字符文件编码

kevinluo

Contents

contents

1 参考链接

libiconv gnu 官方 itro&download

浅析 windows 下字符集和文件编码存储/utf8/gbk

UNICODE 编码 UTF-16 中的 BigEndian（FEFF）和 LittleEndian（FFFE）形象描述

2 字符编码小知识

中文字集进化，GB2312->GBK 通称他们叫做”DBCS”（Double Byte Charecter Set 双字节字符集）。

中文 windows notepad 存盘默认用的 ansi 编码，也就是对应 gbk 字符集。

2.1 1，字符集

这里主要讲两种字符集，DBCS 和 UCS

UCS 规定如何编码，

UTF 规定如何传输、保存这个编码。UTF8、UTF7、UTF16 都是被广泛接受的方案。

2.2 2，BOM

BOM 是在一个文本文件之前，用来标记改文件编码方式的一种记录方式，windows 下是这样做的，linux 不知道。

UCS 编码中”ZERO WIDTH NO-BREAK SPACE” 的字符，它的编码是 FEFF。而 FFFE 在 UCS 中是不存在的字符。

FEFF，就表明这个字节流是 Big-Endian 的 FFFE，就表明这个字节流是 Little-Endian 的。

UTF8 不需要 BOM 来表明字节顺序，但可以用 BOM 来表明编码方式。EFBBBF，就知道这是 UTF8 编码。

假如文件用 UTF8 无 BOM 格式来保存文件，那就不能靠 BOM 头来判断是否是 utf8 编码的，而要对文件中的数据进行简单的编码分析来确定文件的编码格式，也就是对文件的二进制进行分析，和对应编码的字符集进行匹配，最终确定其编码格式。