|  |  |
| --- | --- |
| From: | honda@luckstar.com.tw |
| Date: | 2016-02-06 |
| Subject: | 文字檔案內碼 |

以下說明文字檔案內碼儲存格式:

# 文字編輯器Windows記事本NotePad

輸入: 123中文ABC, 按[Enter], 存檔為不同的檔案格式.



Windows 記事本NotePad存檔格式選項為: ANSI, Unicode, Unicode big endian(正序: Byte Order Mark=Big Endian), UTF-8.

選擇ANSI(預設)時,若為英文版WINDOWS, 則存為ASCII編碼, 繁體中文版存為Big5編碼, 簡體中文版存為GB2312編碼.

|  |  |  |
| --- | --- | --- |
| 記事本NotePad存檔格式 | Windows 版本 | 內碼 |
| ANSI | 英文版 | ASCII |
| 繁體中文版 | Big5 |
| 簡體中文版 | GB2312 |
| Unicode |  | Uincode |
| Unicode big endian |  | Unicode BOM(正序: Byte Order Mark=Big Endian) |
| UTF-8 |  | UTF-8 BOM(正序: Byte Order Mark=Big Endian) |
| ANSI |  | ANSI 或 UTF-8-Little Endian.  此格式需靠讀取中文2碼或3碼判別是ANSI還是UTF-8 |

# 檔案格式

# ANSI

ANSI格式無前置識別碼



# UTF-8即UTF-8-BOM

前置識別碼3bytes = ef bb bf

Windows Notepad存檔為UTF-8格式, 實際上為UTF-8-BOM(正序: Byte Order Mark=Big Endian), 中文內碼為3碼長度. 此格式常容易與UNIX系統程序如shell解析衝突.





# UTF-8-Little Endian

無前置識別碼. 需靠中文編碼(2碼或3碼)識別是ANSI還是UTF-8 Little Endian ?

UTF-8-NOBOM格式與ANSI格式相同, 無前置識別碼. 需靠中文編碼(2碼或3碼)識別. 此格式與Unix文字檔相容性較高.

UTF-8: 中文內碼長度為3碼



ANSI: 中文內碼長度為2碼



# Unicode

前置識別碼2bytes = FF FE



# Unicode Big Endian

前置識別碼2bytes = FE FF

