File Formats <

- Character Encoding
- JSON
- XML

Code Space   0~10FFFF for Unicode, 0~7F for ASCII

Code Point   a value of a character in code space

e.g. 'X' = U+0058 (hex)

Code Unit   8-bit for UTF-8, 16bit for UTF-16.

* UTF: Unicode Transformation Format

Code Space > Code Point > Code Unit

In UTF8, 1~4个 code units = 1 code point 即 1×8bit ~ 4×8 bit

In UTF16, 1~2个 code units = 1 code point 即 1×16bit ~ 2×16bit

PPT 15 可知 1 code unit = 1 byte.

- U+0000 到 U+007F 只需 1 byte 即 1 code unit

  Format: 0xxx xxxx   7个有效 bit

- U+0080 到 U+07FF 需 2byte 即 2 code unit   +5 bit

  Format: 110x xxxx | 10xx xxxx   11个有效 bit

- U+0800 到 U+FFFF 需 3 byte 即 3 code unit   +5bit

  Format: 1110 xxxx | 10xx xxxx | 10xx xxxx   16个有效 bit

- U+10000 到 U+10FFFF 需 4 byte 即 4code unit   +5 bit

  Format: 1111 0xxx | 10xx xxxx | 10xx xxxx | 10xx xxxx

  21个有效 bit

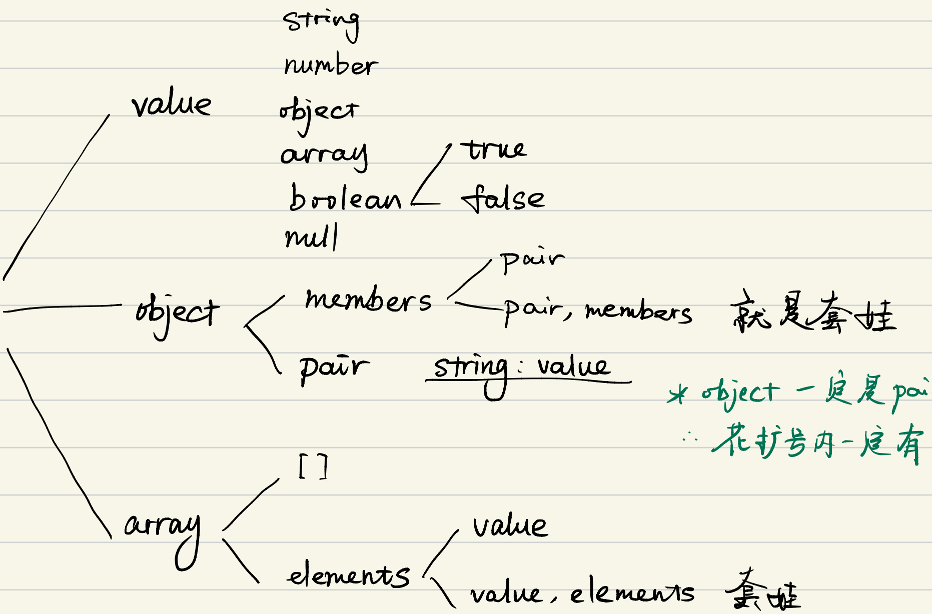
考题: '∈'   U+2208

2208 hex = 0010 0010 0000 1000   It's 16bit.

encode into UTF-8: 1110 0010 1000 1000 1000 1000 (binary)

= e2 88 88 (hex)

Syntax of JSON



* object 一定是pair 构成的
∴ 花扩号内一定有 key : value.

Python → JSON

list → JSON

[1, 2]	[1, 2]
[3, 'abc', True, None]	[3, "abc", true, null]

tuple → JSON

(1, 'abc')	[1, "abc"]

dict → JSON

{'name': 'john', 25: 'age'}	["name": "john", "25": "age"]

object → JSON

['foo', {'bar': ('baz', None, 1.0, 2)}]	
	["foo", {"bar": ["baz", null, 1.0, 2]}]
{(1, 2): 5}	Error
{(2) : 5}	["2": 5]