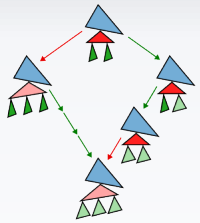


# Лабораторная работа 4: Группы захвата в регексах

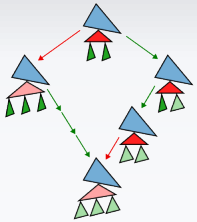
Вариант определяется последней цифрой в зачётке.

(Номера 0,1,3,4,9) Регексы с захватом групп и захватом строк

(Номера 2,5,6,7,8) Регексы с захватом групп и опережающими проверками

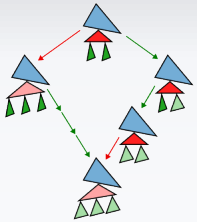
Данная задача стоит 15 баллов базово. При этом может быть выполнена частично:

- На 4 балла — только парсинг регекса с определением корректности его синтаксиса согласно ограничениям.
- На 8 баллов — проверка корректности регекса + построение каркасной КС-грамматики.
- На 10 баллов — построение атрибутной грамматики, но без парсинга.



# Техническое задание

- На вход подаётся regex (расширенное регулярное выражение) в синтаксисе, соответствующем варианту.
- Необходимо построить по нему атрибутную грамматику и запросить тестовый цикл выполнения, проверяющий, входят ли введённые пользователем слова в язык, определяемый выражением.



# Входные данные

Данная грамматика описывает синтаксис расширенных регулярных выражений. NB: грамматика носит чисто описательный характер и является неоднозначной.

$$
\begin{array}{rcl}
[\mathrm{rg}] & ::= & [\mathrm{rg}][\mathrm{rg}] \mid [\mathrm{rg}]|[\mathrm{rg}] \\
 & & \mid ([\mathrm{rg}]) \mid (?:[\mathrm{rg}]) \mid [\mathrm{rg}]* \\
 & & \mid (?[\mathrm{num}]) \mid [a-z] \\
[\mathrm{num}] & ::= & [1-9] \\
\text{(только вариант 1)}[\mathrm{rg}] & ::= & \backslash[\mathrm{num}] \\
\text{(только вариант 2)}[\mathrm{rg}] & ::= & (?=[\mathrm{rg}])
\end{array}
$$

Дополнительные ограничения:

- Количество групп захвата (обычных круглых скобок) в выражении не превышает 9.
- Внутри опережающих проверок нет групп захвата и нет опережающих проверок.
-

# Группы захвата

Выражения в обычных круглых скобках — это захваченные группы. Они нумеруются соответственно порядку вхождения в выражение левых скобок, ограничивающих их.

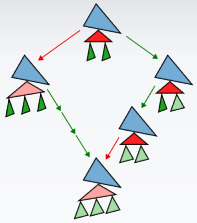
(?[num]) — это ссылка на выражение, определяемое группой захвата с номером [num]. \[num] — это ссылка на строку, захваченную группой захвата с номером [num].

Например, выражение (aa|bb)(?1) распознаёт язык {aabb, aaaa, bbaa, bbbb}, выражение (aa|bb)\1 распознаёт язык {aaaa, bbbb}.

Считаем, что выражение некорректно, если в нём встречается хотя бы один путь разбора, в котором ссылка на строку используется, но не была инициализирована. Например, (a|(bb))(a|\2) некорректно (к моменту обращения ссылка на bb могла быть не инициализирована).

Ссылки на выражения всегда считаются инициализированными, если внутри этих выражений используются только инициализированные к моменту обращения к ним строки. Например, (a|(bb))(a|(?3)) корректно, (a|(?2))(a|(bb\1)) некорректно (группа 1 ещё не дочитана до конца к моменту обращения к ней).

Рекурсивные ссылки на выражения возможны: см. (a(?1)b|c) — регекс для языка $\{a^n c b^n \mid n \in \mathbb{N}\}$.



# Построение атрибутной грамматики

- Проверить, соответствует ли регулярное выражение указанному синтаксису и корректно ли относительно существующих ограничений.
- Построить контекстно-свободную грамматику, распознающую «каркас» регулярного выражения (в котором нет опережающих проверок, а ссылки на захваченные строки заменяются свежими нетерминалами).
- Добавить в КС-грамматику атрибуты, соответствующие ограничениям, накладываемым расширенным синтаксисом. Атрибуты могут иметь произвольные типы, разрешается любой тип атрибутного наследования.