Chapitre 1 - Ensembles de mots

Benjamin WACK (cours) - Mica MURPHY (note) - Antoine SAGET (note)

Lundi 1er Octobre 2018

0) Introduction

Discret est l'opposé de continu, et il peut y avoir un nombre fini ou infini de valeurs. On ne fera ni de géométrie ni d'analyse de fonctions (dérivées, etc.).

1) Mots

a) Alphabets et mots

Définition. Un **alphabet** est un ensemble fini de symboles.

Exemples.

- alphabet de 26 lettres
- code ASCII
- notes de musique

Définition. un **mot** sur un alphabet A est une suite ordonnée finie de symboles de A.

L'ordre des lettres est important : abba est différent de baab. Il peut y avoir des répétitions.

Si x_1, x_2, x_n sont des symboles de A ; on peut parler du mot $x = x_1x_2...x_n$

Cas particulier. Le **mot vide** à 0 symboles noté ϵ.

ϵ n'est pas un symbole de A

On note A^n l'ensemble des mots sur A formés de n symboles et A^* l'ensemble de tous les mots sur A.

Définition. On appelle **longueur d'un mot** le nombre de symboles qui le composent.

$$lg(x_1x_2...x_n) = n$$
$$lg(\epsilon) = 0$$

Dans A^* on retrouve chaque symbole de A sous la forme d'un mot de longueur 1.

Exemples.

- alphabet latin à 26 lettres

  Toute suite de lettres est appelée **mot** (même s'il n'est pas dans le dictionnaire)

- alphabet binaire $B = \{0, 1\}$ Il y a 2^n mots binaires de longueur n.

- alphabet des chiffres $\{0, 1, 2, \ldots, 9\}$ un mot sur cet alphabet représente un nombre entier

Définitions. On appelle **langage** sur A un ensemble (fini ou infini) de mots sur A, autrement dit une partie de A^*.

Exemples.

- Les mots du dictionnaire *Larousse 2018*
- Les suites de chiffres qui ne commencent pas par un 0.
- Le langage d'un seul mot $\{u\}$
- $\{\epsilon\}$
- Le langage vide : $\{\emptyset\}$ (à ne pas confondre avec ϵ !)
- A^*

b) Préfixe, suffixe, facteur

Concaténation

Soient $u = u_1 u_2 \dots u_n$ et $v = v_1 v_2 \dots v_p$ alors **le concaténé** de u et v noté simplement uv est le mot $u_1 u_2 \dots u_n v_1 v_2 \dots v_p$

Exemple. Si $u = 1011$ et $v = 010$ alors $uv = 1011010$

Préfixe, suffixe, facteur

Soient u et v deux mots sur A. On dit que u **est un préfixe de** v si il existe un mot w tel que $v = uw$

> w peut être le mot vide.

On note $u \sqsubseteq v$ le fait que u est préfixe de v $u \sqsubset v$ le fait que u est préfixe strict de v (cas où $w \neq \epsilon$)

Autre caractérisation : si $u = u_1 u_2 \dots u_n$, $v = v_1 v_2 \dots v_p$ alors $u \sqsubseteq v$ si et seulement si $u_1 = v_1, u_2 = v_2, \dots, u_n = v_n$ et $n \leq p$

Propriété. Si $u \sqsubseteq v$ et $v \sqsubseteq u$ alors $u = v$

Propriété. Si $u \sqsubseteq v$ alors $lg\ u \leq lg\ v$ et si $u \sqsubset v$ alors $lg\ u < lg\ v$

On dit que u est un :

- **suffixe** de v s'il existe un mot w tel que $v = wu$.
- **facteur** de v si il existe 2 mots x et y tels que $v = xuy$

Exemples. Soit le mot *baaca* :

- ses préfixes sont ϵ, b, ba, baa, $baac$, $baaca$.
- ses suffixes sont ϵ, a, ca, aca, $aaca$, $baaca$
- ses facteurs sont ϵ, b, ba, baa, $baac$, $baaca$, a, aa, aac, $aaca$, ac, aca, c, ca

Propriété. Si u est un mot de longueur n, il admet exactement $n+1$ préfixes distincts, $n+1$ suffixes distincts et au moins $n+1$ facteurs (souvent plus).

Propriétés.

- $lg(uv) = lg(u) + lg(v)$
- $lg(u^n) = n \times lg(u)$ (où u^n est le mot u répeté n fois)
- $u^0 = \epsilon$

  Soit P : "$w = uv$" et Q : "$lg(w) = lg(u) + lg(v)$" on a $P \Rightarrow Q$.

  La réciproque ($Q \Rightarrow P$) n'est pas vraie : :white_check_mark: Si $w = uv$ alors $lg(w) = lg(u) + lg(v)$:negative_squared_cross_mark: Si $lg(w) = lg(u) + lg(v)$ alors $W = uv$ Contre-exemple : $u = a$, $v = b$, $w = aa$

  En revanche, la contraposée ($!Q \Rightarrow !P$) est vraie : Si $lg(w) \neq lg(u) + lg(v)$ alors $w \neq uv$

c) Distance entre mots

Soient u et v deux mots sur A de même longueur La **distance** de u à v est le nombre de symboles de u qu'il faut modifier pour obtenir v.

Exemples.

- $u = arbre$, $v = aller$, $d(u, v) = 4$ (seul le a est identique aux 2)
- $u = 0101110$, $v = 0011101$, $d(u, v) = 4$ (seuls 3 sur 7 caractères sont identiques aux 2)

Propriétés. (qui disent que d est bien un distance)

- $d(u, v) = 0$ ssi $u = v$
- $d(u, v) = d(v, u)$
- inégalité triangulaire : $\forall u, v, w,$

$$d(u, v) \leq d(u, w) + d(w, v)$$

Preuve. $d(u, v) = \sum_{i=1}^{n} d(u_i, v_i)$, d'où $d(u, w) + d(w, v) = \sum_{i=1}^{n} (d(u_i, w_i) + d(w_i, v_i))$. On peut donc se focaliser sur un seul symbole à la fois : - si $u_i = v_i$ alors $d(u_i, v_i) = 0 \leq d(u_i, w_i) + d(w_i, v_i)$ - si $u_i \neq v_i$ alors $d(u_i, v_i) = 1$ et w_i est différent d'au moins un des deux. $d(u_i, w_i) + d(w_i, v_i) = 1 + 0$ ou $0 + 1$ ou $1 + 1$