

# Регулярные выражения

13.02.2021

Во время изучения  
чего-то нового,  
я самозабвенно  
выдумываю  
невероятные  
ситуации, в которых  
это умение поможет  
мне спасти мир

О нет! Убийца должно  
быть последовал  
за ней в отпуск!



Но чтобы узнать где он, нам нужно  
прочитать 200 Мб писем в поисках  
чего-то схожего по формату с адресом!



Это безнадежно!

Всем расступиться



Я знаю регулярные  
выражения



Регулярные выражения можно использовать во многих языках

программирования, в текстовых редакторах (Notepad++, Sublime, Geany и др.), в

специальных программах для поиска файлов и т. п.

- выяснить, содержится ли регулярное выражение в строке
- найти часть строки, подходящую под регулярное выражение
- заменить часть строки по регулярному выражению

Установить один из текстовых редакторов для своей операционной системы:

1) *Windows*: notepad++ <https://notepad-plus-plus.org/download/v7.4.2.html>

2) *Mac OS, Linux*: sublime <https://www.sublimetext.com/3>

<http://regexr.com/>

<https://regex101.com/>

### Настройка для Notepad++:

Вид - Отображение символов - отображать все символы

### Настройка для Sublime Text:

SublimeText - Preferences - Settings - User

"draw\_white\_space": "all"

# Некоторые особенности

- \s - любой пробельный символ (пробел, новая строка, табуляция)
- \S - не пробел
- \w - слово
- \W - не слово
- \n - новая строка
- \t - табуляция
- \b - граница слова
- \B - не граница слова
- \d - цифра

Регулярные выражения (англ. regular expressions) — формальный язык для работы с подстроками в тексте.

валидация и поиск

Операторы:

. - любой символ (**з**. -- **за**, **зо**, но не **з**)

**?** - символ может быть, а может не быть (**п?ол** -- **пол**, **ол**, но не **пл**)

| - либо один, либо другой (**сон|ром**) соответствует **сон**, **ром**, но не **сор**)

() - группировка (**пл(ов|аст)** соответствует **плов** **пласт**, но не **плав**)

# Операторы

$*$  — предыдущий символ или группа, повторённые любое количество раз

(включая 0)

$+$  — предыдущий символ или группа, повторённые любое положительное

количество раз

соответственно,  $.*$  — любое количество любых символов

# Пример

КО

Кошка, или домашняя кошка (лат. *Felis silvestris catus*)

— домашнее животное, одно из наиболее популярных

(наряду с собакой) «животных-компаньонов»



# Пример

ко

Кошка, или домашняя кошка (лат. *Felis silvestris catus*)

— домашнее животное, одно из наиболее популярных

(наряду с собакой) «животных-компаньонов»

# Пример

и(б|в)?ол

Кошка, или домашняя кошка (лат. *Felis silvestris catus*)

— домашнее животное, одно из наиболее популярных

(наряду с собакой) «животных-компаньонов»

# Пример

и(б|в)?ол

Кошка, или домашняя кошка (лат. *Felis silvestris catus*)

— домашнее ж**ИВОТ**ное, одно из на**иБОЛЕ**е популярных

(наряду с собакой) «ж**ИВОТ**НЫХ-компаньонов»

# Пример

до.\*e+

Кошка, или домашняя кошка (лат. *Felis silvestris catus*)

— домашнее животное, одно из наиболее популярных

(наряду с собакой) «животных-компаньонов»

# Пример

до.\*е+

Кошка, или домашняя кошка (лат. *Felis silvestris catus*)

— домашнее животное, одно из наиболее популярных

(наряду с собакой) «животных-компаньонов»

# Повторение

[...] — один из перечисленных символов (например: [абв])

[...-...] — один символ из диапазона (например: [а-я])

можно комбинировать: [а-яА-Яbq]

экранирование метасимволов: \[ \] \( \) \. \\*

# Начало и конец строки, инвертирование

<sup>^</sup> и <sup>\$</sup> — начало и конец строки

[<sup>^</sup>абв] — инвертированный класс символов (все символы, кроме перечисленных)

внутри квадратных скобок не работают круглые скобки, <sup>\*</sup> <sup>.</sup>

# Кванторы

$\{5\}$  — предыдущее выражение должно повториться ровно 5 раз

$\{3,5\}$  — предыдущее выражение должно повториться от 3 до 5 раз

$\{,5\}$  — max 5 раз,  $\{3,\}$  — min 3 раза



# Классы символов

[A-Z] — символы верхнего регистра

[a-z] — символы нижнего регистра

[A-Za-z] — символы верхнего и нижнего регистра

[A-Za-z0-9] — символы верхнего и нижнего регистра и цифры

[A-Za-z0-9\_] или \w — символы верхнего и нижнего регистра, цифры и \_

[^A-Za-z0-9\_] или \W — все, кроме символов верхнего и нижнего регистра, цифр и \_

группа – часть выражения, заключённая в скобки:

`abc(def)gh`

группы нумеруются по открывающей скобке:

`(abc(def(g|h).*)(ijk))(lm)`

(1) (2) (3) (4) (5)

`\2` – группа с номером 2

# Жадные квантификаторы

\* + ? {min,max} — жадные квантификаторы (greedy)

они пытаются съесть как можно больше символов:

к.\*от

КОТ ВЫПИЛ КОМПОТ

# Жадные квантификаторы

\* + ? {min,max} — жадные квантификаторы (greedy)

они пытаются съесть как можно больше символов:

к.\*от

КОТ ВЫПИЛ КОМПОТ

# Нежадные (ленивые) квантификаторы

`*? +? ?? {min,max}?` — нежадные квантификаторы

ОНИ ПЫТАЮТСЯ СЪЕСТЬ КАК МОЖНО МЕНЬШЕ СИМВОЛОВ:

`к.*?от`

КОТ ВЫПИЛ КОМПОТ

# Нежадные (ленивые) квантификаторы

\*? +? ?? {min,max}?

Они пытаются съесть как можно меньше символов:

к.\*?от

кот выпил компот

*Датский философ Søren Kierkegaard известен не только своими работами, но и количеством вариантов написания фамилии на русском языке.*

Киркегор - датский философ, богослов и писатель, один из предшественников экзистенциализма.

С. Кьеркегор окончил теологический факультет Копенгагенского университета в 1840 году. Степень магистра получил в 1841 году, защитив диссертацию “О понятии иронии, с постоянным обращением к Сократу”, посвященную концепциям иронии у древнегреческих авторов и романтиков. Работы С. Кьеркегора отличаются исключительной психологической точностью и глубиной. Вклад в развитие философии, сделанный Кьеркегаардом. неоценим.

Сёрен Киркегаард: немецкое издание Сёрена Киркегаарда.

Спецкурс “С. Керкегор и история христианства в XIX в.” посвящен датскому философу Серену Керкегору.