Codage de l'information

Rakotoarimalala Tsinjo Tony

Cours 3
IT University - 2022

Définition

- Un ***alphabet*** est un ensemble non vide de *caractères* ou *lettres* ou *symboles*.
- Un ***mot*** u sur un alphabet $\mathcal{A}$ est une suite finie de symboles de $\mathcal{A}$. la longueur de u, notée $|u|$, est le nombre de symboles composants u.
- On définit le ***mot vide***, noté ϵ le mot tel que $|\epsilon| = 0$
- On définit l'opération de ***concaténation*** de deux mots u et v d'un alphabet $\mathcal{A}$, notée $\boldsymbol{u}$ **.** $\boldsymbol{v}$ comme la construction d'un troisième mot w qui est une suite des symboles de u suivis de ceux de v.
  Alors
  $$|u.v| = |v.u| = |u| + |v|$$

Exemples

Considérons par exemple le cas de l'alphabet binaire (c'est-à-dire à deux symboles) : $\mathcal{A} = \{0, 1\}$.

- la suite $(S_n)_{n \geqslant 0}$ de mots de Fibonacci sur $\mathcal{A}$ est définie par $S_1 = 1$ et $S_2 = 0$ et pour $n > 2$:
$$S_n = S_{n-1}S_{n-2}$$
- On donc $S_3 = 01, S_4 = 010, S_5 = 01001, S_6 = 01001010, \dots$
- Il est clair donc que $|S_n| = |S_{n-1}| + |S_{n-2}|$.
Ici par exemple $|S_5| = |S_4| + |S_3| = 5$
- On peut définir un mot infini à partir de la suite de Fibonacci et ce dernier comme donc par $\underbrace{0100101001001}_{S_7}\dots$

Propriétés

- la concaténation :
  - ∗ est associative

$$(u.v).w = u.(v.w) = u.v.w = uvw$$

  - ∗ admet comme élément neutre ϵ

$$u.\epsilon = \epsilon.u = u$$

  - ∗ est non commutative. En général

$$u.v \neq v.u$$

- u^n est la puissance n^{ieme} de u

$$u^n = \underbrace{uuu \dots u}_{n \ fois}$$

Spécialement $u^0 = \epsilon$

Définition

- On dit que v est ***préfixe*** (*resp* ***suffixe***) de u s'il existe un mot w (éventuellement vide) tel que $u = v.w$ (*resp* $u = w.v$)
  Si $u = 01001$ alors l'ensemble de préfixe de u noté $Pref(u)$ est

$$Pref(u) = \{\epsilon, 0, 01, 010, 0100, 01001\}$$

- On définit par $\mathcal{A}^n$ l'ensemble de tous les mots de longueur n de $\mathcal{A}$. Et $\mathcal{A}^*$ l'ensemble de tous les mots de $\mathcal{A}$. Donc

$$\mathcal{A}^* = \bigcup_{n \geqslant 0} \mathcal{A}^n$$

- On appelle ***langage*** sur un alphabet $\mathcal{A}$ tout sous-ensemble de $\mathcal{A}^*$

Exemples

Restons sur l'alphabet binaire. Donnons quelques exemples de langages.

- L défini par l'ensemble de mots binaires sans deux 1 consécutifs.

$$L = \{\epsilon, 0, 1, 00, 01, 10, 000, 001, 010, 100, 101, 0000, 0001, 0010, \ldots\}$$

- L_1 défini par $\{0, 01\}^*$ est l'ensemble des mots obtenus par la concaténation de 0 et/ou 01 (autant de fois qu'on veut)

$$L_1 = \{\epsilon, 0, 00, 01, 001, 010, 0000, 0001, 0101, \ldots\}$$

- L_2 défini par l'ensemble des mots de la forme $\{0^n 1^n, n \geqslant 0\}$

$$L_2 = \{\epsilon, 01, 0011, 000111, \ldots\}$$

Factorisation d'un mot dans un langage

- Soient L un langage sur l'alphabet $\mathcal{A}$ et u un mot de $\mathcal{A}$. On dit que u est factorisable sur L s'il existe une suite de mots $u_i \in L, i \in 1, \ldots p$ (pour un certain p) tel que

$$u = u_1 u_2 \ldots u_p$$

- On dit alors que $u_1, u_2, \ldots u_p$ forment une factorisation de u dans L
- La factorisation peut ne pas être unique, et un mot peut ne pas être factorisable dans L

Exemple

Si $L = \{0, 10, 01\}$.

- Le mot $u = 00110$ admet comme factorisation $u = 0.01.10$
- Le mot $v = 010$ admet comme factorisation $v = 0.10$ et $v = 01.0$
- Le mot $w = 1100$ n'admet pas de factorisation dans L

CODES

Définition

Un ***code*** est un langage dans lequel tous les mots ne possèdent au plus qu'une seule factorisation.

Exercices

- Montrer que les langages suivants ne sont pas des codes

$$L = \{0, 10, 01\}, L_1 \text{ un langage contenant le mot vide}$$

- Montrer que les langages suivants sont des codes:

$$L_0 = \mathcal{A}, L = \{0, 01\}, L_1 = \{10^n, n \in \mathbb{N}\}$$

Codages

Définition

Soient $\mathcal{S}$ et $\mathcal{A}$ deux alphabets. Un codage est un application μ

$$\mu : \mathcal{S}^* \to \mathcal{A}^*$$

tel que:

i. μ est injective (deux mots différents ont des codages différents)

$$\forall u, v \in \mathcal{S}^*, u \neq v \Rightarrow \mu(u) \neq \mu(v)$$

ii. μ est compatible avec l'opération concaténation

$$\forall u, v \in \mathcal{S}^*, \mu(u.v) = \mu(u).\mu(v)$$

iii. c transforme le mot vide en lui même

$$\mu(\epsilon) = \epsilon$$

Remarques

- Si μ est un codage, et d'après la compatibilité d'un codage à la concaténation, on peut déduire que μ est complétement caracterisé par mots associés (les images) aux symboles de $\mathcal{S}$.
- On appelle alors code associé C à μ l'ensemble défini par

$$C = \{\mu(x) | x \in \mathcal{S}\}$$

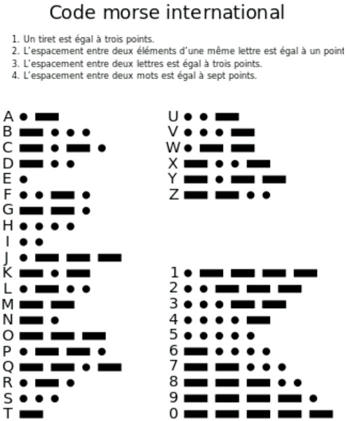
- L'injectivité de μ est exigé pour pouvoir décoder une information codée

Exemple 1: Codage Morse

Le codage Morse est un codage $\nu : \mathcal{A}^* \to M^*$ avec

- $\mathcal{A}$ est l'alphabet composé de l'alphabet latin, des chiffres, signes de ponctuation et de symboles
- $$M = \{\bullet, -, \_\}$$
- "$\bullet$" est appelé ***ti*** et "$-$" ***taah*** et "$\_$" est un caractère d'espacement.

Code morse pour les lettres



- Ici les blancs représentent l'espacement "_"
- Le codage de ***ITU*** est donc

$$\underbrace{\bullet\ \bullet}_{I}\ \underbrace{\qquad}_{3\ espaces}\ \underbrace{-}_{T}\quad \underbrace{\bullet\ \ \bullet\ \ -}_{U}$$

- 1 espace entre deux symboles de M
- 3 espaces pour séparer deux lettres codées
- 5 espaces pour séparer deux mots codés

Exemple 2: Codage ASCII

Le codage ASCII ou American Standard Code for Information Interchange est un codage $\phi : \mathcal{A}^* \rightarrow \mathbb{B}_7^*$ avec

- $\mathcal{A}$ est l'alphabet composé de 128 caractères dont 95 imprimables : les chiffres arabes de 0 à 9, les lettres minuscules et capitales de A à Z, et des symboles mathématiques et de ponctuation.
- $\mathbb{B}_7 = \{0, 1\}^7$: nombre binaire de longueur 7 donc de 0000000 à 1111111
- Donc pour coder ITU on a (voir extrait du code ASCII ci-dessous):

$$\underbrace{1001001}_{I}\,\underbrace{1010100}_{T}\,\underbrace{1010101}_{U}$$

Extrait Code ASCII

1001001	I	Lettre latine capitale I
1001010	J	Lettre latine capitale J
1001011	K	Lettre latine capitale K
1001100	L	Lettre latine capitale L
1001101	M	Lettre latine capitale M
1001110	N	Lettre latine capitale N
1001111	O	Lettre latine capitale O
1010000	P	Lettre latine capitale P
1010001	Q	Lettre latine capitale Q
1010010	R	Lettre latine capitale R
1010011	S	Lettre latine capitale S
1010100	T	Lettre latine capitale T
1010101	U	Lettre latine capitale U

Autres exemples de codages

- Le codage **Baudot**
- Le codage **Manchester**
- Le codage **Miller**
- Le codage **ISO-8859**
- Le codage **UTF-8**

Exercices

1. Codage binaire de l'alphabet $\mathcal{S}$
   On veut coder chaque lettre de $\mathcal{S}$ par un mot de $\mathcal{A}^*$ en utilisant un codage de longueur fixe avec $\mathcal{A} = \{0, 1\}$.
   - Soit $\mathcal{S}$ l'alphabet latin (26 lettres).
     Quelle est la longueur minimale des mots du code ?
   - Soit $\mathcal{S}$ maintenant un alphabet quelconque avec n lettres.
     Quelle est la longueur minimale des mots du code ?
2. Nombre de mots dans l'alphabet binaire
   - Combien y a-t-il de mots de 5 lettres commençant par 0 et terminant par 1 ?
   - Combien y a-t-il de mots de 10 lettres contenant au moins trois 0 et deux 1 ?

Exercices (suite)

1. Montrer que tout langage préfixe autre que $\{\epsilon\}$ est un code. (Un langage préfixe est un langage tel que aucun de mot de ce langage n'est préfixe d'un autre mot de ce langage).
2. Montrer que $L = \{00, 01, 110, 001\}$ est un code.