Содержание

Введение

1 Сведения из теории формальных языков

Данный раздел посвящён минимально необходимым для реализации лексического анализа сведениям из теории формальных языков и конечных автоматов.

1.1 Определение алфавитов и языков

Прежде всего приведём определение алфавита и языка.

Алфавитом называется любое конечное множество некоторых символов. При этом понятие символа не определяется, поскольку оно в теории формальных языков является базовым.

Как правило, алфавит будем обозначать заглавными греческими буквами (например, буквой Σ), возможно, с нижними индексами.

Приведём примеры алфавитов:

1) $\{0, 1\}$ — алфавит Σ_1, состоящий из нуля и единицы;
2) $\{A, B, \ldots, Z\}$ — алфавит Σ_2, состоящий из заглавных латинских букв;
3) $\{\text{А}, \text{Б}, \text{В}, \text{Г}, \text{Д}, \text{Е}, \text{Ё}, \ldots, \text{Я}\}$ — алфавит Σ_3, состоящий из заглавных русских букв;
4) {**int**, **void**, **return**, *, '(', ')', '{', '}', ';', main, *number*} — алфавит Σ_4, состоящий из ключевых слов **int**, **void**, **return** языка C, идентификатора main, звёздочки, круглых скобок, фигурных скобок, точки с запятой, и целых чисел *number* (синтаксис целых чисел — как в языке C);
5) $\{a_1, a_2, a_3, a_4\}$ — алфавит Σ_5, состоящий из каких–то четырёх символов.

Из символов алфавита можно составлять **строки**, то есть конечные последовательности символов. Если строка состоит из символов алфавита Σ, то её называют **строкой над алфавитом** Σ. **Длиной строки** x называется количество символов в этой строке. Длину строки x будем обозначать $|x|$. Строка, вообще не содержащая символов, называется **пустой строкой** и будет обозначаться ε.

Приведём примеры строк:

1) 0111001 — строка над алфавитом Σ_1;

2) ENGLISH, INTEL — строки над алфавитом Σ_2;

3) МОСКВА, ГОРЬКИЙ, АЛЁШКОВО — строки над алфавитом Σ_3;

4) **int** main (**void**)'{' **return** *number*';' '}' — строка над алфавитом Σ_4;

5) $a_1a_3a_2a_2a_4$ — строка над алфавитом Σ_5.

Далее потребуется операция **сцепления** (иногда говорят **конкатенации**) строк. Эта операция заключается в приписывании одной строки в конец другой. Например, если строки α и β таковы, что $\alpha = \text{abc}$, $\beta = \text{defg}$, то конкатенация строк α и β обозначается $\alpha\beta$, и представляет собой строку abcdefg.

Множество всех строк над алфавитом Σ обозначается Σ^*. Скажем, если $\Sigma = \{0, 1\}$, то $\Sigma^* = \{\varepsilon, 0, 1, 00, 01, 10, 11, 000, \ldots, 1010, \ldots\}$. Ясно, что множество всех строк над заданным алфавитом — бесконечно, а точнее — счётно.

Любой набор строк над некоторым алфавитом называется **языком** (ещё называют **формальным языком**, чтобы отличать от естественных языков). Допустим, из всевозможных строк над алфавитом $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, '.', -\}$ можно выбрать те, которые являются корректной записью некоторого вещественного числа: $L = \{0, -1.5, 1002.123345, 777, \ldots\}$. Языки обозначаются заглавными латинскими буквами, возможно с нижними индексами. Язык может являться конечным множеством строк: если L — язык над алфавитом $\{a, b\}$, содержащий лишь строки короче трёх символов, то $L = \{\varepsilon, a, b, aa, ab, ba, bb\}$.

1.2 Операции над языками и регулярные выражения

Поскольку язык — это некоторое множество строк, то нужно уметь это множество как–то описывать. Одним из способов описания являются так называемые регулярные выражения. Прежде чем определить, что такое регулярное выражение, нужно определить операции над языками. Операции над языками, которые нам потребуются, собраны в приводимой ниже табл.1.2.

Таблица 1. Операции над языками.

Операция	Определение и обозначение операции
Объединение L и M	$L \cup M = \{s : s \in L \text{ или } s \in M\}$
Сцепление L и M	$LM = \{st : s \in L \text{ и } t \in M\}$
Замыкание Клини́ языка L	$L^* = \bigcup_{i=0}^{\infty} L^i$
Положительное замыкание языка L	$L^+ = \bigcup_{i=1}^{\infty} L^i$

В этой таблице L и M — некоторые языки

2 Примеры реализации лексического анализа

В настоящем разделе мы приводим примеры реализации лексического анализа для простых ситуаций.