi_enc
[3] http://www.crifan.com/how_to_show_current_file_encoding_in_sublime/

第 4 章 常见系统中的对于不同编码的支持

4.1. Windows系统中的字符编码

Windows中系统中的CodePage

cmd的默认编码是ANSI

参考书目

[1] 【转】字符编码笔记：ASCII，Unicode和UTF-8[1]

[1] http://www.crifan.com/switch_character_encoding_notes_ascii_unicode_and_utf-8/