SC4002 CE4045 CZ4045 Natural Language Processing

Introduction to UTF-8

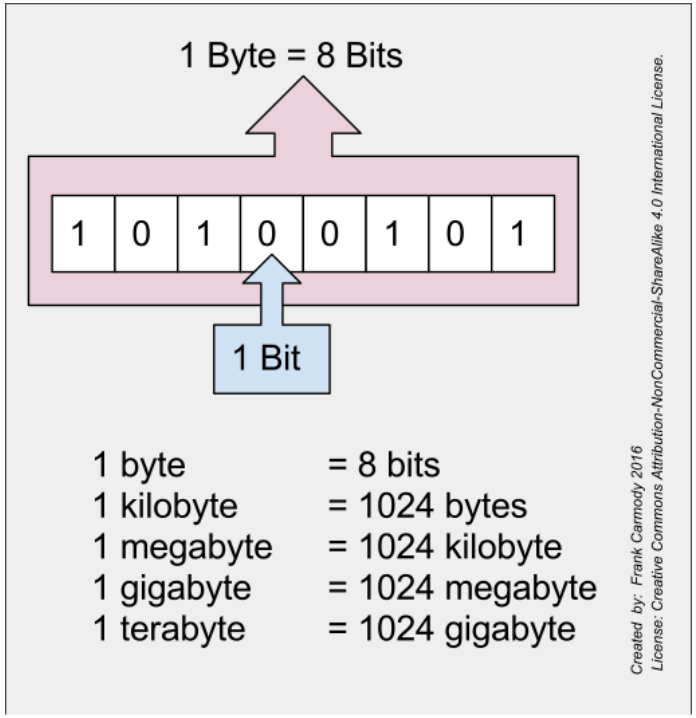
Dr. Sun Aixin

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"

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



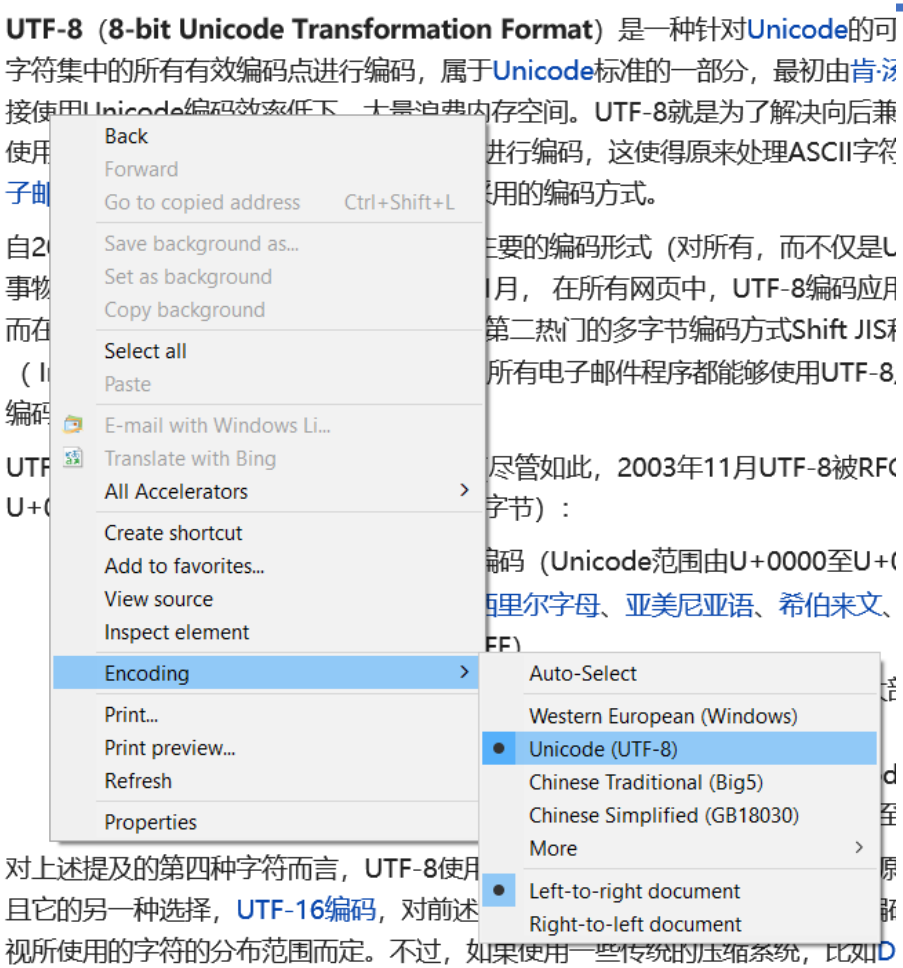
- Encoding scheme: a way to represent characters in binary
  - Unicode
  - Non-Unicode

Unicode

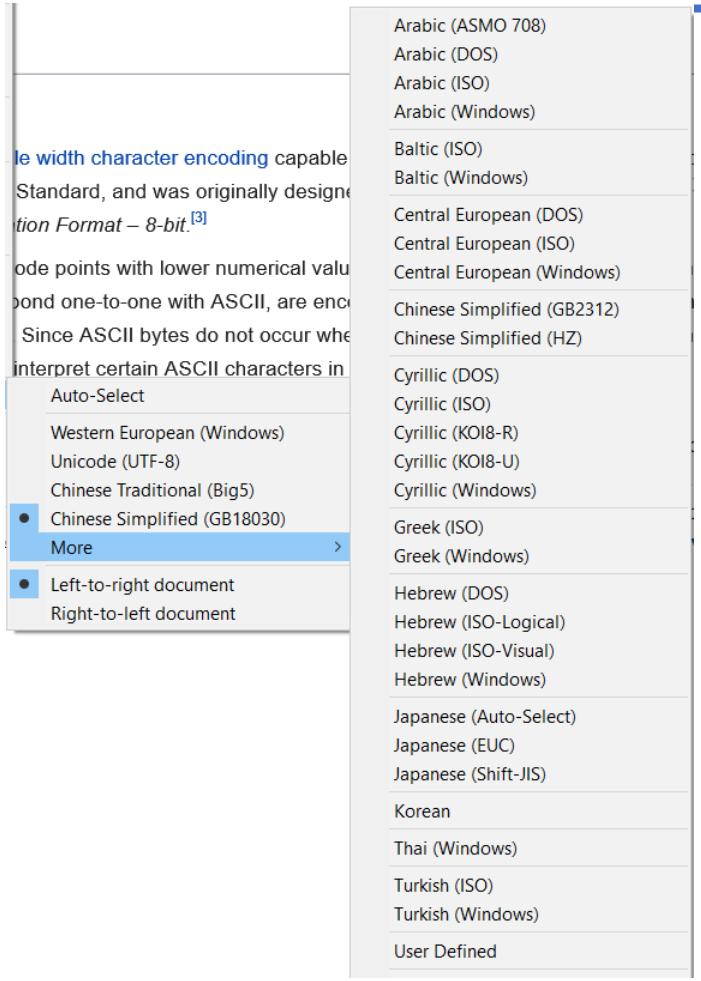
- Unicode is a computing industry standard for the consistent encoding, representation, and handling of text expressed in most of the world's writing systems.
  - The standard is maintained by the Unicode Consortium
  - Unicode 12.1, contains a repertoire of 137,994 characters, covering 150 modern and 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

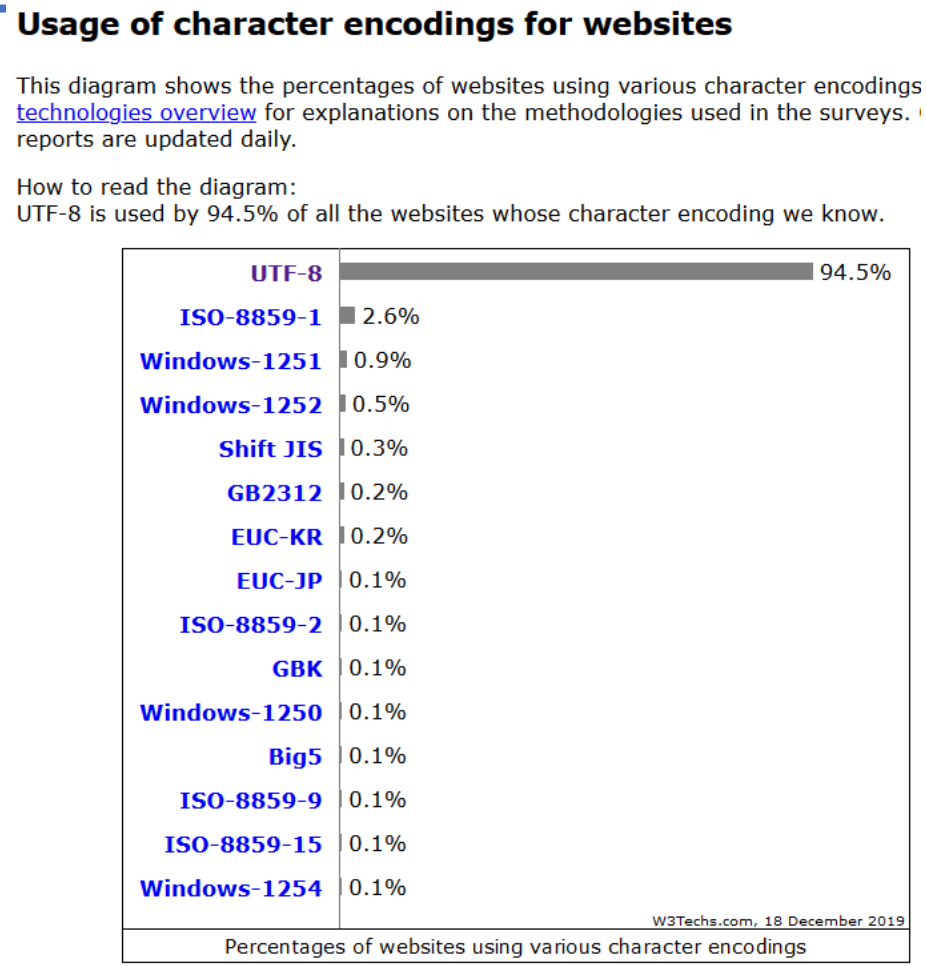
- **Each character** is assigned **a unique integer code**, called "code points", usually in hexadecimal base
  - Code point is in the form of U+<hex-code>, from U+0000 to U+10FFFF.
  - Characters in English, Chinese, or other languages
  - Currency symbols, Mathematical symbols
  - Emojis e.g., 🐶 U+1F436

Display with different encodings



Unicode Transformation Format (UTF)





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
0xxxxxxx				7	007F hex (127)
110xxxxx	10xxxxxx			(5+6)=11	07FF hex (2047)
1110xxxx	10xxxxxx	10xxxxxx		(4+6+6)=16	FFFF hex (65535)
11110xxx	10xxxxxx	10xxxxxx	10xxxxxx	(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).