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 $x = x_1 x_2 ... x_n$

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.

$$lg(x_1 x_2 ... x_n) = n$$
$$lg(\epsilon) = 0$$

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 $u = u_1u_2 \dots u_n$ et $v = v_1v_2 \dots v_p$ alors **le concaténé** de u et v noté simplement uv est le mot $u_1u_2 \dots u_nv_1v_2 \dots v_p$

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 $u = u_1u_2 \dots u_n$, $v = v_1v_2 \dots v_p$ 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. $d(u, v) = \sum\limits_{i=1}^{n} d(u_i, v_i)$, d'où $d(u, w) + d(w, v) = \sum\limits_{i=1}^{n} (d(u_i, w_i) + d(w_i, v_i))$. 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* $1 + 1$

2) Ordre lexicographique

Idée : comme l'odre du dictionnaire.

Soit A un alphabet quelcoqnue, $A = \{a_1, a_2, \ldots, a_k\}$

On dit que A est **ordonné** si on fixe un ordre $<$ sur les symboles, par exemple, $a_1 < a_2 < \cdots < a_k$

Soient u et v deux mots de A^*, u est **avant** v **dans l'odre lexicographiqe** (noté $u \leq_{\text{lex}} v$) si : - u est un préfixe de v OU - il existe un mot w et deux symboles $x < y$ tels que $wx \sqsubseteq u$ et $wy \sqsubseteq v$

Autrement dit si $u = u_1u_2 \ldots u_n$, $v = v_1v_2 \ldots v_p$: - $n \leq p$ et $u_1 = v_1, \cdots, u_n = v_n$ OU - $\exists k$ tel que $u_1 = v_1, \ldots u_k = v_k$ e $u_{k+1} < v_{k+1}$

Remarque. Si u et v sont de longueur 1, les ordres $\leq$ sur A et $\leq_{\text{lex}}$ sur A^* coïncident.

Exemple. Sur $B = \{0, 1\}$ avec $0 < 1$, rangeons tous les mots de longueur ≤ 3: 1. ϵ 2. 0, 1 3. 00, 01, 10, 11 4. 000, 001, 010, 011, 101, 110, 111

$$\epsilon \leq_{\text{lex}} 0 \leq_{\text{lex}} 00 \leq_{\text{lex}} 000 \leq_{\text{lex}} 001 \leq_{\text{lex}} 01 \leq_{\text{lex}} 010 \leq_{\text{lex}} 011$$
$$\leq_{\text{lex}} 1 \leq_{\text{lex}} 10 \leq_{\text{lex}} 100 \leq_{\text{lex}} 101 \leq_{\text{lex}} 11 \leq_{\text{lex}} 110 \leq_{\text{lex}} 111$$

Propriété. $\leq_{\text{lex}}$ est **un ordre total** : quels que soient u et $v \in A^*$ on a toujours $u <_{\text{lex}}$ ou $u >_{\text{lex}}$ ou $u = v$. Par exemple, ce n'est pas le cas pour $\sqsubseteq$.

> *Remarque.* L'odre lexicographique n'est pas commode à définir, par contre on peut écrire un **algorithme** pour décider si $u \leq_{\text{lex}} v$ *(cf. TD3)*

3) Ensembles et dénombrement

a) Notion d'ensemble, fini ou infini

Un ensemble E est une collection d'éléments sans ordre ni répétition. Si E est fini, on peut le décrire explicitement par exemple $\{a, b, c\}$ ou encore $\{c, a, b\}$. Cette notation est limitée : on écrit vite des ensembles comme $\{a, b, c, \cdots\}$, ce qui est ambigü.

En général on décrit plutôt {la forme générale} de éléments de l'ensemble ou {les propriétés}.

Exemple. $\{2k + 1 | k \in \mathbb{N}\} = \{1, 3, 4, 5, \dots\}$ $\{k \in \mathbb{N} | k \textit{ impair}\}$

Notations. - $x \in E$: l'élément x **appartient** à lensemble E - $A \subseteq B$: l'ensemble A est contenu / inclus dans B, aautrement dit tout élément de A appartient à B - $A \subset B$: inclusion stricte (si $A \subseteq B$ et $A \neq B$) - l'ensemble vide $\emptyset$ ne contient aucun élément : $\{\}$

Exemples. - $1 \in 0, 1$ - $-5 \notin \mathbb{N}$ - $-5 \in \mathbb{Z}$ - $\mathbb{N} \subset \mathbb{Z}$ - $\mathbb{Z} \subseteq \mathbb{Z}$ - $\{\epsilon\} \neq \emptyset$ - $\{-1, 1\} \subsetneq \mathbb{N}$

> *Remarque.* $x \in E$ si et seulement si $\{x\} \subseteq E$

Définition. Le **cardinal** d'un ensemble **fini** est le nombre d'éléments qui le composent, noté card E, $\#E$ ou $|E|$.

Exemples. - card $(\{a, b, \dots, z\}) = 26$ - card $(\{0, 1, \dots, 9\}) = 10$ - card $(\{\epsilon\}) = 1$ - card $(\emptyset) = 0$

> **Attention.** Par convention, dans tout ce cours, si on parle du cardianl d'un ensemble, celui-ci est fini

Proprietés. - Si $X \subseteq Y$ alors card $X \leq$ card Y - Si $X \subset Y$ alors card $X <$ card Y

(Notez le parallèle entre $\subseteq$, $\leq$ et $\sqsubseteq$ et entre $\subset$, $<$ et $\sqsubset$)

b) Opérations entre ensembles

Union et intersection

L'**union** de X et Y est l'ensemble des éléments présents dans X ou Y.

$$X \cup Y = \{x | x \in X \text{ ou } x \in Y\}$$

L'**intersection** de X et de Y est l'ensemble des éléments présents la fois dans X et dans Y.

$$X \cap Y = \{x | x \in X \text{ et } x \in y\}$$

> **Attention** : - X et Y sont **différents** si il existe un élément présent dans l'un mais pas dans l'autre - X et Y sont **disjoints** s'ils n'ont aucun élément commun : $X \cap Y = \emptyset$ (disjoint est "plus fort" que différent)

Exemples. - $\{0, 1\} \cup \{1, 2, 3\} = \{0, 1, 2, 3\}$ - $\{-0, 2, 4, \dots\} \cup \{1, 3, 5, \dots\} = \mathbb{N}$ - $\{0, 1\} \cap \{1, 2, 3\} = \{1\}$: ils sont différents mais pas disjoints - $\{0, 2, 4, \dots\} \cap \{1, 3, 5, \dots\} = \emptyset$: ils sont disjoints

Proprieté. card $(X \cup Y)$ + card $(X \cap Y)$ = card X+ card Y

Diagramme de Venn avec union et intersection

D'où card $X \cup Y \leq$ card $X+$ card Y et : on a l'égalité ssi X et Y sont disjoints. Dans ce cas on note l'union disjointe $X \cup +Y$ ou $X+Y$ alors $\boxed{\text{card}(X+Y) = \text{card}(X) + \text{card}(Y)}$

Différence

$$X \backslash Y = \{x \in X | x \notin Y\}$$
$$\text{card}(X \backslash Y) \leq \text{card} X$$

Partition

Définition. Soit un ensemble X (fini ou non). On appelle **partition finie** de X : n sous-ensembles $X_1, X_2, \ldots, X_n$ deux à deux disjoints (= **exclusivité**) et dont l'union forme X.

Autrement dit : tout élément de X fait partie d'un X_i et d'un seul

Exemple. $\mathbb{N}$ est partitionné en nombres pairs et nombres impairs.

$$X = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

$$X = \bigcup_{i=1}^{n} X_i \text{ et } \forall i, j \text{ disjoints } X_i \cap X_j = \emptyset$$

Exemples. - L'ensemble des élèves d'INFO3 rangés par année de naissance. - $\mathbb{Z}$ est partitionné en 5 parties ($\mathbb{Z}/5\mathbb{Z}$) : $\{5k | k \in \mathbb{Z}\}$, $\{5k+1 | k \in \mathbb{Z}\}$, $\{5k+2 | k \in \mathbb{Z}\}$, $\{5k+3 | k \in \mathbb{Z}\}$, $\{5k+4 | k \in \mathbb{Z}\}$

Complémentaire

Soient X et Y deux ensembles tels que $Y \subseteq X$. Alors $X - Y$ ou $C_x Y$ est le supplémentaire de Y dans X. $\{x \in X | x \notin Y\}$: cas particulier de différence

Alors : - Y et $X - Y$ forment 2 partitions de X :

$$X = (X - Y) + Y$$

- card $X =$ card $(X - Y) +$ card Y - $\boxed{\text{card } (X - Y) = \text{card } X + \text{card } Y}$ si $Y \subseteq X$