Université Paris Cité | UFR Informatique
Automates et Analyse Lexicale | Année 2023–2024

Handout 1

Mots et Langages

1 Mots

1.1 Définitions de base

Un *alphabet* est un ensemble fini de *lettres* (ou *symboles*).

Un *mot* sur un alphabet Σ est une séquence de lettres de l'alphabet Σ. On note

- Σ^* pour l'ensemble de tous le mots sur l'alphabet Σ ;
- $|w|$ pour la longueur du mot w. Pour tous les mots w, $|w|$ est un nombre entier non négatif qui peut être 0.
- ϵ (prononcé "epsilon") pour le *mot vide.* C'est la séquence de longueur 0 qui ne contient aucune lettre.

1.2 Mesures sur des mots

Nous avons déjà vu la notion de *longueur* $|w|$ d'un mot w. Si $w \in \Sigma^*$ et $a \in \Sigma$, alors $|w|_a$ est le *nombre d'occurrences de a dans w*, c'est-à-dire le nombre de positions dans le mot w où il y la lettre a. Évidement, si $w \in \Sigma^*$, on a que

$$|w| = \Sigma_{a\in\Sigma}|w|_a$$

1.3 Opérations sur des mots

Soit un alphabet Σ. La *concaténation* de deux mots $v = v_1 \dots v_n \in \Sigma^*$ et $w = w_1 \dots w_m \in \Sigma^*$, est le mot $v_1 \dots v_n w_1 \dots w_m$ (attention les mots v et w peuvent être vides). On note $v \cdot w$ pour la concaténation de v et w, très souvent on n'écrit pas l'opérateur $\cdot$, et on écrit simplement vw à la place de $v \cdot w$.

Par exemple, $abab \cdot cd = ababcd$

Propriétés importantes de la concaténation :

- pour tous mots w : $\epsilon \cdot w = w$
- pour tous mots w : $w \cdot \epsilon = w$
- pour tous mots u, v, w : $u \cdot (v \cdot w) = (u \cdot v) \cdot w$

Si w est un mot et n un nombre non négatif, alors nous écrivons w^n pour

$$\underbrace{w \cdots w}_{n \text{ fois}}$$

Si $w = w_1 \dots w_n$ est un mot de longueur n, alors son *miroir*, noté $\overline{w}$, est le mot $\overline{w} = w_n \cdots w_1$.

1.4 Relations entre des mots

Soit Σ un alphabet. On a les relations suivantes entres des mots sur Σ :

- v est un *préfixe* de w s'il existe un mot u tel que $w = v \cdot u$. Il s'agit d'un *préfixe propre* quand $u \neq \epsilon$.
- v est un *suffixe* de w s'il existe un mot u tel que $w = u \cdot v$. Il s'agit d'un *suffixe propre* quand $u \neq \epsilon$.
- v est un *facteur* de w s'il existe deux mots x, y tels que $w = x \cdot v \cdot y$.
- v est un *sous-mot* de w quand v est obtenu de w en supprimant certains positions de w. Autrement dit, v est un sous-mot de w quand il y a une décomposition de w en facteurs

  $$w = w_0 v_1 \cdots w_{n-1} v_n w_n$$

  tel que

  $$v = v_1 \cdots v_n$$

2 Langages

Si X est un ensemble, nous écrivons $\mathcal{P}(X)$ pour l'ensemble des parties de X.

Un *langage* sur un alphabet Σ est simplement un sous-ensemble de Σ^*. Autrement dit, un langage sur Σ est un ensemble de mots sur Σ.

Toutes les notions ensemblistes s'appliquent donc aux langages : $\emptyset$ est le langage vide, $L_1 \cup L_2$ est l'union des deux langages L_1 et L_2, et $L_1 \cap L_2$ est leur intersection. Quand on parle du complément d'un langage L il faut être précis sur l'alphabet sous-jacent, car le complément du langage L par rapport à l'alphabet Σ est

$$L^{comp} = \{w \in \Sigma^* \mid w \notin L\}$$

Plus intéressant sont les opérations qui sont spécifiques aux langages :

La *concaténation* de deux langages L_1 et L_2 est définie par

$$L_1 \cdot L_2 = \{w_1 \cdot w_2 \mid w_1 \in L_1, w_2 \in L_2\}$$

Cela nous permet de définir, pour un langage L et $n \geq 0$:

$$L^n = \underbrace{L \cdot \ldots \cdot L}_{n \text{ fois}}$$

et finalement

$$L^* = \bigcup_{n \geq 0} L^n$$

L'opérateur $*$ est appelé *l'étoile de Kleene*, en honneur du mathématicien Stephen C. Kleene.