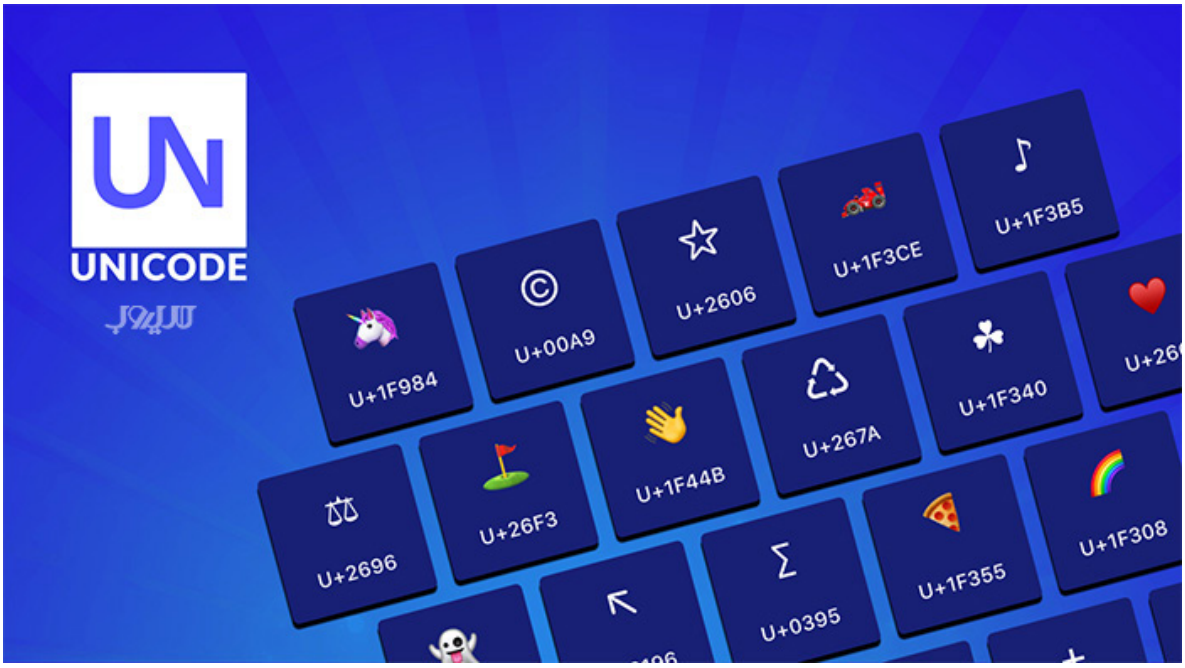
از unicode چه می‌دانید UTF-8 چیست؟

نویسنده: مهران منصوری فر

یونیکد چیست؟ از unicode چه می‌دانید UTF-۸ چیست؟ در این مقاله شما را با یونیکد و روش های کد گذاری و رایجترین روش کد گدگذاری آشنا خواهیم کرد با ما همراه باشید.

وقتی که شما کاراکتری را در یک برنامه ویرایش متن و یا یک اپلیکیشن وب قرار می‌دهید، کامپیوتر این داده‌ها و اطلاعات را آنگونه که هستند نمی‌تواند پردازش کند. کامپیوترها تنها قادر به پردازش اطلاعات و داده‌هایی هستند که به صورت اعداد و ارقام باشند. از این رو نیاز است که برای قابل فهم کردن اطلاعات و داده‌ها برای کامپیوترها، از روش‌های کدگذاری استفاده کنیم. حال سوال این است که کدگذاری چیست؟ روش‌های کدگذاری کدامند؟ کدام روش گزینه‌ای مناسب و بهینه است؟ یونی‌کد یا همان unicode چیست؟ UTF-۸ چیست و چرا محبوب شده؟ برای پاسخ دادن به این دسته از سوالات با ادامه متن همراه شوید تا بیشتر با مفهوم یونیکد و UTF-۸ آشنا شوید.



کدگذاری در کامپیوترها

همه ما می‌دانیم که کامپیوترها تنها با اعداد و ارقام سروکار دارند و تمام اطلاعات نوشتاری، صوتی و تصویری را به صورت اعداد و ارقام پردازش و ذخیره می‌کنند. حروف، اعداد و علایمی که در اپلیکیشن‌های وب مورد استفاده قرار می‌گیرند، به آن شکلی که شما آنها را می‌بینید در کامپیوتر مدیریت نمی‌شوند. برای قابل فهم کردن اطلاعات برای کامپیوتر لازم است برای هر حروف از الفبا، یک عددی اختصاص دهیم. حروف و کاراکترها به مجموعه‌ای از ۰ و ۱ تبدیل می‌شود تا مدیریت آنها برای کامپیوتر ساده‌تر باشد. اختصاص این کدها به اطلاعات توسط سیستم‌های کدگذاری انجام خواهد شد. برای این منظور صدها نوع سیستم کدگذاری برای قابل فهم کردن زبان‌های مختلف برای کامپیوترها به وجود آمد.

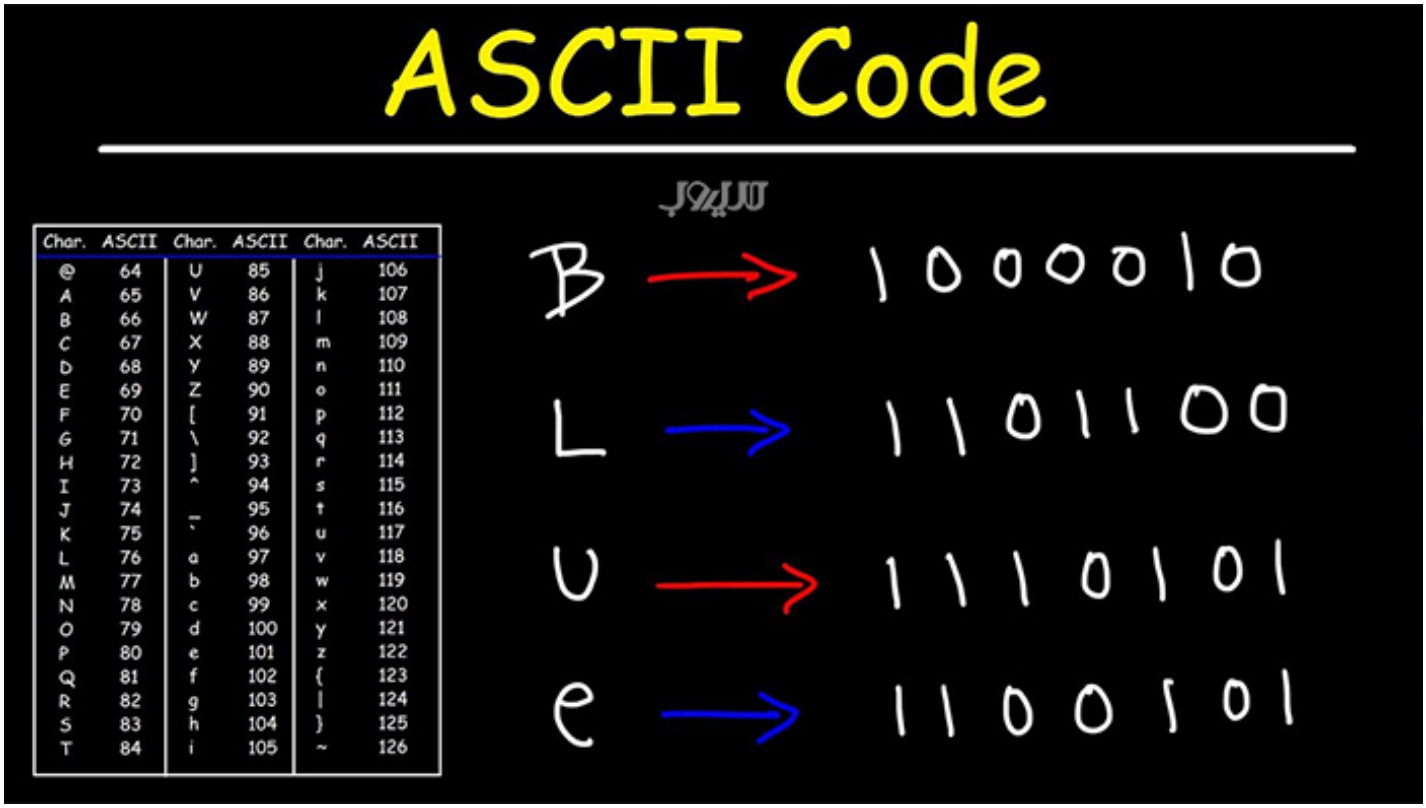
برای زبان فارسی هم تعداد زیادی سیستم‌های کدگذاری به وجود آمد. هر شرکت نرم‌افزاری یک سیستم کدگذاری مخصوص به خودش را داشت. البته وجود تعداد زیاد سیستم‌های کدگذاری تنها مختص به زبان فارسی نبوده و بیشتر زبان‌های دیگر هم با این مشکل روبرو بودند.

کد اسکی یا ASCII چیست؟

انجمن استانداردهای آمریکا در سال ۱۹۶۰ روش کدگذاری ۷ بیتی ASCII را معرفی کرد ASCII مخفف عبارت American Standard Code for Information Interchange است که در آن زمان شامل ۱۲۸ کاراکتر یا ۷ بیت تعریف شد. این استاندارد در آن زمان بیشتر برای زبان‌های لاتین کاربرد داشت. پس از آن در دهه ۱۹۸۰ تصمیم گرفتند که این استاندارد به جای استفاده از ۷ بیت، از یک بایت کامل استفاده

کند. یک بایت کامل شامل ۸ بیت و ۲۵۶ کاراکتر است. از این رو زبان‌های دیگر نیز می‌توانستند از این استاندارد استفاده کنند.

ASCII به روشنی مشخص نکرده که مقادیر بین ۱۲۸ تا ۲۵۵ به چه چیزی اختصاص دارد. در بین زبان دیگر استاندارد واحدی وجود نداشت و هر زبانی الفبای خود را با کد مختص به الفبای خود نشان می‌داد. پس در این زمان به استاندارد واحدی که با تمامی زبان‌ها سازگار باشد و برای هر کاراکتر کد مختص به خود را داشته باشد، نیاز بود. برای حل این مشکل سازندگان رایانه‌ها سعی کردند از صفحه‌های کد (Code Pages) استفاده کنند. اما باز هم این روش کارساز نبود. تا زمانی که افرد از کد صفحه‌های یکسانی استفاده کنند، همه چیز خوب پیش می‌رود. و اما اگر کد صفحه‌ها برای افراد یکسان نباشد، همه چیز به هم می‌ریزد.



Char.	ASCII	Char.	ASCII	Char.	ASCII
@	64	U	85	j	106
A	65	V	86	k	107
B	66	W	87	l	108
C	67	X	88	m	109
D	68	Y	89	n	110
E	69	Z	90	o	111
F	70	[	91	p	112
G	71	\	92	q	113
H	72	]	93	r	114
I	73	^	94	s	115
J	74	_	95	t	116
K	75	`	96	u	117
L	76	a	97	v	118
M	77	b	98	w	119
N	78	c	99	x	120
O	79	d	100	y	121
P	80	e	101	z	122
Q	81	f	102	{	123
R	82	g	103	|	124
S	83	h	104	}	125
T	84	i	105	~	126

وجود یک استاندارد واحد برای کدگذاری در بین زبان‌های مختلف

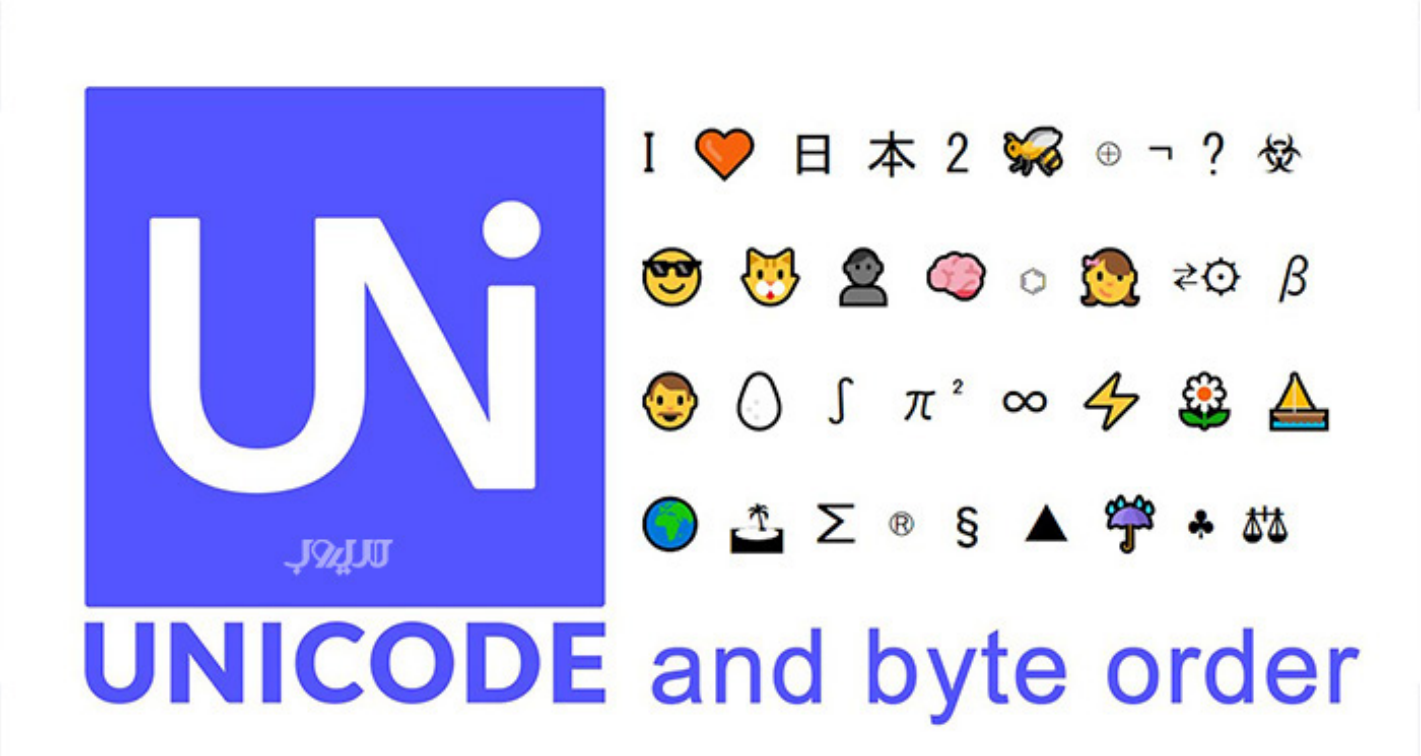
کلید حل این مشکل وجود یک استاندارد واحد بود. بر این اساس مشخص می‌شود که هر کدام از این اعداد چه کاراکترهایی را نمایش می‌دهند. در ابتدا دو استاندارد برای ایجاد مجموعه کاراکترهای واحد صورت گرفت. اولی ISO-۱۰٦٤٦ و دیگری Unicode بود. اما وجود دو استاندارد باز هم مشکل را به صورت کامل حل نکرد. بر این اساس ISO و Unicode تصمیم گرفتند در سال ۱۹۹۱ به یکدیگر بپیوندند. از این رو با معرفی یونیکد (unicode) این مشکل حل شد. حال سوال این است که یونیکد چیست؟ با ادامه متن همراه شوید تا با این استاندار آشنا شوید.

یونیکد یا Unicode چیست؟

یونیکد یا همان UNIVERSAL CHARACTER SET TRANSFORMATION FORMAT یک استاندارد بین‌المللی است که برای تبادل اطلاعات چندزبانه مورد استفاده قرار می‌گیرد. Unicode مستقل از سیستم عامل و یا برنامه و زبان خاصی، به هر یک از حروف یک کد یکتا اختصاص می‌دهد. Unicode می‌تواند تمام حروف زبان‌های مختلف دنیا را در خود جای دهد. یونیکد می‌تواند برای وبسایت‌ها و برنامه‌ها بسیار مفید باشد. از این رو می‌توان گفت که مهم نیست کاربران از چه وبسایت و یا چه مرورگری استفاده می‌کنند؛ تنها کافی است از Unicode پشتیبانی کند.

امروزه اکثر شرکت‌های بزرگ دنیای کامپیوتر از این استاندار استفاده می‌کنند و همچنین می‌توان گفت که تقریبا تمام برنامه‌های کاربردی جدید با این استاندارد کدگذاری شده‌اند. گسترش استاندارد Unicode موجب شده تا تمامی فارسی

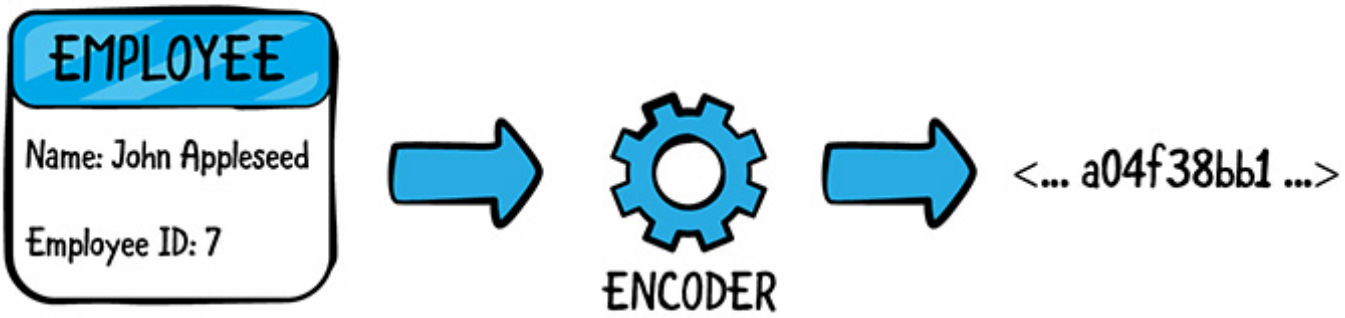
زبان‌ها هم بتوانند در دنیای اینترنت مطالب خود را عرضه کنند. یونیکد موجب شده تا فرایند ایجاد وبسایت‌ها و برنامه‌های فارسی بسیار آسان‌تر و کم هزینه‌تر باشد. یونیکد در واقع مجموعه‌ای از کاراکترست (charset) با اعداد منحصر به فرد است که به آنها در اصطلاح پوینت کد (Point Code) گفته می‌شود. هر Point Code کاراکتر واحدی را نمایش می‌دهد.



انکودینگ یا همان Encoding چیست؟

تبدیل داده‌ها به صورتی که سیستم توانایی خواندن و استفاده از آن را داشته باشد Encoding گویند. کدهای یکتا به روش‌های متفاوتی در کامپیوتر ذخیره می‌شوند؛ این روش‌ها را کدگذاری یا Encoding می‌گویند. می‌توان گفت که اینکودینگ فرآیند تبدیل داده‌ها به فرمت‌های مورد نیاز است. این رمزگذاری شامل تدوین

برنامه‌ها، اجرای برنامه انتقال و ذخیره‌سازی داده‌ها و همچنین پردازش داده‌های برنامه است.



روش‌های کدگذاری یونیکد

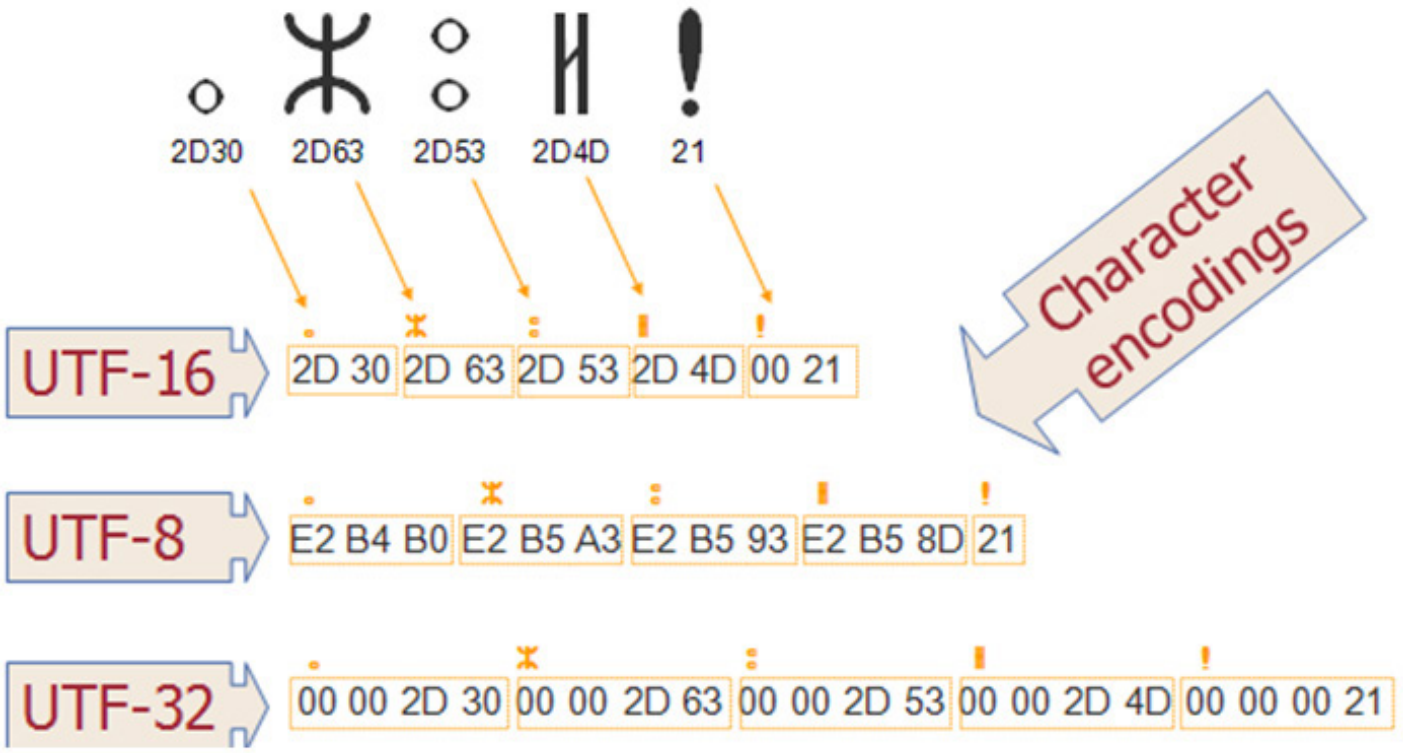
یونیکد به سه روش مختلف کدگذاری می‌شود؛ UTF-۸، UTF-۱۶ و UTF-۳۲. حال سوال این است که UTF چیست؟ تفاوت این روش‌های کدگذاری در چیست؟ UTF مخفف عبارت Unicode Transfer Format است که به معنی «فرمت تحول یونیکد» است. UTF روش کدگذاری است که زیر مجموعه‌ای از استاندارد یونیکد به شمار می‌رود. در ادامه بیشتر با روش‌های کدگذاری یونیکد و تفاوت‌های آنها آشنا خواهید شد.

مقایسه روش‌های کدگذاری UTF-۸، UTF-۱۶ و UTF-۳۲

از تفاوت‌های این سه روش کدگذاری می‌توان به نحوه ارائه حروف، اعداد و علایم در بین زبان‌های مختلف اشاره کرد. می‌توان گفت نحوه ارائه کاراکترها در یک

کشور با کشور دیگر متفاوت است. روش‌های کدگذاری UTF-۸ و UTF-۱۶ هر دو دارای عرض متغیر هستند و می‌توانند از حداکثر ٤ بایت برای رمزگذاری استفاده کنند. اما وقتی به حداقل رسید، UTF-۸ فقط از یک بایت (معادل ۸ بیت) استفاده می‌کند ولی UTF-۱۶ از ۲ بایت (معادل ۱٦ بیت) استفاده می‌کند. این تفاوت تاثیر زیادی در اندازه پرونده‌های رمزگذاری شده دارد. به زبانی دیگر می‌توان گفت که یک فایل رمزگذاری شده با UTF-۱۶ تقریبا دو برابر بزرگتر از پرونده رمزگذاری شده با UTF-۸ است. UTF-۳۲ برخلاف دو روش قبلی، طول ثابتی دارد و بیشترین فضا را اشغال می‌کند.

از سوی دیگر می‌توان گفت که UTF-۸ با ASCII سازگار است اما روش رمزگذاری UTF-۱۶ با ASCII ناسازگار است. روش کدگذاری UTF-۸ بایتگراست و با شبکه‌ها و پرونده‌های بایتگرا مشکلی ندارد؛ اما UTF-۱۶ بایتگرا نیست و برای کار با شبکه‌های بایتگرا نیاز به سفارش بایت دارد. همچنین می‌توان گفت که UTF-۸ در بازیابی از خطاها در مقایسه با UTF-۱۶ بهتر عمل می‌کند. در این مواقع UTF-۸ می‌تواند بایت غیر فاسد بعدی را رمزگشایی کند. UTF-۱۶ هم در صورت خراب شدن بایت‌ها همین کار را می‌کند اما زمانی که برخی از بایت‌ها گم شدند، مشکل وجود دارد. بایت گمشده ترکیب‌های بایت را با هم مخلوط می‌کند و نتیجه نهایی هدر می‌شود.



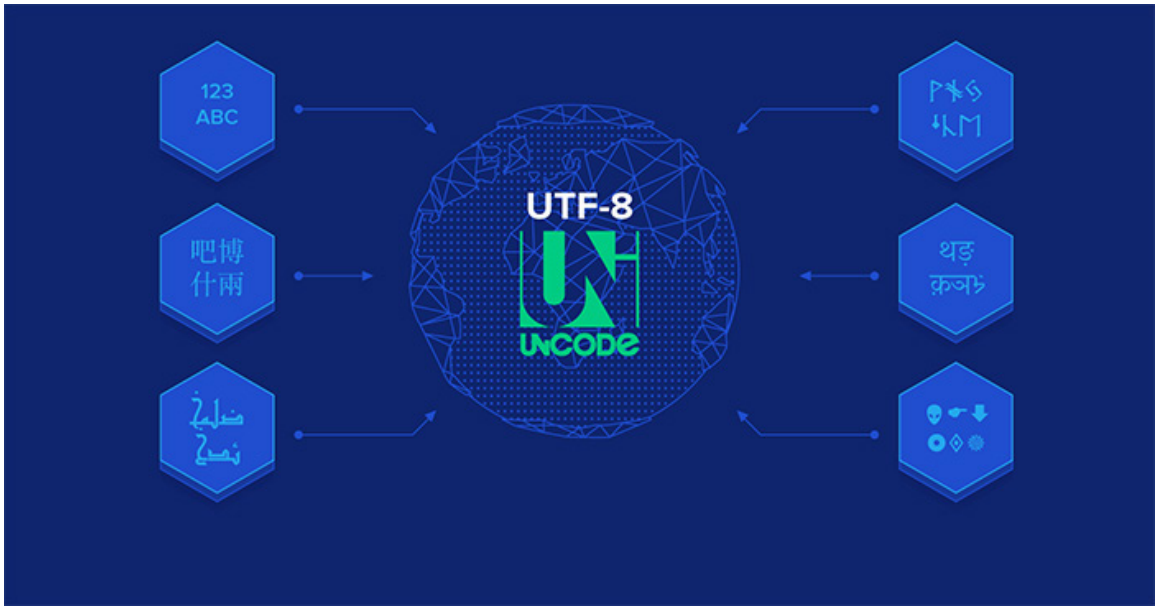
۸-utf چیست؟

۸-UTF مخفف عبارت Unicode Transformation Format ۸-bit به معنای فرمت تبدیل یونیکد ۸ بیتی است. ۸-UTF یکی از روش‌های رمزگذاری یک بایتی (معادل ۸ بیت) با عرض متغییر است که برای ارتباط الکترونیکی استفاده می‌شود. در کنفرانس USENIX در سال ۱۹۹۳، ۸-UTF به طور رسمی معرفی شد. ۸-UTF پرکاربردترین و رایجترین روش برای نمایش متن یونیکد در صفحات وب است و همیشه باید هنگام ایجاد صفحات وب و پایگاه داده خود از ۸-UTF استفاده کنید. ۸-UTF کدگذاری غالب برای شبکه جهانی وب (و فناوری‌های اینترنت) است که تا سال ۲۰۲۲، ۹۸٪ از کل صفحات وب و تا ۱۰۰٪ برای برخی از زبان‌ها را شامل می‌شود.

در این روش کدگذاری هر کاراکتر با یک تا چهار بایت نمایش داده می‌شود. ۸-UTF

با ASCII سازگار است و می‌تواند هر کاراکتر استاندارد یونیکد را نشان دهد. این استاندارد رمزگذاری قادر است همه‌ی کد کاراکترها معتبر در یونیکد را با استفاده از یک تا چهار واحد کد یک بایتی (۸ بیتی) رمزگذاری کند. UTF-8 یکی از روش‌های رمزگذاری است که توسط سازمان بین المللی استاندارد (ISO) در ISO ۱۰۶۴۶ تعریف شده است. این کد می‌تواند حداکثر ۲,۰۹۷,۱۵۲ نقطه کد (۲۱۸۲) را نشان دهد که بیش از اندازه کافی برای پوشش ۱,۱۱۲,۰۶۴ کاراکتر یا پوینت کد فعلی است.

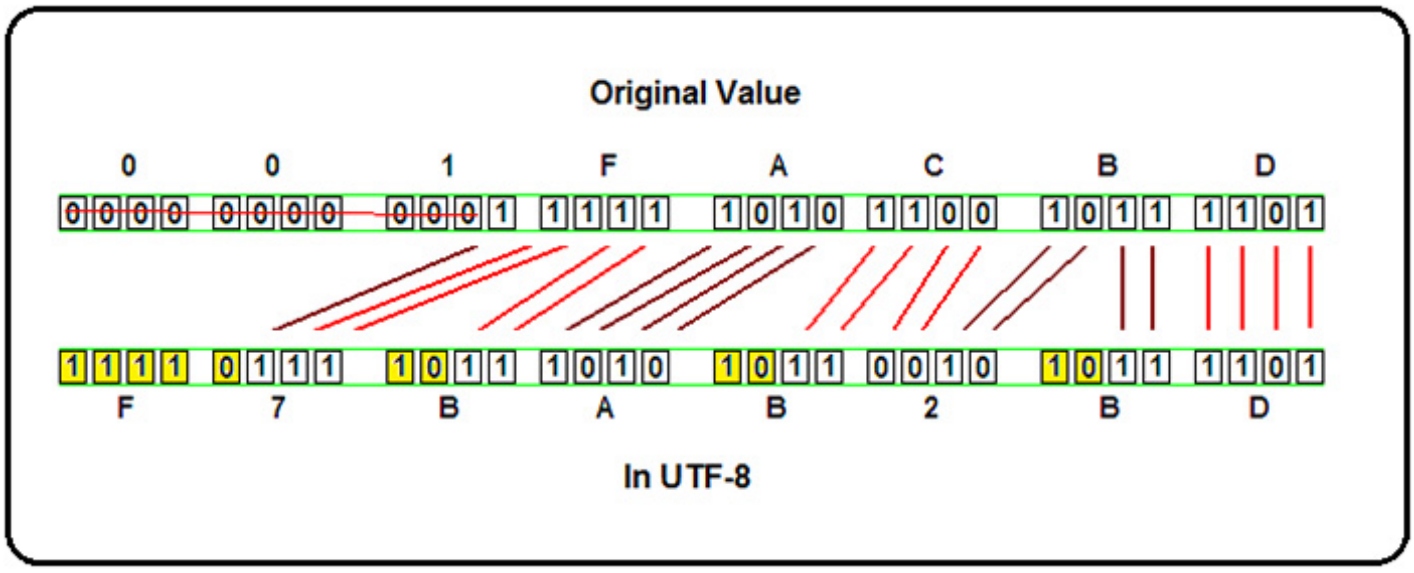
همان طور که گفته شد، UTF-8 یک استاندارد رمزگذاری «با عرض متغیر» است. حال سوال این است که طول متغییر به چه معنا است؟ این بدان معنی است که هر نقطه کد را با تعداد متفاوتی از بایت‌ها، بین یک تا چهار بایت رمزگذاری می‌کند. این کار برای صرفه جویی در فضا بسیار مناسب است. نقاط کد رایج مورد استفاده معمولا با بایت‌های کمتری نسبت به نقاط کد که به ندرت مورد استفاده قرار می‌گیرد، کدگذاری می‌شود. . UTF-8 الگوریتمی است که اعداد مربوط به پوینت‌کدها را به باینری تبدیل می‌کند. از این رو می‌توان آنها را بر روی دیسک ذخیره کرد.



چرا utf-۸ رایج‌ترین و پرکاربردترین روش کدگذاری است؟

همان طور که به آن اشاره کردیم، UTF-۸ به دلیل وجود ویژگی‌ها و مزایای خوبی که دارد، یکی از رایج‌ترین و پرکاربردترین روش‌های کدگذاری تا به امروز است. از جمله مزایای این روش کدگذاری می‌توان به موارد زیر اشاره کرد.

- یکی از مهمترین مزایای UTF-۸ می‌توان به عرض متغییر اشاره کرد؛ اگر در عرض هر کاراکتر یونیکد با چهار بایت نمایش داده می‌شد، یک فایل متنی که به زبان انگلیسی نوشته شده بود چهار برابر اندازه همان فایل رمزگذاری شده با UTF-۸ خواهد بود.
- از دیگر مزایای آن می‌توان به سازگاری با ASCII اشاره کرد. این روش رمزگذاری از کدهای ۰ تا ۱۲۷ برای کاراکترهای اسکی استفاده می‌کند. برای نمایش کدهای اسکی، UTF-۸ نیازی به افزایش حجم ندارد.
- UTF-۸ بایتگراست و با شبکه‌ها و پرونده‌های بایتگرا مشکلی ندارد.
- UTF-۸ در بازیابی از خطاها بسیار خوب عمل می‌کند. اگر بایت‌ها به دلیل وجود خطا و یا مشکلی از بین بروند، UTF-۸ کاراکتر معتبر بعدی را پیدا می‌کند و پردازش را شروع می‌کند.
- UTF-۸ از عملیات ساده بیتی استفاده می‌کند و به عملیات ریاضی مانند ضرب و تقسیم نیازی ندارد.
- UTF-۸ نیازی به BOM یا شاخص کدگذاری ندارد.
- UTF-۸ یکی از روش‌های کدگذاری است که قادر است هر کارکتر یونیکد را کدگذاری کند. UTF-۸ قادر است بدون اینکه مجبور باشند فونت درستی را انتخاب کنند، با اسکریپت‌های متفاوت به درستی فایل‌ها را نمایش دهد.

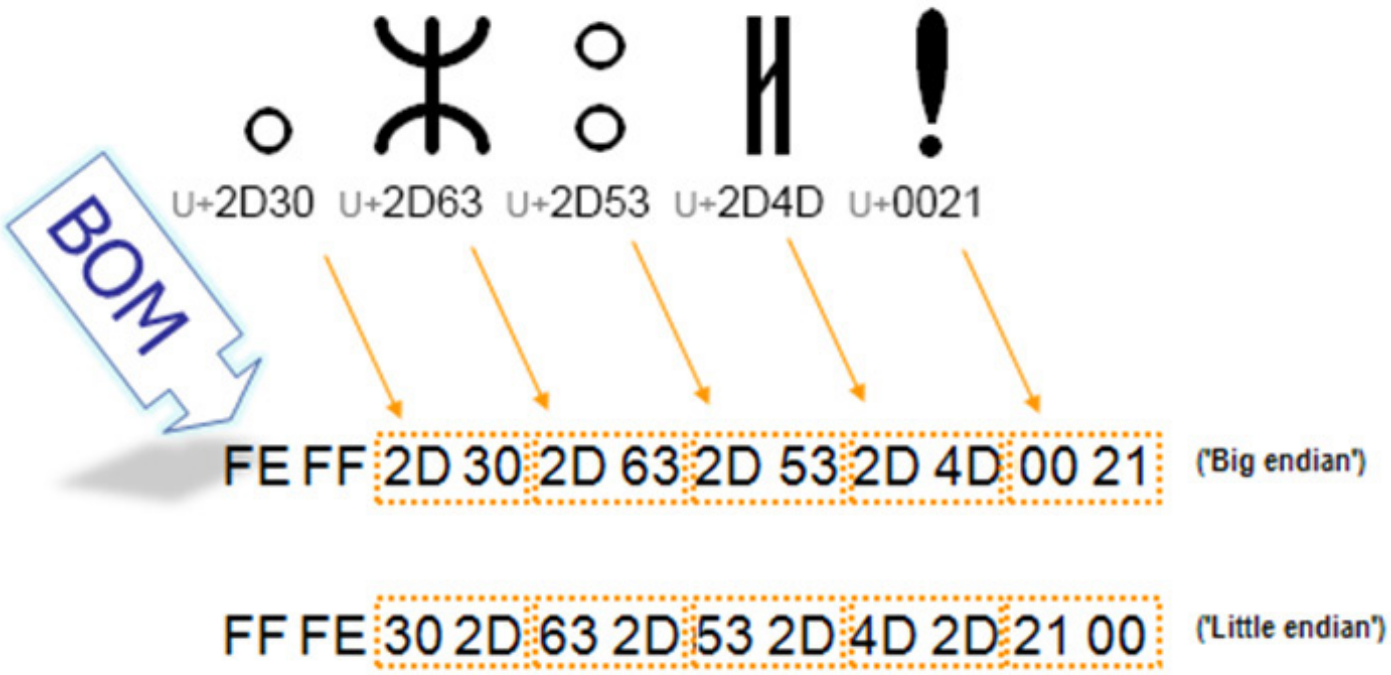


معایب استفاده از روش کدگذاری UTF-۸

استفاده از UTF-۸ چندین معایب دارد که در زیر به برخی از آنها اشاره می‌کنیم.

- شما نمی‌توانید تعداد بایت‌های متن UTF-۸ را از تعداد کاراکترهای UNICODE تعیین کنید زیرا UTF-۸ از یک رمزگذاری طول متغیر استفاده می‌کند.
- UTF-۸ برای آن دسته از کاراکترهای غیر لاتین به ۲ بایت نیاز دارد. این کاراکترها تنها با ۱ بایت در ASCII کدگذاری می‌شوند.
- کدگذاری با UTF-۸ نسبت به Encoding چند بایته که برای یک زبان خاص طراحی شده، حجم بالاتری دارد. در روش کدگذاری چندبایته مختص به یک زبان، برای هر کاراکتر به دو بایت حجم نیاز است، اما در UTF-۸ به سه بایت نیاز هست.
- کدگذاری با UTF-۸ برخی از نرم‌افزارهایی مانند ویرایشگر متن را نمی‌تواند نمایش دهد یا ترجمه کند. البته اگر متن با یک BOM شروع شود این مشکل حل می‌شود.

- کاراکترهایی که در روش‌های کدگذاری ISO-۸۸۵۹ و WINDOWS-۱۲۵۲ تنها با یک بایت نمایش داده می‌شوند، در UTF-۸ به ۲ بایت حجم برای نمایش نیاز دارند.
- می‌توان گفت که متون کدگذاری شده با UTF-۸، بجز برای کاراکترهای اسکی، به حجم بالاتری نسبت به سیستم‌های دیگر نیاز دارد.



جمع‌بندی

همان طور که گفته شد کامپیوترها برای اینکه بتوانند اطلاعات نوشتاری، صوتی و تصویری را پردازش کنند به کدهایی که به صورت اعداد و ارقام باشد نیاز دارد. برای این کدگذاری روش‌های مختلفی از جمله اسکی وجود دارد. یکی از روش‌های استاندارد و مشترک در بین زبان‌های مختلف جهان می‌توان به یونیکد اشاره کرد. یونیکد هم برای کدگذاری از سه روش مختلف استفاده کرده است که UTF-۸

رایج‌ترین و کاربردی‌ترین است. دلیل محبوبیت بالای این روش کدگذاری سازگاری با اسکی است. UTF-۸ تمامی کاراکترهای اسکی را تنها در یک بیت قرار می‌دهد. پس می‌توان گفت که UTF-۸ هم با نسخه‌های قدیمی کدگذاری سازگار است و هم برای زبان‌های انگلیسی و دیگر زبان‌های اروپایی بهینه‌تر است.