Unicode 编码 Emoji CJK 中文 汉字 过滤正则

玉兔是我啊 关注
2021.05.20 22:01:04 字数 1,635 阅读 343

Unicode

- 平面0 (0000–FFFF): 基本多文种平面（Basic Multilingual Plane, BMP）
- 平面1 (10000–1FFFF): 多文种补充平面（Supplementary Multilingual Plane, SMP）
- 平面2 (20000–2FFFF): 表意文字补充平面（Supplementary Ideographic Plane, SIP）
- 平面3 (30000–3FFFF): 表意文字第三平面（Tertiary Ideographic Plane, TIP）
- 平面4 to 13 (40000–DFFFF)尚未使用
- 平面14 (E0000–EFFFF): 特别用途补充平面（Supplementary Special-purpose Plane, SSP）
- 平面15 (F0000–FFFFF)保留作为私人使用区（Private Use Area, PUA）
- 平面16 (100000–10FFFF)，保留作为私人使用区（Private Use Area, PUA）

Unicode 13.0 Character Code Charts

Unicode区段

CJK

一、全部范围

范围	说明
2E80-A4CF	CJK部首补充、康熙部首、表意文字描述符、CJK符号和标点、日文平假名、日文片假名、注音字母、谚文兼容字母、象形字注释标志、注音字母扩展、CJK笔画、日文片假名语音扩展、带圈CJK字母和月份、CJK兼容、CJK统一表意文字扩展A、易经六十四卦符号、CJK统一表意文字、彝文音节、彝文字根
F900-FAFF	CJK兼容表意文字
FE10-FE1F	竖排符号
FE30-FE4F	CJK兼容符号（竖排符号）

范围	说明
FF00-FFEF	全角ASCII、全角中英文标点、半宽片假名、半宽平假名、半宽韩文字母
20000-2A6DF	CJK统一表意文字扩展B
2A700-2EBE0	CJK统一表意文字扩展C-F
2F800-2FA1F	CJK兼容表意文字扩展
300000~31344A	CJK统一表意文字扩展G

二、标点符号

字符集	定义范围	说明
CJK Symbols and Punctuation	3000-303F	CJK标点符号
Vertical Forms	FE10-FE1F	竖排符号
CJK Compatibility Forms	FE30-FE4F	CJK兼容符号（竖排符号）
Halfwidth and Fullwidth Forms	FF00~FFEF	全角ASCII、全角中英文标点、半宽片假名、半宽平假名、半宽韩文字母

三、汉字

范围	说明
2E80-2EFF	CJK部首补充、康熙部首、表意文字结构

范围	说明
3100-312F	注音字母
31A0-31EF	注音字母扩展、CJK笔画
3400-4DBF	CJK统一表意文字扩展A
4E00-9FFF	CJK统一表意文字
F900-FAFF	CJK兼容表意文字
20000-2A6DF	CJK统一表意文字扩展B
2A700-2EBE0	CJK统一表意文字扩展C-F
2F800-2FA1F	CJK兼容表意文字扩展
30000-3134A	CJK统一表意文字扩展G

字符集	定义范围	实际范围v13
基本汉字	4E00-9FFF	4E00-9FFC
扩展A	3400-4DBF	3400-4DBF
扩展B	20000-2A6DF	20000-2A6DD
扩展C	2A700-2B73F	2A700-2B734
扩展D	2B740-2B81F	2B740-2B81D
扩展E	2B820-2CEAF	2B820-2CEA1
扩展F	2CEB0-2EBE0	2CEB0-2EBE0
扩展G	30000-3134A	30000-3134A
兼容汉字	F900-FAFF	F900-FAD9
兼容扩展	2F800-2FA1F	2F800-2FA1D
部首扩展	2E80-2EFF	2E80-2EF3
康熙部首	2F00-2FDF	2F00-2FD5
汉字结构	2FF0-2FFF	2FF0-2FFB
汉语注音	3100-312F	3105-312F
注音扩展	31A0-31BF	31A0-31BA
汉字笔画	31C0-31EF	31C0-31E3
自定义扩展	E000-F8FF	E000-F8FF
〇	3007	3007

过滤正则

定义范围 是Unicode指定的字符区间， 实际范围 是当前版本真正使用的区间，没使用的区间在后续版本更新会被使用，所以过滤规则已定义范围为准。

〇 虽然在符号区但属于汉字。

易经六十四卦符号不属于汉字。

包含兼容和扩展字符

过滤内容	正则
CJK 汉字和符	[\u3400-\u4DBF\u4E00-\u9FFF\uF900-\uFAFF\uFF10-\uFF1F\uFF20-\uFF4F\uFF00-\uFFEF]

过滤内容	正则
CJK 标点符号	[\u3000-\u3006\u3008-\u303F\uFE10-\uFE1F\uFE30-\uFE4F\uFF00-\uFFEF]
中文汉字和符号	[\u2E80-\u2FFF\u3000-\u303F\u3100-\u312F\u31A0-\u31EF\u3400-\u4DBF\u4E00-\u9FFF\uF900-\uFAFF\uFE10-\uFE1F\uFE30-\uFE4F\uFF00-\uFFEF]
仅中文汉字	[\u3007\u2E80-\u2FFF\u3100-\u312F\u31A0-\u31EF\u3400-\u4DBF\u4E00-\u9FFF\uF900-\uFAFF]

常用其它过滤判断

```
CJK 常用汉字和符号(无全角内容)
[\u2E80-\uA4CF\uF900-\uFAFF\uFE10-\uFE1F\uFE30-\uFE4F]

CJK 汉字和符号(无竖排符号)
[\u2E80-\uA4CF\uF900-\uFAFF\uFF00-\uFFEF]

CJK 汉字和符号(无竖排符号和全角)
[\u2E80-\uA4CF\uF900-\uFAFF]

CJK 汉字(无符号和全角)
[\u3007\u2E80-\u2FFF\u3040-\uA4CF\uF900-\uFAFF]

中文汉字和符号(无全角内容)
[\u2E80-\u2FFF\u3000-\u303F\u3100-\u312F\u31A0-\u31EF\u3400-\u4DBF\u4E00-\u9FFF\uF900-\uFAFF\u
```

不含兼容和扩展字符

过滤内容	正则
CJK 标点符号	[\u3000-\u3006\u3008-\u303F\uFF00-\uFFEF]
中文汉字和符号	[\u3000-\u303F\u4E00-\u9FFF\uFF00-\uFFEF]
仅中文汉字	[\u3007\u4E00-\u9FFF]

大于4字不同语言符处理方式不同，可根据需要决定是否添加

```
#| 20000-2A6DF | CJK统一表意文字扩展B |
#| 2A700-2EBE0 | CJK统一表意文字扩展C-F |
#| 2F800-2FA1F | CJK兼容表意文字扩展 |
#| 30000~3134A | CJK统一表意文字扩展G |

#OC
[\U00020000-\U0002A6DF\U000A700-\U0002EBE0\U0002F800-\U0002FA1F\U00030000-\U0003134A]

#Java
[\x{20000}-\x{2A6DF}\x{2A700}-\x{2EBE0}\x{2F800}-\x{2FA1F}\x{30000}-\x{3134A}]

#JavaScript
[\u{20000}-\u{2A6DF}\u{2A700}-\u{2EBE0}\u{2F800}-\u{2FA1F}\u{30000}-\u{3134A}]
```

emoji

参考emoji-regex的正则分为3种标准 RGI标准 、旧标准 、旧标准+文字类型 。
但是这里 `文字类型(无彩色Icon)` 的emoji 把 `#*0-9` 也算在内并不正确。
修改后最终的规则可以参考这里emoji_regex.dart。

Full Emoji List
emoji history index
emoji-test.txt

编码	说明
\u00A0	不间断空格NDSP
\u0020	半角空格SP
\u3000	全角空格IDSP
\u200F	右至左符号
\uFE0E	文本变体选择器
\uFE0F	emoji变体选择器

上標和下標数字
上標和下標字母

参考

中文字符集Unicode 编码范围 - 千千秀字
中文在unicode中的编码范围
Unicode 编码范围和中文编码范围
Regular Expressions Unicode