Регулярные выражения

ЛКШ.2013.Зима, параллель C+

Антон Полднев
lksh@poldnev.ru

31 декабря 2013 г.,
5 января 2014 г.

Что такое регулярное выражение?

Регулярное выражение — строка-шаблон, представляющая собой формальную запись для множества строк.

Примеры:

- `Winter` — соответствует лишь строке `Winter`
- `Goo+gle` — соответствует строкам `Google`, `Gooogle`, `Goooogle`...

Поддержка в языках программирования

- *Python:* модуль `re`
- *C++:* библиотека `<regex>` (C++11, gcc $\geqslant$ 4.9)
- *Perl, Javascript:* встроена в язык

Пример: поиск совпадения с шаблоном

Пытаемся найти подстроку, соответствующую шаблону `l[aio]st`, в строках `listing`, `the last day` и `Lost`:

- `listing`
- `the last day`
- `Lost`

Пример: поиск совпадения с шаблоном

Пытаемся найти подстроку, соответствующую шаблону `l[aio]st`, в строках `listing`, `the last day` и `Lost`:

- `listing`
- `the last day`
- `Lost`

Пример: поиск совпадения с шаблоном

Пытаемся найти подстроку, соответствующую шаблону `l[aio]st`, в строках `listing`, `the last day` и `Lost`:

- `listing`
- `the last day`
- `Lost`

Пример: поиск совпадения с шаблоном

Пытаемся найти подстроку, соответствующую шаблону `l[aio]st`, в строках `listing`, `the last day` и `Lost`:

- `listing`
- `the last day`
- `Lost`: нет совпадений, так как регистр имеет значение

Пример: поиск совпадения с шаблоном (Python)

```
>>> import re
>>> re.search(r'l[aio]st', 'listing')
<_sre.SRE_Match object at 0x02AB60C8>
>>> re.search(r'l[aio]st', 'the last day')
<_sre.SRE_Match object at 0x02AB6090>
>>> re.search(r'l[aio]st', 'Lost')
>>> bool(re.search(r'l[aio]st', 'listing'))
True
>>> bool(re.search(r'l[aio]st', 'the last day'))
True
>>> bool(re.search(r'l[aio]st', 'Lost'))
False
```

Пример: поиск совпадения с шаблоном (C++)

```
#include <iostream>
#include <regex>

using namespace std;

int main() {
    regex r(R"(l[aio]st)");

    if (regex_match("listing", r)) {
        cout << "Case 1: matched" << endl;  // printed
    }
    if (regex_match("the last day", r)) {
        cout << "Case 2: matched" << endl;  // printed
    }
    if (regex_match("Lost", r)) {
        cout << "Case 3: matched" << endl;  // not printed
    }
    return 0;
}
```

Пример: поиск совпадения с шаблоном (Javascript)

```
>>> /l[aio]st/.test('listing')
true
>>> r = /l[aio]st/
/l[aio]st/
>>> r.test('listing')
true
>>> r.test('the last day')
true
>>> r.test('Lost')
false
```

Специальные символы

- Проблема: `1+2` не найдётся в строке `41+29`, так как `+` имеет специальное значение

Специальные символы

- Проблема: `1+2` не найдётся в строке `41+29`, так как `+` имеет специальное значение
- Экранируем специальные символы с помощью `\`

Специальные символы

- Проблема: `1+2` не найдётся в строке `41+29`, так как `+` имеет специальное значение
- Экранируем специальные символы с помощью `\`
- `1\+2` найдётся в строке `41+29`

Специальные символы

- Проблема: `1+2` не найдётся в строке `41+29`, так как `+` имеет специальное значение
- Экранируем специальные символы с помощью `\`
- `1\+2` найдётся в строке `41+29`
- Специальные символы: `(`, `)`, `[`, `]`, `|`, `\`, `/`, `?`, `*`, `+`, `^`, `$`, `.`, `{`, `}`

Якоря начала и конца строки: ^ и $

- `^` в начале регулярного выражения означает, что искать нужно в начале строки
- Где найдётся `^eni`? `eniki`, `beniki`

Якоря начала и конца строки: ^ и $

- `^` в начале регулярного выражения означает, что искать нужно в начале строки
- Где найдётся `^eni`? `eniki`, `beniki`

Якоря начала и конца строки: ^ и $

- `^` в начале регулярного выражения означает, что искать нужно в начале строки
- Где найдётся `^eni`? `eniki`, `beniki`

Якоря начала и конца строки: ^ и $

- `^` в начале регулярного выражения означает, что искать нужно в начале строки
- Где найдётся `^eni`? `eniki`, `beniki`
- `$` в конце регулярного выражения означает, что искать нужно в конце строки
- Где найдётся `s$`? `ship`, `ships`

Якоря начала и конца строки: ^ и $

- `^` в начале регулярного выражения означает, что искать нужно в начале строки
- Где найдётся `^eni`? `eniki`, `beniki`
- `$` в конце регулярного выражения означает, что искать нужно в конце строки
- Где найдётся `s$`? `ship`, `ships`

Якоря начала и конца строки: ^ и $

- `^` в начале регулярного выражения означает, что искать нужно в начале строки
- Где найдётся `^eni`? `eniki`, `beniki`
- `$` в конце регулярного выражения означает, что искать нужно в конце строки
- Где найдётся `s$`? `ship`, `ships`

Якоря начала и конца строки: ^ и $

- Как с помощью функции поиска подстроки, соответствующей регулярному выражению, просто проверить, соответствует ли *вся* строка шаблону?

Якоря начала и конца строки: ^ и $

- Как с помощью функции поиска подстроки, соответствующей регулярному выражению, просто проверить, соответствует ли *вся* строка шаблону?
- Используем `^` и `$`

Якоря начала и конца строки: ^ и $

- Как с помощью функции поиска подстроки, соответствующей регулярному выражению, просто проверить, соответствует ли *вся* строка шаблону?
- Используем `^` и `$`
- Ищем `^str$`: `str`, `substr`

Якоря начала и конца строки: ^ и $

- Как с помощью функции поиска подстроки, соответствующей регулярному выражению, просто проверить, соответствует ли *вся* строка шаблону?
- Используем `^` и `$`
- Ищем `^str$`: `str`, `substr`

Якоря начала и конца строки: ^ и $

- Как с помощью функции поиска подстроки, соответствующей регулярному выражению, просто проверить, соответствует ли *вся* строка шаблону?
- Используем `^` и `$`
- Ищем `^str$`: `str`, `substr`

Любой символ: .

- `.` в регулярном выражении означает один любой символ (как правило, кроме перевода строки `\n`)
- Где найдётся `b.d`?

  `abcde`, `1b8d3`, `abde`, `abccde`, `bbdd`

Любой символ: .

- `.` в регулярном выражении означает один любой символ (как правило, кроме перевода строки `\n`)
- Где найдётся `b.d`?

  `abcde`, `1b8d3`, `abde`, `abccde`, `bbdd`

Любой символ: .

- `.` в регулярном выражении означает один любой символ (как правило, кроме перевода строки `\n`)
- Где найдётся `b.d`?

  `abcde`, `1b8d3`, `abde`, `abccde`, `bbdd`

Любой символ: .

- `.` в регулярном выражении означает один любой символ (как правило, кроме перевода строки `\n`)
- Где найдётся `b.d`?

  `abcde`, `1b8d3`, `abde`, `abccde`, `bbdd`

Любой символ: .

- `.` в регулярном выражении означает один любой символ (как правило, кроме перевода строки `\n`)
- Где найдётся `b.d`?

  `abcde`, `1b8d3`, `abde`, `abccde`, `bbdd`

Любой символ: .

- `.` в регулярном выражении означает один любой символ (как правило, кроме перевода строки `\n`)
- Где найдётся `b.d`?

  `abcde`, `1b8d3`, `abde`, `abccde`, `bbdd`

Любой символ: .

- `.` в регулярном выражении означает один любой символ (как правило, кроме перевода строки `\n`)
- Где найдётся `b.d`?

  `abcde`, `1b8d3`, `abde`, `abccde`, `bbdd`

Альтернатива: |, круглые скобки

- Хотим, чтобы регулярное выражение соответствовало лишь двум строкам: `test` и `twist`

Альтернатива: |, круглые скобки

- Хотим, чтобы регулярное выражение соответствовало лишь двум строкам: `test` и `twist`
- Используем символ `|`: `test|twist`

Альтернатива: |, круглые скобки

- Хотим, чтобы регулярное выражение соответствовало лишь двум строкам: `test` и `twist`
- Используем символ `|`: `test|twist`
- Для группировки частей можно использовать круглые скобки: `t(e|wi)st` соответствует тем же строкам

Альтернатива: |, круглые скобки

- Где найдётся `twist|test`?

  `twist`, `antitest`

Альтернатива: |, круглые скобки

- Где найдётся `twist|test`?

  `twist`, `antitest`

Альтернатива: |, круглые скобки

- Где найдётся `twist|test`?

  `twist`, `antitest`

Альтернатива: |, круглые скобки

- Где найдётся `twist|test`?

  `twist`, `antitest`
- Хотим, чтобы находилось соответствие только целой строке $\longrightarrow$ `^twist|test$`

Альтернатива: |, круглые скобки

- Где найдётся `twist|test`?
  `twist`, `antitest`
- Хотим, чтобы находилось соответствие только целой строке → `^twist|test$`
- В чём проблема?

Альтернатива: |, круглые скобки

- Где найдётся `twist|test`?
  `twist`, `antitest`
- Хотим, чтобы находилось соответствие только целой строке → `^twist|test$`
- В чём проблема?
- `antitest`

Альтернатива: |, круглые скобки

- Где найдётся `twist|test`?
  `twist`, `antitest`
- Хотим, чтобы находилось соответствие только целой строке → `^twist|test$`
- В чём проблема?
- `antitest`
- Исправляем с помощью скобок:
  `^(twist|test)$`

Квантификаторы ?, *, +, {}

Квантификатор после группы символов определяет, сколько раз она может повторяться:

- `?`: 0 или 1 раз
- `*`: 0 и более раз
- `+`: 1 и более раз
- `{2,5}`: от 2 до 5 раз
- `{2,}`: 2 и более раз
- `{,5}`: от 0 до 5 раз
- `{2}`: ровно 2 раза

Квантификаторы ?, *, +, {}

- `^Goo+gle$`: `Gogle`, `Google`, `Goooogle`
- `^G(oo)?gle$`: `Gogle`, `Google`, `Ggle`

Квантификаторы ?, *, +, {}

- `^Goo+gle$`: `Gogle`, `Google`, `Goooogle`
- `^G(oo)?gle$`: `Gogle`, `Google`, `Ggle`
- `^Y(a|o)+ndex$`:

Квантификаторы ?, *, +, {}

- `^Goo+gle$`: `Gogle`, `Google`, `Goooogle`
- `^G(oo)?gle$`: `Gogle`, `Google`, `Ggle`
- `^Y(a|o)+ndex$`: `Yandex`,

Квантификаторы ?, *, +, {}

- `^Goo+gle$`: `Gogle`, `Google`, `Goooogle`
- `^G(oo)?gle$`: `Gogle`, `Google`, `Ggle`
- `^Y(a|o)+ndex$`: `Yandex`, `Yooondex`,

Квантификаторы ?, *, +, {}

- `^Goo+gle$`: `Gogle`, `Google`, `Goooogle`
- `^G(oo)?gle$`: `Gogle`, `Google`, `Ggle`
- `^Y(a|o)+ndex$`: `Yandex`, `Yooondex`, `Yoaondex`

Квантификаторы ?, *, +, {}

- `^Goo+gle$`: `Gogle`, `Google`, `Goooogle`
- `^G(oo)?gle$`: `Gogle`, `Google`, `Ggle`
- `^Y(a|o)+ndex$`: `Yandex`, `Yooondex`, `Yoaondex`

Квантификаторы ?, *, +, {}

- `^Goo+gle$`: `Gogle`, `Google`, `Goooogle`
- `^G(oo)?gle$`: `Gogle`, `Google`, `Ggle`
- `^Y(a|o)+ndex$`: `Yandex`, `Yooondex`, `Yoaondex`
- `^Y.*ndex$`: `Yndex`, `Yundex`, `Y1i883ef-ndex`, `Yande`

Квантификаторы ?, *, +, {}

- `^Goo+gle$`: `Gogle`, `Google`, `Goooogle`
- `^G(oo)?gle$`: `Gogle`, `Google`, `Ggle`
- `^Y(a|o)+ndex$`: `Yandex`, `Yooondex`, `Yoaondex`
- `^Y.*ndex$`: `Yndex`, `Yundex`, `Y1i883ef-ndex`, `Yande`
- `Y.*ndex`: `0Yandexandex0`

Квантификаторы ?, *, +, {}

- `^Goo+gle$`: `Gogle`, `Google`, `Goooogle`
- `^G(oo)?gle$`: `Gogle`, `Google`, `Ggle`
- `^Y(a|o)+ndex$`: `Yandex`, `Yooondex`, `Yoaondex`
- `^Y.*ndex$`: `Yndex`, `Yundex`, `Y1i883ef-ndex`, `Yande`
- `Y.*ndex`: `0Yandexandex0`
- Как видим, квантификаторы по умолчанию *жадные*, то есть «захватывают» максимально возможное число символов

Жадные и ленивые квантификаторы

Ленивые квантификаторы «захватывают» минимально возможное число символов:

- `*?`
- `+?`
- `{3,}?`

Жадные и ленивые квантификаторы

Ленивые квантификаторы «захватывают» минимально возможное число символов:

- `*?`
- `+?`
- `{3,}?`
- `Yaa+?..`: `Yaaandex`

Жадные и ленивые квантификаторы

Ленивые квантификаторы «захватывают» минимально возможное число символов:

- `*?`
- `+?`
- `{3,}?`
- `Yaa+?..`: `Yaaandex`

Классы символов: []

- `[AEIOUaeiou]`: одна любая гласная буква латинского алфавита
- `[0123456789]` или `[0-9]`: любая цифра
- `[^a-zA-Z]`: любой символ, кроме буквы латинского алфавита
- `[-ad]`, `[ad-]` или `[a\-d]`: один символ `a`, `d` или `-`

Специальные классы символов: \w, \d и пр.

- `\d`: любая цифра
- `\D`: любой символ, кроме цифры
- `\w`: то же, что `[a-zA-Z0-9_]`
- `\W`: то же, что `[^a-zA-Z0-9_]`
- `\s`: любой пробельный символ (`[ \f\n\r\t\v]`)
- `\S`: любой непробельный символ

Примеры

1. Проверить, что строка выглядит как номер телефона: `+7 916 520-76-49`

Примеры

1. Проверить, что строка выглядит как номер телефона: `+7 916 520-76-49`

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

(здесь видно, что пробел не проглатывается)

Примеры

1. Проверить, что строка выглядит как номер телефона: `+7 916 520-76-49`

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

   (здесь видно, что пробел не проглатывается)

2. Проверить, что строка является корректной записью времени:

   `22:34:04`

Примеры

1. Проверить, что строка выглядит как номер телефона: `+7 916 520-76-49`

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

(здесь видно, что пробел не проглатывается)

2. Проверить, что строка является корректной записью времени:

`22:34:04`

```
^([01]\d|2[0-3])(:[0-5]\d){2}$
```

Избыточность языка

1. Можно оставить только круглые скобки, якоря `^` и `$`, `|` и `*`

Избыточность языка

1. Можно оставить только круглые скобки, якоря `^` и `$`, `|` и `*`
2. Как обойтись без `?`? `ab?c` $\Leftrightarrow$

Избыточность языка

1. Можно оставить только круглые скобки, якоря `^` и `$`, `|` и `*`
2. Как обойтись без `?`? `ab?c` $\Leftrightarrow$ `a(|b)c`

Избыточность языка

1. Можно оставить только круглые скобки, якоря `^` и `$`, `|` и `*`
2. Как обойтись без `?`? `ab?c` $\Leftrightarrow$ `a(|b)c`
3. Как обойтись без `+`? `ab+c` $\Leftrightarrow$

Избыточность языка

1. Можно оставить только круглые скобки, якоря `^` и `$`, `|` и `*`
2. Как обойтись без `?`? `ab?c` $\Leftrightarrow$ `a(|b)c`
3. Как обойтись без `+`? `ab+c` $\Leftrightarrow$ `abb*c`

Избыточность языка

1. Можно оставить только круглые скобки, якоря `^` и `$`, `|` и `*`
2. Как обойтись без `?`? `ab?c` $\Leftrightarrow$ `a(|b)c`
3. Как обойтись без `+`? `ab+c` $\Leftrightarrow$ `abb*c`
4. На что заменить `{,2}`? `ab{,2}c` $\Leftrightarrow$

Избыточность языка

1. Можно оставить только круглые скобки, якоря `^` и `$`, `|` и `*`
2. Как обойтись без `?`? `ab?c` $\Leftrightarrow$ `a(|b)c`
3. Как обойтись без `+`? `ab+c` $\Leftrightarrow$ `abb*c`
4. На что заменить `{,2}`? `ab{,2}c` $\Leftrightarrow$ `a(|b|bb)c`

Избыточность языка

1. Можно оставить только круглые скобки, якоря `^` и `$`, `|` и `*`
2. Как обойтись без `?`? `ab?c` $\Leftrightarrow$ `a(|b)c`
3. Как обойтись без `+`? `ab+c` $\Leftrightarrow$ `abb*c`
4. На что заменить `{,2}`? `ab{,2}c` $\Leftrightarrow$ `a(|b|bb)c`
5. На что заменить `[0-9]`?

Избыточность языка

1. Можно оставить только круглые скобки, якоря `^` и `$`, `|` и `*`
2. Как обойтись без `?`? `ab?c` $\Leftrightarrow$ `a(|b)c`
3. Как обойтись без `+`? `ab+c` $\Leftrightarrow$ `abb*c`
4. На что заменить `{,2}`? `ab{,2}c` $\Leftrightarrow$ `a(|b|bb)c`
5. На что заменить `[0-9]`?
   На `0|1|2|3|4|5|6|7|8|9`

Избыточность языка

1. Можно оставить только круглые скобки, якоря `^` и `$`, `|` и `*`
2. Как обойтись без `?`? `ab?c` $\Leftrightarrow$ `a(|b)c`
3. Как обойтись без `+`? `ab+c` $\Leftrightarrow$ `abb*c`
4. На что заменить `{,2}`? `ab{,2}c` $\Leftrightarrow$ `a(|b|bb)c`
5. На что заменить `[0-9]`? На `0|1|2|3|4|5|6|7|8|9`
6. Вместо `.` можно перечислить все символы, но лень

Сохраняющие круглые скобки

- Сохраняющие круглые скобки запоминают, какая часть строки соответствует части регулярного выражения в круглых скобках
- По умолчанию все круглые скобки сохраняющие
- `(\d{2}):(\d{2}):(\d{2})` : `54:18:33`

Сохраняющие круглые скобки

- Сохраняющие круглые скобки запоминают, какая часть строки соответствует части регулярного выражения в круглых скобках
- По умолчанию все круглые скобки сохраняющие
- `(\d{2}):(\d{2}):(\d{2})`: `54:18:33`
- `(\d{2}:?){3}` и `54:18:33`: `33` и больше ничего (сохранилось первое вхождение, но могло сохраниться и последнее)

Сохраняющие круглые скобки

- Сохраняющие круглые скобки запоминают, какая часть строки соответствует части регулярного выражения в круглых скобках
- По умолчанию все круглые скобки сохраняющие
- `(\d{2}):(\d{2}):(\d{2})`: `54:18:33`
- `(\d{2}:?){3}` и `54:18:33`: `33` и больше ничего (сохранилось первое вхождение, но могло сохраниться и последнее)
- Несохраняющие круглые скобки: `(?:...)`
- `(?:(\d\d)\.(\d\d)\.)?(\d{4})`: `31.12.2013`, `2014`

Пример: сохраняющие круглые скобки (Python)

```
>>> r = r'(?:(\d\d)\.(\d\d)\.)?(\d{4})'
>>> re.search(r, '31.12.2013').groups()
('31', '12', '2013')
>>> re.search(r, '2014').groups()
(None, None, '2014')
```

Пример: сохраняющие круглые скобки (Javascript)

```
>>> r = /(?:(\d\d)\.(\d\d)\.)?(\d{4})/
>>> r.exec('31.12.2013')
["31.12.2013", "31", "12", "2013"]
>>> r.exec('2014')
["2014", undefined, undefined, "2014"]
```

Пример: сохраняющие круглые скобки (C++)

```
string s("31.12.2013");
regex r(R"((?:(\d\d)\.(\d\d)\.)?(\d{4}))");
smatch sm;
regex_match(s, sm, r);

for (unsigned i = 0; i < sm.size(); ++i) {
    cout << sm[i] << endl;
}
// "31.12.2013", "31", "12", "2013"
```

Где используются регулярные выражения?

- Проверка корректности данных в HTML-формах (e-mail, номер телефона, количество денег...)
- Простейший парсинг HTML-кода (`<b>[^<]*<\/b>`)
- Поиск в тексте слов определённого вида или заданной длины
- Попытка угадать часть речи для слова (`ed$` $\Rightarrow$ глагол в прошедшем времени, `est$` $\Rightarrow$ прилагательное в превосходной степени)
- ...

Обратная связь

- Можно сослаться на сохранённую группу по номеру с помощью `\1`, ..., `\9`
- `(\d+)-\1=0`: `54-981=0`, `493-493=0`, `589-89=0`

Обратная связь

- Можно сослаться на сохранённую группу по номеру с помощью `\1`, ..., `\9`
- `(\d+)-\1=0`: `54-981=0`, `493-493=0`, `589-89=0`

Обратная связь

- Можно сослаться на сохранённую группу по номеру с помощью `\1`, ..., `\9`
- `(\d+)-\1=0`: `54-981=0`, `493-493=0`, `589-89=0`

Обратная связь

- Можно сослаться на сохранённую группу по номеру с помощью `\1`, ..., `\9`
- `(\d+)-\1=0`: `54-981=0`, `493-493=0`, `589-89=0`

Задачи

1. Стишок

Задачи

1. Стишок
2. Стишок с рифмовкой слога

Задачи

1. Стишок
2. Стишок с рифмовкой слога
3. Строки вида `aabbaa`