字符文件編碼

kevinluo

Contents

contents

1 參考鏈接

libiconv gnu 官方 itro&download

淺析 windows 下字符集和文件編碼存儲/utf8/gbk

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

2 字符編碼小知識

中文字集進化，GB2312->GBK 通稱他們叫做”DBCS”（Double Byte Charecter Set 雙字節字符集）。

中文 windows notepad 存盤默認用的 ansi 編碼，也就是對應 gbk 字符集。

2.1 1，字符集

這裏主要講兩種字符集，DBCS 和 UCS

UCS 規定如何編碼，

UTF 規定如何傳輸、保存這個編碼。UTF8、UTF7、UTF16 都是被廣泛接受的方案。

2.2 2，BOM

BOM 是在一個文本文件之前，用來標記改文件編碼方式的一種記錄方式，windows 下是這樣做的，linux 不知道。

UCS 編碼中”ZERO WIDTH NO-BREAK SPACE” 的字符，它的編碼是 FEFF。而 FFFE 在 UCS 中是不存在的字符。

FEFF，就表明這個字節流是 Big-Endian 的 FFFE，就表明這個字節流是 Little-Endian 的。

UTF8 不需要 BOM 來表明字節順序，但可以用 BOM 來表明編碼方式。EFBBBF，就知道這是 UTF8 編碼。

假如文件用 UTF8 無 BOM 格式來保存文件，那就不能靠 BOM 頭來判斷是否是 utf8 編碼的，而要對文件中的數據進行簡單的編碼分析來確定文件的編碼格式，也就是對文件的二進制進行分析，和對應編碼的字符集進行匹配，最終確定其編碼格式。