Algorithmique du texte - définitions de base

Annie Chateau [Sylvain Daudé]

IIIème millénaire

Introduction

Le texte est l'une des représentations de l'information la plus simple et naturelle.

Les données à traiter se présentent souvent comme une suite de caractères (fichiers textes par exemple).

Les *textes* sont l'objet central du traitement de texte sous toutes ses formes.

Mot

Definition (Alphabet)

Un alphabet Σ est un ensemble, non vide, fini ou infini de *symboles*.

Exemple

Pour $k \geq 2$, l'alphabet $\Sigma_k = \{0, 1, \ldots, k-1\}$.

Definition (Mot)

Un mot (ou chaîne de caractères) de l'alphabet Σ est une liste de symboles issus de Σ.

Exemple

3425 est un mot de l'alphabet Σ_6.

Mot

Remarques :

1. un mot peut être fini ou infini
2. un mot fini de longueur n peut être vu comme une application de $\{1, \ldots, n\}$ vers Σ.
3. un mot de longueur $n = 0$ est le mot *vide*, noté ϵ.

Definition (Ensemble des mots d'un alphabet)

L'ensemble des mots finis d'un alphabet Σ est noté Σ^*.
L'ensemble des mots finis *non vides* de l'alphabet Σ est noté Σ^+.

Mot

Exemple

Si $\Sigma = \{a, b\}$, alors $\Sigma^* = \{\epsilon, a, b, aa, ab, bb, aaa, aab, \ldots\}$.

Definition (Longueur d'un mot)

Si w est un mot fini, sa *longueur* (i.e. le nombre de symboles contenus dans w) est notée $|w|$.

Exemple

Le mot `cinq` est de longueur 4.

Remarque : $|\epsilon| = 0$.

Mot

Definition (Occurrence d'un symbole)

Si $a \in \Sigma$ et $w \in \sigma^*$, alors $|w|_a$ désigne le nombre d'occurrences du symbole a dans le mot w.

Exemple

$|\texttt{occurrence}|_{\texttt{r}} = 2$
$|\texttt{occurrence}|_{\texttt{c}} = 3\ldots$

Concaténation

Definition (Concaténation de deux mots)

La concaténation de deux mots finis w et x est la juxtaposition des symboles de w et des symboles de x, notée wx.

Exemple

Si $w = \texttt{titi}$ et $x = \texttt{tata}$ alors $wx = \texttt{tititata}$.

Remarque : La concaténation n'est pas *commutative* $wx \neq xw$, mais elle est *associative* : $(xy)z = x(yz)$.

Concaténation

Remarque : la concaténation est notée comme la multiplication, c'est-à-dire que w^n désigne $w \dots w$ (n fois).

L'ensemble Σ^* muni de la concaténation est un monoïde, avec comme élément identité le mot vide ϵ.

Facteur

Definition (Facteur)

On dit qu'un mot y est un *facteur* d'un mot w s'il existe des mots x et z tels que $w = xyz$.
Le mot x est un *préfixe* (resp. propre) du mot w s'il existe un mot y (resp. $\neq \epsilon$) tel que $w = xy$.
Le mot z est un *suffixe* (resp. propre) du mot w s'il existe un mot y (resp. $\neq \epsilon$) tel que $w = yz$.

Exemple

On considère $w =$ `barbapapa`. Le mot $x =$ `bar` est un préfixe de w, $y =$ papa est un suffixe de w, et $y = rbapa$ est un sous-mot de w.

Sous-séquence

Si $w = a_1 a_2 \dots a_n$ alors pour $i \in \{1, \dots n\}$, on définit : $w[i] = a_i$.

Si $i \in \{1, \dots n\}$ et $i - 1 \leq j \leq n$, on définit : $w[i \dots j] = a_i \dots a_j$.

Remarque : $w[i \dots i] = a_i$ et $w[i \dots i-1] = \epsilon$.

Palindromes

Definition (Mot miroir)

Si $w = a_1a_2 \ldots a_n$ est un mot fini de Σ, on appelle *mot miroir* de w, noté $\overline{w}$, le mot $a_na_{n-1} \ldots a_1$.

Exemple

Le mot `siort` est le mot miroir du mot `trois`.

Definition (Palindrome)

On appelle *mot palindrome* un mot w fini qui est identique à son mot miroir.

Exemple

Le mot `kayak` est un palindrome.

Période

Definition (Période)

Une *période* d'un mot w est un entier $0 < p \leq |w|$ tel que

$$\forall i \in \{1, \ldots, |x| - p\}\ x[i] = x[i + p].$$

On note $period(x)$ la plus petite période de x.

Exemple

Les périodes de aabaaabaa (de longueur 9) sont 4, 7, 8 et 9.

Bord

Definition (Bord)

Un bord du mot x est un sous-mot de x qui est à la fois un préfixe et un suffixe de x.

Exemple

Les bords du mot aabaaabaa sont aabaa, aa, a et ϵ.

Remarque : bord et période sont des notions duales.

Mots de Fibonacci

On considère l'alphabet $\Sigma = \{a, b\}$. On définit les mots de Fibonacci par :

$$\begin{cases} Fib_0 & = & \epsilon \\ Fib_1 & = & b \\ Fib_2 & = & a \\ Fib_n & = & Fib_{n-1}Fib_{n-2} \text{ pour tout } n \geq 2 \end{cases}$$

Mots de Fibonacci

Les longueurs des mots de Fibonacci sont exactement les termes de la suite de Fibonacci...

On définit g_n le préfixe de Fib_n de longueur $|Fib_n| - 2$. On montrera que g_n vérifie les conditions requises pour prouver l'optimalité du lemme de périodicité...

Période

Proposition

Soit x un mot non vide et p un entier tel que $0 < p \leq |x|$. Chacune des conditions équivalentes suivantes définit une période :

1. *x est un facteur d'un mot y^k avec $|y| = p$ et $k > 0$,*
2. *x peut être écrit sous la forme $(uv)^k u$ avec $|uv| = p$, v non vide et $k > 0$,*
3. *il existe des mots y, z et w tels que $x = yw = wz$ et $|y| = |z| = p$.*

Bord

Definition (Bord maximal)

Soit x un mot non vide. On note $Border(x)$ le plus grand bord propre de x (*i.e.* différent de x). On dit que x est *sans bord* si $Border(x) = \epsilon$.

Remarque : Le bord d'un mot est le plus long overlap non trivial quand on essaye de faire coïncider x avec lui-même. Puisque $Border(x)$ est strictement plus petit que x, si on itère le processus on finit par tomber sur le mot vide ϵ.

Bord

Proposition

Soit x un mot et $k \geq 0$ le plus petit entier tel que $Border^k(x)$ est vide. Alors

$$(x, Border(x), Border^2(x), \ldots, Border^k(x))$$

est la suite de tous les bords de x ordonnés par longueur décroissante, et

$$(|x| - |Border(x)|, |x| - |Border^2(x)|, \ldots, |x| - |Border^k(x)|)$$

est l'ensemble de toutes les périodes de x en ordre croissant.

Remarque : On a exactement $period(x) = |x| - |Border(x)|$.

Bord

Les résultats suivants sont utilisés dans les preuves combinatoires sur les mots.

Lemma (Périodicité faible)

Soient p et q deux périodes d'un mot x. Si $p + q \leq |x|$, alors $\mathrm{pgcd}(p, q)$ est aussi une période de x.

Lemma (Périodicité)

Soient p et q deux périodes d'un mot x. Si $p + q - \mathrm{pgcd}(p, q) \leq |x|$, alors $\mathrm{pgcd}(p, q)$ est aussi une période de x.