

2024

КЫК и его друзья

17 мая 2024 г.

Содержание

I КЯК парсер 1

II LL1 парсер 3

Часть I

КЯК парсер

Рассмотрим пример грамматики ниже. Возьмём для разбора строку *baaba*.

$$S \rightarrow AB \mid BC$$

$$A \rightarrow BA \mid a$$

$$B \rightarrow CC \mid b$$

$$C \rightarrow AB \mid a$$

Суть КЯК в рассмотрении подстрок от самых маленьких к целой строке. По итогу нам нужно узнать, является ли строка каким-то из нетерминалов (т.е. словом языка, порождённого грамматикой). Если совсем кратко - принадлежит ли строка языку.

Приступим к разбору. Ещё раз напомним, что мы его ведём снизу вверх - от терминалов к нетерминалам и так пока не охватим всю строку.

$$B \rightarrow b$$

$$A \rightarrow a$$

$$C \rightarrow a$$

Идём дальше. Это были подстроки длины 1. *baaba* также имеет подстроки длины 2, которые формируются из подстрок длины 1 (\times - декартово произведение):

$$ba = b \times a = B \times A = BA = A$$

$$ba = b \times a = B \times C = BC = S$$

Пока у нас есть $A \rightarrow ba, S \rightarrow ba$. Которые получаются, как мы выяснили, вот так:

$$A \rightarrow BA \rightarrow bA \rightarrow ba$$

$$S \rightarrow BC \rightarrow bC \rightarrow ba$$

Тупо последовательно заменяем нетерминалы, пока не доберёмся до терминалов. К сожалению, дальше мы упоремся и запутаемся, если будем так продолжать. Это была всего лишь первая 2-подстрока, а я уже устала.

Видимо, кто-то из авторов тоже устал, и придумал лестницу КЯК или как она там. Выглядит она следующим образом (цифры - длины подстрок, НТ - нетерминалы):

5	HT				
4	HT	HT			
3	HT	HT	HT		
2	HT	HT	HT	HT	
1	HT	HT	HT	HT	HT
	b	a	a	b	a

Эта таблица показывает, из каких нетерминалов какую строку мы можем построить. Пока что для нас она выглядит так:

5	HT				
4	HT	HT			
3	HT	HT	HT		
2	HT	HT	HT	HT	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

И, учитывая что мы уже успели проанализировать первое ba :

5	HT				
4	HT	HT			
3	HT	HT	HT		
2	S,A	HT	HT	HT	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Теперь давайте заполнять эту таблицу. Для следующих подстрок мы получим:

$$aa = a \times a = A, C \times A, C = AA, CA, AC, CC = B$$

$$ab = a \times b = A, C \times B = AB, CB = S, C$$

$$bb = b \times b = B \times B = \emptyset$$

5	HT				
4	HT	HT			
3	HT	HT	HT		
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Строки длины 3.

$$baa = b \times aa = B \times B = BB = \emptyset$$

$$baa = ba \times a = S, A \times A, C = SA, AA, SC, AC = \emptyset$$

$$aab = a \times ab = A, C \times S, C = AS, CS, AC, CC = B$$

$$aab = aa \times b = B \times B = \emptyset$$

$$aba = a \times ba = A, C \times S, A = AS, CS, AA, CA = \emptyset$$

$$aba = ab \times a = S, C \times A, C = SA, CA, SC, CC = B$$

5	HT				
4	HT	HT			
3	\emptyset	B	B		
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Строки длины 4.

$$baab = b \times aab = B \times B = BB = \emptyset$$

$$baab = ba \times ab = S, A \times S, C = SS, AS, SC, AC = \emptyset$$

$$baab = baa \times b = \emptyset \times B = \emptyset$$

$$aaba = a \times aba = A, C \times B = AB, CB = S, C$$

$$aaba = aa \times ba = B \times S, A = BS, BA = A$$

$$aaba = aab \times a = B \times A, C = BA, BC = A, S$$

5	HT				
4	∅	S,A,C			
3	∅	B	B		
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Строки длины 5.

$$baaba = b \times aaba = B \times S, A, C = BS, BA, BC = A, S$$

$$baaba = ba \times aba = S, A \times B = SA, AB = S, C$$

$$baaba = baa \times ba = \emptyset \times S, A = \emptyset$$

$$baaba = baab \times a = \emptyset \times A, C = \emptyset$$

5	S,A,C				
4	∅	S,A,C			
3	∅	B	B		
2	S,A	B	S,C	S,A	
1	B	A,C	A,C	B	A,C
	b	a	a	b	a

Таким образом, строка может быть получена путем последовательного раскрытия (развертки) нетерминалов S, A, C.

УПРАЖНЕНИЕ. Убедиться, что все указанные нетерминалы преобразуются в искомую строку.

Часть II

LL1 парсер

Всё начинается с таблицы парсинга. Для составления таблицы необходимо вначале посчитать множества символов FIRST и FOLLOW для каждого нетерминала NT.

Заметим, что первое правило грамматики для стартового нетерминала, если не указано иное. После стартового нетерминала может идти конец строки, для остальных в общем случае это неверно. Рассмотрим следующую грамматику:

$$A \rightarrow CB$$

$$B \rightarrow +CB \mid \epsilon$$

$$C \rightarrow ED$$

$$D \rightarrow *ED \mid \epsilon$$

$$E \rightarrow id \mid (A)$$

FIRST(NT) - множество символов на первой позиции строки NT (aka begin()):

$$FIRST(A) = FIRST(C)$$

$$FIRST(B) = ' + ' \cup \epsilon$$

$$FIRST(C) = FIRST(E)$$

$$FIRST(D) = ' * ' \cup \epsilon$$

$$FIRST(E) = ' id ' \cup ' ('$$

FOLLOW(NT) - множество символов на первой после последней позиции строки NT (aka end()) (т.е.

1) если видим $P \rightarrow \dots NT Q$, то добавляем $FIRST(Q)$ (если Q непусто) и $FOLLOW(P)$

2) если Q может быть ϵ , то добавляем $FOLLOW(Q)$

3) ϵ не может быть в FOLLOW, если появляется, то его надо убрать)

$$FOLLOW(A) = ') ' \cup ' \$ '$$

$$FOLLOW(B) = FOLLOW(A)$$

$$FOLLOW(C) = FIRST(B) \cup FOLLOW(A) \cup FOLLOW(B) \setminus \epsilon$$

$$FOLLOW(D) = FOLLOW(C) \setminus \epsilon$$

$$FOLLOW(E) = FIRST(D) \cup FOLLOW(C) \cup FOLLOW(D) \setminus \epsilon$$

Получим:

NT	FIRST	FOLLOW
A	id, (), \$
B	+, ϵ), \$
C	id, (+,), \$
D	*, ϵ	+,), \$
E	id, (*, +,), \$

Теперь, таблица парсинга. Она содержит производящие правила на пересечении терминалов и нетерминалов. Заполняется она так для каждого нетерминала NT:

1. Если $\epsilon \in FIRST(NT)$, то пишем правило, которым получился ϵ , во всех строках из $FOLLOW(NT)$.
2. Для всех прочих символов (не \$) из $FIRST(NT)$ в соответствующих столбцах пишется правило, которым этот символ был получен.

NT/T	id	(+)	*	\$
A	$A \rightarrow CB$	$A \rightarrow CB$				
B			$B \rightarrow +CB$	$B \rightarrow \epsilon$		$B \rightarrow \epsilon$
C	$C \rightarrow ED$	$C \rightarrow ED$				
D			$D \rightarrow \epsilon$	$D \rightarrow \epsilon$	$D \rightarrow *ED$	$D \rightarrow \epsilon$
E	$E \rightarrow id$	$E \rightarrow (A)$				

Удивительно, но алгоритм предельно прост. Мы всего лишь инициализируем стек стартовым нетерминалом, а в конец рассматриваемой строки добавляем конечный символ \$.

Далее происходит следующее. Мы достаём символ из строки и нечто со стека. Если нечто - терминал и совпало с символом, то *pop* и двигаем текущий символ. Если не совпало - ошибка. (При любой ошибке выходим из цикла.) Если же нечто - нетерминал, то мы ищем в таблице запись на пересечении нечто и текущего символа строки. Если запись не найдена, то ошибка. Иначе *pop*. Правая часть записи добавляется в стек, так что первый её элемент наверху.

Так повторяется, пока текущий символ не станет концом строки или не будет получена ошибка. Если ошибок не было, то строка принадлежит языку.

Рассмотрим пример. Грамматика

$$S' \rightarrow S\$$$

$$S \rightarrow xYzS \mid a$$

$$Y \rightarrow xYz \mid y$$

и строки *xyyza* и *xyyzz*.

Таблица 1:

NT	FIRST	FOLLOW
S'	x,a	\$
S	x,a	\$
Y	x,y	z,\$

Таблица 2:

NT/T	a	x	y	z	\$
S'	$S' \rightarrow S\$$	$S' \rightarrow S\$$			
S	$S \rightarrow a$	$S \rightarrow xYzS$			
Y		$Y \rightarrow xYz$	$Y \rightarrow y$		

Начинаем разбор строки *xyyza*:

Стек: S'
 Строка: *xyyza*\$
 Символ: x
 Нечто - S'

Нечто - нетерминал, и потому ищем пересечение S' и x . Находим $S' \rightarrow S\$$.
Стек: $\$S$

Строка: $xyzza\$$

Символ: x

Нечто - S

Нечто - нетерминал, и потому ищем пересечение S и x . Находим $S \rightarrow xYzS$.

Стек: $\$SzYx$

Строка: $xyzza\$$

Символ: x

Нечто - x

Нечто - терминал, и он совпал с текущим символом.

Стек: $\$SzY$

Строка: $xyzza\$$

Символ: x

Нечто - Y

Нечто - нетерминал, и потому ищем пересечение Y и x . Находим $Y \rightarrow xYz$.

Стек: $\$SzzYx$

Строка: $xyzza\$$

Символ: x

Нечто - x

Нечто - терминал, и он совпал с текущим символом.

Стек: $\$SzzY$

Строка: $yzza\$$

Символ: y

Нечто - Y

Нечто - нетерминал, и потому ищем пересечение Y и y . Находим $Y \rightarrow y$.

Стек: $\$Szzzy$

Строка: $yzza\$$

Символ: y

Нечто - y

Нечто - терминал, и он совпал с текущим символом.

Стек: $\$Szzz$

Строка: $zza\$$

Символ: z

Нечто - z

Нечто - терминал, и он совпал с текущим символом.

Стек: $\$Sz$

Строка: $za\$$

Символ: z

Нечто - z

Нечто - терминал, и он совпал с текущим символом.

Стек: $\$S$

Строка: $a\$$

Символ: a

Нечто - S

Нечто - нетерминал, и потому ищем пересечение S и a . Находим $S \rightarrow a$.

Стек: $\$a$

Строка: $a\$$

Символ: a

Нечто - a

Нечто - терминал, и он совпал с текущим символом.

Стек: \$

Строка: \$

Символ: \$

Нечто - \$

Нечто - \$, и он совпал с текущим символом.

Стек:

Строка:

Стек опустел, значит цикл закончился. Ошибки не было, посему строка *xyzza* принадлежит языку.

Теперь строка *xyzzz*. Здесь всё то же самое до момента:

Стек: \$S z

Строка: zz \$

Символ: z

Нечто - z

Нечто - терминал, и он совпал с текущим символом.

Стек: \$S

Строка: z \$

Символ: z

Нечто - S

Нечто - нетерминал, и потому ищем пересечение S и z . Не нашли. ОШИБКА. Выходим.

Таким образом, строка *xyzzz* не принадлежит языку.