

Character Encoding

Nguyễn Văn Vũ

07/12/2012

Nội dung

- Các ví dụ liên quan đến character encoding
- Nguyên nhân & cách giải quyết
- Các vấn đề khác liên quan đến character encoding
- Cơ sở của việc mã hóa Unicode-UTF8
- Sự tương quan giữa dữ liệu (nhìn thấy được) và dữ liệu được lưu trữ (datafile)

Những cái nhìn đầu tiên...

- Cuộc đời vẫn đẹp sao....nà ná na nà ná na...

Buổi mai hôm ấy. Một buổi mai đầy sương thu và gió lạnh.↓

Mẹ tôi âu yếm dắt tôi đi trên con đường dài và hẹp.↓

Con đường này tôi đã đi lại mấy lần.↓

Nhưng hôm nay tôi bỗng thấy lạ: Hôm nay tôi đi học.↓

- Siu nhưn GAO =))

- Tình yêu chớm phai màu....ố ô ố ô ố ô - Còn tuổi nào cho em???

Buá»•i mai hă' m á°¥y. Má»™t buá»•i mai Ä'á°Sy sÆ°Æ;ng thu vă giĂ³ lá°;nh.↓
 Má°¹ tĂ' i Äcư yá°;m dá°-t tĂ' i Ä' i trĂ°n con Ä'Æ°á»ng dă i vă há°¹p.↓
 Con Ä'Æ°á»ng nă y tĂ' i Ä'Ä£ Ä' i lá°;i má°¥y lá°Sn.↓
 NhÆ°ng hă' m nay tĂ' i bá»-ng thá°¥y lá°;: Hă' m nay tĂ' i Ä' i há»c.↓
 ↓

- Vì sao nên nổi???? Sai font...???

- Và còn đâu lời thề non hẹn biển...@#\$%#@#%
 - Người ra đi, lệ sầu úa trên mi gầy...

Bu?i mai hôm ?y. M?t bu?i mai d?y suong thu và gió l?nh.↓
M? tôi âu y?m d?t tôi đi trên con du?ng dài và h?p.↓
Con du?ng này tôi đã đi l?i m?y l?n.↓
Nhưng hôm nay tôi b?ng th?y l?: Hôm nay tôi đi h?c.↓

- Nên nổi vì sao??? Sai font, bla bla bla...???

Cái thấy được và Cái lưu trữ

- “Cái thấy được” quyết định trên cơ sở nào?
 - Font
 - Bảng mã
 - Editor/Viewer
- “Cái lưu trữ” quyết định trên cơ sở nào?
 - Bảng mã
 - Mã hóa

Từ những tìm hiểu đầu tiên...

- TCVN-2, TCVN-3(ABC), VNI Windows
- ISO-8859-1, latin1, cp1252,
- Unicode-UTF8, Unicode-UTF16
-
- ➔ what the hell???

Khám phá...

- Cái thấy được và Cái lưu trữ

42	75	E1	BB	95	69	20	6D	61	69	20	68	C3	B4	6D	20	E1	B	u	á	»	•	i	m	a	i	h	Ả	'	m	á
BA	A5	79	2E														°	¥	y	.										

```
>>> s = "Ễ"
```

```
>>> s
```

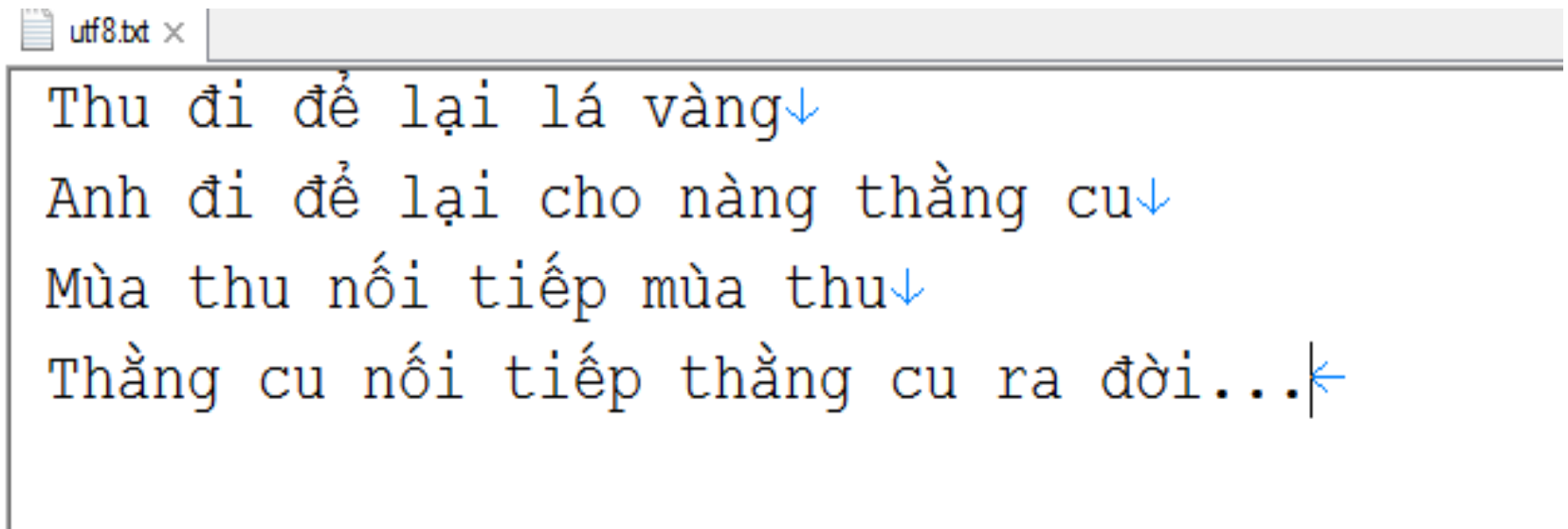
```
'Ễ'
```

```
>>> s.encode()
```

```
b'\xe1\xbb\x84'
```


Editor & Unicode (UTF-8)

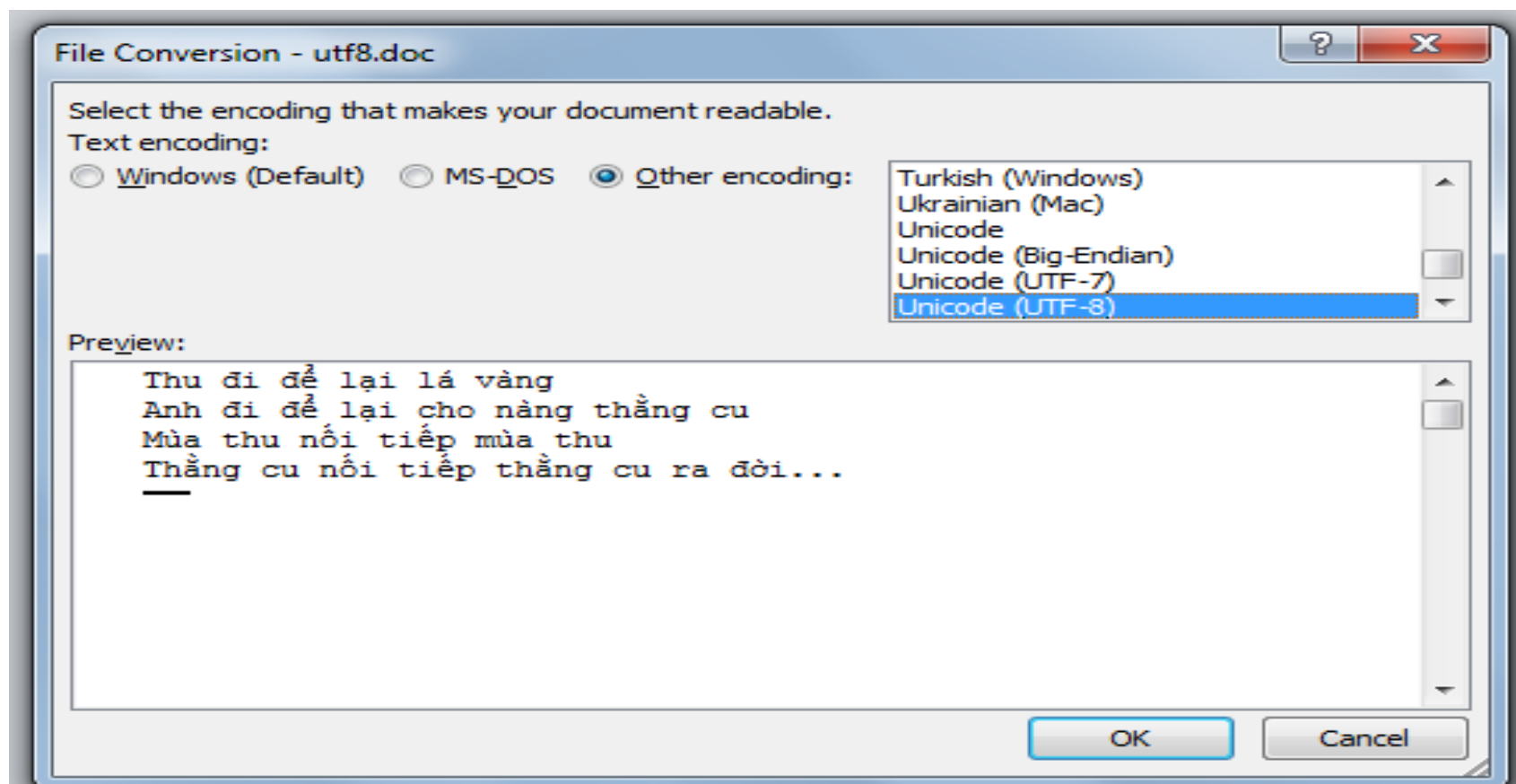
- Text



Thu đi để lại lá vàng↓
Anh đi để lại cho nàng thằng cu↓
Mùa thu nối tiếp mùa thu↓
Thằng cu nối tiếp thằng cu ra đời...↵

MS-Word & Unicode (UTF8)

- MS-Word hỗ trợ Unicode (UTF8) w/ or wo/ BOM



MS-Excel & Unicode

	A	B	C	D	E
1	Thu Ä'í Ä'á»f láºì lÄi vÄ ng				
2	Anh Ä'í Ä'á»f láºì cho nÄ ng tháº±ng cu				
3	MÄ¹a thu ná»'í tiáº¿p mÄ¹a thu				
4	Tháº±ng cu ná»'í tiáº¿p tháº±ng cu ra Ä'á»i...				
5					

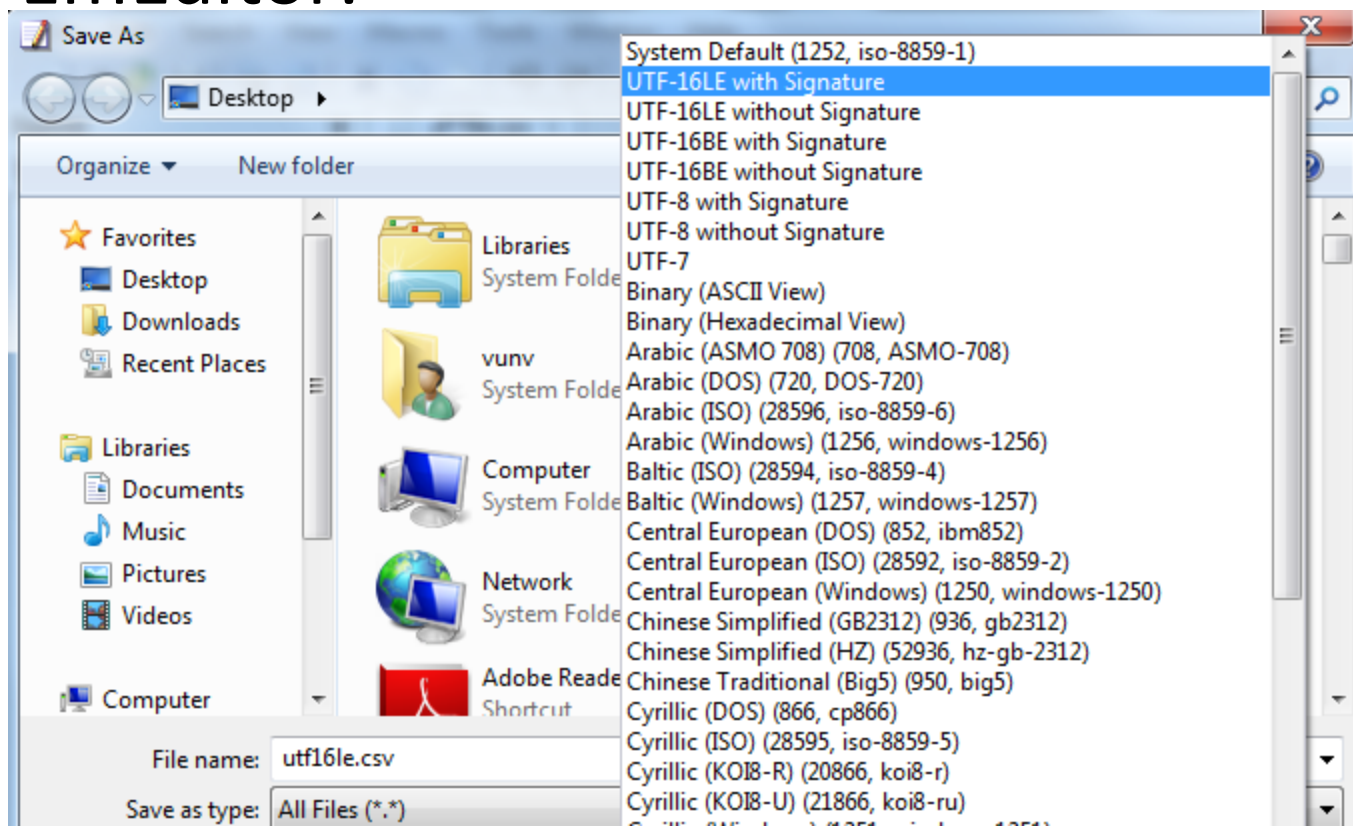
- MS-Excel không hỗ trợ Unicode UTF8, chỉ hỗ trợ Unicode UTF16-LE → đối với dữ liệu được encode ở UTF8 cần chuyển đổi sang UTF16-LE w/BOM

MS-Excel & Unicode

- PHP:

`$value = chr(255).chr(254).mb_convert_encoding($value,'UTF-16LE', 'UTF-8');`

- EmEditor:



Lưu trữ: Tập tin (File)

- UTF8 (BOM) - removed
- Line-feed (End Of File – EOF)
 - MAC: CR
 - Linux/Unix: LF
 - Windows: CR+LF

Cái lưu trữ: cơ sở dữ liệu

- Binary
- Text
- Collation
 - latin1_swedish_ci
 - latin1_swedish_cs
 - utf8_unicode_ci
 - utf8_unicode_cs
 -
- Và còn cái gì nữa????

Import & Export: cơ sở dữ liệu

- Import
 - Encoding của nguồn dữ liệu(dữ liệu): latin1, tcvn3, utf8,....
 - Encoding của database(data file): latin1, utf8,jis,...
- Export
 - Encoding của dữ liệu
 - Encoding của table

Import & Export: cơ sở dữ liệu

- Import
 - LOAD DATA LOCAL INFILE *file* INTO TABLE *table*
CHARACTER SET '*utf8*'
- Export
 - mysql -e "" -default-character-set=*charset_name*
 - *charset_name==table_charset==datafile_charset*

Ảnh hưởng của collation

- latin1_swedish_ci
- utf8_unicode_ci
- ➔ Select ← thực hiện được không?
- ➔ Join ← cái này thì sao ???
- <http://dev.mysql.com/doc/refman/5.1/en/charset-repertoire.html>

MySQL & Unicode(UTF8)

- MySQL thật sự không hỗ trợ UTF8 một cách đầy đủ
- Các vấn đề gặp phải: primary key (text)
- Giải pháp để có 1 db hỗ trợ utf8 dùng được:
 - Tạo db gốc với encoding latin1
 - Alter table change encoding to utf8

MySQL & Unicode (UTF8)

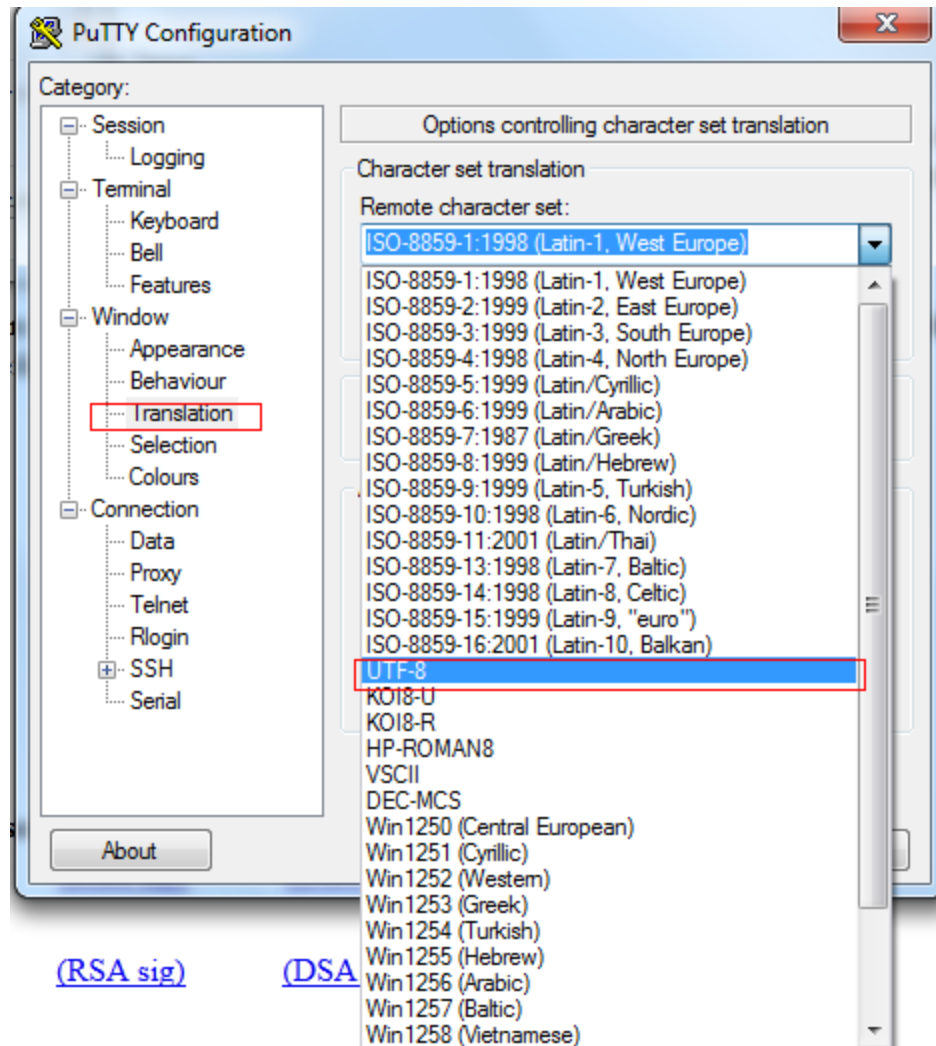
- Có thể load 1 file data với encoding là TCVN3 vào 1 table/database có encoding là UTF8 hay không?
- Trong trường hợp này cần giải pháp như thế nào?
- Table encoding latin1 + data UTF8 ← điều gì sẽ xảy ra?
- Table encoding utf8 + data utf8

Các thứ linh tinh khác

- Keyboard, keycode, các bộ gõ ở đâu trong bức tranh này????
 - Unikey
 - Vietkey
- Các bộ chuyển đổi (converter)
 - Iconv
 - Uvconv

- Character encoding bị ảnh hưởng/thay đổi như thế nào khi được chuyển từ nơi này sang nơi khác?
- Khi copy & paste (buffer) giữa các trình soạn thảo (editor)?
- Khi chuyển đổi qua lại giữa local và server?
- Sự ảnh hưởng/tác động của `connection_charset`

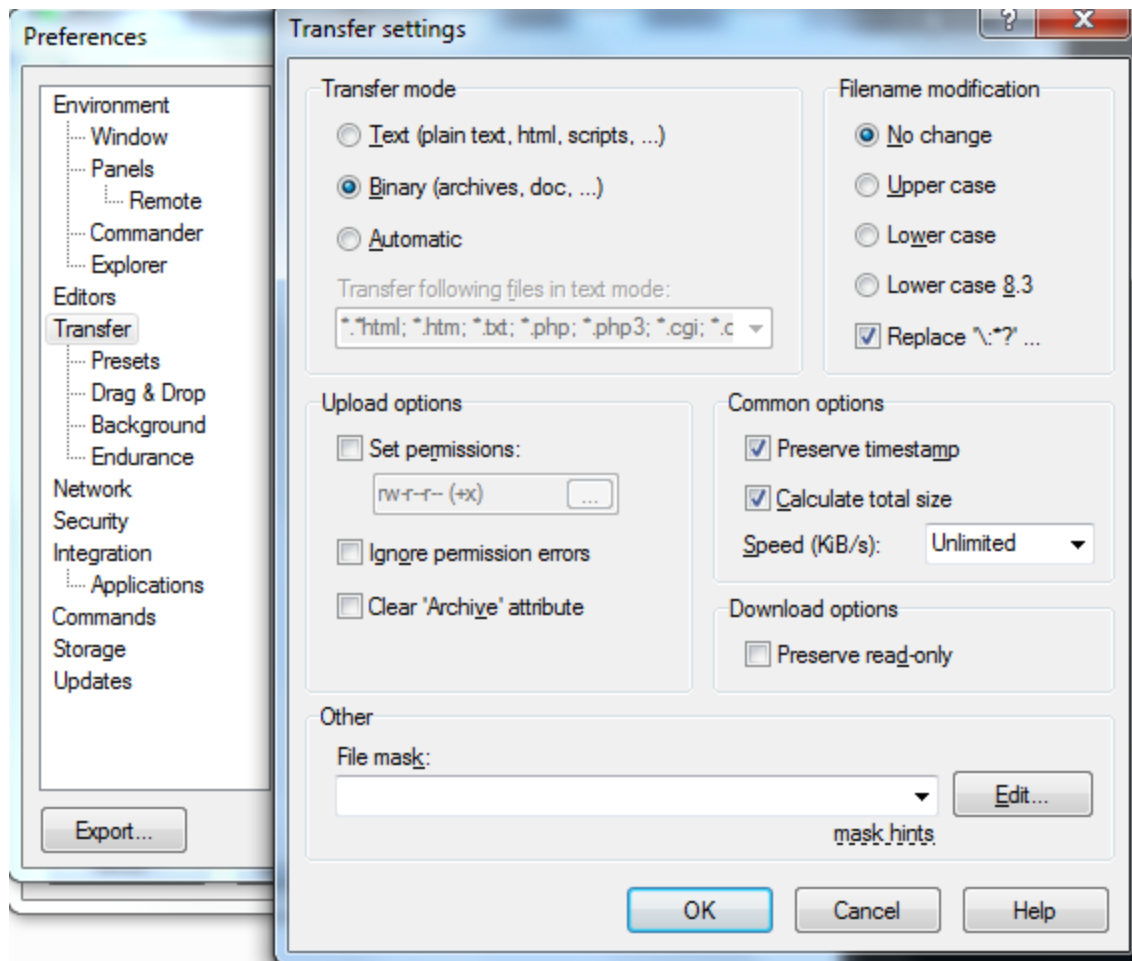
Xác lập trên các công cụ thường dùng



(RSA sig)

(DSA

- `export LANG=en_US.UTF8`



Hiện thực mã hóa Unicode-UTF8

- $a = 0x1EA7$
- $a1 = a \gg 12$
 $a2 = (a \gg 6) \& 0b0000111111$
 $a3 = a \& 0b0000000000111111$
- $c1 = 0b11100000 \mid a1$
 $c2 = 0b10000000 \mid a2$
 $c3 = 0b10000000 \mid a3$
- $d1 = c1 \& 0xf$
 $d2 = c2 \& 0x3f$
 $d3 = c3 \& 0x3f$
- $e1 = d1 \ll 12$
 $e2 = d2 \ll 6$
 $e3 = d3$
- $e = e1 \mid e2 \mid e3$
- $e == a$
- <http://home.tiscali.nl/t876506/utf8tbl.html>