Languages et expressions régulières

Ambre Le Berre

2025/2026

MP option info

Agenda

1. Vocabulaire
2. Expression régulière, languages réguliers

Vocabulaire

Alphabet et mots

- Un alphabet est un ensemble Σ de lettres. Par exemple, $\Sigma = \{a, b\}$ est un alphabet.
-
-
-
-

Alphabet et mots

- Un alphabet est un ensemble Σ de lettres. Par exemple, $\Sigma = \{a, b\}$ est un alphabet.
- Un mot sur un alphabet est une suite finie de lettres de cet alphabet. Par exemple, *abba* est un mot sur Σ.
-
-
-

Alphabet et mots

- Un alphabet est un ensemble Σ de lettres. Par exemple, $\Sigma = \{a, b\}$ est un alphabet.
- Un mot sur un alphabet est une suite finie de lettres de cet alphabet. Par exemple, *abba* est un mot sur Σ.
- On note ε le mot vide, qui ne contient aucune lettre.
-
-

Alphabet et mots

- Un alphabet est un ensemble Σ de lettres. Par exemple, $\Sigma = \{a, b\}$ est un alphabet.
- Un mot sur un alphabet est une suite finie de lettres de cet alphabet. Par exemple, $abba$ est un mot sur Σ.
- On note ε le mot vide, qui ne contient aucune lettre.
- Pour deux mots u et v, on note uv leur concaténation.
- On note $|u| \in \mathbb{N}$ la longueur d'un mot.

Vocabulaires sur les mots

Soit $m = m_1 m_2 ... m_n \in \Sigma^*$ un mot.

- un préfixe de m est un mot $m_1 ... m_i$ avec $0 \leq i \leq n$.
-
-
-
-

Vocabulaires sur les mots

Soit $m = m_1 m_2 ... m_n \in \Sigma^*$ un mot.

- un préfixe de m est un mot $m_1 ... m_i$ avec $0 \leq i \leq n$.
- un suffixe de m est un mot $m_{i+1} ... m_n$ avec $0 \leq i \leq n$.
-
-
-

Vocabulaires sur les mots

Soit $m = m_1 m_2 ... m_n \in \Sigma^*$ un mot.

- un préfixe de m est un mot $m_1 ... m_i$ avec $0 \leq i \leq n$.
- un suffixe de m est un mot $m_{i+1} ... m_n$ avec $0 \leq i \leq n$.
- un facteur de m est un mot $m_i ... m_j$ avec $0 \leq i, j \leq n$.
-
-

Vocabulaires sur les mots

Soit $m = m_1 m_2 ... m_n \in \Sigma^*$ un mot.

- un préfixe de m est un mot $m_1 ... m_i$ avec $0 \leq i \leq n$.
- un suffixe de m est un mot $m_{i+1} ... m_n$ avec $0 \leq i \leq n$.
- un facteur de m est un mot $m_i ... m_j$ avec $0 \leq i, j \leq n$.
- un sous-mot de m est un mot $m_{i_1} m_{i_2} ... m_{i_k}$ avec $i_1 < i_2 < ... < i_k \in [\![1, n]\!]$
-

Vocabulaires sur les mots

Soit $m = m_1 m_2 ... m_n \in \Sigma^*$ un mot.

- un préfixe de m est un mot $m_1 ... m_i$ avec $0 \leq i \leq n$.
- un suffixe de m est un mot $m_{i+1} ... m_n$ avec $0 \leq i \leq n$.
- un facteur de m est un mot $m_i ... m_j$ avec $0 \leq i, j \leq n$.
- un sous-mot de m est un mot $m_{i_1} m_{i_2} ... m_{i_k}$ avec $i_1 < i_2 < ... < i_k \in [\![1, n]\!]$
- ε est préfixe, suffixe, facteur et sous-mot de n'importe quel mot.

Languages

- Un language est un ensemble $L \subseteq \Sigma^*$ de mots sur un alphabet
-

Languages

- Un language est un ensemble $L \subseteq \Sigma^*$ de mots sur un alphabet
- On note $\emptyset$ le language vide (à ne pas confondre avec $\{\varepsilon\}$)

Expression régulière, languages réguliers

Opérations régulières sur les languages

On défini plusieurs opérations dites régulières sur les languages. Soit L_1, L_2 deux languages sur un alphabet Σ.

- $L_1 \cup L_2 = \{u \mid u \in L_1 \text{ ou } u \in L_2\}$ est l'union des languages.
-
-
-

Opérations régulières sur les languages

On défini plusieurs opérations dites régulières sur les languages. Soit L_1, L_2 deux languages sur un alphabet Σ.

- $L_1 \cup L_2 = \{u \mid u \in L_1 \text{ ou } u \in L_2\}$ est l'union des languages.
- $L_1 L_2 = \{uv \mid u \in L_1, v \in L_2\}$ est la concaténation des languages.
-
-

Opérations régulières sur les languages

On défini plusieurs opérations dites régulières sur les languages. Soit L_1, L_2 deux languages sur un alphabet Σ.

- $L_1 \cup L_2 = \{u \mid u \in L_1 \text{ ou } u \in L_2\}$ est l'union des languages.
- $L_1 L_2 = \{uv \mid u \in L_1, v \in L_2\}$ est la concaténation des languages.
- $L_1^* = \{u_1 u_2 ... u_n \mid n \in \mathbb{N}, u_1 ... u_n \in L_1\}$ l'étoile de Kleene d'un language (ou juste étoile).
-

Opérations régulières sur les languages

On défini plusieurs opérations dites régulières sur les languages. Soit L_1, L_2 deux languages sur un alphabet Σ.

- $L_1 \cup L_2 = \{u \mid u \in L_1 \text{ ou } u \in L_2\}$ est l'union des languages.
- $L_1 L_2 = \{uv \mid u \in L_1, v \in L_2\}$ est la concaténation des languages.
- $L_1^* = \{u_1 u_2 ... u_n \mid n \in \mathbb{N}, u_1 ... u_n \in L_1\}$ l'étoile de Kleene d'un language (ou juste étoile).
- L'intersection et la soustraction ensembliste ne sont pas des opérations régulières !

Expressions régulières

On utilise expressions régulières pour définir une partie des languages.

Définition inductive

- Cas de bases :
  - ε est une expression régulière représentant le language $\{\varepsilon\}$
  - si $a \in \Sigma$, a est une expression régulière représentant le language $\{a\}$.
-

Expressions régulières

On utilise expressions régulières pour définir une partie des languages.

Définition inductive

- Cas de bases :
  - ε est une expression régulière représentant le language $\{\varepsilon\}$
  - si $a \in \Sigma$, a est une expression régulière représentant le language $\{a\}$.
- Cas inductifs : si e_1, e_2 sont des expressions regulières qui représentent des languages L_1 et L_2, alors
  - e_1^* représente le language L_1^*
  - $e_1 \mid e_2$ représente le language $L_1 \cup L_2$
  - $e_1 e_2$ représente le language $L_1 L_2$

Expressions régulières : raccourcis

On utilise souvent les syntaxes suivantes par facilité :

- e^+ pour désigner ee^*
- e^n pour désigner $eee...e$ (n fois)
- $e?$ pour désigner $e \mid \varepsilon$

Quelques exemples

Décrivez, en quelques mots, les languages décrits par les expressions régulières suivantes sur l'alphabet $\Sigma = \{a, b\}$:

- $(a|b)^*$
- $(a|b)^+$
- $(b^*ab^*)^5$
- $(ab \mid b)^*$

Dans l’autre sens

Donnez des expressions régulières pour les languages suivants :

- les mots qui contiennent “abba”
- les mots qui contiennent d’abord que des b (au moins 1), puis que des a (au moins 1)
- les mots dont la longueur est multiple de 3
- les mots dont “abba” est un sous-mot

Language régulier / rationnel

Un language régulier ou rationnel est un language qui peut être représenté par une expression régulière.

Q : Est ce que tous les languages sont réguliers ?

Language régulier / rationnel

Un language régulier ou rationnel est un language qui peut être représenté par une expression régulière.

Q : Est ce que tous les languages sont réguliers ? Non !

Un exemple de language non régulier

Le language $L_0 = \{a^n b^n \mid n \in \mathbb{N}\}$ n'est pas régulier.

Un exemple de language non régulier

Le language $L_0 = \{a^n b^n \mid n \in \mathbb{N}\}$ n'est pas régulier.

Intuition : les expressions régulières ont toujours une "mémoire" finie. Entre autre, elle ne peuvent pas "compter" n "a" pour attendre ensuite le même nombre de "b".

Lemme de l'étoile (première version)

Soit L un language régulier. Alors il existe N, tel que pour tout mot $u \in L$ avec $|u| \geq N$, il existe une décomposition $u = xyz$, telle que :

- $|xy| \leq N$
- $y \neq \varepsilon$
- $xy^*z \subseteq L$

Montrer qu'un language n'est pas rationnel

Montrons que L_0 n'est pas rationnel par l'absurde : si L_0 est rationnel, alors, soit $N \in \mathbb{N}$ suffisemment grand pour appliquer le lemme de l'étoile.

On prend le mot $u = a^{N+1}b^{N+1} \in L_0$. Supposons qu'il existe $xyz = u$ tel que

$$|xy| \leq N, y \neq \varepsilon \text{ et } xy^*z \subseteq L$$

Montrer qu'un language n'est pas rationnel

Montrons que L_0 n'est pas rationnel par l'absurde : si L_0 est rationnel, alors, soit $N \in \mathbb{N}$ suffisemment grand pour appliquer le lemme de l'étoile.

On prend le mot $u = a^{N+1}b^{N+1} \in L_0$. Supposons qu'il existe $xyz = u$ tel que

$$|xy| \leq N, y \neq \varepsilon \text{ et } xy^*z \subseteq L$$

Alors, comme $|xy| \leq N$, x et y sont de la forme a^p et a^k, et z est de la forme $a^l b^{p+k+l}$.

Donc, $a^p\left(a^k\right)^* a^l b^{p+k+l} \subseteq L_0$. Entre autre, $a^{p+2k+l}b^{p+k+l} \in L_0$. Absurde !

Exercice

Montrer que le language $L_p = \{u_1 u_2 ... u_k \mid u_1 u_2 ... u_k = u_k ... u_2 u_1\}$ le language des palindromes, n'est pas rationnel.

Exercice

Montrer que le language $L_p = \{u_1 u_2 ... u_k \mid u_1 u_2 ... u_k = u_k ... u_2 u_1\}$ le language des palindromes, n'est pas rationnel.

Idée : considérer le mot $a^{N+1} b a^{N+1}$

Expressions régulière en pratique

La plupart des éditeurs de texte vous permettent de rechercher des expressions régulières :

- remplacer `(\n\s*= .*)\n` par `$1.\n`

Expressions régulière en pratique

Syntaxe :

- `.` → n'importe que caractère (sauf `\n`)
- `\s` → n'importe quel espace / tab / etc
- `\w` → n'importe quelle lettre
- `[abfg ]` → un des caractère parmis "a" "b" "f" "g" et " ".
- `[^ab ]` → tout sauf un des caractères parmis "a" "b" " ".
- `*`, `+`, `?`, `|` → comme ce qu'on a vous

Expressions régulière en pratique

Remplacement :

- `$0` → tout
- `$1` → la première partie entre parenthèse
- `$2` → la deuxième, etc.