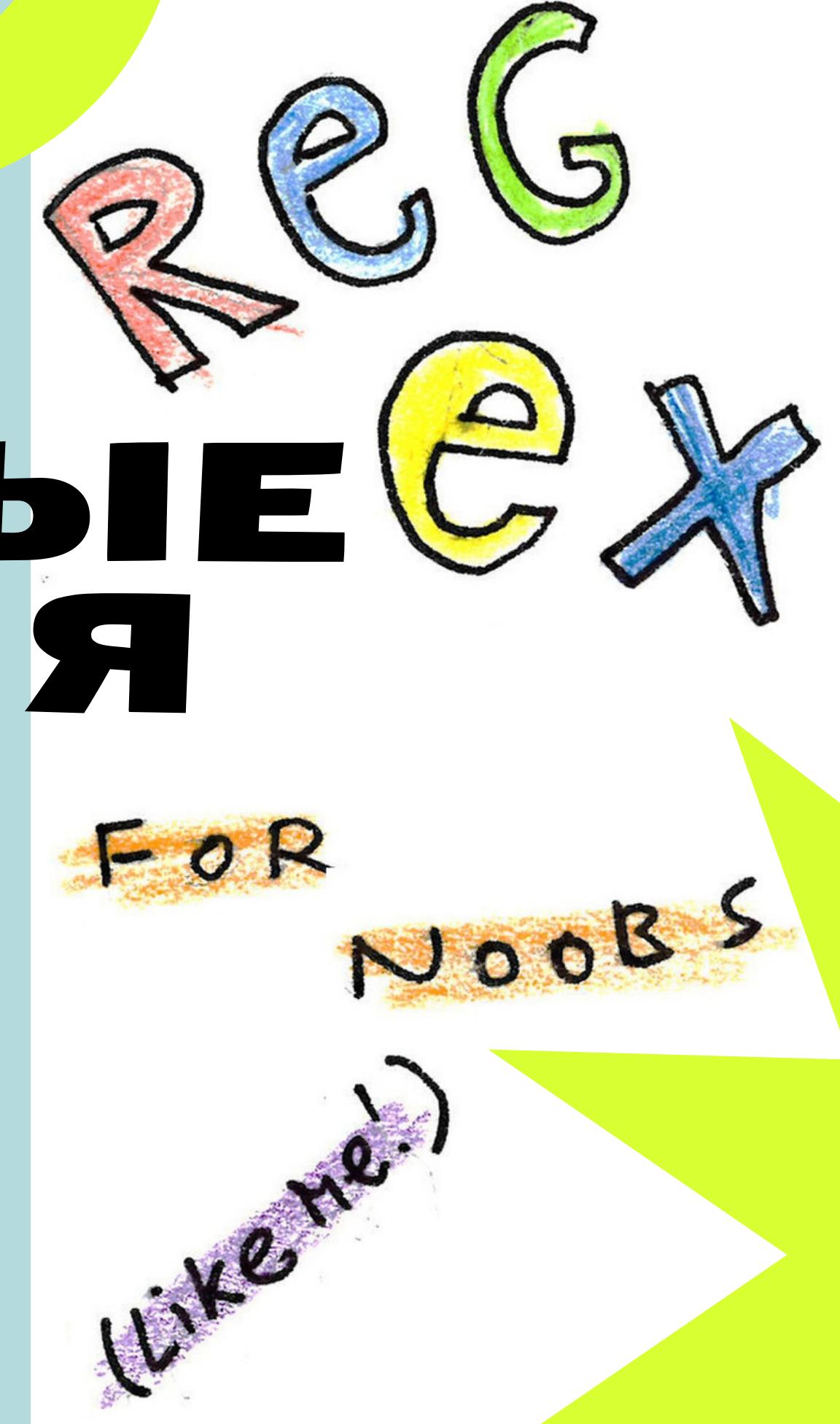


Системные технологии 2021

РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ

КАЛЕНКОВИЧ ВИОЛА
АХРАМОВИЧ ИРЕНА
КАРЧМИТ ВЛАД



ЧТО РЕГУЛЯРКА ДЕЛАЕТ ЧЕСТИРОВЩИКИ?

Регулярные выражения
(англ. regular expressions,
сокр. RegExp, RegEx, жарг.
регэкспы или регексы) —
ЭТО ...

А) ФОРМАЛЬНЫЙ ЯЗЫК
ПОИСКА И
ОСУЩЕСТВЛЕНИЯ
МАНИПУЛЯЦИЙ С
ПОДСТРОКАМИ В ТЕКСТЕ,
ОСНОВАННЫЙ НА
ИСПОЛЬЗОВАНИИ
МЕТАСИМВОЛОВ
(СИМВОЛОВ-ДЖОКЕРОВ,
АНГЛ. WILDCARD
CHARACTERS)

HOW TO REGEX

STEP 1: OPEN YOUR FAVORITE EDITOR

@GARABATOKID



STEP 2: LET YOUR CAT PLAY ON YOUR KEYBOARD

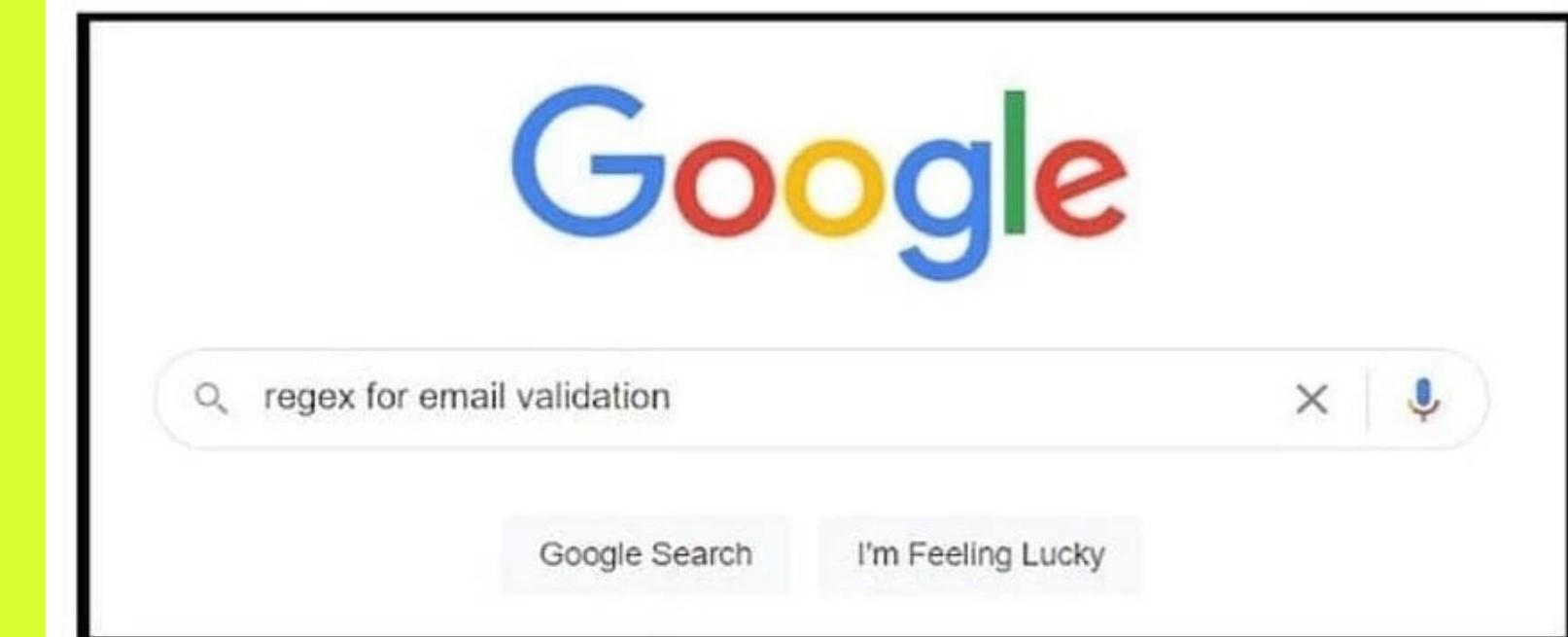


В) СТРОКА-ОБРАЗЕЦ
(АНГЛ. PATTERN, ПО-
РУССКИ ЕЁ ЧАСТО
НАЗЫВАЮТ «ШАБЛОНОМ»,
«МАСКОЙ»), СОСТОЯЩАЯ
ИЗ СИМВОЛОВ И
МЕТАСИМВОЛОВ И
ЗАДАЮЩАЯ ПРАВИЛО
ПОИСКА

DAY1 OF PROGRAMMING



10 YEARS OF PROGRAMMING



для чего это нужно?

Для проверки
корректности
пользователь-
ского ввода

Для поиска
и/или
удаления
чего-либо

Для замены
одной части
подстрок
другими

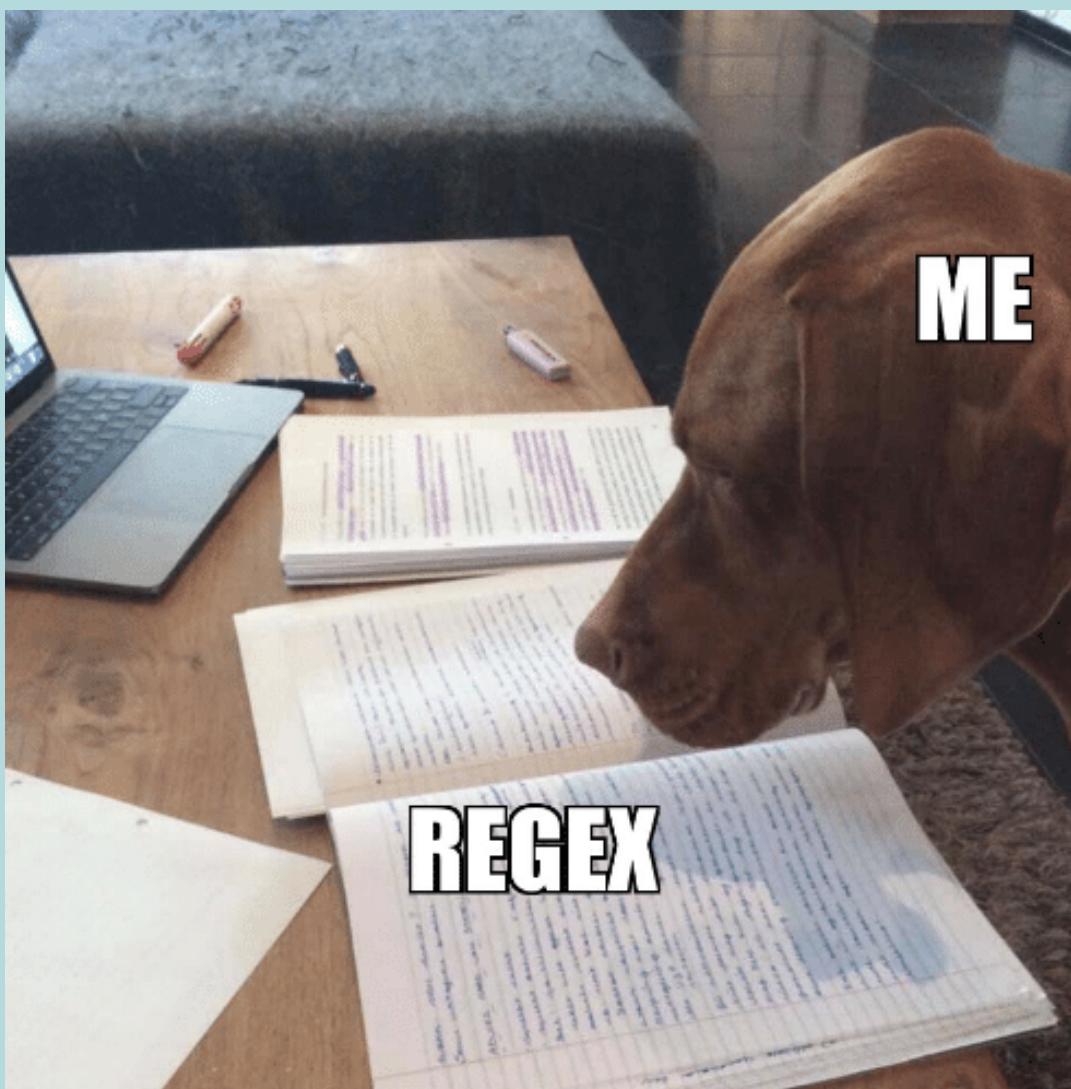
ЧЕМ РЕГУЛЯРНЫЕ ВЫРАЖЕНИЯ ЛУЧШЕ ПРОСТОГО ОПИСКА?

Тем, что позволяют задать шаблон.



NOTE PAD ++

Search Mode →
Regular expression



Screenshot of Notepad++ showing the 'Find' dialog box. The search term 'LT' is entered in the 'Find what:' field. The 'Search Mode' dropdown is set to 'Regular expression', which is highlighted with a red box. Other options in the 'Search Mode' dropdown include 'Normal' and 'Extended (\n, \r, \t, \0, \x...)'. The 'Find Next' button is highlighted with a blue box. The main window shows a file named 'change.log' containing a large amount of text.

*C:\Program Files\Notepad++\change.log - Notepad++ [Administrator]

File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?

ЭпИУС_390303.csv change.log

1 Notepad++ v8.1.9 bug-fixes:
2
3 ST ST ST ST ST ST ST ST
4 T ST ST ST ST S T ST ST ST ST S
5 ST
6 ST ST ST ST ST S T ST ST ST ST ST
7 ST
8 ST
9 ST
10 ST
11 ST S
12 ST
13 T ST ST ST ST ST ST ST ST ST S
14 ST
15 T ST ST ST ST ST ST ST ST ST S
16 ST
17 T ST ST ST ST ST ST ST ST S
18 ST
19 T ST ST ST S
20 ST
21 T ST ST S
22 ST
23 T S

Find

Find Replace Find in Files Find in Projects Mark

Find what: LT Find Next

In selection Count

Find All in Current Document

Find All in All Opened Documents

Close

Backward direction

Match whole word only

Match case

Wrap around

Search Mode

Normal

Extended (\n, \r, \t, \0, \x...)

Regular expression . matches newline

Transparency

On losing focus

Always

Normal text file length : 887 lines : 23 Ln : 23 Col : 1 Pos : 888 Windows (CR LF) UTF-8 INS

СПЕЦСИМВОЛЫ



Точка используется для поиска одного символа.

Большинство символов в регулярном выражении представляют сами себя за исключением специальных символов: [] \ / ^ \$. | ? * + () { }

Если вы хотите найти один из этих символов внутри вашего текста, его надо экранировать символом \ (обратный слеш).



ПРАВИЛО ПОИСКА ДЛЯ ТОЧКИ:

- Поиск любого символа
- ＼ - Поиск точки

[КВАДРАТНЫЕ СКОБКИ]

Внутри квадратных скобок мы указываем набор допустимых символов перечислением или диапазоном с помощью дефиса -

Обратите внимание — если мы перечисляем возможные варианты, мы не ставим между ними разделителей! Ни пробел, ни запятую — ничего.

[а-я] — все русские буквы в нижнем регистре от «а» до «я» (кроме «ё»)

[А-Яа-яЁё] — все русские буквы

[нл] — только «н» и «л»

[0-9] — любая цифра

[А-ГО-Р] — буквы от «А» до «Г» и от «О» до «Р»

[абв] — только «а», «б» или «в»

[а, б, в] — «а», «б», «в», пробел или запятая (что может привести к нежелательному результату)

КАРЕТА

внутри квадратных скобок означает исключение:



[^0-9] — любой символ, кроме цифр

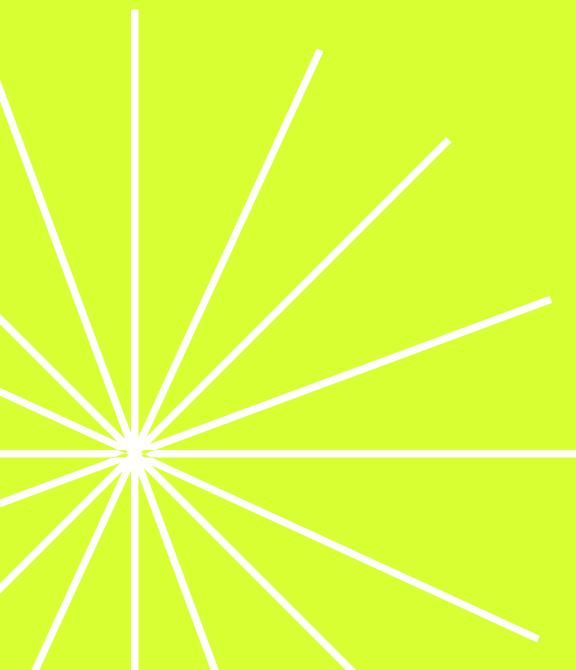
[^ёЁ] — любой символ, кроме буквы «ё» в любом регистре

[^а-в8] — любой символ, кроме букв «а», «б», «в» и цифры 8

**один символ! не два
или сто, а один!**

[1-31]. Нет, это не диапазон от 1 до 31, эта запись читается так:

- Диапазон от 1 до 3
- И число 1



ВЕРТИКАЛ ЬНАЯ | ЧЕРТА И (КРУГ ЛЫЕ СКОБКИ)

Вертикальной чертой обозначается логический оператор ИЛИ.
А при помощи скобок можно группировать части регулярных выражений.



Если мы хотим указать допустимые значения:

- Одного символа — используем []
- Нескольких символов или целого слова — используем | и ()

АН|ЛЯ

АНя
АЛя
муляж
банан

А(Н|Л)Я

АНя
АЛя
баня
халява

(<ДЕНЬ>)\.|<МЕСЯЦ>)\.|<ГОД>)

ОБЩЕЕ
ВЫРАЖЕНИЕ для
ДАТЫ

ВОПРОСИТЕЛ ЬНЫЙ ЗНАК (?), ПЛЮС (+) И ЗВЕЗДОЧКА (*)

Вопросительный знак (?) означает, что предшествующий ему символ может присутствовать или отсутствовать в строке.

Плюс (+) означает, что предшествующий символ должен присутствовать и может повторяться несколько раз подряд.

Звездочка (*) означает, что предшествующий символ может присутствовать, отсутствовать или повторяться несколько раз подряд.

10?

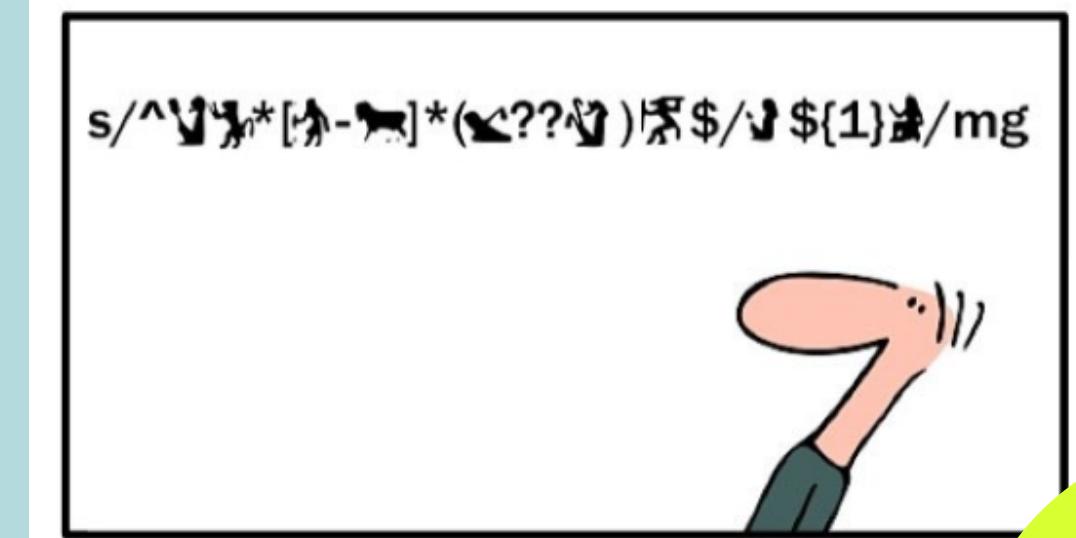
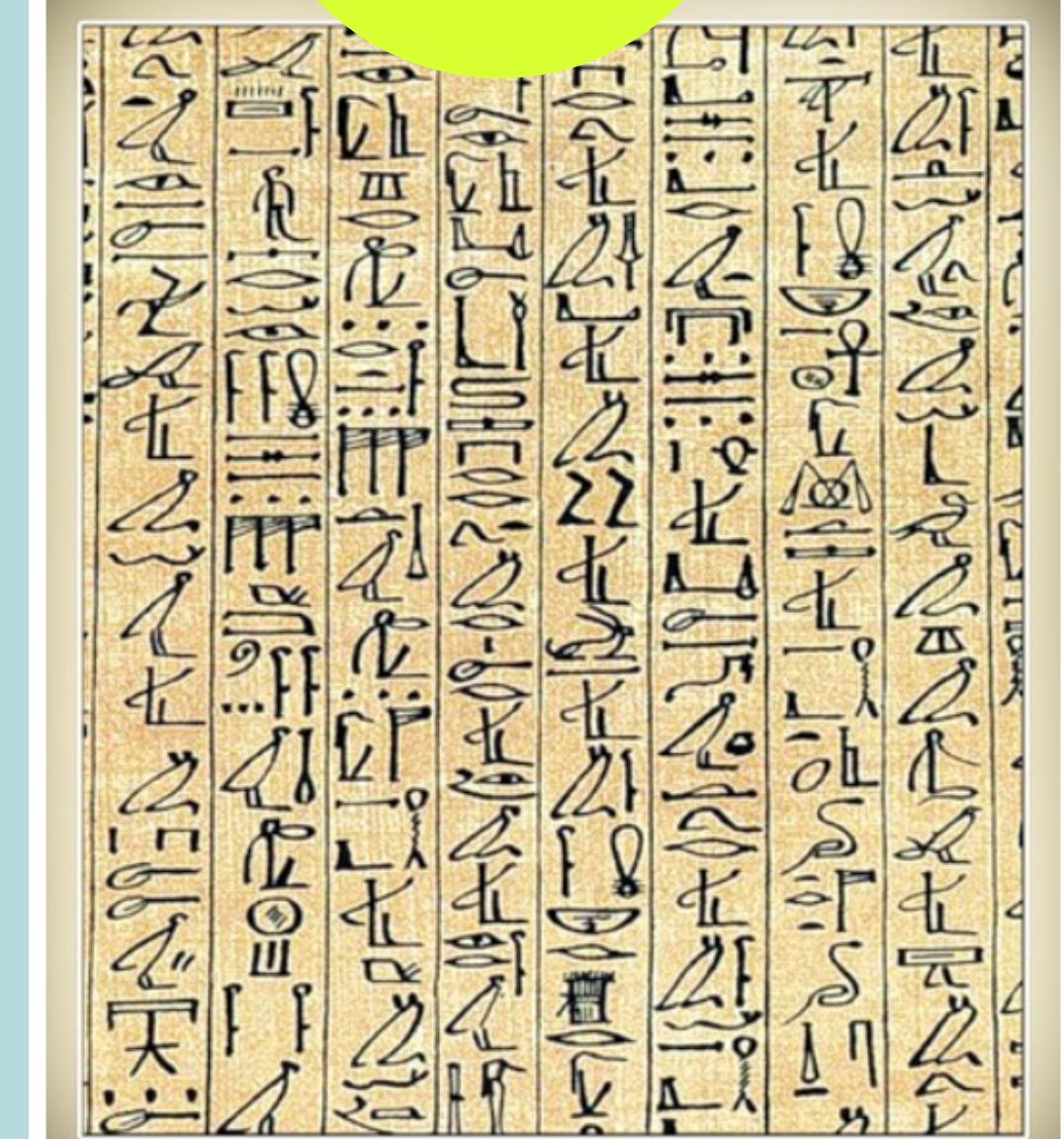
- 1
- 10

10+

- 10
- 100
- 1000 ...и т. д.

10*

- 1
- 10
- 100
- 1000 ...и т. д.



Древнеегипетское регулярное выражение

ФИГ УРНЫЕ ЖСКОБКИ

используются для указания
конкретное количество повторений

{N} РОВНО N РАЗ

{M,N} ОТ M ДО N ВКЛЮЧИТЕЛЬНО

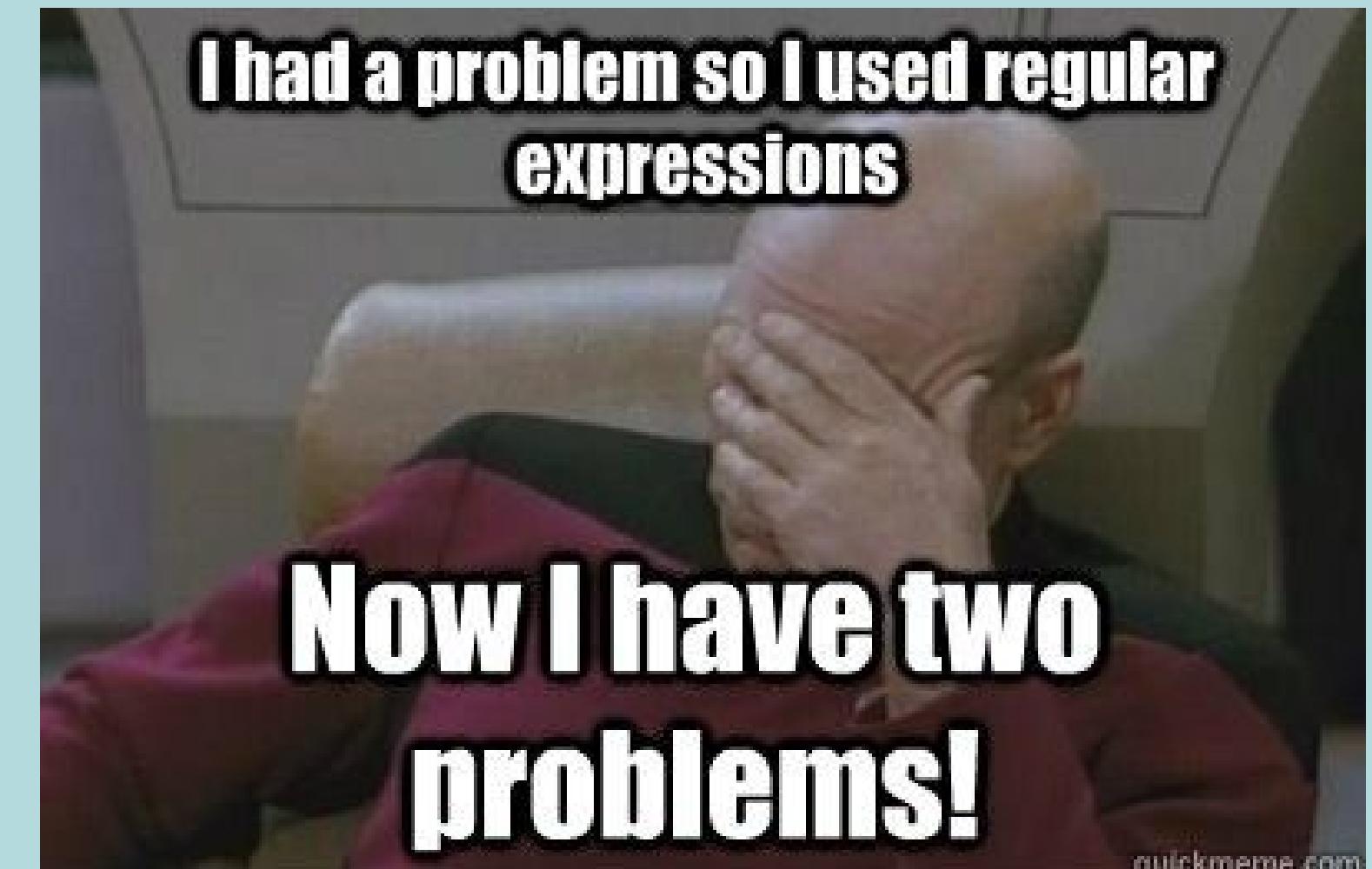
{M,} НЕ МЕНЕЕ M

{,N} НЕ БОЛЕЕ N

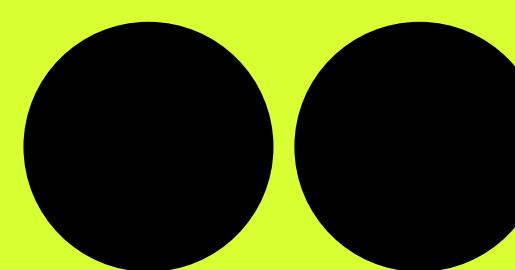
Позиция внутри строки



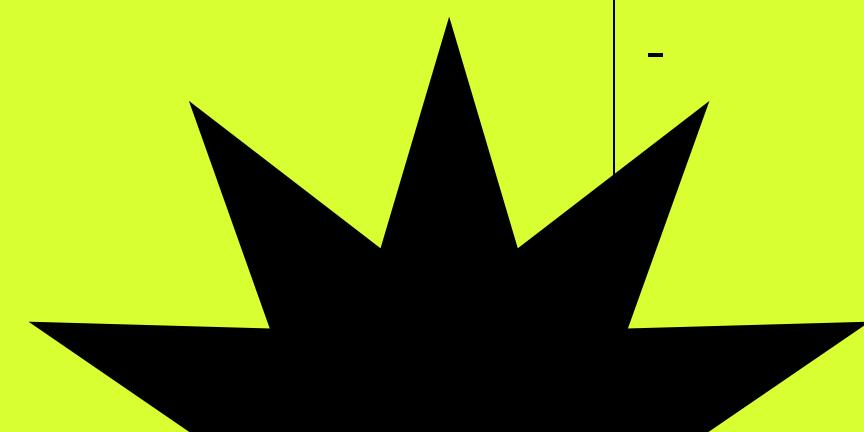
- \b Граница слова
- \B Не граница слова
- ^ Начало текста (строки)
- \$ Конец текста (строки)



МЕТАСИМВОЛЫ



Символ	Пояснение	Эквивалент
(пробел)	Пробел	-
\r	Возврат каретки (Carriage return, CR)	-
\n	Перевод строки (Line feed, LF)	-
\t	Табуляция (Tab)	-
\v	Вертикальная табуляция (vertical tab)	-
\f	Конец страницы (Form feed)	-
[\b]	Возврат на 1 символ (Backspace)	-
\d	Цифровой символ	[0-9]
\D	Нецифровой символ	[^0-9]
\s	Пробельный символ	[\f\n\r\t\v]
\S	Непробельный символ	[^ \f\n\r\t\v]
\w	Буквенный или цифровой символ или знак подчёркивания	[[:word:]]
\W	Любой символ, кроме буквенного или цифрового символа или знака подчёркивания	[^[:word:]]
.	Вообще любой символ	-



болят глаза?



ДАТА АД.ММ.ГГГГ

- 2 цифры дня
- точка
- 2 цифры месяца
- точка
- 4 цифры года

[0-9][0-9]\.[0-9][0-9]\.[0-9][0-9][0-9]

Найдет:

01.01.1999
05.08.2015

Тоже найдет:

08.08.8888
99.99.2000

Пробуем ограничить:

- День месяца может быть максимум 31 — первая цифра [0-3]
- Максимальный месяц 12 — первая цифра [01]
- Год или 19.., или 20.. — первая цифра [12], а вторая [09]

[0-3][0-9]\.[0-1][0-9]\.[12][09][0-9][0-9]

Не найдет:

08.08.8888
99.99.2000

Но найдет:

33.01.2000
01.19.1999
05.06.2999



Регулярки

Расписываем день подробнее

Если первая цифра:

- 0 — вторая может от 1 до 9 (даты 00 быть не может)
- 1, 2 — вторая может от 0 до 9
- 3 — вторая только 0 или 1

Составим регулярные выражения на каждый пункт:

- 0[1-9]
- [12][0-9]
- 3[01]

Получаем для дня: 0[1-9]|[12][0-9]|3[01]

По аналогии разбираем месяц и год

0[1-9]|1[0-2]\.19[0-9][0-9]|20[0-2][01]

Собираем вместе

0[1-9]|1[0-2]\.19[0-9][0-9]|20[0-2][01]

Как читается это выражение?

- ИЛИ 0[1-9]
- ИЛИ 1[0-2]
- ИЛИ 1[0-2]\.19[0-9][0-9]|20[0-2][01]



ПРОБЛЕМА!

Система не знает, что перебор вариантов | закончился на точке после дня.

(0[1-9]|1[0-2]\.19[0-9][0-9]|20[0-2][01])\.
(0[1-9]|1[0-2])\.
(19[0-9][0-9]|20[0-2][01])

EMAIL-АДРЕС

Из чего состоит:

- Буквы / цифры / _
- Потом @
- Снова буквы / цифры / _
- Точка
- Буквы

\w+@\w+\.\w+

Найдет:

test@mail.ru

vladyusha99@gmail.com

pupsik_9_and_kotik66@yandex.megamozg

Символ * часто используют с точкой — когда нам неважно, какой идет текст до интересующей нас фразы, мы заменяем его на «.*» — любой символ ноль или более раз.

@.\..*

Найдет:

test@mail.ru

olga31@gmail.com

pupsik_99@yandex.ru

Но также найдет:

@yandex.ru

test@.ru

test@mail.



ПРИМЕРЫ В СТ

E-mail (пример: «nick@mail.com»):

```
^([a-zA-Z_-]+\.)*[a-zA-Z_-]+@[a-zA-Z_-]+(\.[a-zA-Z_-]+)*\.[a-zA-Z]{2,6}$
```

URL (пример: «http://www.my-site.com»):

```
^((https?|ftp)\:\w)?([a-zA-Z]{1})((\.[a-zA-Z-])|([a-zA-Z-]))*\.( [a-zA-Z]{2,6})(\w?)$
```

Номера телефона (пример: «+375(29)555-55-55»):

```
^+\d{3}(\d{2})\d{3}-\d{2}-\d{2}$
```



для вас
выступали

Каленкович
Виола

Kalankovich_VA@st.by

Ахрамович
Ирена

Akhramovich_II@st.by

Карчмит
Влад

Karchmit_US@st.by

контактная
информация