

SC4002 CE4045 CZ4045 Natural Language Processing

Introduction to UTF-8

Dr. Sun Aixin

UTF-8 topics are unexaminable



Before we talk about language

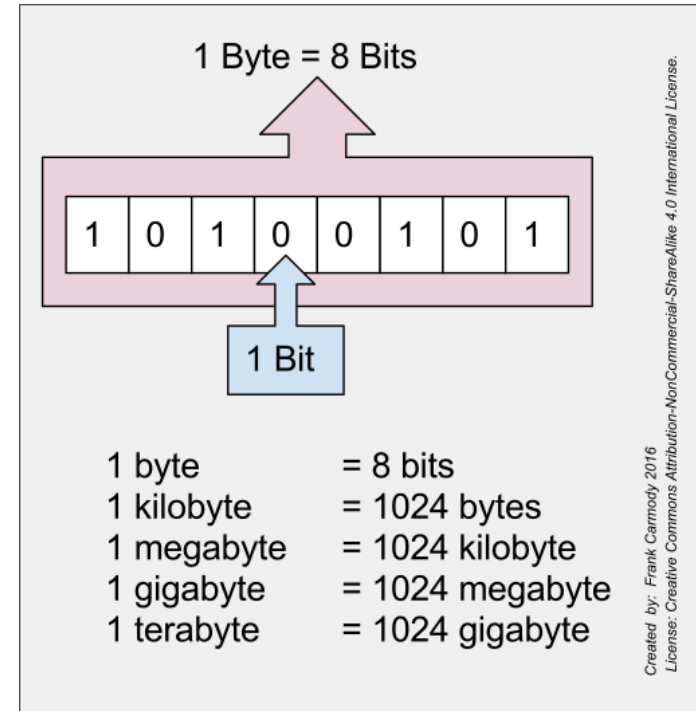
➤ Computer recognizes and stores **0** and **1**

- Bits, bytes

➤ How does computer store text and symbols?

- “Hello World”
- ☺ ☹
- “自然语言”

<i>ε</i> 1D700	<i>υ</i> 1D710	<i>E</i> 1D720	<i>Υ</i> 1D730	<i>λ</i> 1D740	<i>ε</i> 1D750	<i>Λ</i> 1D760	<i>α</i> 1D770	<i>ρ</i> 1D780
<i>ζ</i> 1D701	<i>φ</i> 1D711	<i>Z</i> 1D721	<i>Φ</i> 1D731	<i>μ</i> 1D741	<i>ϑ</i> 1D751	<i>M</i> 1D761	<i>β</i> 1D771	<i>ς</i> 1D781
<i>η</i> 1D702	<i>χ</i> 1D712	<i>H</i> 1D722	<i>X</i> 1D732	<i>ν</i> 1D742	<i>κ</i> 1D752	<i>N</i> 1D762	<i>γ</i> 1D772	<i>σ</i> 1D782
<i>θ</i> 1D703	<i>ψ</i> 1D713	<i>Θ</i> 1D723	<i>Ψ</i> 1D733	<i>ξ</i> 1D743	<i>φ</i> 1D753	<i>Ξ</i> 1D763	<i>δ</i> 1D773	<i>τ</i> 1D783



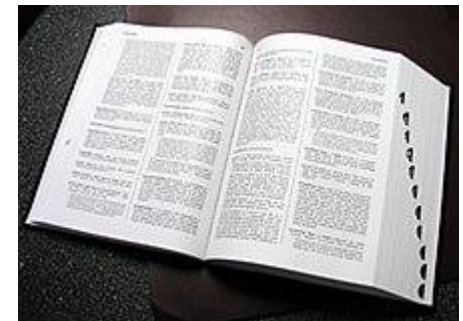
➤ Encoding scheme: a way to represent characters in binary

- Unicode
- Non-Unicode

Unicode

- Unicode is a computing industry standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems.
 - The standard is maintained by the Unicode Consortium
 - Unicode 12.1, contains a repertoire of 137,994 characters, covering 150 modern and historic scripts, and multiple symbol sets and emoji.
 - Unicode 14.0 was released in September 2021, Unicode 15.0 will be released in September 2022

- **Each character is assigned a unique integer code**, called “code points”, usually in hexadecimal base
 - Code point is in the form of U+<hex-code>, from U+0000 to U+10FFFF.
 - Characters in English, Chinese, or other languages
 - Currency symbols, Mathematical symbols
 - Emojis e.g., 🐶 U+1F436

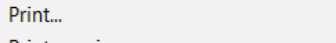


Display with different encodings

UTF-8 (8-bit Unicode Transformation Format) 是一种针对Unicode的可变长字符集中的所有有效编码点进行编码，属于Unicode标准的一部分，最初由Ken Thompson设计。它使用8位字节进行编码，这使得原来处理ASCII字符的硬件和软件可以无修改地使用。UTF-8就是为了解决向后兼容而设计的编码方式。

UTF-8 尽管如此，2003年11月UTF-8被RFC 3629 (Unicode 3.1 字节)：

[Add to favorites...](#)
[View source](#)
[Inspect element](#)



对上面提及的第四种字符而言，UTF-8使用且它的另一种选择，UTF-16编码，对前述视所使用的字符的分布范围而定。不过，如果使用一些传统的压缩系统，比如

UTF-8鑄?8-bit Unicode Transformation Format鑄爰櫛涓?紈確擴漢?Unicode
互鑒尤鐳鑄冲涑涓 砮鐳倣 Unicode瀛� 閫�囡鑄勒堅鏈爰沿鑄鞣纒鑄倣倣
鋒堡鑄 ??鍐?湘橋鑄路悞厠鑄懇語鉅?^{[2][3]} 鑄爰落伾富鑄鍊肩琿瑋杓鑄鏡鑄迥鑄

笈铨岢	元閑乞氮路瑰咲瀛悃	L闊淬	伉TF-8灝辨楸涓篆箇璫e	昱錫戔悅鐳煎
漢瑰籜	Back		方繡鏐睬	肩殄錦蒙釜瀛曄媛抄澆
桺海仵	Forward		鉅倣氛姝洵滄江	愴筦銀恨負鑾駁聿

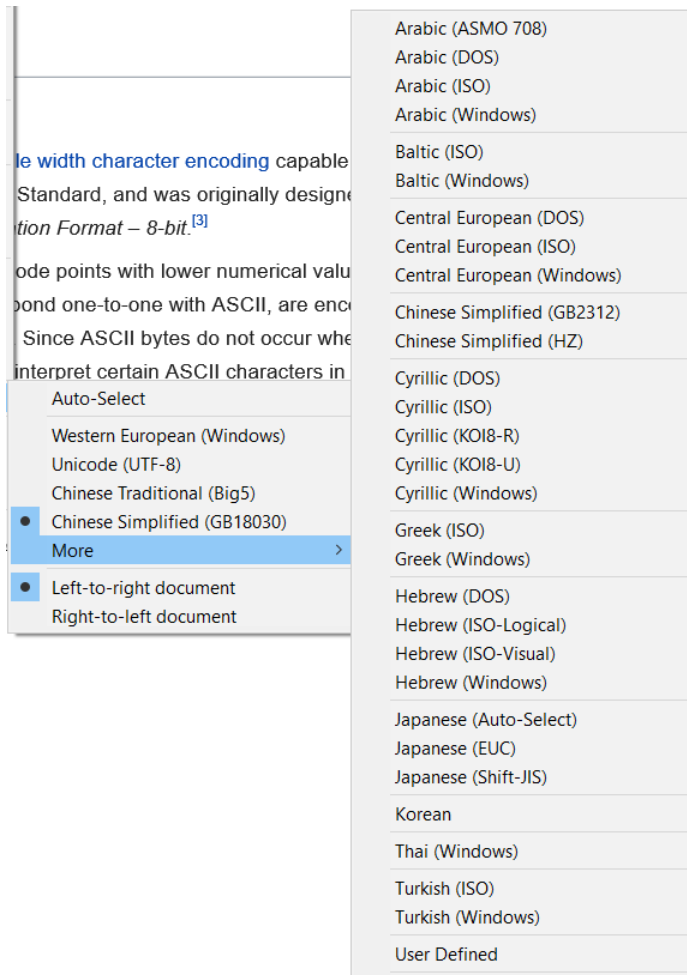
鑪	Save background as...	相敬克鍾?涓昏 鑪動紕鑪佅觀窺氣紙
剝鼓W	Set as background	課兵庫壑鍾變簣鑪?(for all things)鈇
柎鑪臺	Copy background	枳鑪疊SCII律柎鑪饒軒泵涓哄昂鑪底
綉縹典	Select all	嘉編爰律柎鑪經瑠紡Shift JIS鍛燭B 23
	Paste	

Consortium, IMC 铸文缓璁 璁璁委
ML 鑄函欢鍍了 TML 鑄函欢鑄勳規璁よ

UTF-8	Create shortcut	迂行 紕 鐮 恍 紙 灝 界 濡 容 铸 2003
鑿 丿 師	Add to favorites...	00 錄 瘡 + 10FFFF 铸 岬 管 灝 辨 嶽 璇 存 漢 漢

-
1. View source
Inspect element
2. Encoding >
Print...
3. Print preview...
Refresh
4. Properties
- Auto-Select
Western European (Windows)
Unicode (UTF-8)
Chinese Traditional (Big5)
● Chinese Simplified (GB18030)
More >
● Left-to-right document
Right-to-left document
- Unicode
UTF-8

Unicode Transformation Format (UTF)

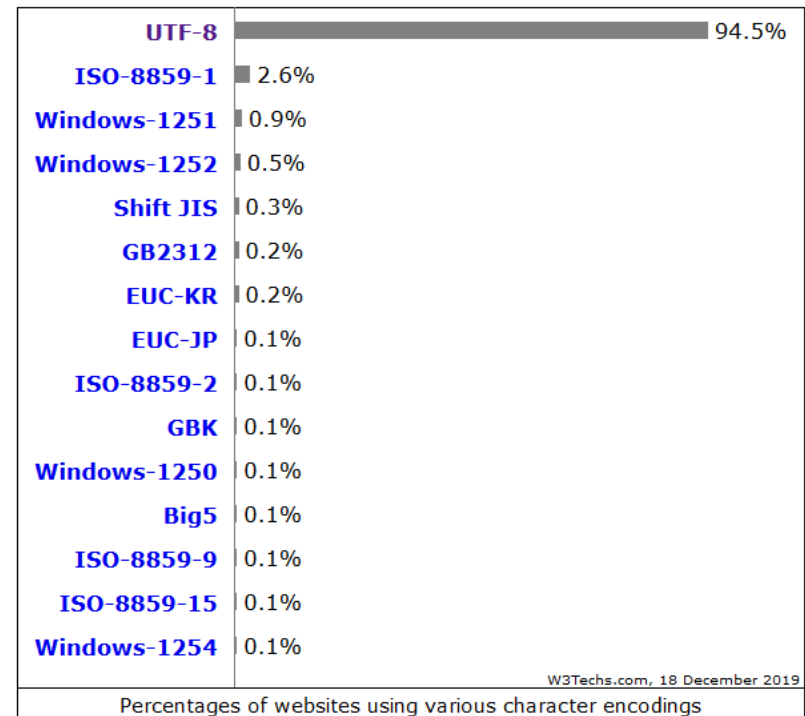


Usage of character encodings for websites

This diagram shows the percentages of websites using various character encodings [technologies overview](#) for explanations on the methodologies used in the surveys. Reports are updated daily.

How to read the diagram:

UTF-8 is used by 94.5% of all the websites whose character encoding we know.



Source: https://w3techs.com/technologies/overview/character_encoding



UTF-8

➤ UTF stands for Unicode Transformation Format

- The '8' means it uses 8-bit blocks to represent a character.

1st Byte	2nd Byte	3rd Byte	4th Byte	No. of Free Bits	Maximum Expressible Unicode Value
0 xxxxxxx				7	007F hex (127)
110 xxxxx	10 xxxxxx			(5+6)=11	07FF hex (2047)
1110 xxxx	10 xxxxxx	10 xxxxxx		(4+6+6)=16	FFFF hex (65535)
11110 xxx	10 xxxxxx	10 xxxxxx	10 xxxxxx	(3+6+6+6)=21	10FFFF hex (1,114,111)

ā

(Latin Small Letter A With Macron)

Unicode: decimal 257, binary 100000001

UTF-8 (binary) 11000100:10000001

<https://www.unicode.org/charts/>



Text processing

- Texts are stored in a continuous bit array of 0 and 1s

```
01001000 01100101 01101100 01101100 01101111 00100000  
01010111 01101111 01110010 01101100 01100100
```

Hello World

- Computer does not know any boundary regarding words or sentences;
- There are many different languages
 - With or without explicit word boundaries
 - Reads from left to right or right to left
 - We mainly focus on **English**



Jieba: 请 南京市 市 长江大桥 先生 致辞
SnowNLP: 请 南京市 市长 江 大桥 先生 致辞
PKUSeg: 请 南京市 市长江 大桥 先生 致辞
THULAC: 请 南京市 市 长江 大桥 先生 致辞
HanLP: 请 南京市 市 长江大桥 先生 致辞
FoolNLTK: 请 南京市 市长 江大桥 先生 致辞
LTP: 请 南京市 市长江 大桥 先生 致辞
CoreNLP: 请 南京市 市 长江 大桥 先生 致辞
BaiduLac: 请 南京市市 长江大桥 先生 致辞
Stanza: 请 南京 市 市 长 江 大桥 先生 致辞



Summary

- A very high-level introduction to Unicode and UTF-8
- There are other encodings, but are less widely used
- Computer stores text in a string of Zeros and Ones
- Computer does not know any boundary regarding words or sentences

Computer stores and display languages,
but does not understand languages (for now).

