Switching character sets with the engine running (on the web)

CP1252 to UTF-8 in 4 easy steps

Character sets

- Map bytes to characters
- To use Unicode terminology: Each character is represented by a “code point”, but you can encode a “code point” in different ways.

The old web default

- latin1, AKA ISO-8859-1
- Except…

“latin I”
is pronounced
“cp1252”

- When a user agent [browser] would otherwise use a character encoding given in the first column [ISO-8859-1, aka latin1] of the following table to either convert content to Unicode characters or convert Unicode characters to bytes, it must instead use the encoding given in the cell in the second column of the same row [windows-1252, aka cp1252].

http://mail.python.org/pipermail/python-list/2012-November/635240.html

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_																
1_																
2_	SP 0020 *32*	! 0021 *33*	" 0022 *34*	# 0023 *35*	$ 0024 *36*	% 0025 *37*	& 0026 *38*	' 0027 *39*	(0028 *40*	) 0029 *41*	* 002A *42*	+ 002B *43*	, 002C *44*	- 002D *45*	. 002E *46*	/ 002F *47*
3_	0 0030 *48*	1 0031 *49*	2 0032 *50*	3 0033 *51*	4 0034 *52*	5 0035 *53*	6 0036 *54*	7 0037 *55*	8 0038 *56*	9 0039 *57*	: 003A *58*	; 003B *59*	< 003C *60*	= 003D *61*	> 003E *62*	? 003F *63*
4_	@ 0040 *64*	A 0041 *65*	B 0042 *66*	C 0043 *67*	D 0044 *68*	E 0045 *69*	F 0046 *70*	G 0047 *71*	H 0048 *72*	I 0049 *73*	J 004A *74*	K 004B *75*	L 004C *76*	M 004D *77*	N 004E *78*	O 004F *79*
5_	P 0050 *80*	Q 0051 *81*	R 0052 *82*	S 0053 *83*	T 0054 *84*	U 0055 *85*	V 0056 *86*	W 0057 *87*	X 0058 *88*	Y 0059 *89*	Z 005A *90*	[005B *91*	\ 005C *92*	] 005D *93*	^ 005E *94*	_ 005F *95*
6_	` 0060 *96*	a 0061 *97*	b 0062 *98*	c 0063 *99*	d 0064 *100*	e 0065 *101*	f 0066 *102*	g 0067 *103*	h 0068 *104*	i 0069 *105*	j 006A *106*	k 006B *107*	l 006C *108*	m 006D *109*	n 006E *110*	o 006F *111*
7_	p 0070 *112*	q 0071 *113*	r 0072 *114*	s 0073 *115*	t 0074 *116*	u 0075 *117*	v 0076 *118*	w 0077 *119*	x 0078 *120*	y 0079 *121*	z 007A *122*	{ 007B *123*	| 007C *124*	} 007D *125*	~ 007E *126*	
8_																
9_																

latin I

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_	NUL 0000 *0*	SOH 0001 *1*	STX 0002 *2*	ETX 0003 *3*	EOT 0004 *4*	ENQ 0005 *5*	ACK 0006 *6*	BEL 0007 *7*	BS 0008 *8*	HT 0009 *9*	LF 000A *10*	VT 000B *11*	FF 000C *12*	CR 000D *13*	SO 000E *14*	SI 000F *15*
1_	DLE 0010 *16*	DC1 0011 *17*	DC2 0012 *18*	DC3 0013 *19*	DC4 0014 *20*	NAK 0015 *21*	SYN 0016 *22*	ETB 0017 *23*	CAN 0018 *24*	EM 0019 *25*	SUB 001A *26*	ESC 001B *27*	FS 001C *28*	GS 001D *29*	RS 001E *30*	US 001F *31*
2_	SP 0020 *32*	! 0021 *33*	" 0022 *34*	# 0023 *35*	$ 0024 *36*	% 0025 *37*	& 0026 *38*	' 0027 *39*	(0028 *40*	) 0029 *41*	* 002A *42*	+ 002B *43*	, 002C *44*	- 002D *45*	. 002E *46*	/ 002F *47*
3_	0 0030 *48*	1 0031 *49*	2 0032 *50*	3 0033 *51*	4 0034 *52*	5 0035 *53*	6 0036 *54*	7 0037 *55*	8 0038 *56*	9 0039 *57*	: 003A *58*	; 003B *59*	< 003C *60*	= 003D *61*	> 003E *62*	? 003F *63*
4_	@ 0040 *64*	A 0041 *65*	B 0042 *66*	C 0043 *67*	D 0044 *68*	E 0045 *69*	F 0046 *70*	G 0047 *71*	H 0048 *72*	I 0049 *73*	J 004A *74*	K 004B *75*	L 004C *76*	M 004D *77*	N 004E *78*	O 004F *79*
5_	P 0050 *80*	Q 0051 *81*	R 0052 *82*	S 0053 *83*	T 0054 *84*	U 0055 *85*	V 0056 *86*	W 0057 *87*	X 0058 *88*	Y 0059 *89*	Z 005A *90*	[005B *91*	\ 005C *92*	] 005D *93*	^ 005E *94*	_ 005F *95*
6_	` 0060 *96*	a 0061 *97*	b 0062 *98*	c 0063 *99*	d 0064 *100*	e 0065 *101*	f 0066 *102*	g 0067 *103*	h 0068 *104*	i 0069 *105*	j 006A *106*	k 006B *107*	l 006C *108*	m 006D *109*	n 006E *110*	o 006F *111*
7_	p 0070 *112*	q 0071 *113*	r 0072 *114*	s 0073 *115*	t 0074 *116*	u 0075 *117*	v 0076 *118*	w 0077 *119*	x 0078 *120*	y 0079 *121*	z 007A *122*	{ 007B *123*	| 007C *124*	} 007D *125*	~ 007E *126*	DEL 007F *127*
8_	€ 20AC *128*		‚ 201A *130*	ƒ 0192 *131*	„ 201E *132*	… 2026 *133*	† 2020 *134*	‡ 2021 *135*	ˆ 02C6 *136*	‰ 2030 *137*	Š 0160 *138*	‹ 2039 *139*	Œ 0152 *140*		Ž 017D *142*	
9_		‘ 2018 *145*	’ 2019 *146*	“ 201C *147*	” 201D *148*	• 2022 *149*	– 2013 *150*	— 2014 *151*	˜ 02DC *152*	™ 2122 *153*	š 0161 *154*	› 203A *155*	œ 0153 *156*		ž 017E *158*	Ÿ 0178 *159*

CP1252

But the web is all Unicode, right?

- That would be…easy.
- But don’t we all love
  - ☃
  - (ﾉ °□°) ﾉ ⁀ ┻━━┻

- When you POST a snowman into a browser on a page served as latin1,
  - the browser submits the ASCII bytes `&#9731;`
- Uh oh…

- How do you tell the difference between having typed ☃ and `&#9731;`?

- You don't. :(

- So don't use latin1, because you'll be mangling user data, and fail to round-trip things in bad ways.

- And you probably didn't use a standard web escaper function, and instead replaced <, >, and " with `<`, `>`, and `&quot;`, so that user text almost looked okay.

- Let's fix it!

Step 1

- If you're using MySQL, and your tables were `charset=latin1`, you're in luck because
- MySQL also pronounces it as "CP1252"!

Step 1

- `ALTER TABLE x DEFAULT CHARSET utf8mb4, MODIFY y VARCHAR(20) CHARACTER SET utf8mb4 COLLATE utf8mb4_unicode_ci`
- (MySQL's utf8 charset is limited to codepoints between U+0000 and U+FFFF; ☃ will work, but 🎲 will not)

Step 1

- If you continue to ONLY connect to the DB using a latin1 connection charset, you don't have to worry about unrepresentable characters!

Step 2

- Upgrade your app to speak UTF-8 on the web, and handle UTF-8 data internally
- Change the DB connection charset to utf8
- (I had to do this for a PHP app, where everything is a bytestring in an unspecified encoding; the implicit charset went from cp1252+entities to utf8)

Step 3

- Trawl the database looking for HTML entities, and convert them to UTF-8.

- Caveats:
  - ' and friends will exist
  -   and friends will exist
  - ☃ and friends will exist
  - 𐀀 and friends will exist
  - Surrogate pairs will exist
  - € means €, not a C1 control code

Step 4

- Stop letting entities through your escaper (use a template language's default escaper!), as all your text is now proper

`text/plain`!

Enjoy the UTF8!

Be sure to read
http://www.joelonsoftware.com/articles/Unicode.html
and
http://nedbatchelder.com/text/unipain/unipain.html

Thanks!
taavi@taaviburns.ca – @jaaaarel
PyCon Canada 2013