Regular language or Regular Expressions

Let, A be a finite alphabet with k elements.

$$W_A := \{(a_i)_{i \leq n} | n \in \mathbb{N}, a_i \in A\}$$

Definition

A formal language in A is a subset $\mathscr{L} \subset W$

Example:

- The set of all C programs that can be compiled using a C compiler are a formal language in the character set of ASCII.
- $A = \{H, T\}$, $\mathfrak{T} = \{(a_i)_{(i \leq n)} | n \in \mathbb{N}, a_i \in A, \#(i \ni a_i = H) = \#(j \ni a_j = T)\}$

RegExs

A RegEx is the smallest formal language R (on character set of A) with the following:

- $$\varepsilon \in R$$
- $$A \subset R$$
- $$\forall u, v \in R, u | v \in R$$
- $$\forall u, v \in R, u \cdot b \in R$$
- $$\forall u \in R, u^* \in R$$

Matching

We'll show matching with the symbol †

$$\forall w \in W, r \in R$$

$$\dagger : W \longrightarrow R \longrightarrow \{\text{false}, \text{true}\}$$

$$\begin{aligned} w \dagger r \quad := & \\ & | \ (w = \varepsilon) \wedge (r = \varepsilon) \\ & | \ (w = r) \wedge w \in A \wedge r \in A \\ & | \ (w \dagger u \vee w \dagger v) \wedge (r = u | v) \\ & | \ (w = w_1 w_2 \wedge w_1 \dagger u \wedge w_2 \dagger v) \wedge (r = u \cdot v) \\ & | \ (w = (w_i)_{i \leq n \in \mathbb{N}} \wedge \forall i, w_i \dagger u) \wedge (r = u^*) \end{aligned}$$

Implementation example: ripgrep on Linux and Unix

Example:

- colo(u|ε)r matches color and colour (($x|\varepsilon$) is commonly written as x?)
- (0|1)*0 matches all binary strings which represent even numbers.

Definition

A language $\mathfrak{L}_u$ is called S-regular if it consists of all the words that match a given regular expression u.