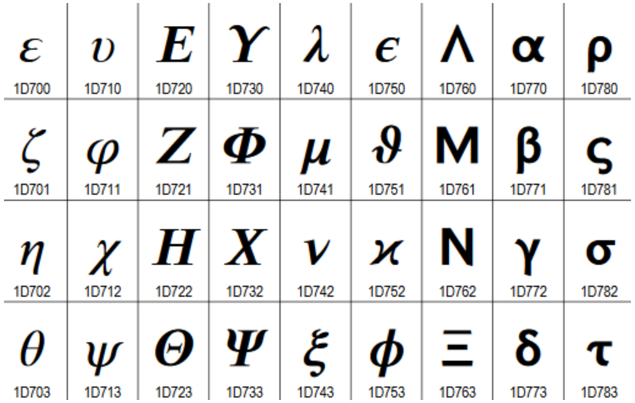
CE4045 CZ4045 SC4002
Natural Language Processing

Introduction to UTF-8

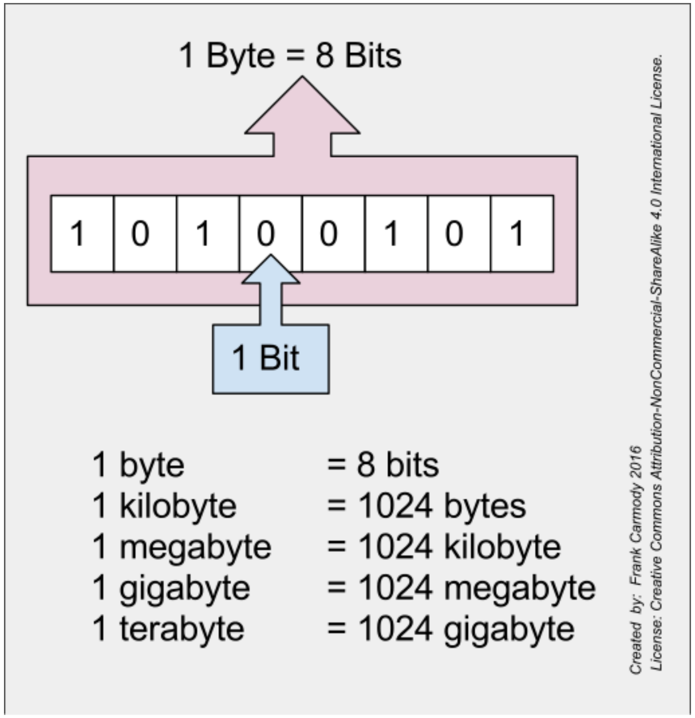
Dr. Sun Aixin

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



1 byte	= 8 bits
1 kilobyte	= 1024 bytes
1 megabyte	= 1024 kilobyte
1 gigabyte	= 1024 megabyte
1 terabyte	= 1024 gigabyte

Created by: Frank Carmody 2016
License: Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

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

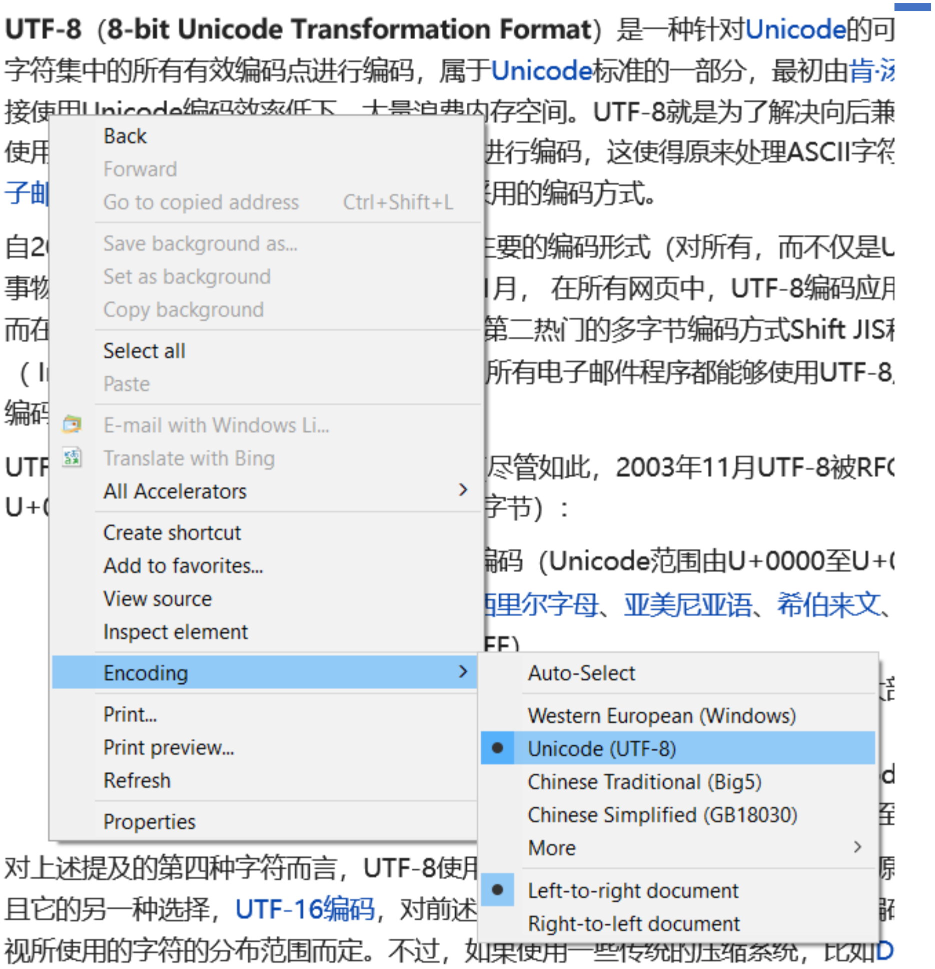
Unicode

For your information only

- 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



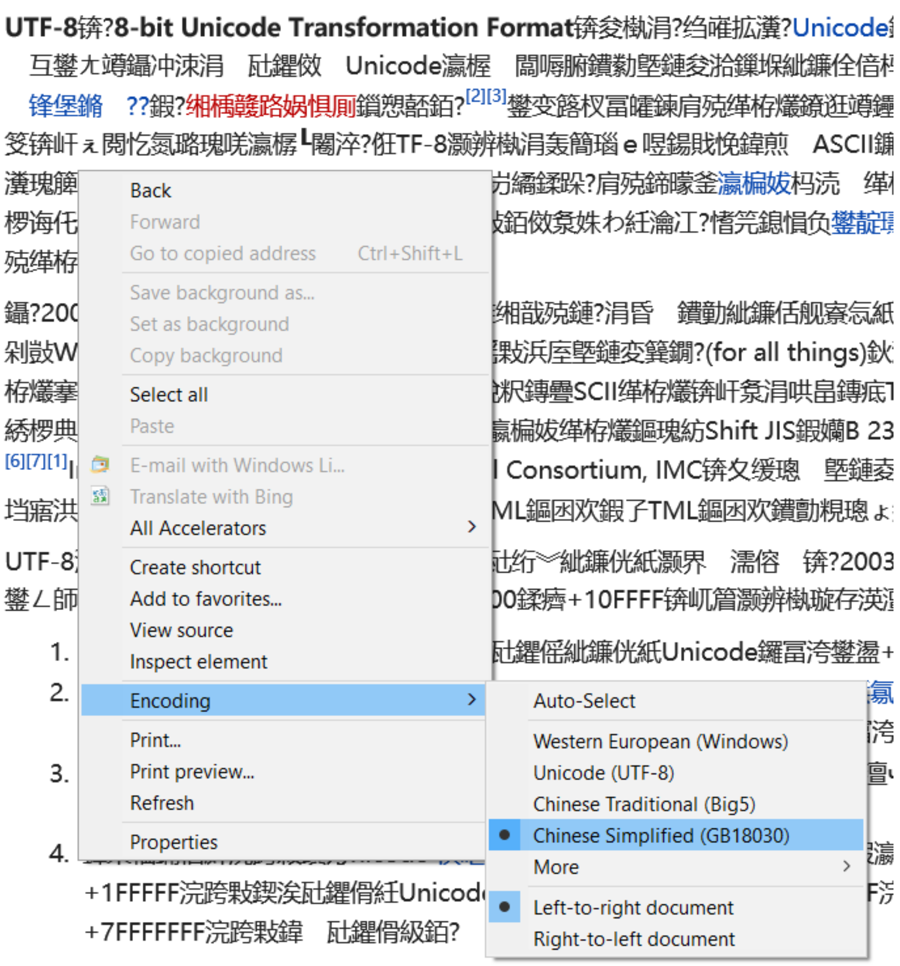
UTF-8（8-bit Unicode Transformation Format）是一种针对Unicode的可
字符集中的所有有效编码点进行编码，属于Unicode标准的一部分，最初由肯·汸
接使用Unicode编码效率低下，大量浪费内存空间。UTF-8就是为了解决向后兼
使用 进行编码，这使得原来处理ASCII字符
子邮 用的编码方式。

自2 主要的编码形式（对所有，而不仅是U
事物 月， 在所有网页中，UTF-8编码应用
而在 第二热门的多字节编码方式Shift JIS和
（I 所有电子邮件程序都能够使用UTF-8,
编码

UTF 尽管如此，2003年11月UTF-8被RFC
U+(字节）：

编码（Unicode范围由U+0000至U+(
西里尔字母、亚美尼亚语、希伯来文、
FF)

对上述提及的第四种字符而言，UTF-8使用 原
且它的另一种选择，UTF-16编码，对前述 弱
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D

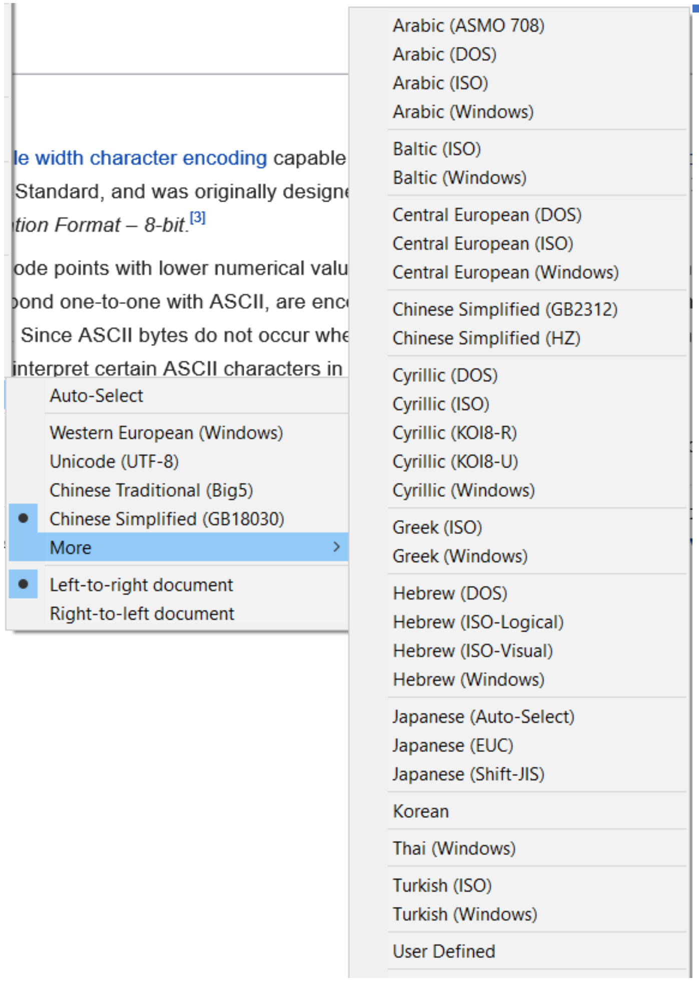


UTF-8锛?8-bit Unicode Transformation Format锛夋槸涓?绉嶉拡瀵?Unicode
互鐢ㄤ竴鑷冲洓涓 瓧鑺傚 Unicode瀛楃 闆嗕腑鐨勬墍鏈夋湁鏁堢紪鐮佺偣杩
锋堡鏅 ??鍜?缃楁媺路娲惧厠鍜岃晶鑲?[2][3]鐢变簬杈冩棭鍚屾椂瀛樺湪鐨勭紪
笅锛屽ぇ閲忔氮璐瑰唴瀛樼┖闂淬?傜TF-8灏辨槸涓轰簡瑙ｅ喅鍚戝悗鍏煎 ASCII鐮
灏瑰簳 岃繖浣跨敤鍘熸潵澶勭悊鐮侀噰鐢ㄧ殑缂栫爜鏂瑰紡 绯
椤诲伅 戠紪鐮佸舰寮忥紙瀵规墍鏈夛紝鑰屼笉浠呮槸
殑缂栫

鑷?200 细哉殑鏈?涓昏 鐨勭紪鐮佸舰寮忥紙瀵
剁敳W 墍鏈夛紝鑰屼笉浠呮槸?(for all things)鈥
杩欑 鐢ㄥ埌SCII缂栫爜锛屽叾涓哄畠鐨勬櫘
绣椤典 瀛楄妭缂栫爜鏂瑰紡Shift JIS鍜孏B 23
[6][7][1] I Consortium, IMC锛夋帹鑽 墍鏈
挡寤洪 ML鏂囦欢鍜孒TML鏂囦欢鐨勬牸鑰ょ

UTF-8 杩樻湁鍏朵粬缂栫爜 濡備 锛?2003
鐢ㄥ師 00鍒癠+10FFFF锛屽叡璁＄紪鐮佹暟瀛涗
1. 鐢ㄥ埌鍏朵粬Unicode鑼冨洿鍐呯殑+
2.
3.
4.
+1FFFFF浣跨敤鍥涘瓧鑺傜紪鐮乁nicod
+7FFFFFFF浣跨敤鍏 瀛楄妭缂栫爜?

Unicode Transformation Format (UTF)



le width character encoding capable

Standard, and was originally designe

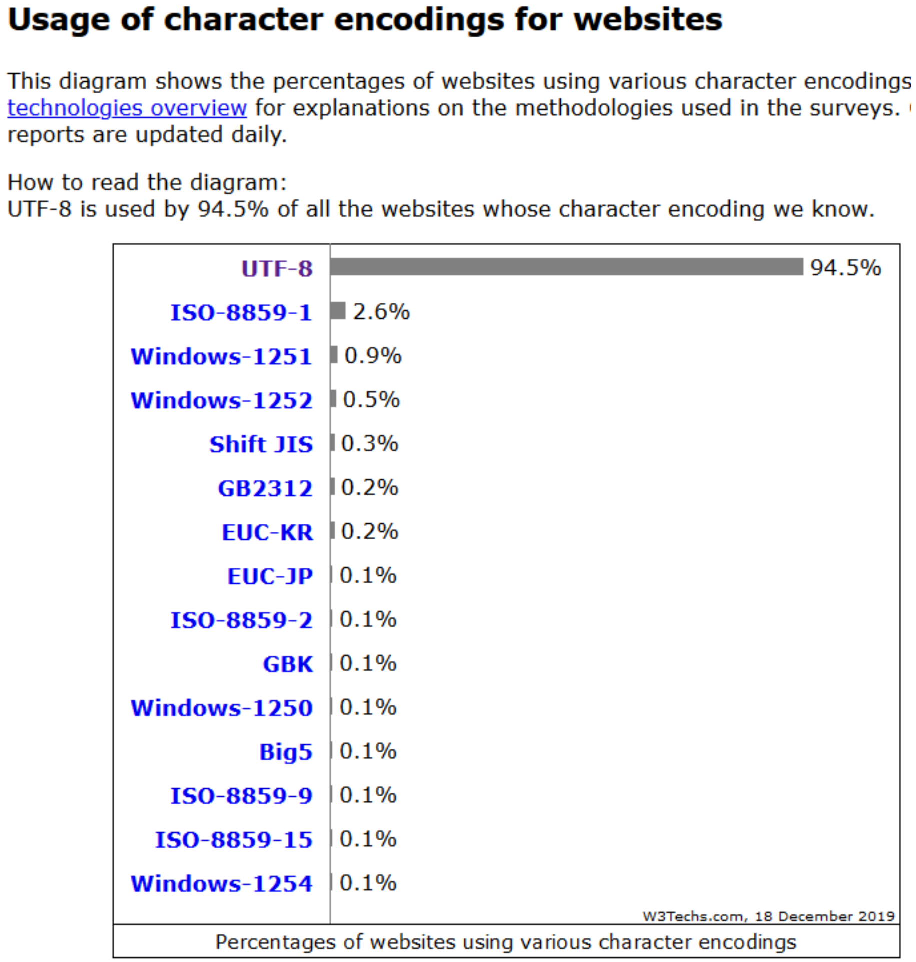
tion Format – 8-bit.[3]

ode points with lower numerical valu

pond one-to-one with ASCII, are enc

Since ASCII bytes do not occur whe

interpret certain ASCII characters in



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.

Percentages of websites using various character encodings

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

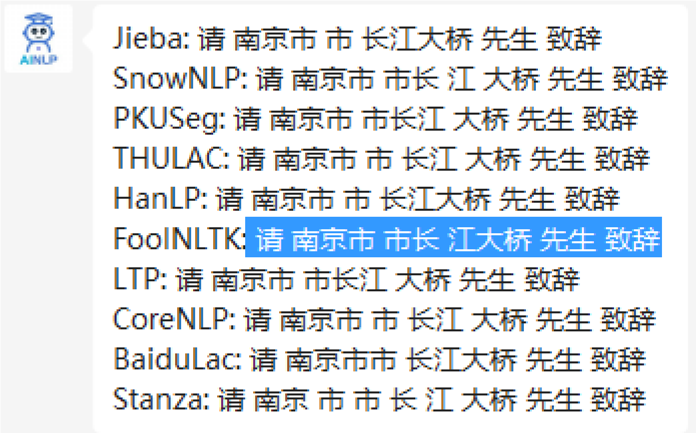
```
01001000 01100101 01101100 01101100 01101111 00100000
  01010111 01101111 01110010 01101100 01100100
```

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).