

Marco Scheurer  
@phink0

Following

2014. From Geneva Airport trains are running to **Genve, Neuchtel and Zrich.**

Reply Retweet Favorited More

Trains au départ de Genève-Aéroport				
Catégories	Heure	Destinations suisses	Voie	
IR	12:53	Genve Nyon Lausanne	Brig	3
ICN	13:09	Genve Neuchtel Biel/Bienne	Basel SBB	4
IR	13:23	Genve Nyon Lausanne	Brig	3
IR	13:53	Genve Nyon Lausanne	Brig	3
ICN	14:09	Genve Olten Zrich HB	St. Gallen	4
IR	14:23	Genve Nyon Lausanne	Brig	3
IR	14:53	Genve Nyon Lausanne	Brig	3
ICN	15:09	Genve Neuchtel Biel/Bienne	Basel SBB	4
IR	15:23	Genve Nyon Lausanne	Brig	3
IR	15:53	Genve Nyon Lausanne	Brig	3
ICN	16:09	Genve Olten Zrich HB	St. Gallen	4
IR	16:23	Genve Nyon Lausanne	Sion	3
IR	16:53	Genve Nyon Lausanne	Brig	3

# Unicode Hacks

Nicolas Seriot

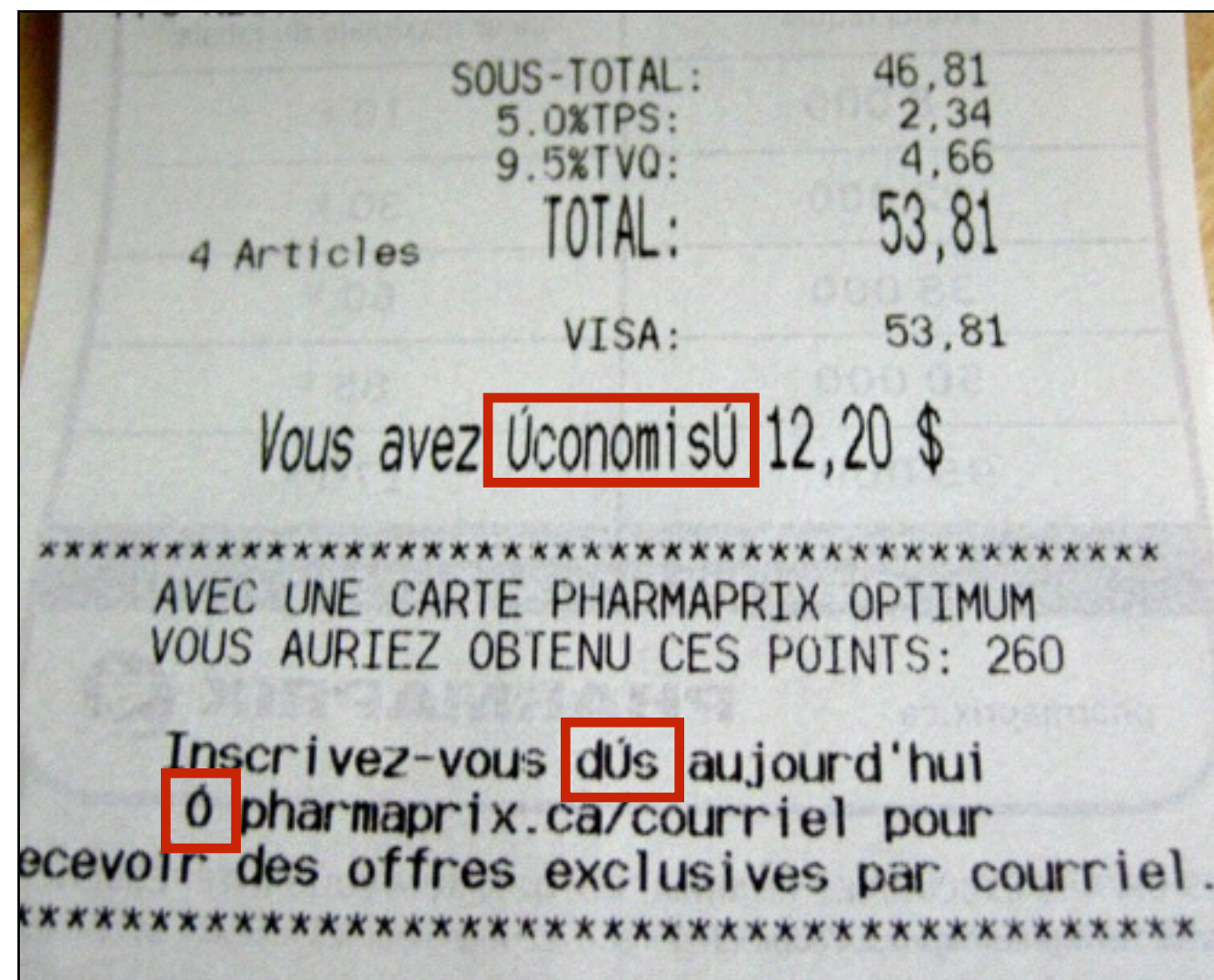
November 6th, 2014



Application Security Forum - 2014

Western Switzerland

5-6 novembre 2014  
Y-Parc / Yverdon-les-Bains



<http://unicode-wall-of-shame.com>


●●●○○ Sunrise E 23:01 49 %

< Settings Personal Hotspot

Personal Hotspot



Now Discoverable.


Other users can look for your shared network using Wi-Fi under the name **'nst 5** ".

Wi-Fi Password

test1234 >



TO CONNECT USING WI-FI

- 1 Choose "nst 5  " from the Wi-Fi settings on your computer or other device.
- 2 Enter the password when prompted.

TO CONNECT USING BLUETOOTH

- 1 Pair iPhone with your computer.
- 2 On iPhone, tap Pair or enter the code displayed on your computer.
- 3 Connect to iPhone from computer.

TO CONNECT USING USB

- 1 Plug iPhone into your computer.








- full presentation at SoftShake
- 10 min. / 38 slides → 15.8 s. / slide
- an article is coming...

reddit PROGRAMMING **comments** related other discussions (1)

↑ **I ? Unicode** (seriot.ch)  
 1149 submitted 9 days ago by nst021  
 ↓ 554 comments share save hide delete nsfw retry thumb

**Hacker News** new | threads | comments | show | ask | jobs | submit

\* **I ? Unicode [pdf]** (seriot.ch)  
 98 points by beefburger 10 days ago | comments

 **Nicolas Seriot**  
 @nst021

Weird, huge stats for my Unicode slides.  
 1052k downloads, 59k unique (ip,day,agent).

← ↻ ★ ...

Top 10 of 59088 Total Sites By KBytes						
#	Hits	Files	KBytes	Visits	Hostname	
1	1170	1114	4378369	7	199.16.156.124	
2	1119	1059	4242725	9	199.16.156.126	
3	1189	1125	4237841	5	199.16.156.125	
4	540	540	4022767	0	ec2-54-185-96-103.us-west-2.compute.amazonaws.com	
5	528	528	3933372	0	ec2-54-202-244-76.us-west-2.compute.amazonaws.com	
6	528	528	3933372	0	ec2-54-244-154-122.us-west-2.compute.amazonaws.com	
7	522	522	3888675	0	ec2-54-184-38-108.us-west-2.compute.amazonaws.com	
8	519	519	3866326	0	ec2-54-212-245-232.us-west-2.compute.amazonaws.com	
9	510	510	3799280	0	ec2-54-245-0-218.us-west-2.compute.amazonaws.com	
10	508	508	3784381	0	ec2-54-212-229-209.us-west-2.compute.amazonaws.com	

slides PDF, October 2014

poster PNG, October 2014

2:38 PM - 31 Oct 2014







Baudot Code

1990's: 8 bit encodings

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL	STX	SOT	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
10	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
20	SP	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
30	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
40	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
50	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
60	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
70	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

80																
90																
A0	NEST	!	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	®	¯	°
B0		±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D0	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F0	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ



BCD

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2	SPC	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

1963: ASCII

ISO/IEC 8859-1 (Latin 1)

	00	01	02	03	04	05	06	07	08	09	0A	0B	0C	0D	0E	0F
00	NUL	STX	SOT	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
10	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
20	SP	!	"	#	\$	%	&	'	(	)	*	+	,	-	.	/
30	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
40	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
50	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
60	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
70	p	q	r	s	t	u	v	w	x	y	z	{		}	~	DEL

80																
90																
A0	NEST	!	¢	£	¤	¥	¦	§	¨	©	ª	«	¬	®	¯	°
B0		±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
C0	À	Á	Â	Ã	Ä	Å	Æ	Ç	È	É	Ê	Ë	Ì	Í	Î	Ï
D0	Ð	Ñ	Ò	Ó	Ô	Õ	Ö	×	Ø	Ù	Ú	Û	Ü	Ý	Þ	ß
E0	à	á	â	ã	ä	å	æ	ç	è	é	ê	ë	ì	í	î	ï
F0	ð	ñ	ò	ó	ô	õ	ö	÷	ø	ù	ú	û	ü	ý	þ	ÿ

ISO/IEC 8859-6 (Arabic)

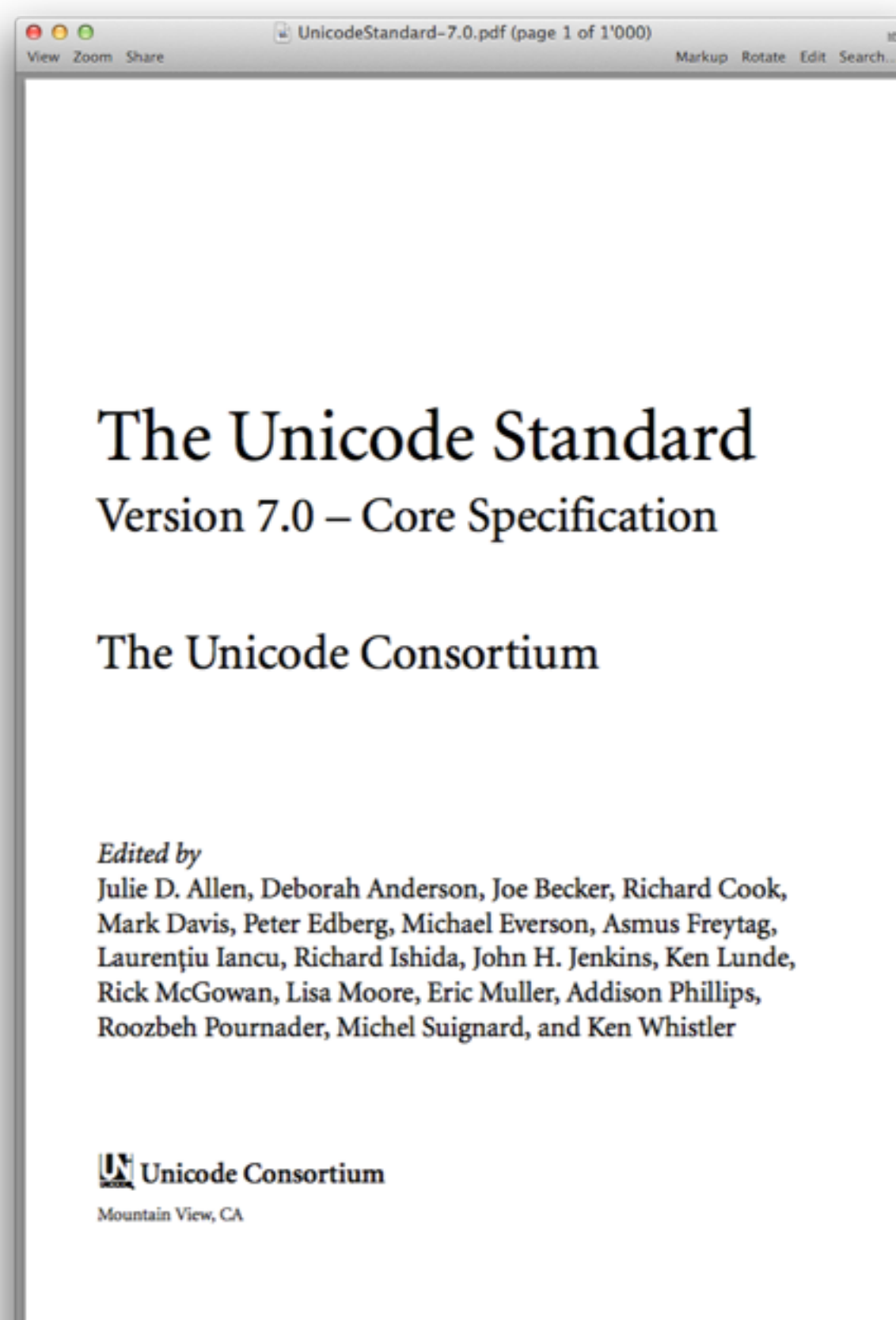


EBCDIC





# The Unicode Consortium

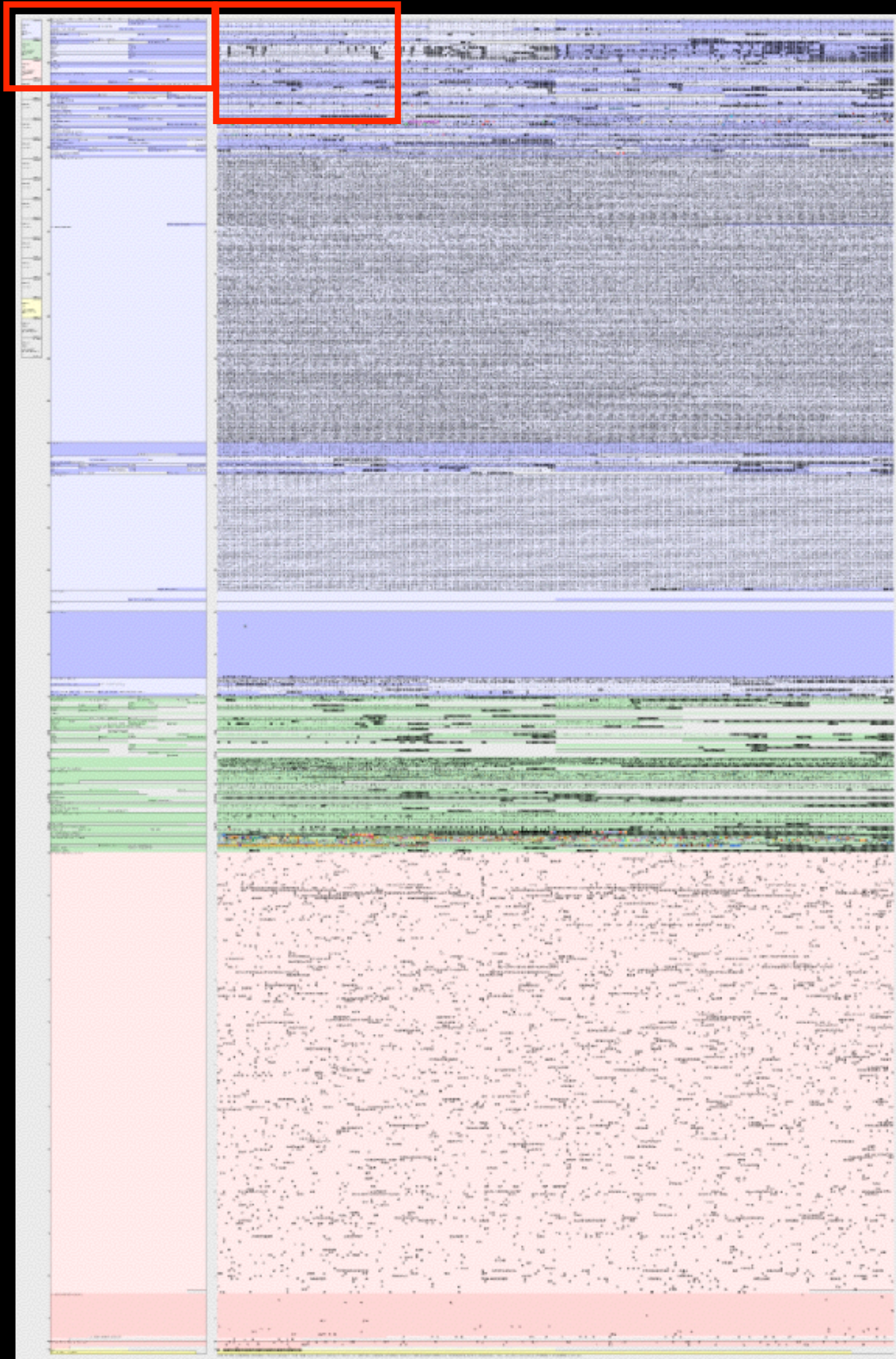


Egyptian Hieroglyphs													
13000													
	1300	1301	1302	1303	1304	1305	1306	1307	1308	1309	130A	130B	130C
0	13000	13010	13020	13030	13040	13050	13060	13070	13080	13090	130A0	130B0	130C0
1	13001	13011	13021	13031	13041	13051	13061	13071	13081	13091	130A1	130B1	130C1
2	13002	13012	13022	13032	13042	13052	13062	13072	13082	13092	130A2	130B2	130C2

Egyptian Hieroglyphs													
13000													
The characters in this block are taken primarily from Alan Gardiner's works on Middle Egyptian.													
A. Man and his occupations													
13000	EGYPTIAN HIEROGLYPH A001	1303A	EGYPTIAN HIEROGLYPH A049										
13001	EGYPTIAN HIEROGLYPH A002	1303B	EGYPTIAN HIEROGLYPH A050										
13002	EGYPTIAN HIEROGLYPH A003	1303C	EGYPTIAN HIEROGLYPH A051										
13003	EGYPTIAN HIEROGLYPH A004	1303D	EGYPTIAN HIEROGLYPH A052										
13004	EGYPTIAN HIEROGLYPH A005	1303E	EGYPTIAN HIEROGLYPH A053										
		1303F	EGYPTIAN HIEROGLYPH A054										
		13040	EGYPTIAN HIEROGLYPH A055										
		13041	EGYPTIAN HIEROGLYPH A056										







	00	10	20	30	40	50	60	70	80	90	A0	B0	C0	D0	E0	F0			
Plane 00 BMP Basic Multilingual Plane 0x000000 0x00FFFF	000	Basic Latin									Latin-1 Supplement								
		Latin Extended-A									Latin Extended-B								
								IPA Extensions			Spacing Modifier Letters								
		Combining Diacritical Marks									Greek and Coptic								
		Cyrillic																	
		Cyrillic Supplement				Armenian						Hebrew							
		Arabic																	
		Syriac				Arabic Supplement				Thaana				Nko					
		Samaritan						Mandaic						Arabic Extended-A					
		Plane 01 SMP Supplementary Multilingual Plane 0x010000 0x01FFFF		Devanagari								Bengali							
Gurmukhi								Gujarati											
Oriya								Tamil											
Telugu								Kannada											
Malayalam								Sinhala											
Thai								Lao											
Tibetan																			
Plane 02 SIP Supplementary Ideographic Plane 0x020000 0x02FFFF	010			Myanmar										Georgian					
				Hangul Jamo															
				Ethiopic															
												Ethiopic Supplement			Cherokee				
		Unified Canadian Aboriginal Syllabics																	
												Ogham			Runic				

	00	10	20	30	40
000					! " # \$ % & ' ( ) * + , - . / 0 1 2 3 4 5 6 7 8 9 : ; < = > ? @ A B C
	À Á Â Ã Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã				ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
	Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã				ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
	Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã				ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
	Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã				ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
	Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã				ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
	Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã				ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
	Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã				ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
	Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã				ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
	Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã				ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
010					À Á Â Ã Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã
					ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
					Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã
					ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
					Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã
					ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
					Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã
					ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
					Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã
					ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
020					À Á Â Ã Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã
					ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
					Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã
					ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
					Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã
					ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
					Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã
					ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã
					Ä Å Æ Ç È É Ê Ë Ì Í Î Ï Ñ Ò Ó Ô Õ Ö × Ø Ù Ú Û Ü Ý Þ ß à á â ã
					ä å æ ç è é ê ë ì í î ï ñ ò ó ô õ ö ÷ ø ù ú û ü ý þ ß à á â ã



glyphs



text rendering engine  
NSLayoutManager

codepoints

U+2603 SNOWMAN

algorithms

normalization

collation

casing

binary representation

E2 98 83 (UTF-8)

fonts

Times New Roman.ttf

Unicode Standard

# Visual Similarities

AA A A Δ A A

www.google.com – U+0067 LATIN SMALL LETTER G

www.google.com – U+0261 LATIN SMALL LETTER SCRIPT G

৪ – U+09EA BENGALI DIGIT FOUR

୨ – U+0B68 ORIYA DIGIT TWO



# Country Flags

U+1F1E6	+	U+1F1E7			
U+1F1E8	+	U+1F1F3			
U+1F1E9	+	U+1F1EA			
U+1F1EA	+	U+1F1F8			
U+1F1EB	+	U+1F1F7			
U+1F1EC	+	U+1F1E7			
U+1F1EE	+	U+1F1F9			
U+1F1EF	+	U+1F1F5			
U+1F1F0	+	U+1F1F7			
U+1F1F7	+	U+1F1FA			
U+1F1FA	+	U+1F1F8			



# Bi-directional Text

```
# U+202E RIGHT-TO-LEFT OVERRIDE  
# double click a .jpg, open an .exe  
$ python3 -c "print('s\u202Egpj.exe')"  
sexe.jpg
```





glyphs



text rendering engine  
NSLayoutManager

codepoints

U+2603 SNOWMAN

algorithms

normalization

collation

casing

binary representation

E2 98 83 (UTF-8)

fonts

Times New Roman.ttf

Unicode Standard





**Nicolas Seriot**

@nst021

Here is a nice little Core Text crasher for OS X:  
\$ python -c "print u'\u0647\u0020\u0488\u0488\u0488'"

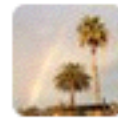


RETWEETS

38

FAVORITES

34



10:49 AM - 25 Mar 2013

\$ gdb Twitter

(gdb) r

Starting program: /Applications/Twitter.app/Contents/MacOS/Twitter

Program received signal EXC\_BAD\_ACCESS, Could not access memory.

Reason: KERN\_INVALID\_ADDRESS at address: 0x00000001084e8008

0x00007fff9432ead2 in vDSP\_sveD ()

(gdb) bt

#0 0x00007fff9432ead2 in vDSP\_sveD ()

#1 0x00007fff934594fe in TStorageRange::SetStorageSubRange ()

#2 0x00007fff93457d5c in TRun::TRun ()

#3 0x00007fff934579ee in CTGlyphRun::CloneRange ()

#4 0x00007fff93466764 in TLine::SetLevelRange ()

#5 0x00007fff93467e2c in TLine::SetTrailingWhitespaceLevel ()

#6 0x00007fff93467d58 in TRunReorder::ReorderRuns ()

#7 0x00007fff93467bfe in TTypesetter::FinishLineFill ()

#8 0x00007fff934858ae in TFramesetter::FrameInRect ()

#9 0x00007fff93485110 in TFramesetter::CreateFrame ()

#10 0x00007fff93484af2 in CTFramesetterCreateFrame ()

...



MAIN MENU

MY STORIES: 25

FORUMS

SUBSCRIBE

JOBS

## INFINITE LOOP / THE APPLE ECOSYSTEM

### Rendering bug crashes OS X, iOS apps with string of Arabic characters (Updated)

CoreText bug crashes any iOS 6 and OS X programs that use the API.

by Andrew Cunningham and Dan Goodin Aug 29 2013, 9:30pm CEST



Share



Tweet

149

LATEST FEATURE STORY



FEATURE STORY (1 PAGE)

# The Register®

Biting the hand that feeds IT

Data Centre

Software

Networks

Security

Business

Hardware

Science

Bootnotes

Video

Forums

Weekend Edition

Search site



Operating Systems

Applications

Developer

Verity Stob

SOFTWARE > OPERATING SYSTEMS

## Anatomy of a killer bug: How just 5 characters can murder iPhone, Mac apps

What evil lurks in the Unicode of Death ... oh, a buffer overrun

By Chris Williams, 4 Sep 2013

93

RELATED STORIES

**Analysis** There has been much sniggering into sleeves after wags found they could upset iOS 6 iPhones and iPads, and Macs running OS X 10.8, by sending a simple rogue text message or email.

A bug is triggered when the CoreText component in vulnerable Apple operating

MOST READ

MOST COMMENTED

YARR! Pirates walk the plank: DMCA magnets sink in Google results

Whisper tracks its users. So we tracked down its LA office. This is what happened next

Xperia Z3: Crikey, Sony – ANOTHER flagship phondleslab?

Ex-US Navy fighter pilot MIT prof: Drones beat humans - I should know

Apple flings iOS 8.1 at world+dog: Our AMAZEBALLS 9-step installation guide



# OS X Finder

```
$ echo -e "\xFF\xFE" > x.txt # UTF-16LE BOM
$ xattr -w com.apple.TextEncoding "utf-16le" x.txt
$ qlmanage -p x.txt # or QuickLook with Finder
```

```
[ERROR] An uncaught exception was raised outside of any generator: *** -[NSConcreteTextStorage attribute:atIndex:longestEffectiveRange:inRange:]: Range or index out of bounds
2014-10-24 10:53:08.474 qlmanage[5268:11f] *** Terminating app due to uncaught exception 'NSRangeException', reason: '*** -[NSConcreteTextStorage attribute:atIndex:longestEffectiveRange:inRange:]: Range or index out of bounds'
*** First throw call stack:
(
    0  CoreFoundation          0x00007fff89ebe25c __exceptionPreprocess + 172
    1  libobjc.A.dylib          0x00007fff87934e75 objc_exception_throw + 43
    2  CoreFoundation          0x00007fff89ebe10c +[NSException raise:format:] + 204
    3  AppKit                   0x00007fff81a83a7a -[NSConcreteTextStorage attribute:atIndex:longestEffectiveRange:inRange:] + 118
    4  AppKit                   0x00007fff81951ded -[NSMutableAttributedString(NSMutableAttributedStringKitAdditions) fixGlyphInfoAttributeInRange:] + 204
    5  AppKit                   0x00007fff81951cd8 -[NSMutableAttributedString(NSMutableAttributedStringKitAdditions) fixAttributesInRange:] + 39
    6  AppKit                   0x00007fff81a838e1 -[NSTextStorage processEditing] + 109
    7  AppKit                   0x00007fff81a7f742 -[NSTextStorage endEditing] + 110
    8  AppKit                   0x00007fff81c5db4f _NSReadAttributedStringFromURLOrData + 14525
    9  AppKit                   0x00007fff81c5e3a5 -[NSAttributedString(NSAttributedStringKitAdditions) initWithURL:options:documentAttributes:
```





glyphs



text rendering engine  
NSLayoutManager

codepoints

U+2603 SNOWMAN

algorithms

normalization

collation

casing

binary representation

E2 98 83 (UTF-8)


fonts

Times New Roman.ttf

Unicode Standard



# Weird Code Points May Bypass Filters

- Non-characters: eg. U+FFFE, U+FFFF, U+1FFFE, U+10FFFF  
Unassigned code points: eg. U+2073
- Must not be **deleted** (as allowed by Unicode < 5.2 C7) but **replaced** with  U+FFFD REPLACEMENT CHARACTER.

```
<a href="java\uFFFFscript:alert("XSS")>
```





# Non-Characters and OS X Bash / HFS+

```
$ mkdir /tmp/test
$ cd /tmp/test
$ touch `printf "a\xef\xbb\xbf" `
# or "a\uFFFEb".encode('utf-8')
# which is a non-character
$ ls a*
a?b
$ touch ab
$ ls a*
a?b
# where did ab go?!
```





# Regex

```
$ python3
>>> import re
>>> reg = re.compile("\d")
>>> gen = ( chr(c) for c in range(0, 0xFFFF) if re.match(reg, chr(c)) )
>>> print(''.join(gen))
0123456789.123456789.1234567891234567891011121314151617181920212223242526272829303132333435363738394041424344454647484950515253545556575859606162636465666768697071727374757677787980818283848586878889909192939495969798991001011021031041051061071081091101111121131141151161171181191201211221231241251261271281291301311321331341351361371381391401411421431441451461471481491501511521531541551561571581591601611621631641651661671681691701711721731741751761771781791801811821831841851861871881891901911921931941951961971981992002012022032042052062072082092102112122132142152162172182192202212222232242252262272282292302312322332342352362372382392402412422432442452462472482492502512522532542552562572582592602612622632642652662672682692702712722732742752762772782792802812822832842852862872882892902912922932942952962972982993003013023033043053063073083093103113123133143153163173183193203213223233243253263273283293303313323333343353363373383393403413423433443453463473483493503513523533543553563573583593603613623633643653663673683693703713723733743753763773783793803813823833843853863873883893903913923933943953963973983994004014024034044054064074084094104114124134144154164174184194204214224234244254264274284294304314324334344354364374384394404414424434444454464474484494504514524534544554564574584594604614624634644654664674684694704714724734744754764774784794804814824834844854864874884894904914924934944954964974984995005015025035045055065075085095105115125135145155165175185195205215225235245255265275285295305315325335345355365375385395405415425435445455465475485495505515525535545555565575585595605615625635645655665675685695705715725735745755765775785795805815825835845855865875885895905915925935945955965975985996006016026036046056066076086096106116126136146156166176186196206216226236246256266276286296306316326336346356366376386396406416426436446456466476486496506516526536546556566576586596606616626636646656666676686696706716726736746756766776786796806816826836846856866876886896906916926936946956966976986997007017027037047057067077087097107117127137147157167177187197207217227237247257267277287297307317327337347357367377387397407417427437447457467477487497507517527537547557567577587597607617627637647657667677687697707717727737747757767777787797807817827837847857867877887897907917927937947957967977987998008018028038048058068078088098108118128138148158168178188198208218228238248258268278288298308318328338348358368378388398408418428438448458468478488498508518528538548558568578588598608618628638648658668678688698708718728738748758768778788798808818828838848858868878888898908918928938948958968978988999009019029039049059069079089099109119129139149159169179189199209219229239249259269279289299309319329339349359369379389399409419429439449459469479489499509519529539549559569579589599609619629639649659669679689699709719729739749759769779789799809819829839849859869879889899909919929939949959969979989991000100110021003100410051006100710081009101010111012101310141015101610171018101910201021102210231024102510261027102810291030103110321033103410351036103710381039104010411042104310441045104610471048104910501051105210531054105510561057105810591060106110621063106410651066106710681069107010711072107310741075107610771078107910801081108210831084108510861087108810891090109110921093109410951096109710981099110011001110021100311004110051100611007110081100911010110111101211013110141101511016110171101811019110201102111022110231102411025110261102711028110291103011031110321103311034110351103611037110381103911040110411104211043110441104511046110471104811049110501105111052110531105411055110561105711058110591106011061110621106311064110651106611067110681106911070110711107211073110741107511076110771107811079110801108111082110831108411085110861108711088110891109011091110921109311094110951109611097110981109911100111001111002111003111004111005111006111007111008111009111001011100111100121100131100141100151100161100171100181100191100201100211100221100231100241100251100261100271100281100291100301100311100321100331100341100351100361100371100381100391100401100411100421100431100441100451100461100471100481100491100501100511100521100531100541100551100561100571100581100591100601100611100621100631100641100651100661100671100681100691100701100711100721100731100741100751100761100771100781100791100801100811100821100831100841100851100861100871100881100891100901100911100921100931100941100951100961100971100981100991101001101011101021101031101041101051101061101071101081101091101101101111101121101131101141101151101161101171101181101191101201101211101221101231101241101251101261101271101281101291101301101311101321101331101341101351101361101371101381101391101401101411101421101431101441101451101461101471101481101491101501101511101521101531101541101551101561101571101581101591101601101611101621101631101641101651101661101671101681101691101701101711101721101731101741101751101761101771101781101791101801101811101821101831101841101851101861101871101881101891101901101911101921101931101941101951101961101971101981101991102001102011102021102031102041102051102061102071102081102091102101102111102121102131102141102151102161102171102181102191102201102211102221102231102241102251102261102271102281102291102301102311102321102331102341102351102361102371102381102391102401102411102421102431102441102451102461102471102481102491102501102511102521102531102541102551102561102571102581102591102601102611102621102631102641102651102661102671102681102691102701102711102721102731102741102751102761102771102781102791102801102811102821102831102841102851102861102871102881102891102901102911102921102931102941102951102961102971102981102991103001103011103021103031103041103051103061103071103081103091103101103111103121103131103141103151103161103171103181103191103201103211103221103231103241103251103261103271103281103291103301103311103321103331103341103351103361103371103381103391103401103411103421103431103441103451103461103471103481103491103501103511103521103531103541103551103561103571103581103591103601103611103621103631103641103651103661103671103681103691103701103711103721103731103741103751103761103771103781103791103801103811103821103831103841103851103861103871103881103891103901103911103921103931103941103951103961103971103981103991104001104011104021104031104041104051104061104071104081104091104101104111104121104131104141104151104161104171104181104191104201104211104221104231104241104251104261104271104281104291104301104311104321104331104341104351104361104371104381104391104401104411104421104431104441104451104461104471104481104491104501104511104521104531104541104551104561104571104581104591104601104611104621104631104641104651104661104671104681104691104701104711104721104731104741104751104761104771104781104791104801104811104821104831104841104851104861104871104881104891104901104911104921104931104941104951104961104971104981104991105001105011105021105031105041105051105061105071105081105091105101105111105121105131105141105151105161105171105181105191105201105211105221105231105241105251105261105271105281105291105301105311105321105331105341105351105361105371105381105391105401105411105421105431105441105451105461105471105481105491105501105511105521105531105541105551105561105571105581105591105601105611105621105631105641105651105661105671105681105691105701105711105721105731105741105751105761105771105781105791105801105811105821105831105841105851105861105871105881105891105901105911105921105931105941105951105961105971105981105991106001106011106021106031106041106051106061106071106081106091106101106111106121106131106141106151106161106171106181106191106201106211106221106231106241106251106261106271106281106291106301106311106321106331106341106351106361106371106381106391106401106411106421106431106441106451106461106471106481106491106501106511106521106531106541106551106561106571106581106591106601106611106621106631106641106651106661106671106681106691106701106711106721106731106741106751106761106771106781106791106801106811106821106831106841106851106861106871106881106891106901106911106921106931106941106951106961106971106981106991107001107011107021107031107041107051107061107071107081107091107101107111107121107131107141107151107161107171107181107191107201107211107221107231107241107251107261107271107281107291107301107311107321107331107341107351107361107371107381107391107401107411107421107431107441107451107461107471107481107491107501107511107521107531107541107551107561107571107581107591107601107611107621107631107641107651107661107671107681107691107701107711107721107731107741107751107761107771107781107791107801107811107821107831107841107851107861107871107881107891107901107911107921107931107941107951107961107971107981107991108001108011108021108031108041108051108061108071108081108091108101108111108121108131108141108151108161108171108181108191108201108211108221108231108241108251108261108271108281108291108301108311108321108331108341108351108361108371108381108391108401108411108421108431108441108451108461108471108481108491108501108511108521108531108541108551108561108571108581108591108601108611108621108631108641108651108661108671108681108691108701108711108721108731108741108751108761108771108781108791108801108811108821108831108841108851108861108871108881108891108901108911108921108931108941108951108961108971108981108991109001109011109021109031109041109051109061109071109081109091109101109111109121109131109141109151109161109171109181109191109201109211109221109231109241109251109261109271109281109291109301109311109321109331109341109351109361109371109381109391109401109411109421109431109441109451109461109471109481109491109501109511109521109531109541109551109561109571109581109591109601109611109621109631109641109651109661109671109681109691109701109711109721109731109741109751109761109771109781109791109801109811109821109831109841109851109861109871109881109891109901109911109921109931109941109951109961109971109981109991110000111000111100021110003111000411100051110006111000711100081110009111000101110001111000121110001311100014111000151110001611100017111000181110001911100020111000211110002211100023111000241110002511100026111000271110002811100029111000301110003111100032111000331110003411100035111000361110003711100038111000391110004011100041111000421110004311100044111000451110004611100047111000481110004911100050111000511110005211100053111000541110005511100056111000571110005811100059111000601110006111100062111000631110006411100065111000661110006711100068111000691110007011100071111000721110007311100074111000751110007611100077111000781110007911100080111000811110008211100083111000841110008511100086111000871110008811100089111000901110009111100092111000931110009411100095111000961110009711100098111000991110010011100101111001021110010311100104111001051110010611100107111001081110010911100110111001111110011211100113111001141110011511100116111001171110011811100119111001201110012111100122111001231110012411100125111001261110012711100128111001291110013011100131111001321110013311100134111001351110013611100137111001381110013911100140111001411110014211100143111001441110014511100146111001471110014811100149111001501110015111100152111001531110015411100155111001561110015711100158111001591110016011100161111001621110016311100164111001651110016611100167111001681110016911100170111001711110017211100173111001741110017511100176111001771110017811100179111001801110018111100182111001831110018411100185111001861110018711100188111001891110019011100191111001921110019311100194111001951110019611100197111001981110019911100200111002011110020211100203111002041110020511100206111002071110020811100209111002101110021111100212111002131110021411100215111002161110021711100218111002191110022011100221111002221110022311100224111002251110022611100227111002281110022911100230111002311110023211100233111002341110023511100236111002371110023811100239111002401110024111100242111002431110024411100245111002461110024711100248111002491110025011100251111002521110025311100254111002551110025611100257111002581110025911100260111002611110026211100263111002641110026511100266111002671110026811100269111002701110027111100272111002731110027411100275111002761110027711100278111002791110028011100281111002821110028311100284111002851110028611100287111002881110028911100290111002911110029211100293111002941110029511100296111002971110029811100299111003001110030111100302111003031110030
```



# Regex

```
$ jsc
>>> /a.c/.test('abc')
true
>>> /a.c/.test('a🐛c')
false
>>> /a...c/.test('a🐛c')
true
```



glyphs



text rendering engine  
NSLayoutManager

codepoints

U+2603 SNOWMAN

fonts

Times New Roman.ttf

algorithms

normalization

collation

casing

binary representation

E2 98 83 (UTF-8)

Unicode Standard





# Normalization

TR#15

é ①  
U+00E9 U+2460

*Canonical decomposition*

*Compatibility decomposition*

e ˆ ①  
U+0065 U+0301 U+2460

NFD

NFKD

é ˆ 1  
U+0065 U+0301 U+0031

*Canonical composition*

é ①  
U+0065 U+2460

NFC

(most common)


NFKC

é 1  
U+00E9 U+0031






# NFC doesn't Always Compose

 HEBREW LETTER  
SHIN WITH DAGESH  
AND SHIN DOT  
U+FB2C

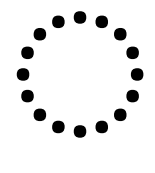
NFC(U+FB2C)



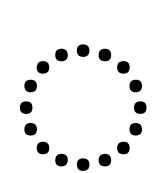
buffer overflow

 HEBREW LETTER  
SHIN WITH DAGESH  
AND SHIN DOT  
U+05E9

+

 HEBREW LETTER  
SHIN  
U+05BC

+

 HEBREW LETTER  
SHIN DOT  
U+05C1





# NFKD Expands Up to 18x



U+FDFA  
ARABIC  
LIGATURE  
SALLALLAHOU  
ALAYHE  
WASALLAM

```
>>> import unicodedata
```

```
>>> s = '\uFDFA'
```

```
>>> len(s)
```

```
1
```

```
>>> s_nfkd = unicodedata.normalize('NFKD', s)
```

```
>>> s_nfkd.encode('unicode-escape')
```

```
b'\\u0635\\u0644\\u0649 \\u0627\\u0644\\u0644\\u0647 \\u0639\\  
\\u0644\\u064a\\u0647 \\u0648\\u0633\\u0644\\u0645'
```

```
>>> len(s_nfkd)
```

```
18
```

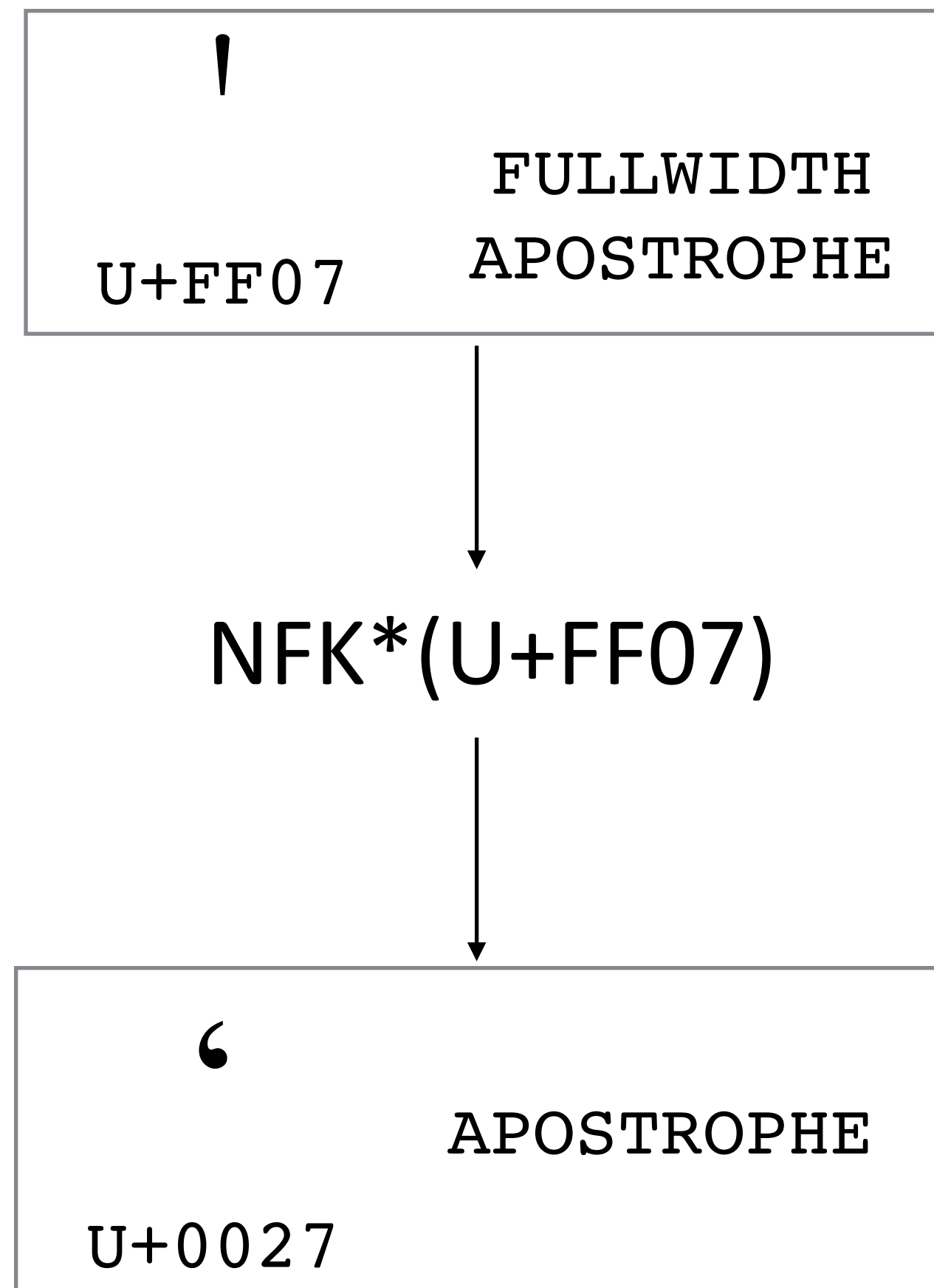


buffer overflow





# NFK\* May Bypass Filters



SQL injection



## Spotify Labs

Think it. Build it. Ship it. Tweak it. Blog it.



[Home](#) [About](#) [Puzzles](#)

Posted on **June 18, 2013** by [Mikael Goldmann](#)

## Creative usernames and Spotify account hijacking

1. Find a user account to hijack. For the sake of this example let us hijack the account belonging to user bigbird.
2. Create a new spotify account with username **BIGBIRD** (in python this is the string `u'\u1d2e\u1d35\u1d33\u1d2e\u1d35\u1d3f\u1d30'`).
3. Send a request for a password reset for your new account.
4. A password reset link is sent to the email you registered for your new account. Use it to change the password.
5. Now, instead of logging in to account with username **BIGBIRD**, try logging in to account with username bigbird with the new password.
6. Success! Mission accomplished.

<https://labs.spotify.com/2013/06/18/creative-usernames/>





glyphs



text rendering engine  
NSLayoutManager

codepoints

U+2603 SNOWMAN

fonts

Times New Roman.ttf

algorithms

normalization

collation

casing

binary representation

E2 98 83 (UTF-8)

Unicode Standard

# Unicode Collation Algorithm – TR#10 (UTS)

- **Text comparison**  
café < cafe ?  
cafe < café ?
- **Usage dependent**  
German dictionary: öf < of  
German phonebook: of < öf
- **Unstable over time**  
Sorted lists should be versioned

German			Swedish
Åkersberga	1	2	Alingsås
Alingsås	2	4	Oskarshamn
Äpplebo	3	7	Utting
Oskarshamn	4	6	Üttfeld
Östersund	5	8	Zwickau
Üttfeld	6	1	Åkersberga
Utting	7	3	Äpplebo
Zwickau	8	5	Östersund

(Steven R. Loomis, Mark Davis)





glyphs



text rendering engine  
NSLayoutManager

codepoints

U+2603 SNOWMAN

fonts

Times New Roman.ttf

algorithms

normalization

collation

casing

binary representation

E2 98 83 (UTF-8)

Unicode Standard

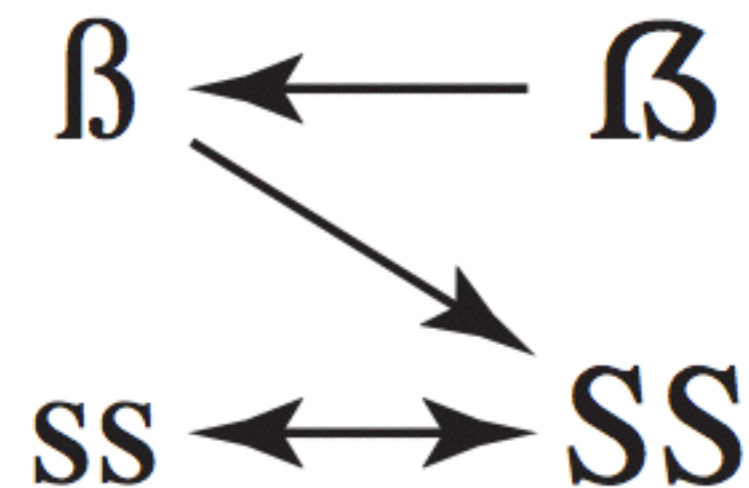
# Case Folding

<http://www.unicode.org/Public/UNIDATA/CaseFolding.txt>

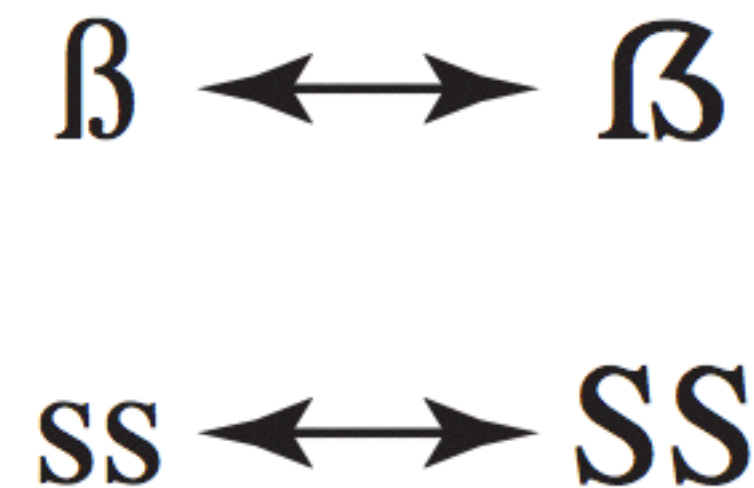
```
# The data supports both implementations that require simple case foldings  
# (where string lengths don't change), and implementations that allow full case folding  
# (where string lengths may grow). Note that where they can be supported, the  
# full case foldings are superior: for example, they allow "MASSE" and "Maße" to match.
```

Figure 5-16. Casing of German Sharp S

Default Casing

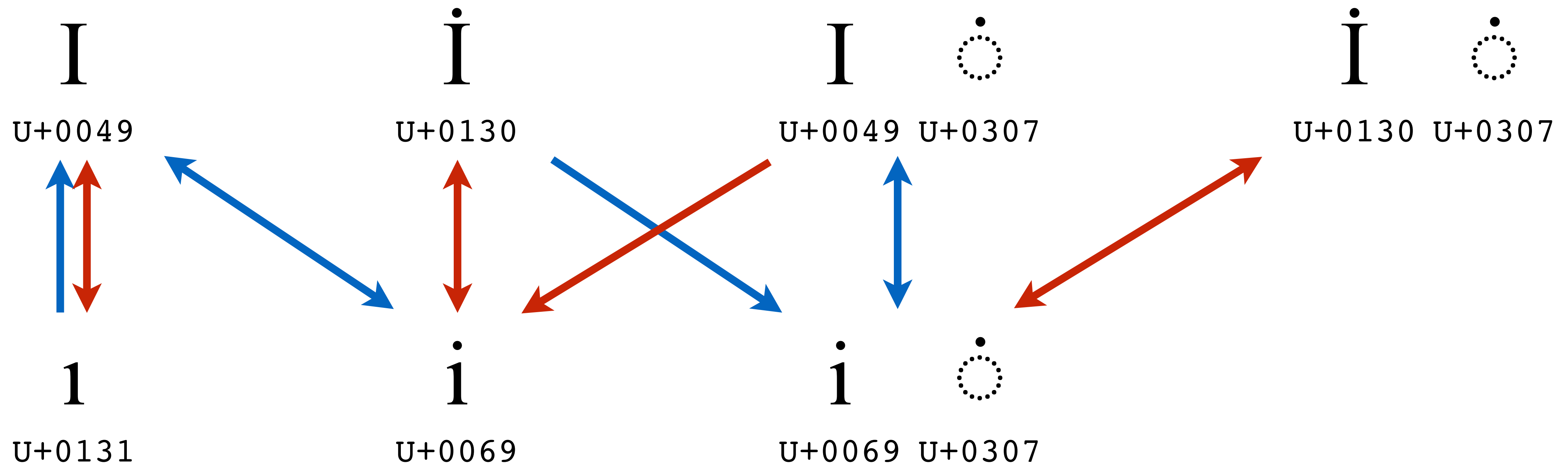


Tailored Casing





# Case Conversion



Posix Locale

Turkish Locale



# Case Conversion – Locale

```
NSString *s = [NSString stringWithFormat:@"İstanbul"];  
NSLocale *locale = [NSLocale localeWithLocaleIdentifier:@"tr_TR"];  
NSString *s2 = [s uppercaseStringWithLocale:locale];  
  
// İSTAMBUL ✓
```





# Python 3

- **✗** Collation: still compare codepoints

```
>>> 'café' < 'caff'  
False
```

- **✗** Case Conversion restricted to 1:1 case mappings

```
>>> 'ß'.upper()  
'ß'
```

- **✗** Case conversion ignores locale

**✗** Additionally, locale is global

```
>>> import locale  
>>> locale.setlocale(locale.LC_ALL, 'tr_TR')  
>>> s = "istanbul"  
>>> s.upper()  
'ISTANBUL'
```



glyphs



text rendering engine  
NSLayoutManager

codepoints

U+2603 SNOWMAN

algorithms

normalization

collation

casing

binary representation

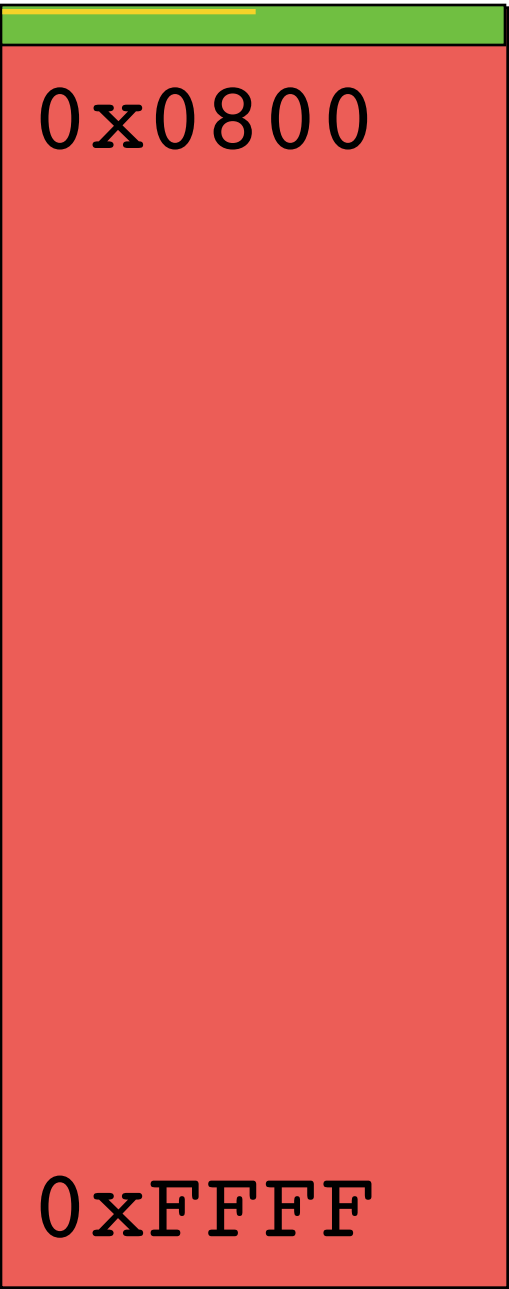
E2 98 83 (UTF-8)

fonts





Times New Roman.ttf

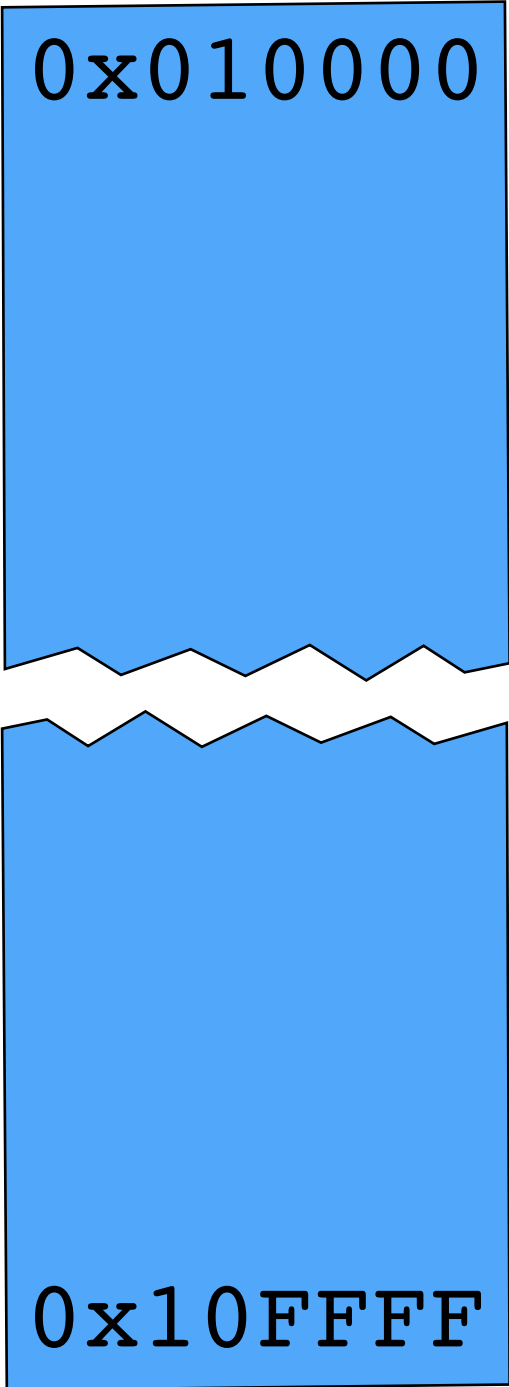
Unicode Standard





# UTF-8

	Bits	Hex Min	Hex Max	Byte Sequence in Binary
 1	7	00000000	0000007f	<b>0</b> vvvvvvvv
 2	11	00000080	000007FF	<b>110</b> vvvvv <b>10</b> vvvvvvv
 3	16	00000800	0000FFFF	<b>1110</b> vvvv <b>10</b> vvvvvvv <b>10</b> vvvvvvv
 4	21	00010000	001FFFFF	<b>11110</b> vvv <b>10</b> vvvvvvv <b>10</b> vvvvvvv <b>10</b> vvvvvvv
5	26	00200000	03FFFFFF	<b>111110</b> vv <b>10</b> vvvvvvv <b>10</b> vvvvvvv <b>10</b> vvvvvvv <b>10</b> vvvvvvv
6	31	04000000	7FFFFFFF	<b>1111110</b> v <b>10</b> vvvvvvv <b>10</b> vvvvvvv <b>10</b> vvvvvvv <b>10</b> vvvvvvv <b>10</b> vvvvvvv



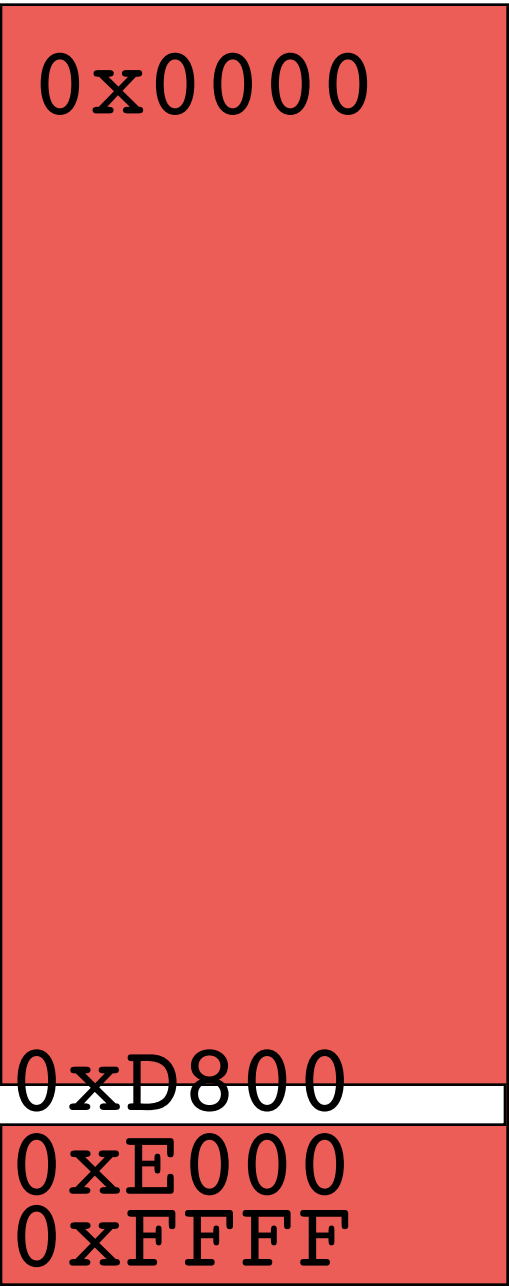
Malformed UTF-8 sequences include:

- overlong encoding, 0x1 on 2 bytes

**11000000 10000001**                      0xC0 0x41

- unexpected continuation byte

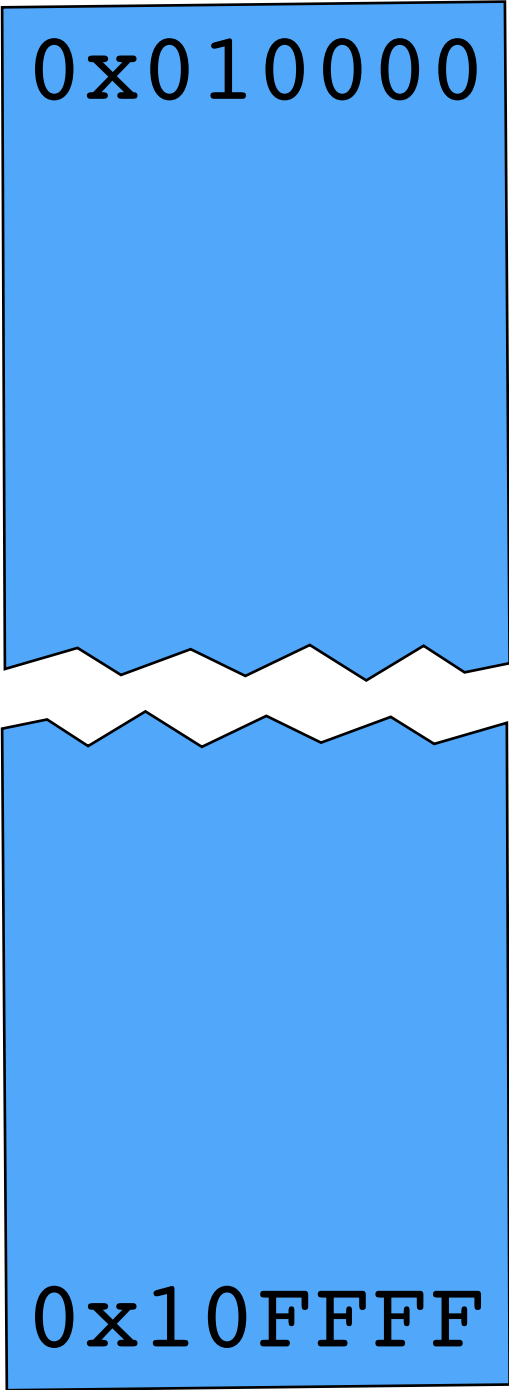
**11000000 00000000**                      0xC0 0x00



# UTF-16

	Bits	Hex Min	Hex Max	Byte Sequence in Binary
	2	16	00000000 0000FFFF	vvvvvvvv vvvvvvvv
	4	21	00010000 001FFFFF	<b>110110</b> ww wwwwww <b>110111</b> ww wwwwww

www.. is (vvv.. - 0x10000) to map a 20 bits value



Malformed sequences include unpaired surrogates such as:

- **110110**ww wwwwww not followed by **110111**ww wwwwww
- **110111**ww wwwwww not preceded by **110110**ww wwwwww



# Wide Characters



- Unicode code points were first defined on 16 bits (UCS-2)
- and now Java **char** / Objective-C **unichar** are 16 bits
- code points  $> 0xFFFF$  defined as a pair of 16 bits values
- **sizeof(wchar\_t)** is generally 16 bits on Windows, 32 bits on Linux

# Objective-C / Cocoa

```
NSString *s1 = @"abc";
NSString *s2 = @"\U0001F600bc";

NSLog(@"s1 %@", s1); // s1 abc
NSLog(@"s2 %@", s2); // s2 😊bc

NSLog(@"s1[0] -> %C", [s1 characterAtIndex:0]); // s1[0] -> a
NSLog(@"s2[0] -> %C", [s2 characterAtIndex:0]);
// nothing printed because
// s2 = [0xD83D, 0xDE00], and U+D83D is a high surrogate
// and NSLog() ignores nil strings
```





# HFS+

**Important:** The terms used in this Q&A, precomposed and decomposed, roughly correspond to Unicode Normal Forms C and D, respectively. However, most volume formats do not follow the exact specification for these normal forms. For example, HFS Plus (Mac OS Extended) uses a variant of Normal Form D in which U+2000 through U+2FFF, U+F900 through U+FAFF, and U+2F800 through U+2FAFF are not decomposed (this avoids problems with round trip conversions from old Mac text encodings). It's likely that your volume format has similar oddities.

Apple Technical Q&A QA1173



# HFS+

```
# what you write...  
$ echo ü; echo ü | xxd  
ü  
00000000: c3bc 0a # NFC  
  
# is not what you read  
$ touch ü; ls; ls | xxd  
ü  
00000000: 75cc 880a # NFD
```

```
# watch your Finder go nuts!!!  
$ cd; touch `printf "\x41\xe9"`  
# NFC("Aé")  
$ open .  
# fixed in OS X 10.10
```





# Conclusion

- Unicode is cool. Unicode is hard. Unicode is ubiquitous.
- How well do you know your framework of choice?
- Everything dealing with Unicode is a bug nest.
- Under-studied topic. Tons of low-hanging fruits.
- See Chris Weber's <http://websec.github.io/unicode-security-guide/>



**« Unicode is just too complex to ever be secure. »**  
– Bruce Schneier, 2000

