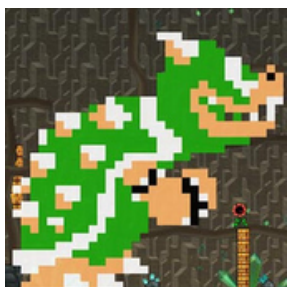


بیژن ابراهیمی

«عضو گروه کاربران لینوکس مشهد»

MashhadLUG.org



@bijan@quitter.se

یونی کد

”The Unicode“

جلسه باز نرم افزاری مشهد،

۴ اسفند ۹۲

پیشگفتار

“The Introduction”

چرا یونی کُد؟

“Why Unicode?”

اگر فکر می کنید ...

“if you think ...”

متن ساده ← ASCII

یونی کُد ← انکدینگ (۲ بایتی)

کیبورد فارسی مایکروسافت مناسب فارسی نویسی است

یا برای فارسی نویسی نیاز به ابزار فرمت بندی است

اگر تا حالا...

“If by now ...”

 بانک ملت bank mellat		
دروازه پرداخت اینترنتی قبض بانک ملت		
1392/11/23 20:17	<div>نام سازمان</div> <div>0a030~0š0† 0,0~0~0¶ 0†0~ 0±0~0³0a0s00!0š 0s0³0a0s0† 0@0±0s0³0s0† 0±0¶0~0š</div>	نام سازمان
	<div>شناسه قبض</div> <div>146074000</div>	شناسه قبض
	<div>شناسه پرداخت</div> <div>222</div>	شناسه پرداخت
	<div>مبلغ قابل پرداخت</div> <div>ریال 22000</div>	مبلغ قابل پرداخت
تراکنش مالی مورد نظر با شماره پیگیری 2689149 با موفقیت انجام شد.		

شرکت به پرداخت ملت ارایه دهنده خدمات پرداخت الکترونیک بانک ملت www.behpardakht.com
در صورت بروز هرگونه مشکل در پرداخت اینترنتی قبض با شماره تلفن 021-27312733 تماس حاصل فرمایید.

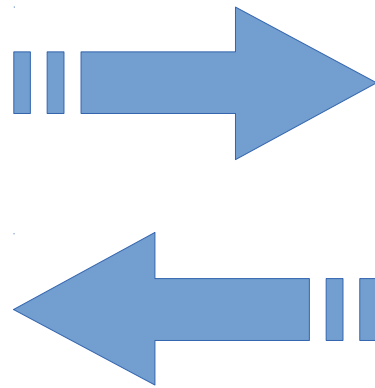
انکدینگ کاراکتر

“The Character Encoding”

انکدینگ کاراکتر چیست؟

“What is The Character Encoding?”

H Q F U
 N I K
 C O < A P B z
 T J s x L w
 ! M Y G D ?
 E v



```

0100110011011001011010
0100100011101010101101
0011001100000110010101
0101000100111101101101
0010110100101010111101
0101010101010101
  
```



تاریخچه یونی کد

“The [unicode](#) History”

انکدینگ آسکی

“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

	7	6	5	4	3	2	1
	x	x	x	x	x	x	x

۱۲۸ کاراکتر اضافی

اسکی و بیت اضافی

“ASCII and the extra bit”

	a	b	c	d	e	f	g	h	i
a			1	9	5	4			6
b	7	4	2		8			3	5
c	6		5	7	2			8	
d		1	8		6	9			4
e		2	9		7	5	1		
f	4	7		1	3			5	9
g	9				1				
h	1		7	8			5		
i		5					8		7

Please enter row:

آنسی و تولد گڈپیجھا

“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	ù	é	â	ä	à	å	ç	ê	ë	è	ï	î	í	Ä	Å
90	É	æ	Æ	ô	ö	ò	û	ü	ÿ	Ö	Ü	Ç	£	¥	Pts	f
a0	á	í	ó	ú	ñ	Ñ	ª	º	¿	¬	½	¼	¡	«	»	
b0	▒	▒	▒													
c0	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣
d0	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣	␣
e0	α	β	γ	π	Σ	σ	μ	τ	Φ	Θ	Ω	δ	∞	φ	ε	η
f0	≡	±	≥	≤			÷	≈	°	.	.	√	n	2	■	

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	,	f	„	...	†	‡	^	%	ث	<	œ	ج	ز	ڈ
90	گی	'	'	"	"	•	-	—	ک	™	ڑ	>	œ			ں
a0		,	ç	£	¤	¥	¦	§	"	©	ھ	«	¬		®	™
b0	°	±	²	³	´	µ	¶	·	¸	¹	º	»	¼	½	¾	¿
c0	ه	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	-	ف	ق	ك
e0	à	ل	â	م	ن	و	ç	è	é	ê	ë	ی	ی	î	ï	
f0					ô		÷		ù		û	ü				ے

CP1256 (Arabic)

مشکلات گُد پیج ها

“Code Pages Problems”

- ✓ چندزبانه سازی عملاً غیرممکن بود
- ✓ مشکل در ارتباط با سیستم های مختلف دیگر
- ✓ نامناسب برای زبان های آسیایی

تفاوت میان کاراکتر کدپیج‌ها

“Difference between `codepages` characters”

```
$ python
>>> print chr(202).decode('cp437')
=
>>> print chr(202).decode('cp1256')
c
>>>
```



استاندارد یونی‌کُد

“The Unicode standard”

فارسی

日本語

English

Slovenščina

العربية

עברית

Հայերեն



Русский

中文

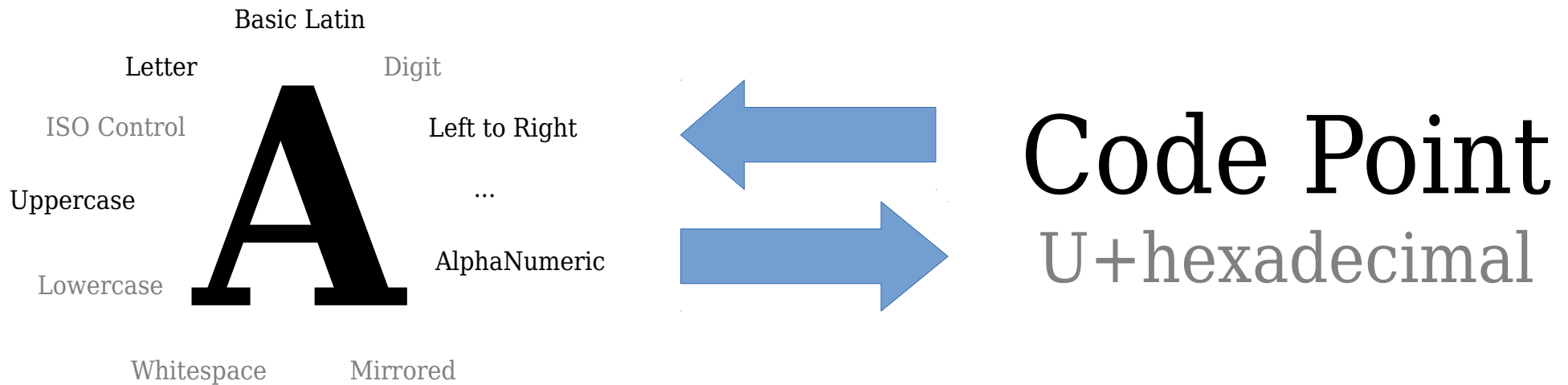
Polski

Български

...

استاندارد یونی کد

“The Unicode standard”



خصوصیات کاراکترهای یونی‌کد

“Unicode characters properties”

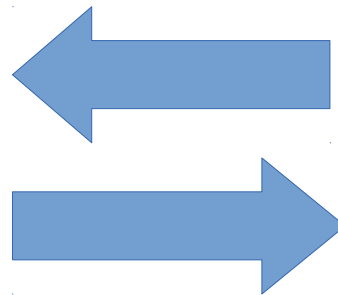
```
$ python
>>> import unicodedata as ud
>>> ud.name(u"ب")
'ARABIC LETTER BEH'
>>> ud.category(u"ب")
'Lo'
>>> ud.numeric(u"٣")
3.0
>>>
```



استاندارد یونی کد

“The Unicode standard”

Code Point
U+hexadecimal



```
0100110011011001011010  
0100100011101010101101  
0011001100000110010101  
0101000100111101101101  
0010110100101010111101  
01010101010101
```

انکدینگ‌های یونی‌کد

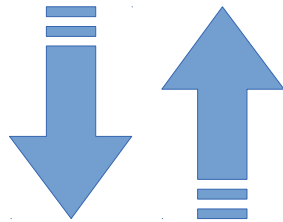
“The Unicode encodings”

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

انکدینگ‌های یونی‌کد

“The Unicode encodings”

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



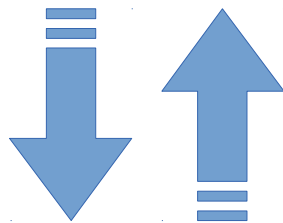
0048 0065 006C 006C 006F

4800 6500 6C00 6C00 6F00

انکدینگ‌های یونی‌کد

“The Unicode encodings”

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



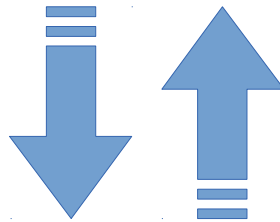
0048 0065 006C 006C 006F low-endian

4800 6500 6C00 6C00 6F00 hi-endian

انکدینگ‌های یونی‌کد

“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

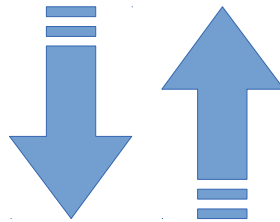
FEFF 4800 6500 6C00 6C00 6F00

hi-endian

انکدینگ‌های یونی‌کد

“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”

طول متغیر

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+3FFFFFFF	5	111110xx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	
31	U+400000	U+7FFFFFFF	6	1111110x	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx

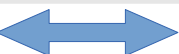
انکدینگ UTF-8

“UTF-8 encoding”

بایت مقدم

بایت‌های ادامه

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 یونی‌کد

“Other **unicode** encodings”

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

دیگر انکدینگ‌های یونی‌کد

“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 انکدینگ‌ها

“Character Encodings **conversion**”

یک نکتہ طلائی

“A Golden note”

«در حافظه چیزی به نام
متن ساده وجود ندارد»

“There's nothing as **plain text** on memory”

«داشتن یک رشته بدون
دانستن نوع اندینگ آن
بی معنی 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

راهکار شناسایی انکدینگ موزیلا

“Mozilla universal charset detection”

✓ شمای گُذ

✓ نسبت توزیع حروف

✓ نسبت توزیع دو حرف متوالی

راهکار شناسایی انکدینگ موزیلا

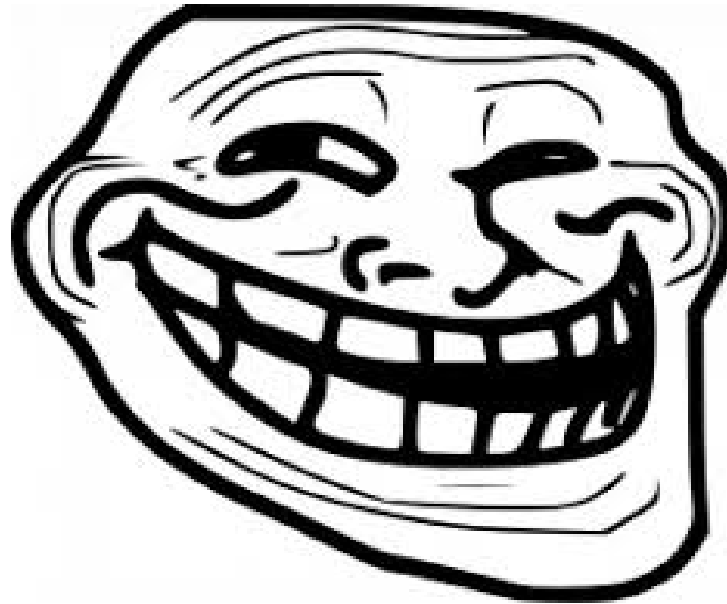
“Mozilla universal charset detection”

```
$ python
>>> import chardet
>>> "hello world".encode("utf16")
'\xff\xfe\x0e\x01\x01\x0o\x00
\x0w\x0o\x0r\x01\x0d\x00'
>>> chardet.detect("\xff\xfe\x0e\x01\x01\x0o°
\x0\x0w\x0o\x0r\x01\x0d\x00")
{'confidence': 1.0, 'encoding': 'UTF-16LE'}
```



۳. حدس زدن نوع انکدینگ!

“Guess the type of encoding!”



استفاده از یونی‌کُد در
همه جا کافی است؟

“is using `unicode` in everywhere enough?”



بانک ملت
bank mellat

دروازه پرداخت اینترنتی قبض بانک ملت

1392/11/23 20:17

0a030^0š0† 0,0^0^0 0†0^0^0000!0š
00000000† 0@0±0000000† 0±0^0^0š

146074000

222

ریال 22000

نام سازمان

شناسه قبض

شناسه پرداخت

مبلغ قابل پرداخت

تراکنش مالی مورد نظر با شماره پیگیری 2689149 با موفقیت انجام شد.

شرکت به پرداخت ملت ارایه دهنده خدمات پرداخت الکترونیک بانک ملت www.behpardakht.com
در صورت بروز هرگونه مشکل در پرداخت اینترنتی قبض با شماره تلفن 021-27312733 تماس حاصل فرمایید.

```
$
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("utf8").encode("ascii", "xmlcharrefreplace")
>>> real_string
'&#1578;&#1587;&#1608;&#1610;&#1607;&#1602;&#1576;&#
1608;&#1590;&#1570;&#1576;'
>>> print real_string
تسويه قبوض آب روستائي استان خراسان رضوي
>>>
```

همیشه شما تنها
تولیدکننده محتوا نیستید!

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

یونی کد و چندزبانی

“Unicode and [multilingual](#)”

متن دوجبهتی

“Bi-Directional Text”

«این یک متن Bi-Directional می باشد»



RTL



LTR



RTL

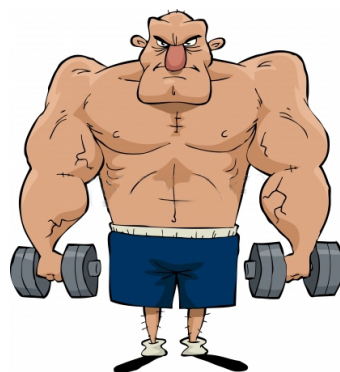
جهت کاراکترها

“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 علائم فارسی بر صفحه کلید 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

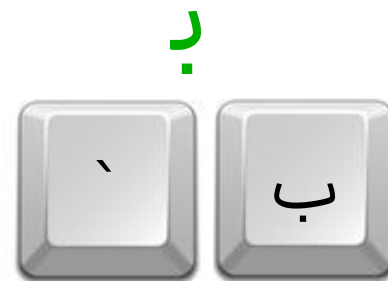
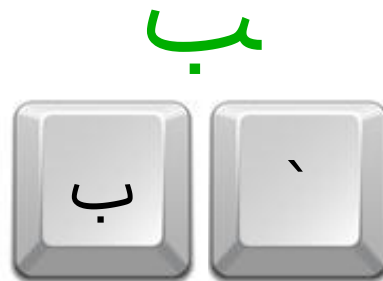
اتصال مجازی	۱	۲	۳	۴	۵	۶	۷	۸	۹	۰	-	=	\	پس بر
جهش	چ	ج	ح	خ	ه	ع	غ	ف	ق	ث	ص	ض		
ورود	گ	ک	م	ن	ت	ا	ل	ب	ی	س	ش			
تبدیل	/	.	و	پ	د	ذ	ر	ز	ط	ظ	تبدیل			
مهار	دگرساز راست	فاصله										دگرساز	مهار	

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



مثال‌هایی از سطح ۱ کیبورد استاندارد فارسی

۱۲



۱۲



کیبورد استاندارد فارسی

“Persian standard keyboard”

حالت با تبدیل

پس بر		+	-	()	*	,	x	%	لل	/	'	!	÷
		}	{]	[~	-	~	~	~	~	°	جهش	
ورود		؛	:	،	ة	آ	أ	ي	ئ	ؤ	قف	تبدیل		
تبدیل		؟	<	>	ء	ء	فاصله مجازی	'	ژ	~	ك	تبدیل		
مهار	دگر ساز راست		فاصله مجازی								دگر ساز		مهار	

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	‌

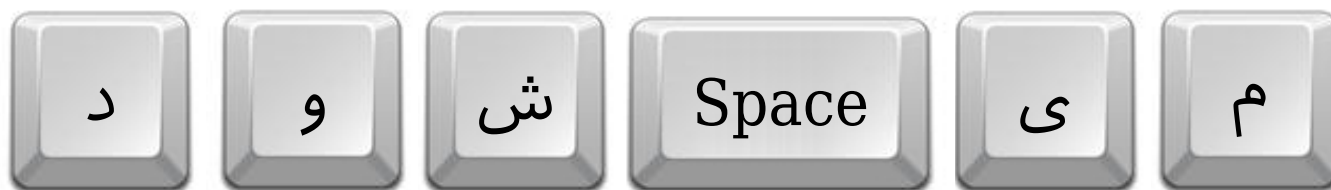
مثال‌هایی از سطح ۲ کیبورد استاندارد فارسی

میشود



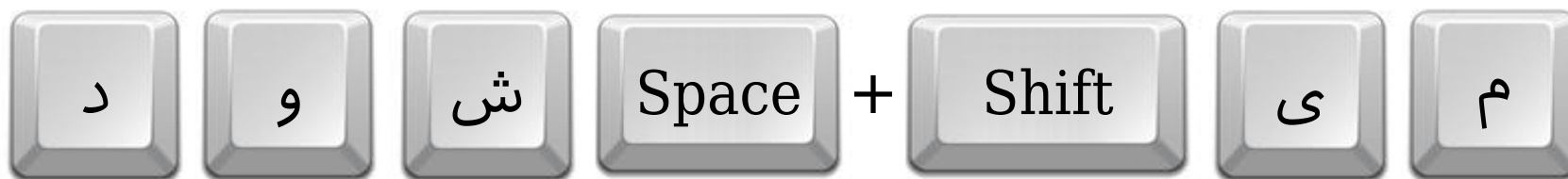
مثال‌هایی از سطح ۲ کیبورد استاندارد فارسی

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

پس بر	-	-	-	نشانه راست به چپ	نشانه چپ به راست	•	&	^	%	\$	#	@	`	~	
زیر متن راست به چپ		زیر متن چپ به راست		پایان زیر متن		زیر متن اکیداً		زیر متن اکیداً		€		°		جهش	
ورود		"		;		↵		↵		ا		د		قفل تبدیل	
تبدیل		?		'		,		...		ء		اتصال مجازی		تبدیل	
مهار	دگرساز راست		فاصله نشکن										دگرساز		مهار

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 ...



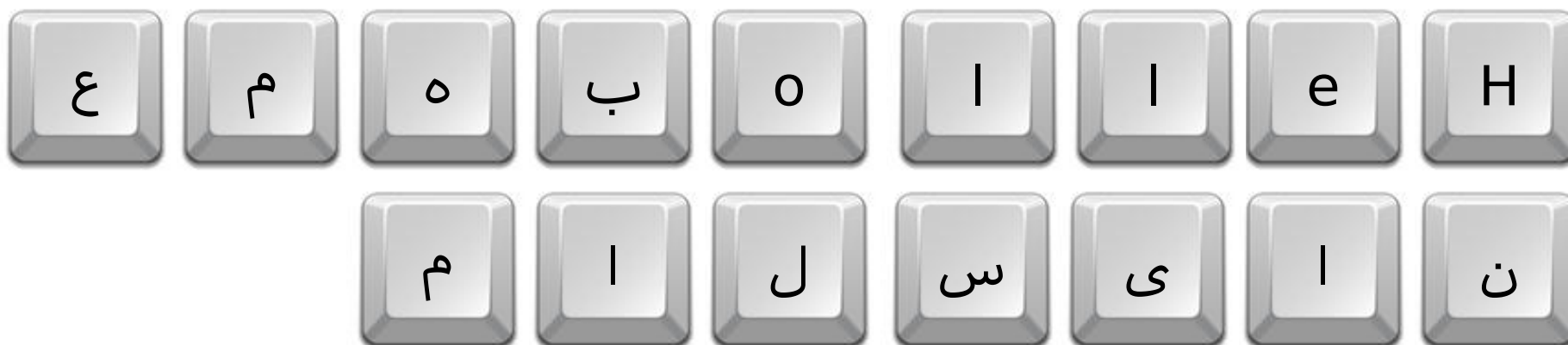
مثال‌هایی از سطح ۳ کیبورد استاندارد فارسی

و ...



مثال‌هایی از سطح ۳ کیبورد استاندارد فارسی

به معنای سلام Hello



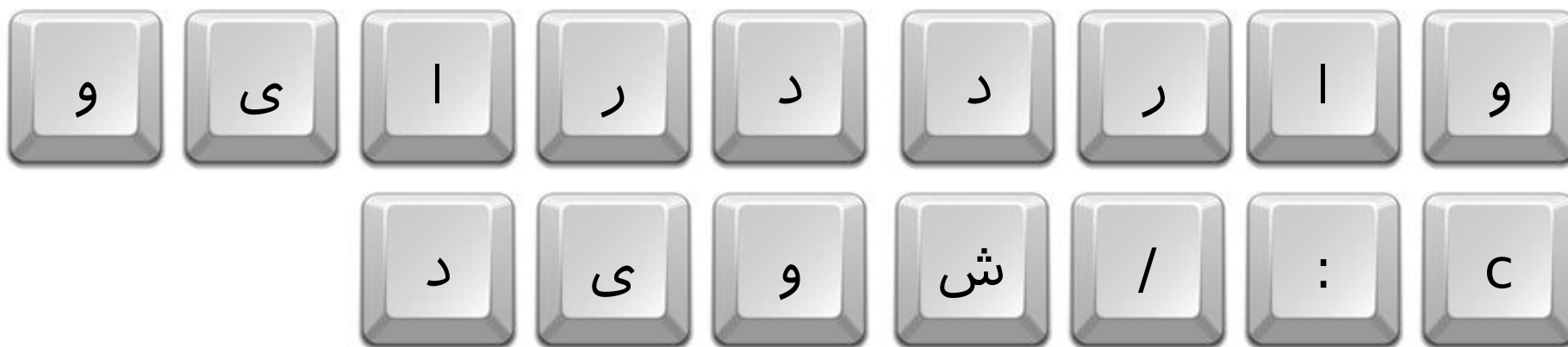
مثال‌هایی از سطح ۳ کیبورد استاندارد فارسی

Hello به معنای سلام



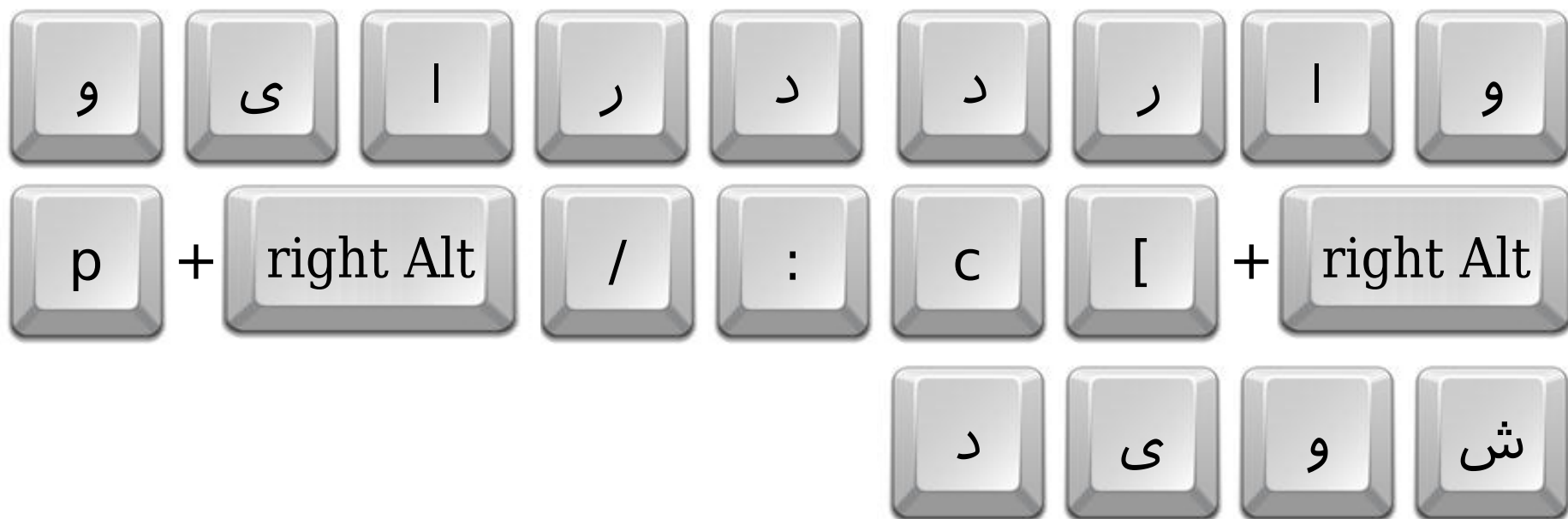
مثال‌هایی از سطح ۳ کیبورد استاندارد فارسی

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”

یونی کڈ

توسط بیژن ابراهیمی