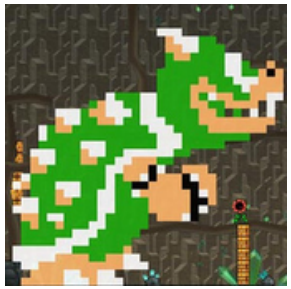
بیژن ابراهیمی

«عضو گروه کاربران لینوکس مشهد»

MashhadLUG.org



@bijan@quitter.se

یونی‌کُد

"The Unicode"

جلسه باز نرم‌افزاری مشهد،

۴ اسفند ۹۲

پیشگفتار

"The Introduction"

چرا یونی کُد؟

“Why Unicode?”

اگر فکر می‌کنید ...

“if you think ...”

متن ساده ⬅ ASCII

یونی‌کُد ⬅ انکدینگ (۲ بایتی)

کیبورد فارسی مایکروسافت مناسب فارسی‌نویسی است

یا برای فارسی‌نویسی نیاز به ابزار فرمت‌بندی است

اگر تا حالا...

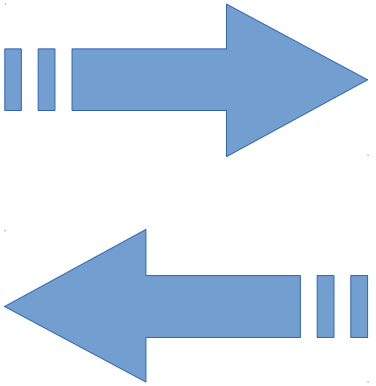
"If by now ..."

انکدینگ کاراکتر

"The Character Encoding"

انکدینگ کاراکتر چیست؟

“What is The Character Encoding?”



01001100110110010110100100010001110101010110100110000011001010101000100111101101101001011101001010101011110101010101

تاریخچه یونی‌کُد

“The unicode History”

انکدینگ اَسکی

"ASCII Encoding"

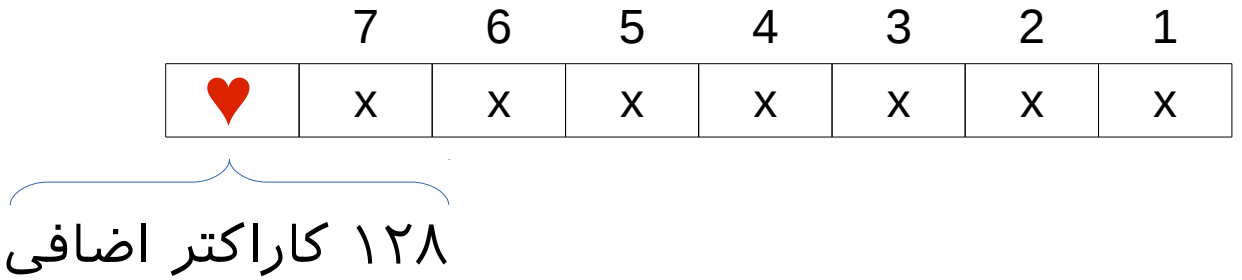
	0	1	2	3	4	5	6	7
0	NUL	DLE	space	0	@	P	`	p
1	SOH	DC1 XON	!	1	A	Q	a	q
2	STX	DC2	"	2	B	R	b	r
3	ETX	DC3 XOFF	#	3	C	S	c	s
4	EOT	DC4	$	4	D	T	d	t
5	ENQ	NAK	%	5	E	U	e	u
6	ACK	SYN	&	6	F	V	f	v
7	BEL	ETB	'	7	G	W	g	w
8	BS	CAN	(	8	H	X	h	x
9	HT	EM	)	9	I	Y	i	y
A	LF	SUB	*	:	J	Z	j	z
B	VT	ESC	+	;	K	[	k	{
C	FF	FS	,	<	L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

- ۷ بیت (۱۲۸ کاراکتر)
- ۳۳ کاراکتر کنترلی
- ۹۵ کاراکتر چاپی
- ۱ بیت اضافی

اسکی و بیت اضافی

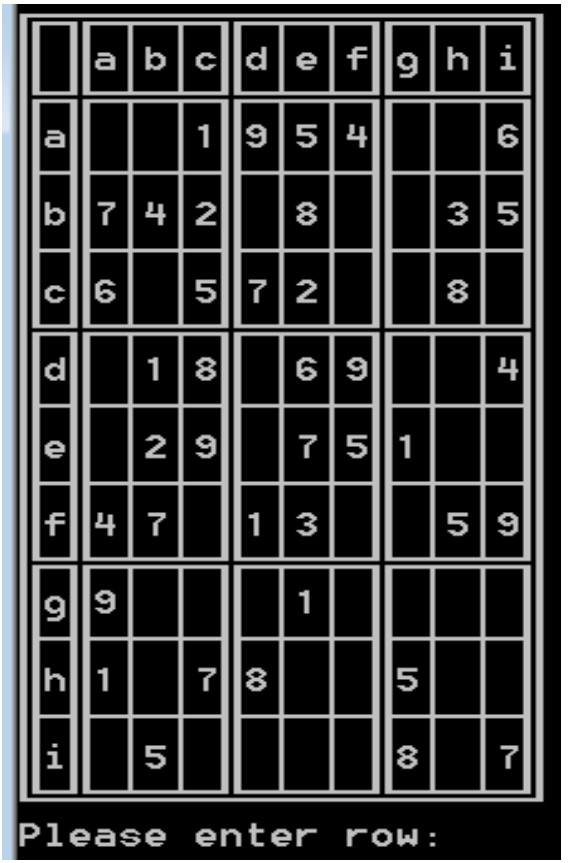
"ASCII and the extra bit"

- گسترش سیستم‌های ۸ بیتی
- بیت مورد علاقه توسعه‌دهندگان



اسکی و بیت اضافی

"ASCII and the extra bit"



آنسی و تولد کُدپیج ها

“ANSI and The birth of the Code Pages”

	00	01	02	03	04	05	06	07	08	09	0a	0b	0c	0d	0e	0f
00	␀															
10																
20		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
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	{	|	}	~	
80	Ç	ü	é	â	ä	à	å	ç	ê	ë	è	ï	î	ì	Ä	Å
90	É	æ	Æ	ô	ö	ò	û	ù	ÿ	Ö	Ü	¢	£	¥	Pts	ƒ
a0	á	í	ó	ú	ñ	Ñ	ª	º	¿	⌐	¬	½	¼	¡	«	»
b0	▒	▒	▓	|	┤	╡	╢	╖	╕	╣	‖	╗	╝	╜	╛	┐
c0	└	┴	┬	├	─	┼	╞	╟	╚	╔	╩	╦	╠	═	╬	╧
d0	╨	╤	╥	╙	╘	╒	╓	╫	╪	┘	┌	▒	▄	▌	▐	▀
e0	α	ß	Γ	π	Σ	σ	µ	τ	Φ	Θ	Ω	δ	∞	φ	ε	∩
f0	≡	±	≥	≤	⌠	⌡	÷	≈	°	∙	·	√	ⁿ	²	■	

Fixed

Codepage dependent

CP437 (IBM)

	00	01	02	03	04	05	06	07	08	09	0a	0b	0c	0d	0e	0f
00	␀															
10																
20		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
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	{	|	}	~	
80	€	پ	‚	ƒ	„	…	†	‡	ˆ	‰	ٹ	‹	Œ	چ	ز	ڈ
90	گ	‘	’	“	”	•	–	—	ک	™	ڑ	›	œ			ں
a0		،	ç	£	¤	¥	¦	§	¨	©	ھ	«	¬		®	¯
b0	°	±	²	³	´	µ	¶	·	¸	¹	؛	»	¼	½	¾	؟
c0	ہ	ء	آ	أ	ؤ	إ	ئ	ا	ب	ة	ت	ث	ج	ح	خ	د
d0	ذ	ر	ز	س	ش	ص	ض	×	ط	ظ	ع	غ	ـ	ف	ق	ك
e0	à	ل	â	م	ن	ه	و	ç	è	é	ê	ë	ى	ي	î	ï
f0					ô			÷		ù		û	ü			ے

CP1256 (Arabic)

مشکلات کُدپیج‌ها

“Code Pages Problems”

- چندزبانه سازی عملاً غیرممکن بود
- مشکل در ارتباط با سیستم‌های مختلف دیگر
- نامناسب برای زبان‌های آسیایی

تفاوت میان کاراکتر کدپیج ها

"Difference between codepages characters"

```
$ python
>>> print chr(202).decode('cp437')
╩
>>> print chr(202).decode('cp1256')
ت
>>>
```

استاندارد یونی‌کُد

“The Unicode standard”

فارسی

日本語

English

Slovenščina

العربية

עברית

Հայերեն

Русский

Polski

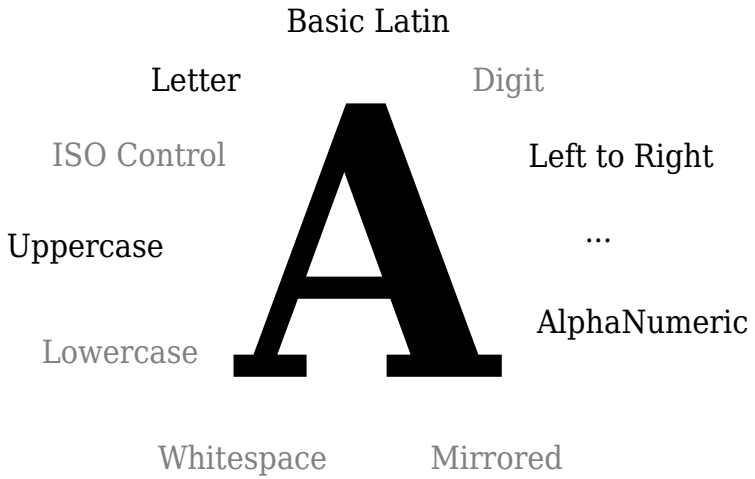
中文

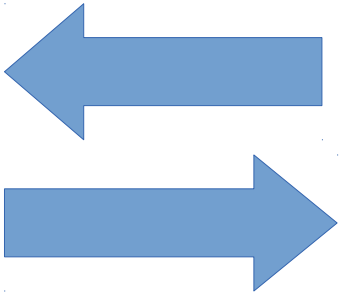
Български

...

استاندارد یونی کُد

"The Unicode standard"





Code Point

U+hexadecimal

خصوصیات کاراکترهای یونی‌کُد

“Unicode characters properties”

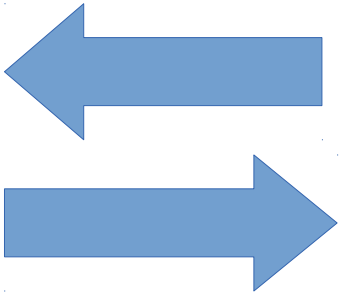
```
$ python
>>> import unicodedata as ud
>>> ud.name(u”ب”)
'ARABIC LETTER BEH'
>>> ud.category(u”ب”)
'Lo'
>>> ud.numeric(u”۳”)
3.0
>>>
```

استاندارد یونی‌کُد

“The Unicode standard”

Code Point

U+hexadecimal



0100110011011001011010
0100100011101010101101
0011001**1000001**10010101
0101000100111101101101
0010110100101010111101
01010101010101

انکدینگ‌های یونی‌کُد

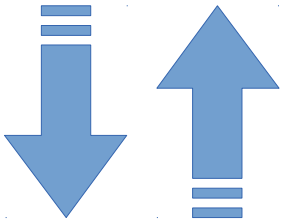
“The Unicode encodings”

“Hello” U+0048 U+0065 U+006C U+006C U+006F

انکدینگ‌های یونی‌کُد

“The Unicode encodings”

“Hello” U+0048 U+0065 U+006C U+006C U+006F



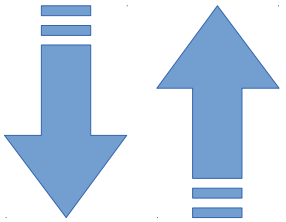
0048 0065 006C 006C 006F

4800 6500 6C00 6C00 6F00

انکدینگ‌های یونی‌کُد

"The Unicode encodings"

"Hello" U+0048 U+0065 U+006C U+006C U+006F



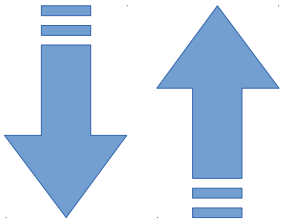
0048 0065 006C 006C 006F low-endian

4800 6500 6C00 6C00 6F00 hi-endian

انکدینگ های یونی کُد

"The Unicode encodings"

"Hello" U+0048 U+0065 U+006C U+006C U+006F



Byte Order Mark (BOM)

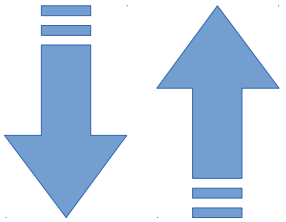
FFFE 0048 0065 006C 006C 006F low-endian

FEFF 4800 6500 6C00 6C00 6F00 hi-endian

انکدینگ‌های یونی‌کُد

"The Unicode encodings"

"Hello" U+0048 U+0065 U+006C U+006C U+006F



Byte Order Mark (BOM)

FFFE	0048 0065 006C 006C 006F	low-endian
FEFF	4800 6500 6C00 6C00 6F00	hi-endian

UCS-2

مشکلات انکدینگ‌های یونی‌کُد

“The Unicode encodings cons”

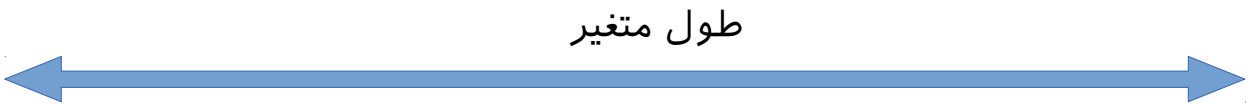
- عدم بهینگی در فضا با نگهداری بیت‌های صفر
- عدم سازگاری با انکدینگ اسکی
- عدم سازگاری با برنامه‌های قدیمی

انکدینگ UTF-8

“The UTF-8 encoding”

انکدینگ UTF-8

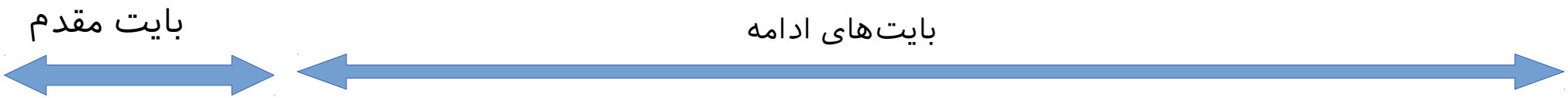
"UTF-8 encoding"



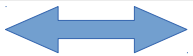
Bites	First	Last	Bytes	Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6
7	U+000	U+007F	1	0xxxxxxx					
11	U+0080	U+07FF	2	110xxxxx	10xxxxxx				
16	U+0800	U+FFFF	3	1110xxxx	10xxxxxx	10xxxxxx			
21	U+10000	U+1FFFFF	4	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx		
26	U+200000	U+3FFFFFF	5	111110xx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	
31	U+400000	U+7FFFFFFF	6	1111110x	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx

انکدینگ UTF-8

"UTF-8 encoding"



Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6
0xxxxxxx					
110xxxxx	10xxxxxx				
1110xxxx	10xxxxxx	10xxxxxx			
11110xxx	10xxxxxx	10xxxxxx	10xxxxxx		
111110xx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	
1111110x	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx



معرف تعداد بایت‌های ادامه

مزایای انکدینگ UTF8

“The UTF8 encoding pros”

- ✓ بهینه در نگهداری فضا
- ✓ قابلیت کشف خطا
- ✓ سازگاری کامل با انکدینگ اسکی

ASCII	“Hello”	48 65 6C 6C 6F
UTF-8	“Hello”	48 65 6C 6C 6F

دیگر انکدینگ‌های یونی‌کُد

“Other unicode encodings”

- UCS-2 (LE-BE) + BOM
- UTF-16 (LE-BE) + BOM
- UTF-32 (LE-BE) + BOM
- UTF-7

دیگر انکدینگ‌های یونی‌کُد

"Other unicode encodings"

```
$ python
>>> unichr(202).encode('utf-16le')
'\xca\x00'
>>> unichr(202).encode('utf-16be')
'\x00\xca'
>>> unichr(202).encode('utf-16')
'\xff\xfe\xca\x00'
>>> unichr(202).encode('utf-32')
'\xff\xfe\x00\x00\xca\x00\x00\x00'
>>> unichr(202).encode('utf-7')
'+AMo-'
```

تبدیل انکدینگ ها

“Character Encodings conversion”

یک نکته طلائی

“A Golden note”

«در حافظه چیزی به نام
متن‌ساده وجود ندارد»

”There's nothing as plain text on memory”

«داشتن یک رشته بدون دانستن نوع اندینگ آن بی‌معنی است»

”It does not make sense to have a string without knowing what encoding it uses”

۱. ارسال نوع انکدینگ

"Sending the encoding type"

- HTTP Content-Type: text/html; charset=UTF-8
- HTML 4 <meta http-equiv="Content-Type" content="text/html; charset=UTF-8">
- HTML 5 <meta charset="UTF-8">
- XML <xml encoding="UTF-8">

۱. ارسال نوع انکدینگ

"Sending the encoding type"

```
$ curl -I http://google.com
HTTP/1.1 301 Moved Permanently
Location: http://www.google.com/
Content-Type: text/html; charset=UTF-8
Date: Mon, 24 Feb 2014 12:32:10 GMT
Expires: Wed, 26 Mar 2014 12:32:10 GMT
Cache-Control: public, max-age=2592000
Server: gws
Content-Length: 219
X-XSS-Protection: 1; mode=block
X-Frame-Options: SAMEORIGIN
$
```

۲. شناسایی نوع انکدینگ

"Detecting the encoding type"

- نوع انکدینگ مشخص نیست
- اطلاعات نوع انکدینگ قابل اطمینان نیست

https://en.wikipedia.org/wiki/Charset_detection

http://www-archive.mozilla.org/projects/intl/UniversalCharsetDetection.html

راهکار شناسایی انکدینگ موزیلا

“Mozilla universal charset detection”

- شِمای کُد
- نسبت توزیع حروف
- نسبت توزیع دو حرف متوالی

https://en.wikipedia.org/wiki/Charset_detection

http://www-archive.mozilla.org/projects/intl/UniversalCharsetDetection.html

راهکار شناسایی انکدینگ موزیلا

"Mozilla universal charset detection"

```
$ python
>>> import chardet
>>> "hello world".encode("utf16")
'\xff\xfeh\x00e\x00l\x00l\x00o\x00
\x00w\x00o\x00r\x00l\x00d\x00'
>>> chardet.detect("\xff\xfeh\x00e\x00l\x00l\x00o
\x00\x00w\x00o\x00r\x00l\x00d\x00")
{'confidence': 1.0, 'encoding': 'UTF-16LE'}
>>>
```

۳. حدس‌زدن نوع انکدینگ!

“Guess the type of encoding!”

استفاده از یونی‌کُد در همه‌جا کافی است؟

“is using unicode in everywhere enough?”

بانک ملت
bank mellat

دروازه پرداخت اینترنتی قبض بانک ملت

نام سازمان	Ø³Ù^ÙŠÙ‡ Ù,Ø¨Ù^Ø¶ Ø¢Ø¨ Ø±Ù^Ø³Ø§Ø¦ÙŠ Ø§Ø³Ø§Ø±Ø§Ù† Ø®Ø²Ø±
شناسه قبض	146074000
شناسه پرداخت	222
مبلغ قابل پرداخت	22000 ریال

1392/11/23 20:17

تراکنش مالي مورد نظر با شماره پيگيري 2689149 با موفقيت انجام شد.

شرکت به پرداخت ملت ارایه دهنده خدمات پرداخت الکترونیک بانک ملت www.behpardakht.com

در صورت بروز هرگونه مشکل در پرداخت اينترنتي قبض با شماره تلفن 021-27312733 تماس حاصل فرماييد.

```
$
str=”&#216;&#170;&#216;&#179;&#217;&#136;&#217;&#138
;&#217;&#135;
&#217;&#130;&#216;&#168;&#217;&#136;&#216;&#182;
&#216;&#162;&#216;&#168;”
$ dec2hex(){ echo “obase=16; $1” | bc }
$ echo $str | grep -o “[0-9]*” | while read num; do
echo -n “\x`dec2hex #num`”; done | chardet
<stdin>: utf-8 (confidence: 0.99)
$ python
>>> real_string =
(chr(216)+chr(170)+chr(216)+chr(179)+...).decode(“ut
f8”).encode(“ascii”, “xmlcharrefreplace”)
>>> real_string
'&#1578;&#1587;&#1608;&#1610;&#1607;&#1602;&#1576;&#
1608;&#1590;&#1570;&#1576;'
>>> print real_string
تسویه قبوض آب روستائي استان خراسان رضوي
>>>
```

همیشه شما تنها
تولیدکننده محتوا نیستید!

“You're not always the only content producer!”

یونی‌کُد و چندزبانی

“Unicode and multilingual”

متن دوجهتی

"Bi-Directional Text"

«این یک متن Bi-Directional می‌باشد»

جهت کاراکترها

“Character's directional”

کاراکترهای خنثی
Neutral Characters

فاصله‌های خال
علایم نگارشی

کاراکترهای قوی
Strong Characters

حروف الفبا

کاراکترهای ضعیف
Weak Characters

اعداد

کیبورد استاندارد فارسی

“Persian standard keyboard”

مؤسسه استاندارد و تحقیقات صنعتی ایران

"Institute of Standards
And Industrial Research of Iran"

استانداردهای ملی
صفحه‌آرایی (کیبورد) فارسی

“National standards of persian layout”

ISIRI 820

«حروف فارسی در ماشین تحریر»

۱۳۵۲

ISIRI 2901

«طرز قرارگرفتن حروف و علایم زبان فارسی بر روی صفحه کلید»

۱۳۷۳

ISIRI 9147

«چیدمان حروف و علایم فارسی بر صفحه‌کلید رایانه»

استاندارد ملی ۹۱۴۷

”ISIRI 9147“

- مبتنی بر استاندارد یونی‌کُد
- سازگار با استاندارد ۲۹۰۱
- نویسه‌های استاندارد شده ۶۲۱۹
- دستور خط فارسی مصوب فرهنگستان زبان و ادب

کیبورد استاندارد فارسی

"Persian standard keyboard"

حالت عادی

کیبورد استاندارد فارسی

“Persian standard keyboard”

نام	کاراکتر یونی‌کُد	کُد پوینت یونی‌کُد	کد HTML
اتصال مجازی	ZERO WIDTH JOINER	U+200d	‍

مثال‌هایی از سطح ۱
کیبورد استاندارد فارسی

بـ بـ بـ

کیبورد استاندارد فارسی

"Persian standard keyboard"

حالت با تبدیل

کیبورد استاندارد فارسی

“Persian standard keyboard”

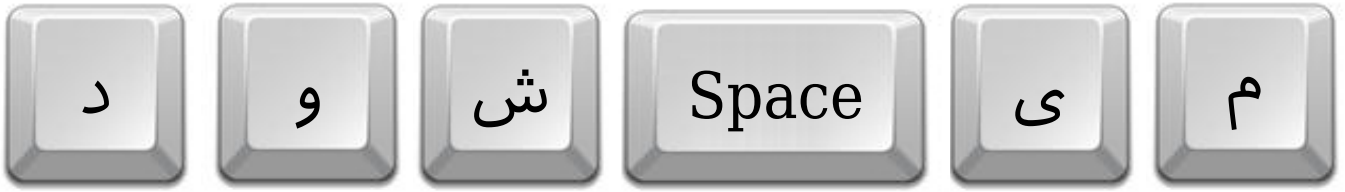
نام	کاراکتر یونی‌کُد	کُد پوینت یونی‌کُد	کد HTML
اتصال مجازی	ZERO WIDTH JOINER	U+200d	‍
فاصله مجازی	ZERO WIDTH NON-JOINER	U+200c	‌

مثال‌هایی از سطح ۲
کیبورد استاندارد فارسی

میشود

مثال‌هایی از سطح ۲
کیبورد استاندارد فارسی

می شود



مثال‌هایی از سطح ۲ کیبورد استاندارد فارسی

می‌شود

مثال‌هایی از سطح ۲
کیبورد استاندارد فارسی

"نقل قول"

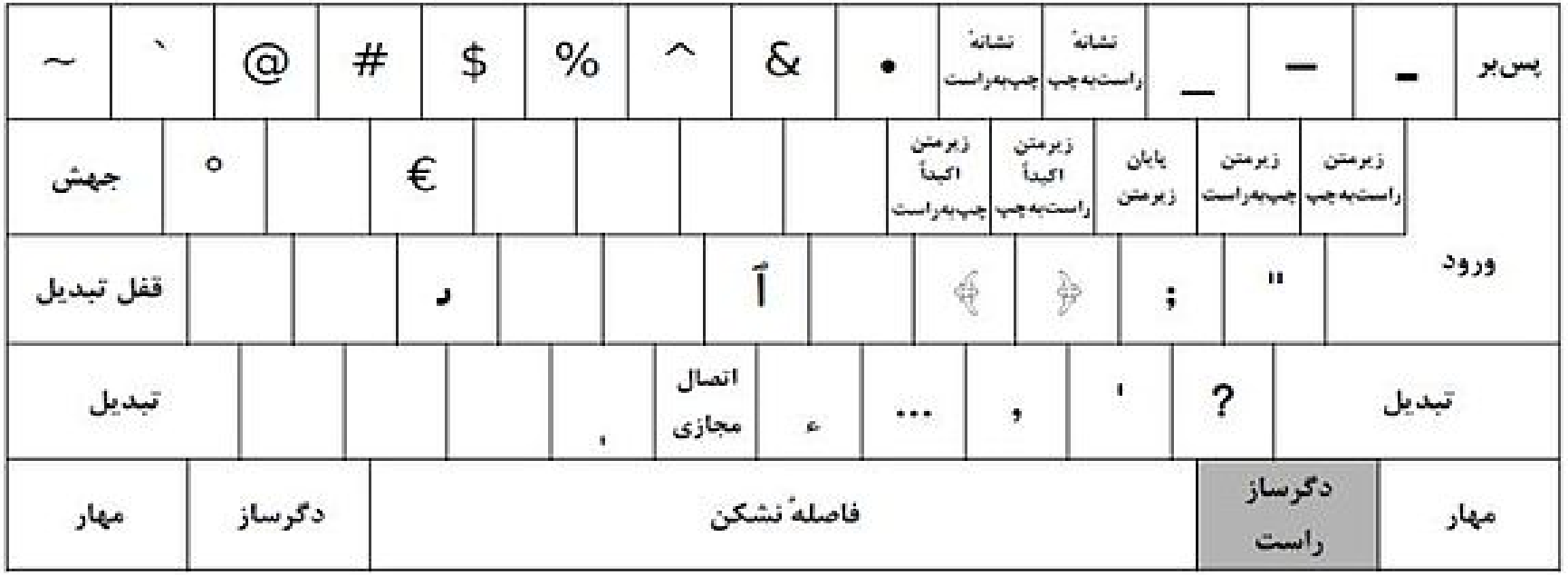
مثال‌هایی از سطح ۲ کیبورد استاندارد فارسی

«نقل قول»

کیبورد استاندارد فارسی

"Persian standard keyboard"

حالت با دگرساز راست



کیبورد استاندارد فارسی

"Persian standard keyboard"

نام	کاراکتر یونی‌کُد	کُد پوینت یونی‌کُد	کد HTML
نشانه راست به چپ	RIGHT-TO-LEFT MARK	U+200f	‏
نشانه چپ به راست	LEFT-TO-RIGHT MARK	U+200e	‎
زیر متن راست به چپ	RIGHT-TO-LEFT EMBEDDING	U+202b	‫
زیر متن چپ به راست	LEFT-TO-RIGHT EMBEDDING	U+202a	‪
پایان زیر متن	POP DIRECTIONAL FORMATTING	U+202c	‬

کیبورد استاندارد فارسی

“Persian standard keyboard”

نام	کاراکتر یونی‌کُد	کُد پوینت یونی‌کُد	کد HTML
زیرمتن اکیدا راست به چپ	RIGHT-TO-LEFT OVERRIDE	U+202e	‮
زیرمتن اکیدا چپ به راست	RIGHT-TO-LEFT OVERRIDE	U+202d	‭

مثال‌هایی از سطح ۳
کیبورد استاندارد فارسی

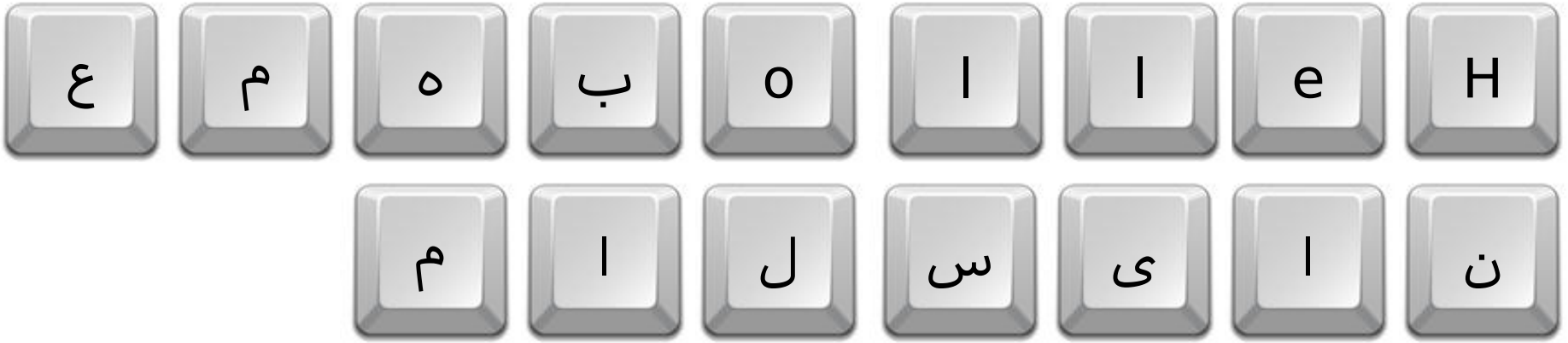
و ...

مثال‌هایی از سطح ۳
کیبورد استاندارد فارسی

و ...

مثال‌هایی از سطح ۳ کیبورد استاندارد فارسی

به معنای سلام Hello

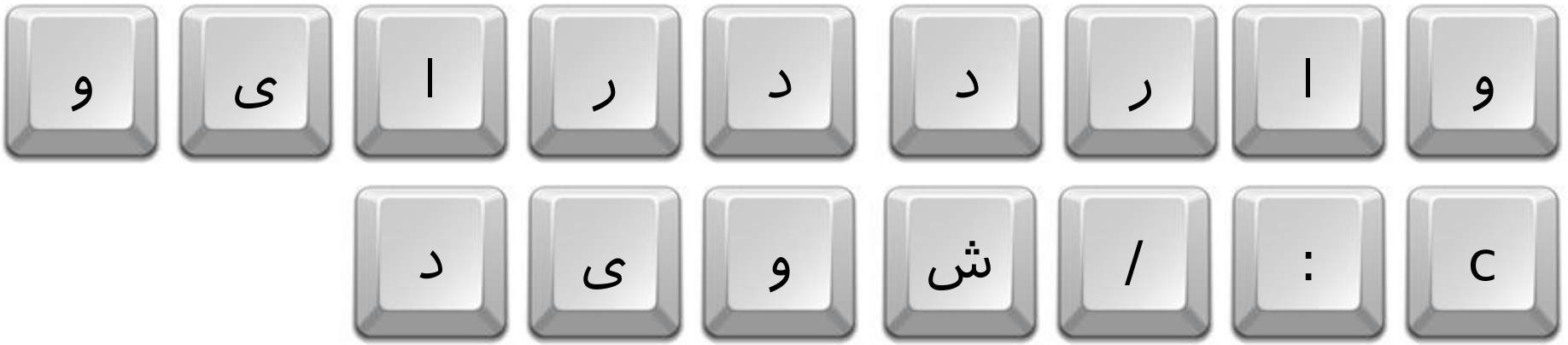


مثال‌هایی از سطح ۳ کیبورد استاندارد فارسی

Hello به معنای سلام

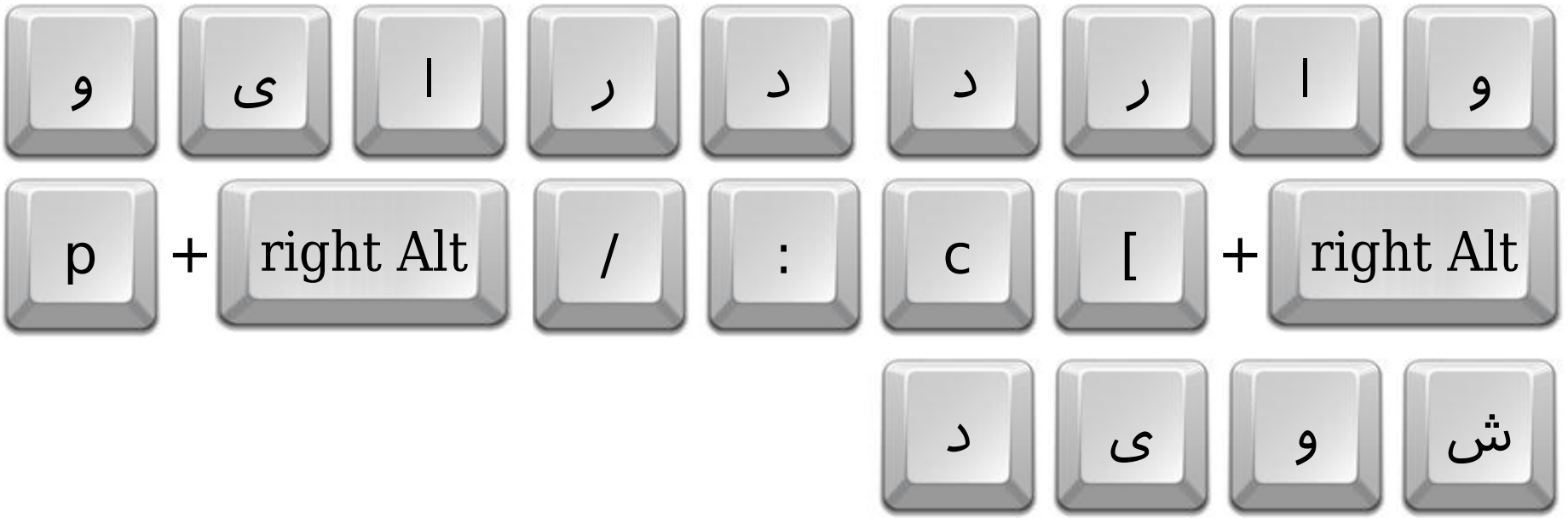
مثال‌هایی از سطح ۳ کیبورد استاندارد فارسی

وارد درایو c:/ شوید



مثال‌هایی از سطح ۳ کیبورد استاندارد فارسی

وارد درایو c:/ شوید

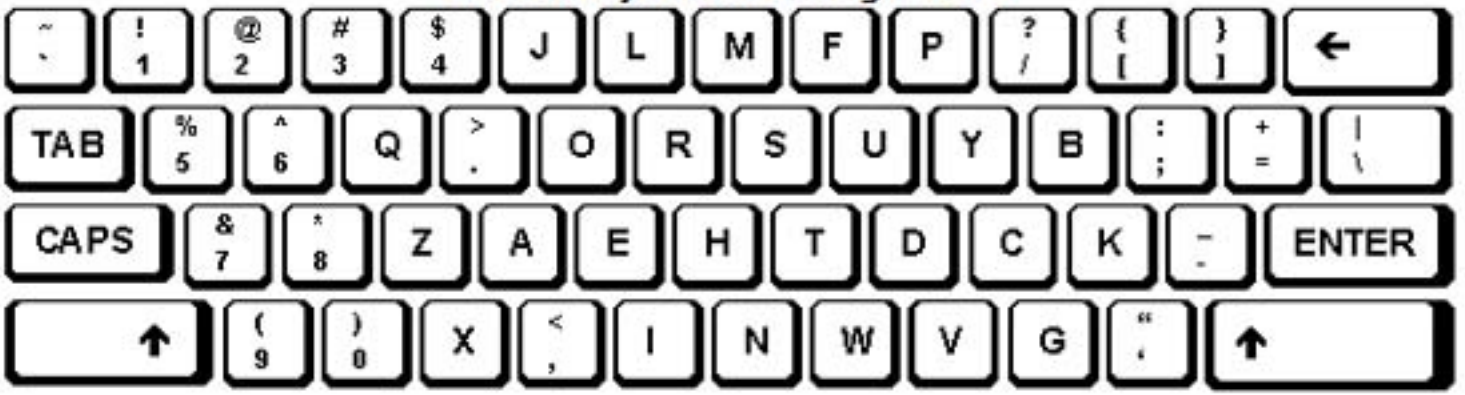


کیبورد استاندارد فارسی و سازگاری با گذشته

Persian Standard keyboard
and backward compatibility

Dvorak layout for two hands

Dvorak layout for the right hand



Dvorak layout for the left hand

با تشکر از وقت و حوصله شما

Thank you for your time and patience

پرسش و پاسخ

"Question/Answer"

یونی‌کُد

توسط بیژن ابراهیمی