



Тим Чаптыков
vk.com/tc
[@chaptykov](https://twitter.com/chaptykov)



постправда

токсичность

СЛОВО ГОДА

по версии Oxford Dictionaries

постправда



токсичность

2016

2017

2018



絵文字

Тим Чаптыков
vk.com/tc
@chapykov

Рефакторинг эмоджи в VK

- Поддержка всех эмоджи
- исправление старых ошибок
- Замена эмоджи в тексте на изображения
- Сборка, спрайты, оптимизация, нейминг

Что нужно знать об эмоджи?



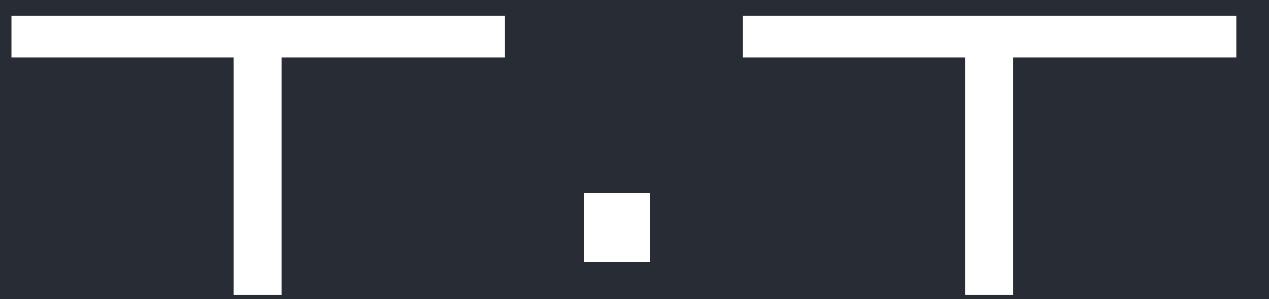
История



Запад



Япония



Корея



Китай

(ω))

Каомоджи

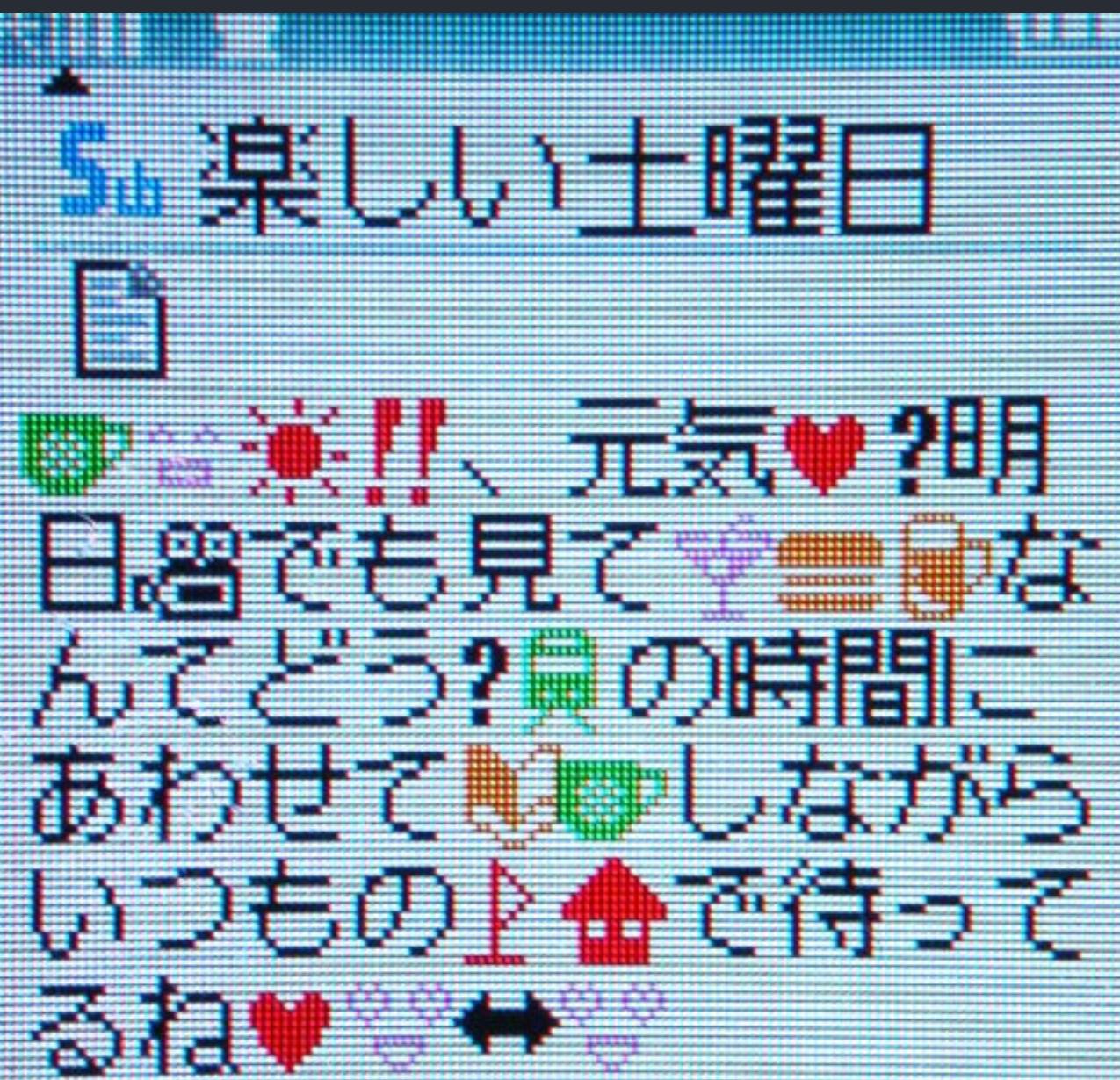
<http://каомоји.ру>

A portrait of Shigetaka Kurita, a Japanese man with dark hair and a warm smile. He is wearing a dark blue ribbed sweater over a blue collared shirt. He is holding a white spiral-bound notepad in his hands, which features a simple line drawing of a house with a chimney and a window.

Шигетака Курита
Дизайнер интерфейсов



Пиктограммы с использованием
эмоджи в i-mode



Текст с использованием
эмоджи в i-mode

Набор эмоджи на телефоне DoCoMo
2008 год





The Original Emoji

Музей современного искусства, Манхэттен



Николя Лауфани
Глава Smiley

	>:-o	Angry
	=D>	Applause
	8-)	Cool
	:'(Cry
	:-[Embarrassed
	>:-}	Evil Grin
	<:-	Foolish
	:-!	Foot in Mouth
	O:-)	Halo
	:-)	Happy
	>:D<	Hug
	:-*	Kiss
	:-D	Laugh
	:-X	Lips Sealed
	:-\$	Money Mouth
	@>---	Rose
	:-("	Sad
	X;{	Sick
	:-P	Stick Out Tongue
	=-O	Surprise
	:-\	Undecided
	:-	Whatever
	;:-)	Wink
	:-O	Yawn

2010

Apple и Google вносят предложение
о внесении эмоджи в Unicode



Japanese
post office

U+1F3E3



European
post office

U+1F3E4



Grinning face

U+1F600



Grinning face
with squinting eyes

U+1F604

Emoji Usage over Time



ЭМОДЖИ – ЭТО ТЕКСТ

Количество символов

- ` .length // 2



Кодировки

1 A

2 B

3 C

4 D

5 E

65 A

66 B

67 C

68 D

69 E

65 :

64 @

65 A

66 B

67 C

68 D

69 E

01111111 05 :
1000000 64 @
1000001 65 A
1000010 66 B
1000011 67 C
1000100 68 D
1000101 69 E

01000000 64 @

01000001 65 A

01000010 66 B

01000011 67 C

01000100 68 D

01000101 69 E

ASCII

128 символов

01001000 0x48 H

01101001 0x69 i

00100001 0x21 !

KOI8

256 символов

КОИ-8 код обмена информацией, 8 бит

Восьмибитовая кодировка для кириллических алфавитов, совместимая с ASCII.

Использовалась, как основная русская кодировка в Unix-совместимых ОС.

КОІ8-Р

Русский

КОІ8-У

Украинский

КОІ8-Т

Таджикский

11110101 0xF5 y

11010010 0xD2 p

11000001 0xC1 a

	0	1	2	3	4	5	6	7	8	9	А	В	С	Д	Е	Ф
8	-		Г	Л	Ц	Т	+	+	+	■	■	■	■	■	■	■
9	☒	☒	☒	☒	☒	•	✓	≈	≤	≥	J	°	2	·	÷	
A	=		F	ё	Г	҃	҂	҄	҅	҆	҇	҈	҉	Ҋ	ҋ	Ҍ
B			‡	Ё		҃	҂	҄	҅	҆	҇	҈	҉	Ҋ	Ҍ	©
C	ю	а	б	ц	д	е	ф	г	х	и	й	к	л	м	н	о
D	п	я	р	с	т	у	ж	в	ъ	ы	з	ш	э	щ	ч	ъ
E	ю	А	Б	Ц	Д	Е	Ф	Г	Х	И	Й	К	Л	М	Н	О
F	П	Я	Р	С	Т	У	Ж	В	ъ	ы	з	ш	э	щ	ч	ъ

	0	1	2	3	4	5	6	7	8	9	А	В	С	Д	Е	Ф
8	-		Г	Л	Ц	Т	+	+	+	■	■	■	■	■	■	■
9	☒	☒	☒	☒	☒	•	✓	≈	≤	≥	J	°	2	·	÷	
A	=		F	ё	Г	҃	҆	҇	҈	҉	Ҋ	ҋ	Ҍ	ҍ	Ҏ	ҏ
B			‡	Ё		҃	҄	҅	҆	҇	҈	҉	Ҋ	ҋ	Ҍ	©
C	ю	а	б	ц	д	е	ф	г	х	и	й	к	л	м	н	о
D	п	я	р	с	т	у	ж	в	ъ	ы	з	ш	э	щ	ч	ъ
E	ю	А	Б	Ц	Д	Е	Ф	Г	Х	И	Й	К	Л	М	Н	О
F	П	Я	Р	С	Т	У	Ж	В	ъ	ы	з	ш	э	щ	ч	ъ

Обратная совместимость

11110000	F0
11010010	D2
11001001	C9
11010111	D7
11000101	C5
11010100	D4
00100000	20
11001110	CE
11000101	C5
11010010	D2
11000100	C4
11000001	C1
11001101	CD

Привет нердам

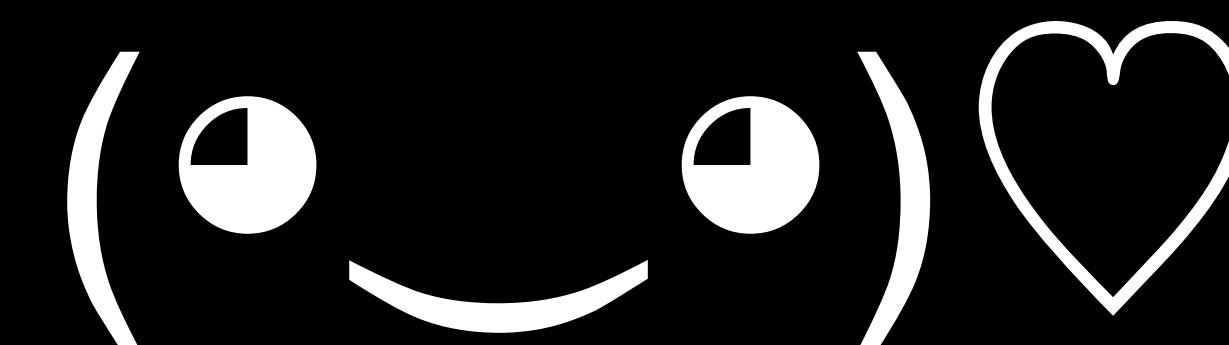
01110000	F0
01010010	D2
01001001	C9
01010111	D7
01000101	C5
01010100	D4
00100000	20
01001110	CE
01000101	C5
01010010	D2
01000100	C4
01000001	C1
01001101	CD

01110000	F0
01010010	D2
01001001	C9
01010111	D7
01000101	C5
01010100	D4
00100000	20
01001110	CE
01000101	C5
01010010	D2
01000100	C4
01000001	C1
01001101	CD

pRIWET NERDAM

01110000	F0
01010010	D2
01001001	C9
01010111	D7
01000101	C5
01010100	D4
00100000	20
01001110	CE
01000101	C5
01010010	D2
01000100	C4
01000001	C1
01001101	CD

pRIWET NERDAM



CP1251

256 символов

CP1251 Windows-1251

Восьмибитовая кодировка для кириллических алфавитов.

Использовалась, как основная русская кодировка в Windows.

0 1 2 3 4 5 6 7 8 9 А В С Д Е Ф

8 Ѓ Г , Ѓ „ … † ‡ € % Ђ < Њ Ђ Ѓ Џ

9 Ѣ ‘ ’ “ ” • – – ™ Ќ > Њ Ђ Ѣ Џ

А ў ў Ј Ѹ Г ; § Ё © € « „ ® Ї

В ° ± І і г ѹ ¶ · ё № € » ј \$ s ѕ

С А Б В Г Д Е Ж З И Й К Л М Н О П

Д Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Ъ Э Ю Я

Е а б в г д е ж з и й к л м н о п

Ф р с т у ф х ц ч ш щ ъ ѫ Ѣ Э Ю Я

0 1 2 3 4 5 6 7 8 9 А В С Д Е Ф

8 Ѓ Г , г „ … † ‡ € % Ђ < Њ Ђ Ѓ ѕ

9 Ѣ ‘ ’ “ ” • – – ™ Ќ > Њ Ђ Ѣ ѕ

А ў ў Ј ѣ Г ; § Ё © € « „ ® Ї

В ° ± І і г ѹ ¶ · ё № € » ј \$ s ѫ

С А Б В Г Д Е Ж З И Й К Л М Н О П

Д Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Ъ Э Ю Я

Е а б в г д е ж з и й к л м н о п

Ф р с т у ф х ц ч ш щ ъ ѫ Ѣ є ю я

0 1 2 3 4 5 6 7 8 9 А В С Д Е Ф

8 Ѓ Ѓ , Ѓ „ … † ‡ € % Ќ < Њ Ђ Ѓ Џ

9 Ѣ ‘ ’ “ ” • – ™ Ќ > Њ Ђ Ѣ Џ

А ў ў Ј ѕ Г ; § Ё © € « „ ® Ї

В ° ± І і г ѹ ¶ · ё № € » ј Ѫ Ѯ ѵ

С А Б В Г Д Е Ж З И Й К Л М Н О П

Д Р С Т У Ф Х Ц Ч Ш Щ Ъ Ы Ъ Э Ю Я

Е а б в г д е ж з и й к л м н о п

Ф р с т у ф х ц ч ш щ ъ ѫ Ѣ ѧ ю я

«Теперь мы можем верстать так, как завещал
А. Лебедев в § 62 своего „Ководства“», —
пронеслось в моей голове...

ВКонтакте и CP1251



ВКонтакте и CP1251

- Контент в БД
- Кодировка страниц
- Некоторые исходники

ВКонтакте и CP1251

- Контент в БД
- Кодировка страниц
- Некоторые исходники

(ノ°益°)ノ

256 символов – мало.

Unicode 1991

65 536 символов

$$2^{16} = 65\ 536$$

01111101 01110101 0x7d75 絵

01100101 10000111 0x6587 文

01011011 01010111 0x5b57 字

65 536 символов – тоже мало.

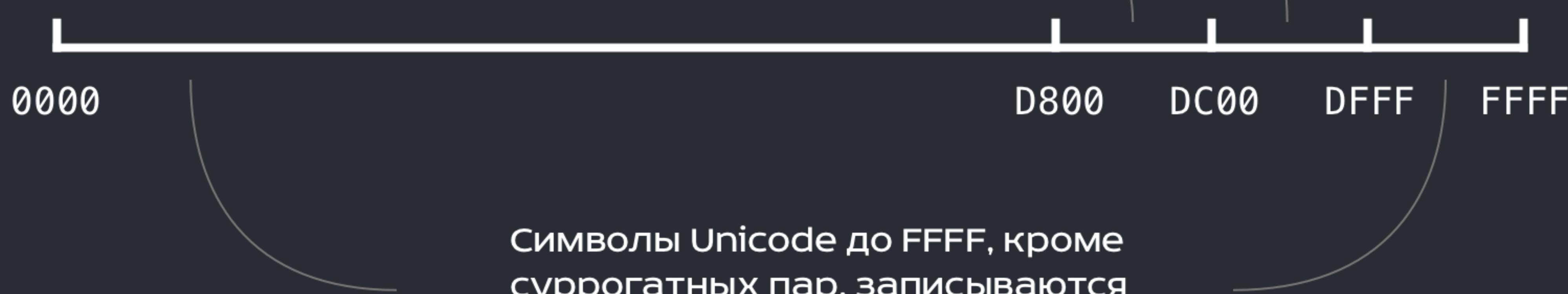
(‘ ~ ‘,)

UTF-16

1 112 064 символа

Нижняя часть суррогатных пар
Low Surrogates

Верхняя часть суррогатных пар
High Surrogates



Принцип кодирования

1F600₁₆

Кодируемый диапазон

10000₁₆...10FFFF₁₆

max-min: 11111111111111111111
20 разрядов

База верхней части

D800₁₆
1101100000000000
10 разрядов

База нижней части

DC00₁₆
1101110000000000
10 разрядов

-10000₁₆

F600₁₆

000011101100000000

$$\begin{array}{r} + \\ \hline D800_{16} \\ \hline D83D_{16} \end{array}$$

$$\begin{array}{r} + \\ \hline DC00_{16} \\ \hline DE00_{16} \end{array}$$

D83D

DE00

$$2^{20} + 2^{16} - 2048 = 1\ 112\ 064$$

Количество символов

```
'😊'.split(' ').map((s) => {  
    return s.charCodeAt(0).toString(16);  
});  
  
// ["d83d", "de00"]
```

UTF-16

`str.split()`

`str.charCodeAt()`

`String.fromCharCode()`

Unicode

`Array.from()`

`str.codePointAt()`

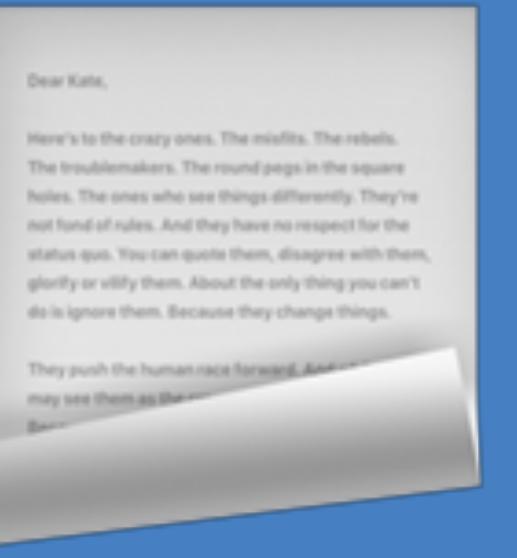
`String.fromCodePoint()`

Количество символов

```
Array.from('😊').length // 1
```

Ho...

Array.from('👩‍👧‍👦').length // 7



Стандарт

Лигатуры

f i

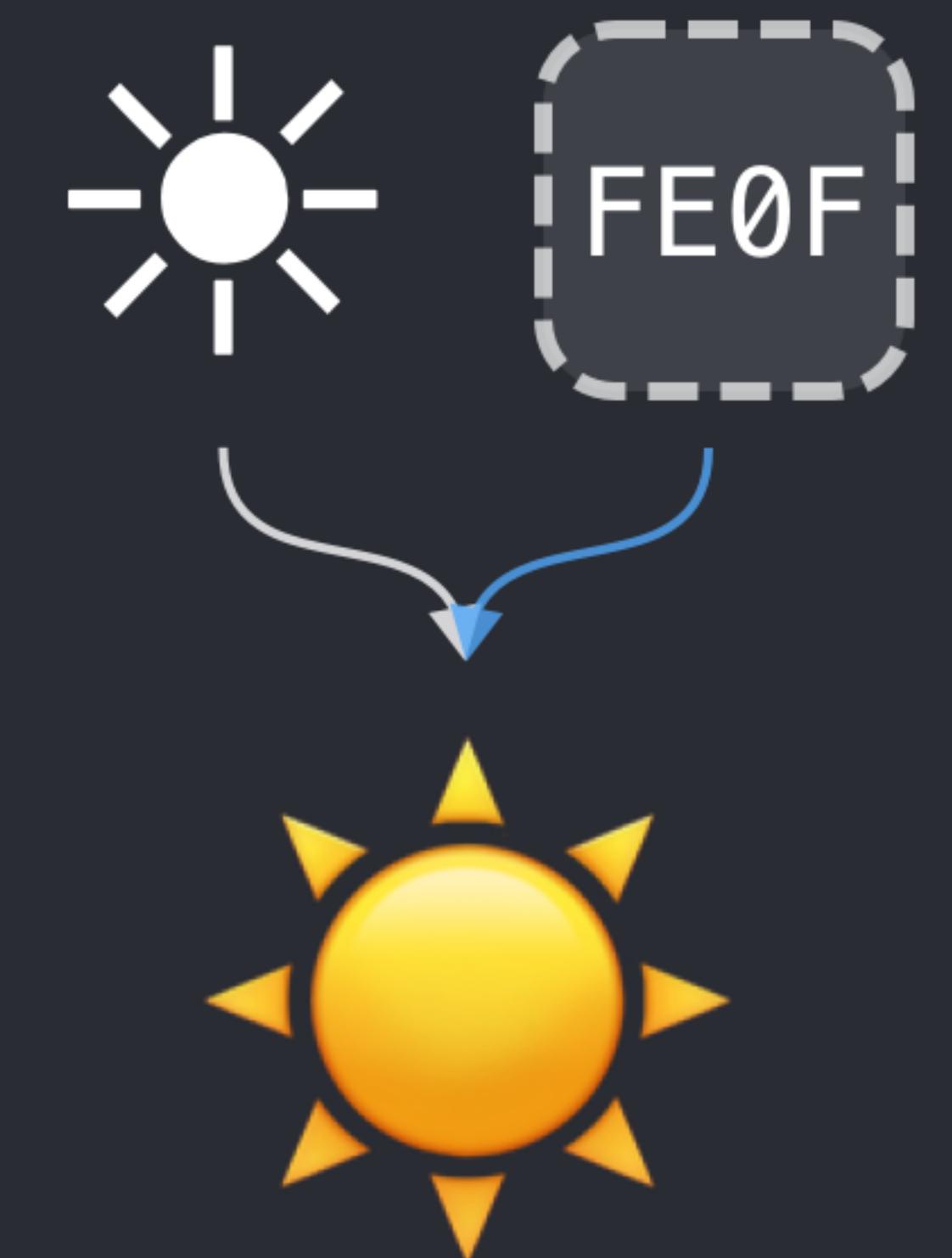


fi

fire

fire

Variations



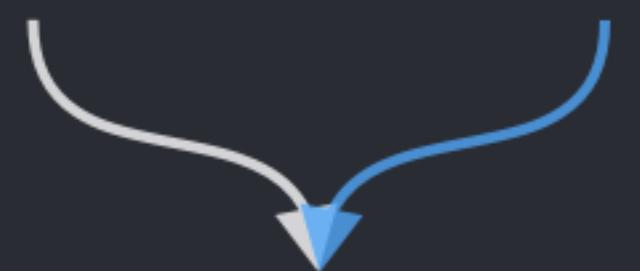
U+2600 U+FE0F

keycaps

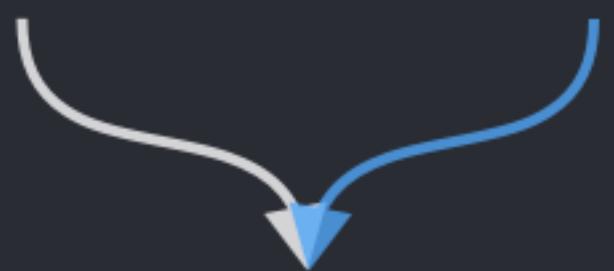


U+0031 U+20E3

1 FE0F



1 20E3



1

0

1

2

3

4

5

6

7

8

9

#

*

Разобрать на байткод

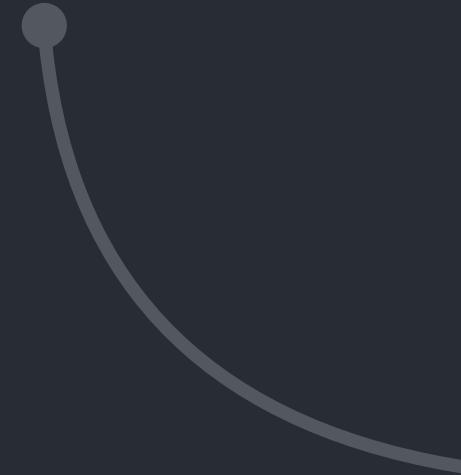
```
encodeURIComponent('😊').replace(/%/ig, '');
```

```
// "F09F988C"
```

Разобрать на байткод

```
encodeURIComponent('💋');  
// "%F0%9F%92%8B"
```

```
encodeURIComponent('✳️');  
// "*%EF%B8%8F%E2%83%A3"
```



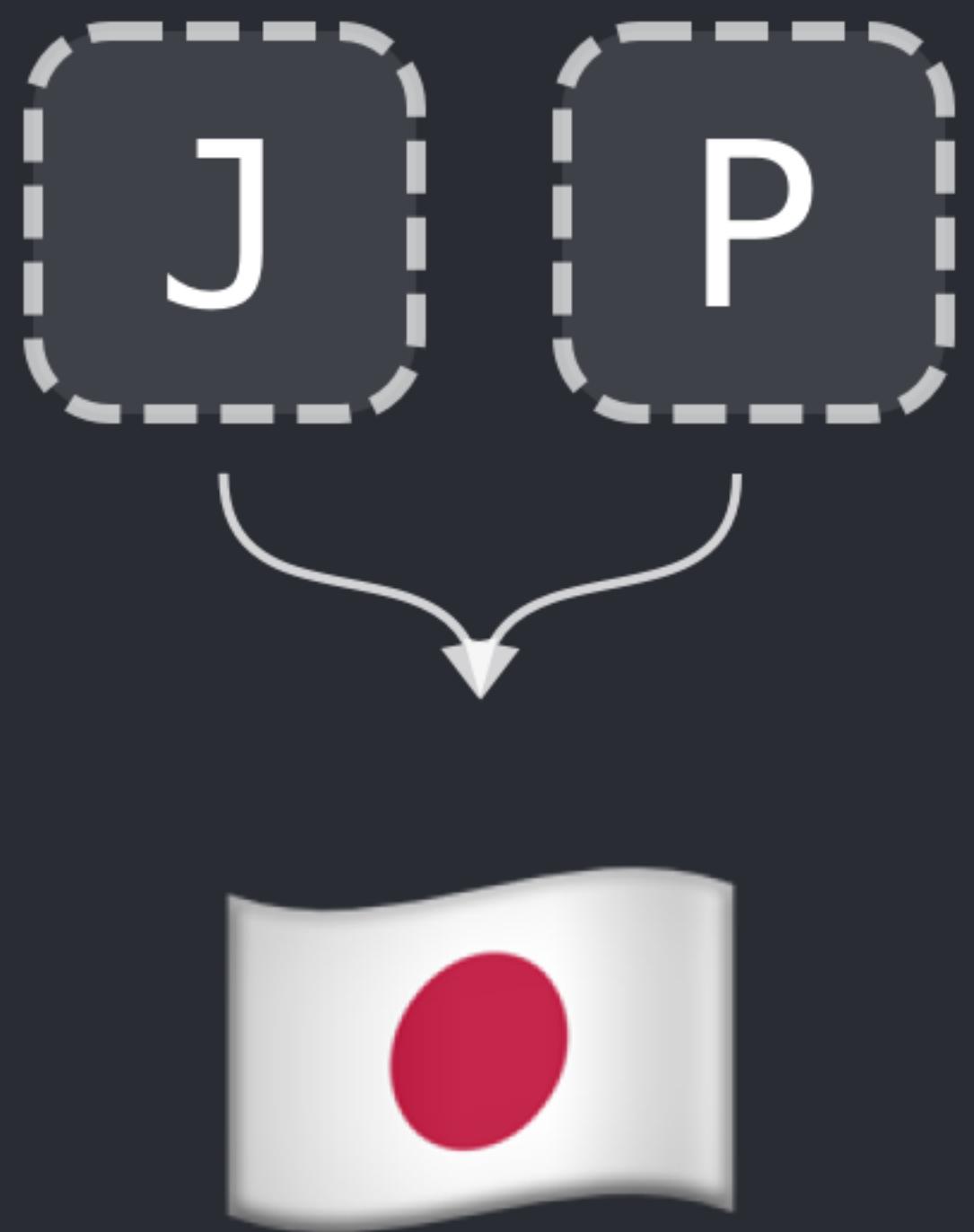
Цифры и * – безопасные символы
для URI и не заменяются

Flags

Enclosed Alphanumeric Supplement

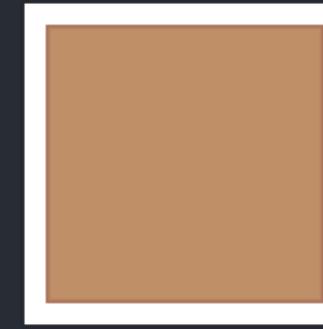
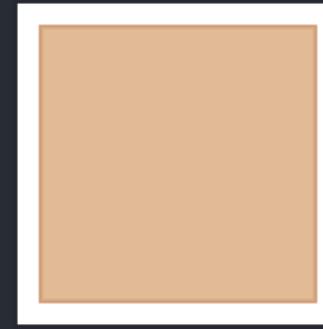
A – Z

U+1F1E6 – U+1F1FF

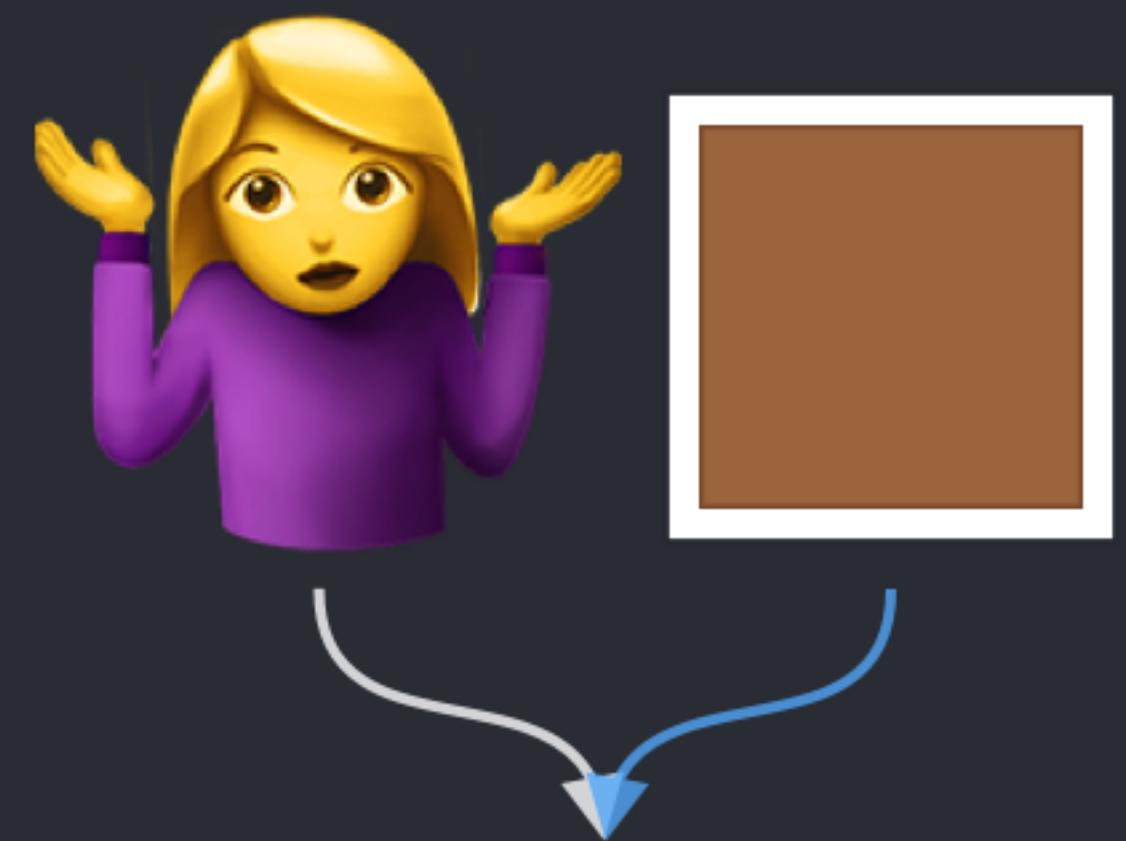


skin tones

Шкала Фитцпатрика



U+1F3FB – U+1F3FF

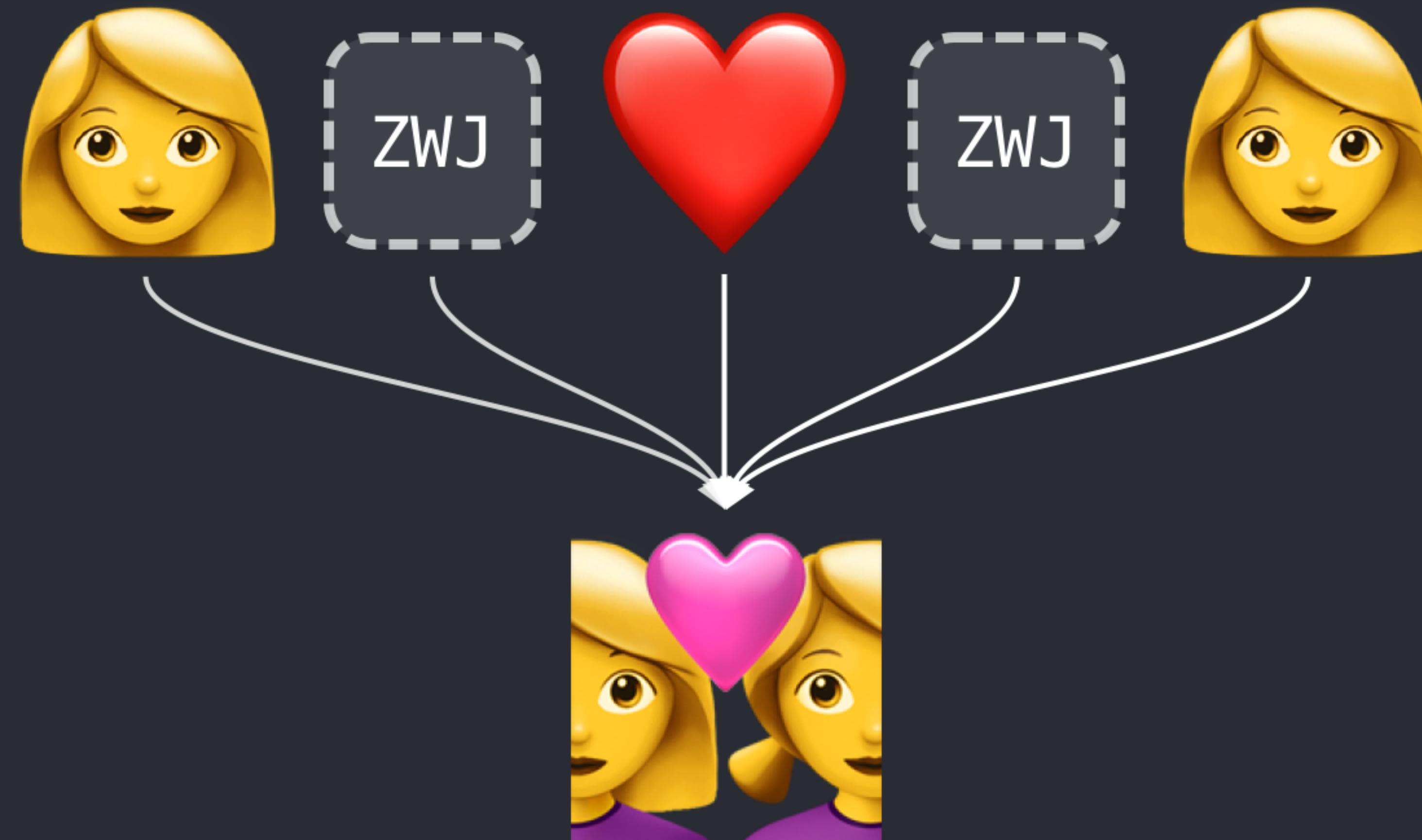


ZWJ Sequences

zero-width joiner



U+200D



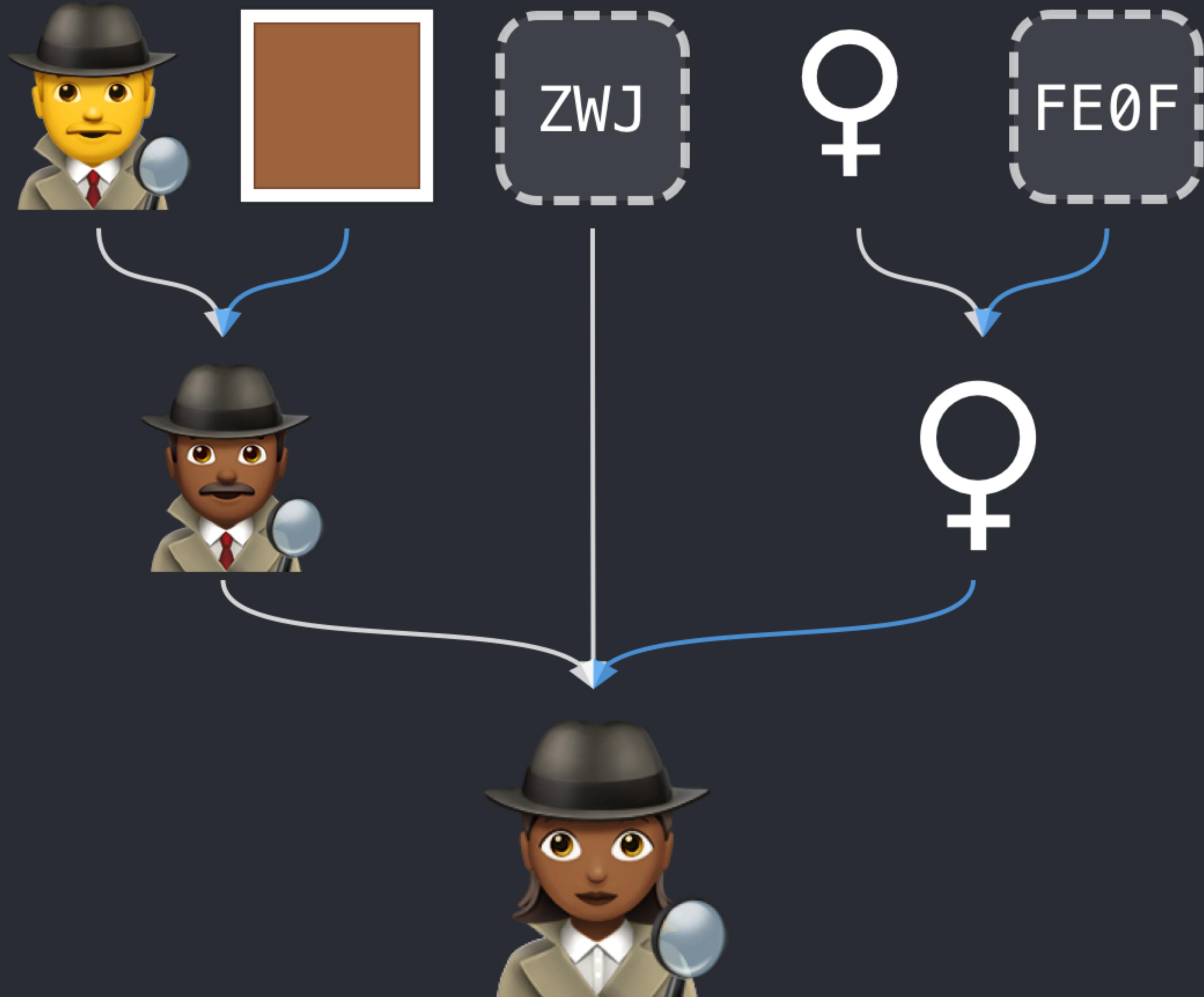


ZWJ

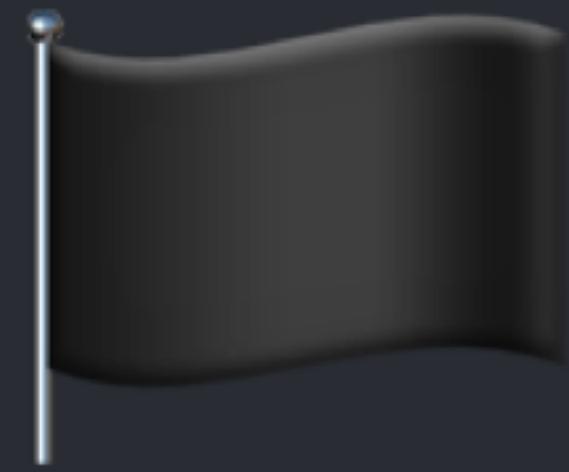
♀

FE0F





Tag sequences



G

B

E

N

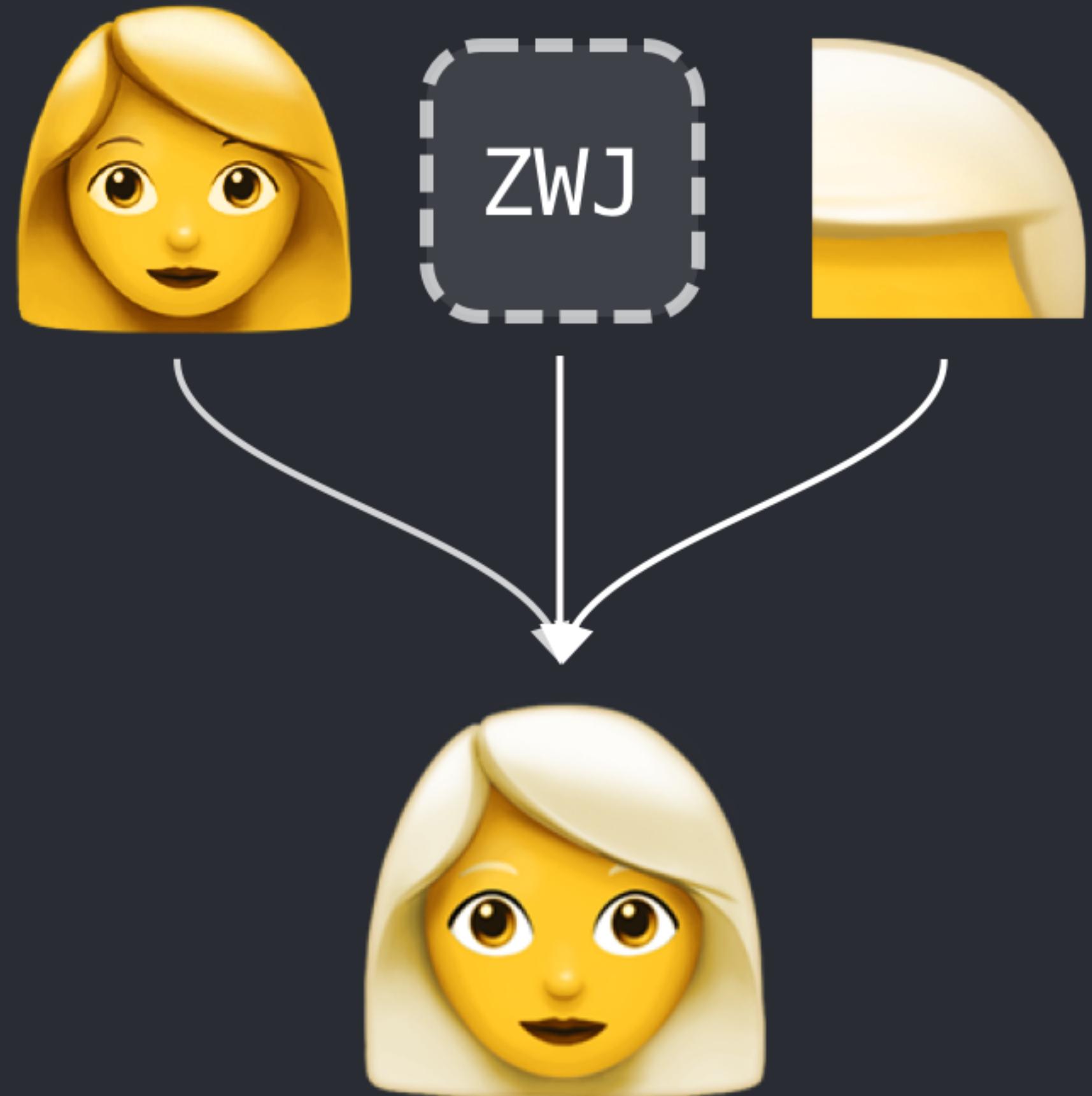
G

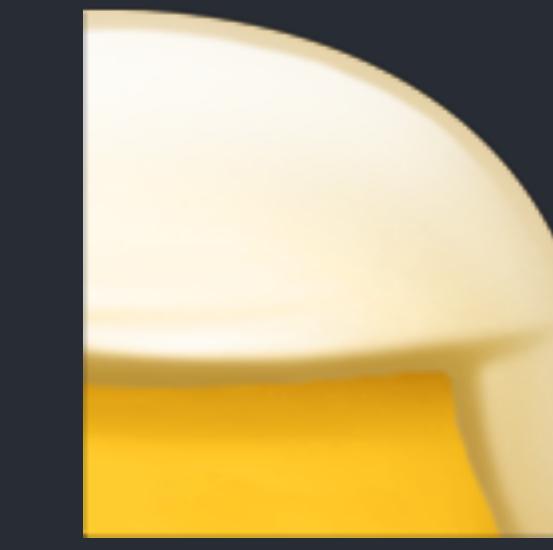
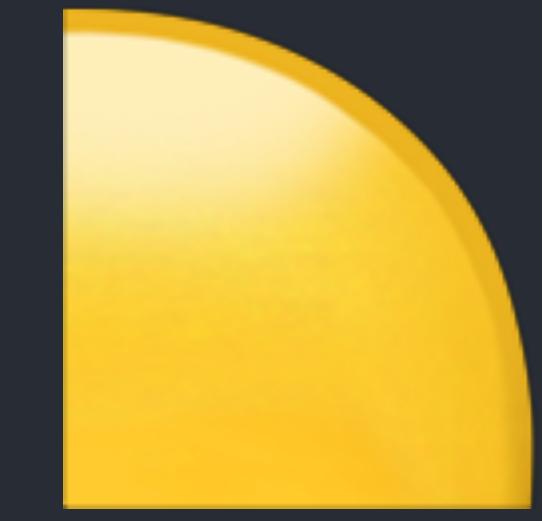
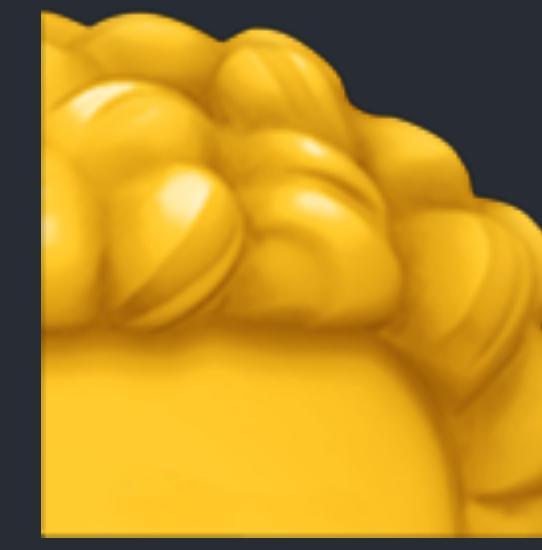
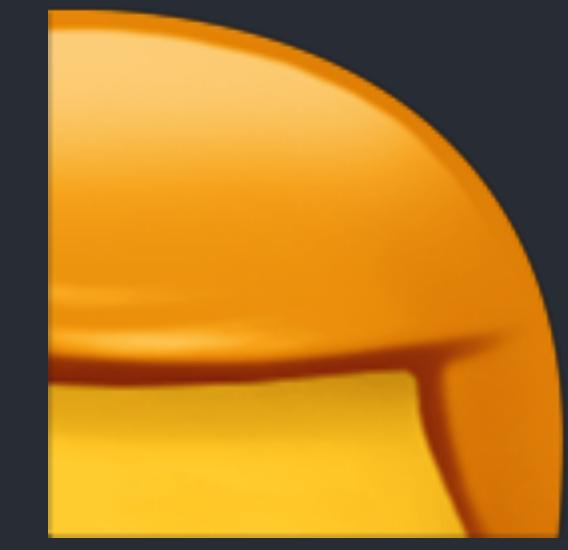
E007F



Hair

2018



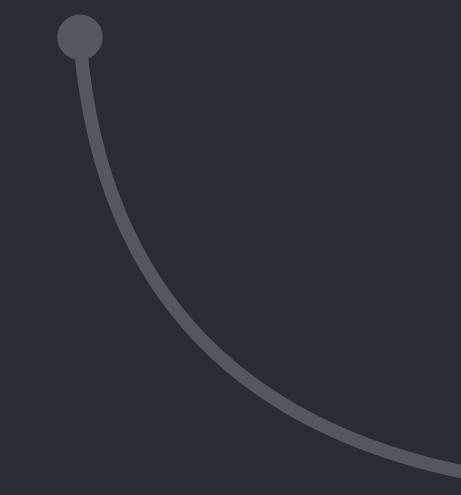


Backspace

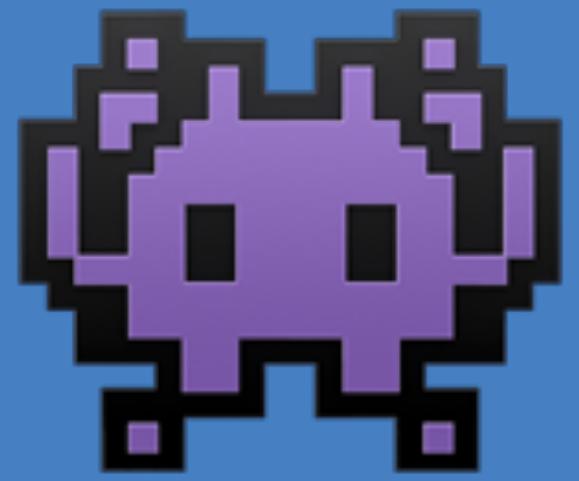
Количество символов

```
Array.from('👩‍👧‍👦')
```

```
// [👩, "", 👩, "", 👧, "", 👧, ""]
```



ZWJ – непечатный символ



Регулярные выражения

Зачем?

- Заменять эмоджи на изображения
- Выводить подсказки для использования стикеров

Специфика

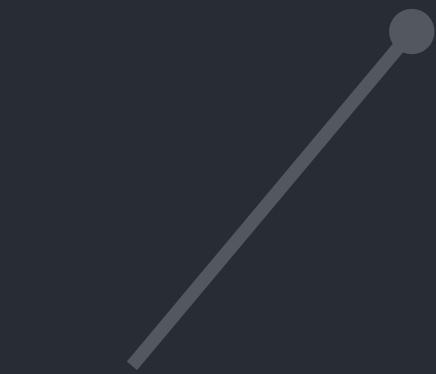
- PHP, JS
поддержка регулярных выражений на JS и PCRE
- CP-1251 в исходниках и данных на сервере
- Производительность
- Толерантность к отступлениям от стандарта
и отсутствие ложных срабатываний

lodash/lodash

./internal/unicodeWords.js

Вариант Lodash

```
const rsEmoji = (  
  `(?:${[rsDingbat, rsRegional, rsSurrogatePair].join(' | ')})${rsSeq}`  
);
```



[\\ud800-\\dbff] [\\dc00-\\uffff]

\u

JavaScript
charCode

\x

PCRE
codePoint

JavaScript RegExp & Unicode

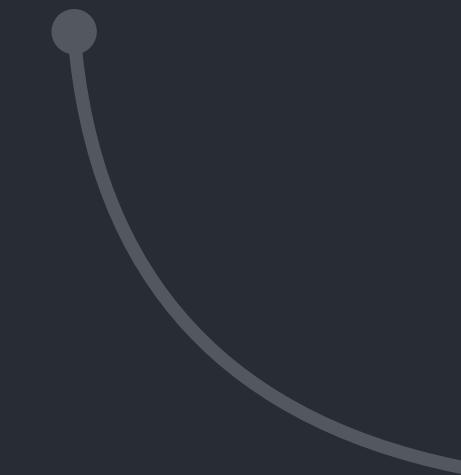
```
/\u{1F600}/u.test('😊'); // true
```

```
/uD83D\uDE00/u.test('😊'); // true
```

JavaScript RegExp & Unicode

```
/\u{1F600}/u.test('😊'); // true
```

```
/uD83D\uDE00/u.test('😊'); // true
```



Не забываем про флаг `/u`
для регулярок с Юникодом

700 мс
на сервере на 1 вызов

[\\ud800-\\udbff] [\\udc00-\\uffff]

JS

$1023 + 1023$

PHP

1023×1023

Выводы

- «Ну, не надо было конвертировать куда попало»
- «Регулярка не для этого была предназначена»

[mathiasbynens/emoji-regex](https://github.com/mathiasbynens/emoji-regex)

Emoji Properties

- Emoji
- Emoji_Presentation
- Emoji_Modifier
- Emoji_Modifier_Base
- Emoji_Component
- Extended_Pictographic

Unicode property escapes

```
const regexGreekSymbol = /\p{Script_Extensions=Greek}/u;  
regexGreekSymbol.test('π'); // true
```

Unicode property escapes in JavaScript regular expressions

mathiasbynens.be/notes/es-unicode-property-escapes

ES2018: Emoji regexp

```
const regex = /\p{Emoji_Modifier_Base}\p{Emoji_Modifier}?
|\p{Emoji_Presentation}|\p{Emoji}\uFE0F/gu;
```

ES2018: Emoji regexp

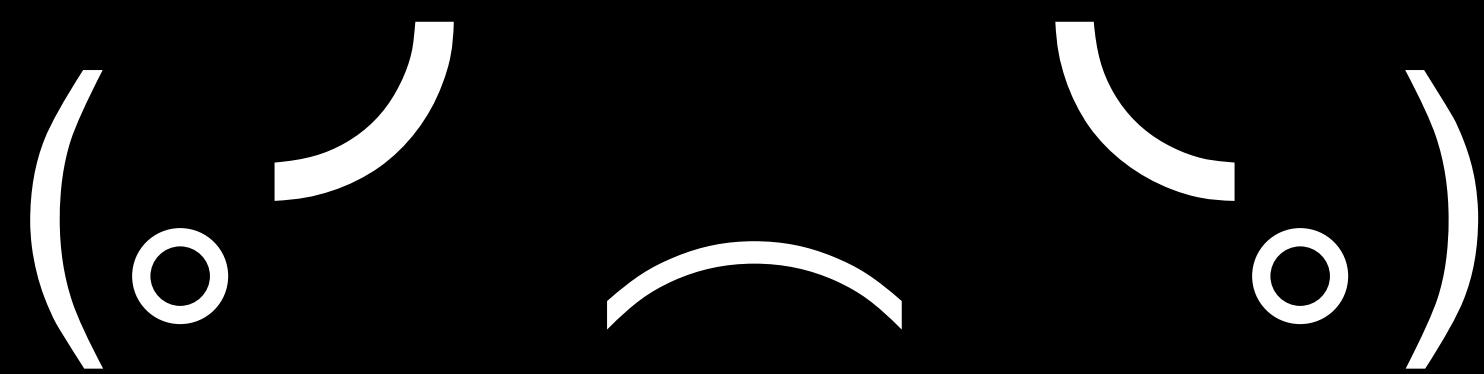
```
const regex = /\p{Emoji_Modifier_Base}\p{Emoji_Modifier}?
|\p{Emoji_Presentation}|\p{Emoji}\uFE0F/gu;
```

Unicode properties + JS RegExp



Поддержка в браузерах

- Chrome 64+
- Safari TP 42



Плагин для Babel

github.com/babel/babel/blob/master/packages/babel-plugin-proposal-unicode-property-regex/README.md

emoji-regex

github.com/mathiasbynens/emoji-regex

Толерантность и U+FE0F

Текстовая версия

github.com/mathiasbynens/emoji-regex/blob/master/text.js

- Некоторые Емоjī используются без U+FE0F и поддерживаются операционными системами
- Если мы игнорируем U+FE0F, то получим неприятные ложные срабатывания

Regex101

regex101.com

Unicode Emoji data files

unicode.org/Public/emoji/12.0

Full Emoji List 12.0

unicode.org/emoji/charts/full-emoji-list.html



Заключение

Ничего страшного



Алгоритм

- Разобрать на UTF-16 символы
- Найти описание UTF-16 символов
- Разобрать на Unicode-символы
- Найти описание Unicode-символов

Разобрать на UTF-16 слова

```
'🔥'.split(' ').map((s) => {  
    return s.charCodeAt(0).toString(16);  
});  
  
// ["d83d", "dd25"]
```



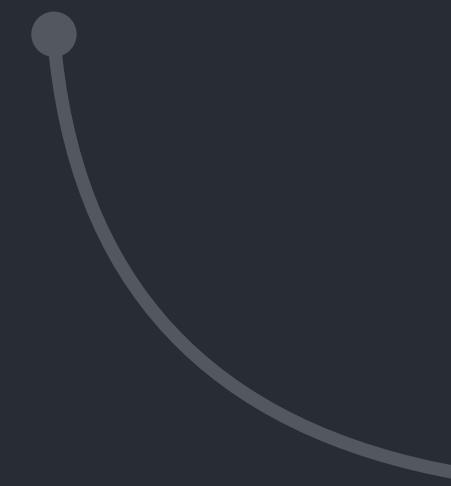
Суррогатная пара, указывающая
на символ 0x1f525

Codepoints

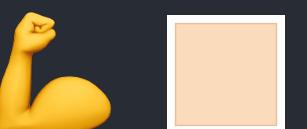
codepoints.net

Разобрать на Unicode-символы

```
Array.from('💪').map((s) => {  
  return s.codePointAt(0).toString(16);  
});  
  
// ["1f4aa", "1f3fb"]
```

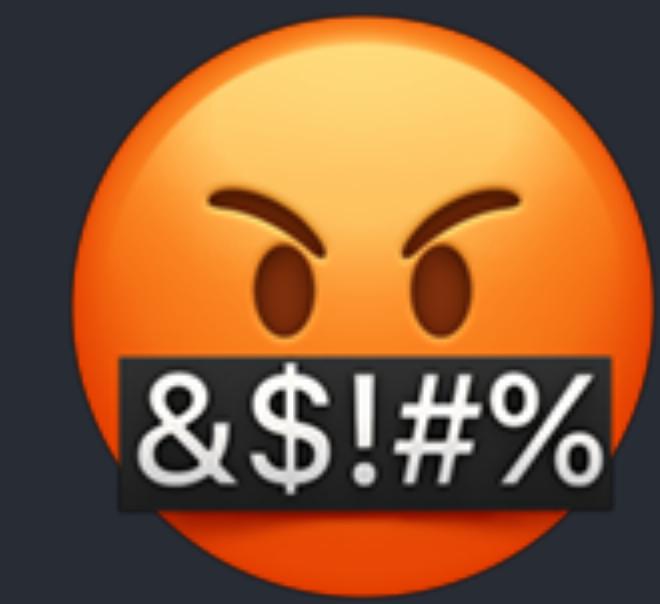


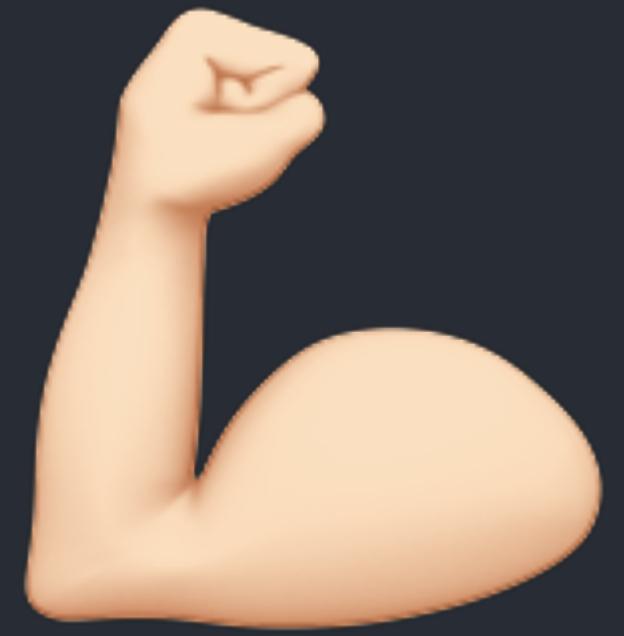
Эмоджи «flexed bicep»
со светлым цветом кожи



ЭМОДЖИПЕДИЯ

emojipedia.org





Unicode Technical Standard: Unicode Emoji

unicode.org/reports/tr51

Презентация об эмоджи

unicode.org/emoji/slides.html

История эмоджи

medium.com/@k4i/japanese-emoji-1-bf150c2825d1

VK Tech

vk.com/tech



Тим Чаптыков
vk.com/tc
[@chaptykov](https://twitter.com/chaptykov)