Lesson 4: Regular expressions

Theme: Regular expressions as alternative description of regular languages.

In the following we fix an alphabet $\Sigma$. Regular expressions (over $\Sigma$) are expressions built inductively as follows.

- $\emptyset$ is a regular expression.
- $a$ is a regular expression, for every symbol $a \in \Sigma$.
- If $e_1, e_2$ are regular expressions, then so are $(e_1 \cdot e_2)$ and $(e_1 \cup e_2)$.
- If $e$ is a regular expression, then so is $(e)^*$.

A regular expression $e$ over $\Sigma$ defines a language, denoted by $L(e)$, over the same alphabet as follows.

- If $e$ is $\emptyset$, then $L(e) = \emptyset$.
- If $e$ is $a$, where $a \in \Sigma$, then $L(e) = \{a\}$.
- If $e$ is of the form $(e_1 \cdot e_2)$, where $e_1$ and $e_2$ are regular expressions, then $L(e) = L(e_1) \cdot L(e_2)$.
- If $e$ is of the form $(e_1 \cup e_2)$, where $e_1$ and $e_2$ are regular expressions, then $L(e) = L(e_1) \cup L(e_2)$.
- If $e$ is of the form $(e_1)^*$, where $e_1$ is a regular expression, then $L(e) = L(e_1)^*$.

Usually, we omit writing $\cdot$ in $(e_1 \cdot e_2)$, and instead, we simply write $(e_1e_2)$. Also, when there is no ambiguity, we will omit writing the brackets and simply write $e_1e_2$ and $e_1^*$, instead of $(e_1e_2)$ and $(e_1)^*$.

The following theorem states that the class of languages defined by regular expressions is exactly the class of regular languages.

**Theorem 4.1** Regular expressions define precisely the class of regular languages. That is, for every regular expression $e$ over $\Sigma$, $L(e)$ is a regular language, and vice versa, for every regular language $L$, there is a regular expression $e$ such that $L(e) = L$.

Combining what we have learnt so far, we obtain three different, but equivalent, characterisations of regular languages, as stated below.

**Corollary 4.2** Let $L$ be a language. The following are equivalent.

- $L$ is accepted by a DFA.
- $L$ is accepted by an NFA.
- $L$ is defined by a regular expression.

**Remark 4.3** The term regular expressions are commonly abbreviated as regex. In most literatures and websites, the term “regex” are used more often than “regular expression.” Due to its widespread applications, many modern programming languages now include libraries for regex. The following are some of them.

- Java: [https://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html](https://docs.oracle.com/javase/7/docs/api/java/util/regex/Pattern.html)
- Python: [https://docs.python.org/2/library/re.html](https://docs.python.org/2/library/re.html)