Automates et analyse lexicale
Le lemme de l'étoile

I) Lemme de l'étoile

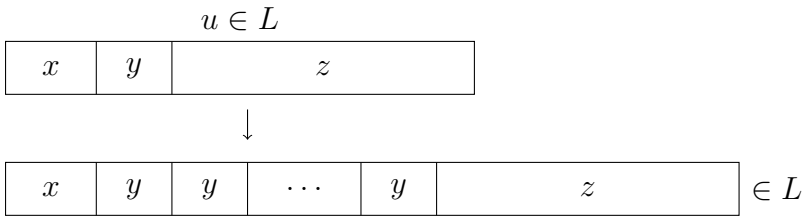
Le lemme de l'étoile (ou lemme d'itération, ou de la pompe, ou de pompage, etc.) est un outil puissant et souvent le seul à notre disposition pour montrer qu'un langage n'est pas reconnaissable. En même temps, ce sujet est souvent la « bête noire » de la plupart des étudiants qui suivent leur premier cours sur la théorie des automates finis et des langages reconnaissables/rationnels.
Il suffit de regarder son énoncé (que pour l'instant on admirera seulement pour la « beauté » de sa structure logique) pour en comprendre la raison :

> Si L est un langage rationnel alors $\exists N \in \mathbb{N}$ tel que $\forall u \in L$ avec $|u| > N, \exists$ un découpage $u = xyz$ avec $y \neq \varepsilon$ et $|xy| <= N$ tel que $\forall k \in \mathbb{N}, xy^k z \in L$

Il semblerait que la difficulté à comprendre cet énoncé vienne des ses quatre quantificateurs logiques imbriqués ($\exists \dots$ tel que $\forall \dots$ $\exists \dots$ tel que $\forall \dots$), alors que généralement notre petit cerveau humain ne peut vraiment comprendre ce qui se passe (ou se dit) que quand il y en a au maximum trois.
Ce document vise à fournir quelques éléments pour mieux comprendre aussi bien l'énoncé du lemme que son application.

II) La notion de facteur itérant

Comprendre la notion de *facteur itérant* peut faciliter la compréhension du lemme de l'étoile, surtout pour ceux qui n'aiment pas trop les quantificateurs.
Definition. Soit L un langage, u un mot de L. Un facteur non vide y de u est dit *itérant* si le nouveau mot obtenu en remplaçant y à l'intérieur de u par un nombre quelconque de ses copies (et donc en l'itérant) est toujours un mot de L.



Si on a donc un mot $u = xyz \in L$ et si le facteur y est itérant alors on doit avoir $u' = xy^k z \in L$ pour tout entier k (y compris $k = 0$, ce qui revient à effacer complètement y).

Per exemple, si $L = a^*b^*$ et $u = a^5b^4$ alors tout facteur de u qui ne contient que des a ou qui ne contient que des b est itérant.

En revanche, le facteur $y = ab$ n'est pas un facteur itérant de u : en effet, on a $w = xyz$ avec $x = a^4$ et $z = b^3$. Mais si on prend $k = 2$, le mot

$$xy^2z = aaaa\mathbf{abab}bbb$$

n'est pas dans $L = a^*b^*$. On peut dire que pour $k = 2$ on est « sorti » du langage. En fait on « sortirait » du langage pour tout entier $k > 1$ (alors que pour $k = 0$ on « reste » dans le langage : $xy^0z = aaaabbb \in L = a^*b^*$).

III) L'énoncé du lemme de l'étoile

Le lemme de l'étoile peut être énoncé comme suit :

> Si L est un langage rationnel alors tout mot de L
> suffisamment long possède au moins un facteur itérant.

Ou, si l'on veut préciser le « suffisamment long » :

> Si L est un langage rationnel alors il existe un entier N tel que tout mot $u \in L$
> de longueur supérieure à N possède un facteur itérant.

De plus, la variante du lemme de l'étoile que l'on a présentée dans ce cours précise que ce facteur itérant doit être trouvé parmi les N premières lettres du mot u, ce qui nous facilitera la tâche par la suite.

IV) À quoi sert le lemme de l'étoile

Appelons P la propriété d'un langage :
« il existe un entier N tel que tout mot $u \in L$ de longueur supérieure à N possède un facteur itérant ».
Il est important de comprendre que la propriété P est une condition *nécessaire* pour que L soit rationnel, mais elle n'est pas *suffisante* (le lemme dit "si", mais il ne dit pas "seulement si"). En effet il existe des langages qui satisfont la propriété P mais qui ne sont pas rationnels.
Mais puisque la condition du Lemme est nécessaire, on déduit que si un langage ne satisfait pas la propriété P alors il n'est pas rationnel.

> On se sert donc du lemme de l'étoile uniquement pour montrer qu'un langage **n'est pas** rationnel.

Pour cela, il faudra donc montrer que L ne satisfait pas la propriété P.

V) La mise en œuvre

Pour montrer qu'un langage L ne satisfait pas P, il faut montrer qu'un tel N n'existe pas, et donc que quelque soit N, il existe un mot $u \in L$ de longueur supérieure à N qui ne possède aucun facteur itérant parmi ses N premiers caractères.
Il faudra donc montrer qu'aucun facteur y se trouvant parmi ses N premiers caractères de ce mot u n'est itérant.
Ce qui revient à montrer que pour tout facteur y se trouvant parmi ses N premiers caractères de $u = xyz$ on peut trouver un entier k qui nous fait sortir du langage, c'est-à-dire tel que $xy^kz \notin L$.
En récapitulant il faudra montrer que :

> Quel que soit un entier N, on peut exhiber un mot u de L tel que
> pour tout découpage $u = xyz$ avec $y \neq \varepsilon$ et $|xy| \leqslant N$,
> on peut exhiber un entier k tel que $xy^kz \notin L$.

Ou, pour ceux qui aiment les quantificateurs :

$$\forall N \in \mathbb{N}, \exists u \in L \text{ avec } |u| > N, \text{ tel que } \forall u = xyz \text{ avec } |y| > 0 \text{ et } |xy| \leqslant N,$$

$$\exists k \in \mathbb{N} \text{ tel que } xy^kz \notin L$$

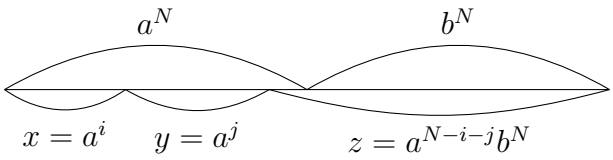
(ce qui est exactement la négation de la propriété dans l'énoncé du lemme de l'étoile donné à la première page).

> **Note.**
> Les seuls éléments de la preuve que l'on peut choisir (et en fait qu'on doit choisir, puisqu'on doit les exhiber) sont ceux qui, selon l'énoncé, sont censés exister et donc le mot u et l'entier k.
> En revanche on ne peut pas choisir l'entier N ni le facteur y (ou de manière équivalente le découpage $u = xyz$) puisqu'on doit faire une démostration valide pour tous les choix possibles de N et y.
> On pourra (et on devra) quand même tenir compte de la contrainte $|xy| \leqslant N$ qui nous permet de limiter les choix pour y uniquement aux facteurs se trouvant parmi les N premiers caractères.

Toutes les démonstrations qu'un langage n'est pas rationnel en utilisant le lemme de l'étoile ont exactement la même structure. Le même schéma pourra donc être appliqué pour toute preuve, à condition de changer (en latin *mutatis mutandis*) les choix opérés pour u et k et éventuellement la justification du fait que le mot xy^kz obtenu n'est pas dans L.
Si on comprend ce schéma, il n'y aura en principe que ces choses à modifier !
D'ailleurs, pour la rédaction des exemples, je me suis servi de la fonction copier-coller et ai modifié seulement les parties nécessaires à partir du texte du premier. Les parties en italique sont des commentaires spécifiques à chaque exemple mais vous pourrez remarquer que la structure reste la même.

a) Un premier exemple : $L_0 = \{a^n b^n | n \geqslant 0\}$

Soit $L_0 = \{a^n b^n | n \geqslant 0\}$ on veut montrer que L_0 n'est pas rationnel.
Soit N un entier quelconque.
Maintenant nous devons choisir un mot de L de longueur supérieure à N. C'est là qu'il faut y mettre du sien. Normalement on choisit le mot de L qui vient naturellement à l'esprit (voir plus loin la section « Sur le choix du mot u »).
Soit le mot $u = a^N b^N$.
Maintenant nous devons étudier tout découpage $u = xyz$ avec $y \neq \varepsilon$ et $|xy| \leqslant N$. Faire un dessin peut aider à ce moment de la preuve car il permet de déduire les formes possibles pour chacun des facteur x, y, z.



Tout découpage $u = xyz$ avec $y \neq \varepsilon$ et $|xy| \leqslant N$ est de la forme :
- $x = a^i$ pour un certain i ;
- $y = a^j$ pour un certain $j \neq 0$;
- $z = a^{N-i-j}b^N$.

Il faut maintenant choisir un entier k qui nous fait sortir du langage, là aussi normalement il ne faut pas chercher trop loin (voir la section « Sur le choix de l'entier k »).
Si on choisit $k = 2$ on a $xy^2z = a^{N+j}b^N \notin L$, parce que la plage des a n'a pas la même longueur que la plage des b ($j > 0$, donc elle est plus longue).
Donc le langage n'est pas rationnel.
Note. Dans ce cas, n'importe quel choix de k (sauf $k = 1$!) nous aurait fait sortir du langage. Par exemple, si on choisit $k = 0$ on a $xy^0z = a^{N-j}b^N \notin L$, parce que la plage des a n'a pas la même longueur que la plage des b ($j > 0$, donc elle est plus courte).

b) Sur le choix du mot u

Contraintes à respecter absolument :
- u doit être pris dans L ;
- u doit avoir une longueur supérieure à N.

Normalement, il ne faut pas chercher loin et il suffit de prendre un mot de longueur supérieure à N et qui tout naturellement appartient à L.
Evidemment, ce mot devra être choisi judicieusement, dans le sens où le but sera de montrer qu'aucun des facteurs se trouvant parmi ses N premiers caractères n'est itérant.
Généralement on choisit un mot pour lequel l'appartenance à L est « fragile à casser » (voir plus loin par exemple le choix du mot $u = a^N b^{N+1}$ pour le langage $L = \{a^n b^m | m > n \geqslant 0\}$).

c) Sur le choix de l'entier k

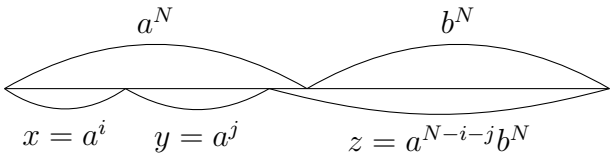
Là aussi généralement il ne faut pas chercher loin.
Une chose est sûre : le choix $k = 1$ est toujours mauvais ! ! ! En effet $xy^1z = xyz = u$ et u est choisi dans L, alors que notre but est de trouver un mot xy^kz qui n'est pas dans L !

Souvent les choix $k = 0$ (qui revient à effacer le facteur y) ou $k = 2$, qui revient à ajouter une (et non deux !) copie de y permettent de trouver un mot au dehors de L. Dans certains cas cependant la recherche d'un k qui fait sortir du langage est un peu plus compliquée (voir plus loin l'exemple du langage $L = \{a^p | p \text{ premier}\}$.)

VI) Exemples

Exemple 2.

Soit $L = \{w \in \{a, b\}^* \text{ tels que } |w]_a = |w|_b\}$ (w a autant de a que de b). On veut montrer que L n'est pas rationnel.
Soit N un entier quelconque.
Dans ce cas, exactement les mêmes choix que pour L_0 conviennent.
Soit le mot $u = a^N b^N$.

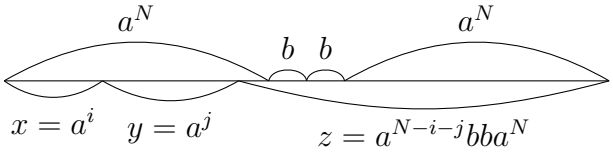


Tout découpage $u = xyz$ avec $y \neq \varepsilon$ et $|xy| \leqslant N$ est de la forme :
- $x = a^i$ pour un certain i ;
- $y = a^j$ pour un certain $j \neq 0$;
- $z = a^{N-i-j}b^N$.

Si on choisit $k = 2$ on a $xy^2z = a^{N+j}b^N \notin L$, parce que ce mot n'a pas autant de a que de b.
Donc le langage n'est pas rationnel.
Note. Dans ce cas aussi, n'importe quel choix de $k \neq 1$ nous aurait fait sortir du langage.

Exemple 3.

Soit $L = \{w \in \{a, b\}^* \text{ tels que } w = v\tilde{v} \text{ avec } v \in \{a, b\}^*\}$ (w est un palindrome de longueur paire). On veut montrer que L n'est pas rationnel.
Soit N un entier quelconque.
Dans ce cas on fait un choix judicieux avec des b qui marquent le centre du mot.
Soit le mot $u = a^N bba^N$.

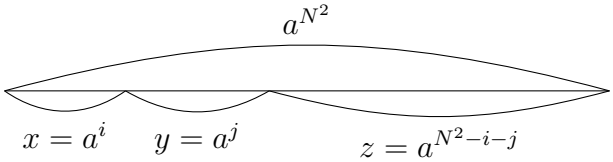


Tout découpage $u = xyz$ avec $y \neq \varepsilon$ et $|xy| \leqslant N$ est de la forme :
- $x = a^i$ pour un certain i ;
- $y = a^j$ pour un certain $j \neq 0$;
- $z = a^{N-i-j}bba^N$.

Si on choisit $k = 2$ on a $xy^2z = a^{N+j}bba^N \notin L$, parce que ce mot n'est pas un palindrome de longueur paire.
Donc le langage n'est pas rationnel.
Note. Dans ce cas aussi, n'importe quel choix de $k \neq 1$ nous aurait fait sortir du langage.

Exemple 4.

Soit $L = \{a^{n^2} \mid n \geqslant 0\}$. On veut montrer que L n'est pas rationnel.
Soit N un entier quelconque.
Soit le mot $u = a^{N^2}$.



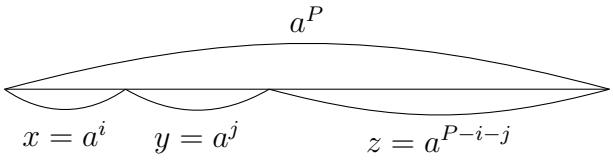
Tout découpage $u = xyz$ avec $y \neq \varepsilon$ et $|xy| < N$ est de la forme :
- $x = a^i$ pour un certain i ;
- $y = a^j$ pour un certain $j \neq 0$;
- $z = a^{N^2-i-j}$.

Si on choisit $k = 2$ on a $xy^2z = a^{N^2+j}$.

Dans ce cas ce n'est pas tout à fait évident que le mot obtenu soit en dehors du langage L. Il faut donner un petit argument pour le justifier que $N^2 + j$ n'est pas un carré.
Puisque $j > 0$, on a $N^2 + j > N^2$. Par ailleurs puisque $j \leqslant N$ on a $N^2 + j \leqslant N^2 + N < N^2 + 2N + 1 = (N + 1)^2$.
Puisque $N^2 + j$ est strictement compris entre N^2 et $(N + 1)^2$ il ne peut pas être un carré.
Donc $a^{N^2+j} \notin L$ et le langage n'est pas rationnel.
Note. Dans ce cas le choix $k = 0$ nous aurait aussi fait sortir du langage.

Exemple 5.

Soit $L = \{a^p \mid p \text{ nombre premier}\}$. On veut montrer que L n'est pas rationnel.
Soit N un entier quelconque.
Dans ce cas on fait un choix judicieux pour être sûr de choisir un mot u qui est bien dans L.
Soit le mot $u = a^P$, ou P est le plus petit nombre premier plus grand de N (P existe certainement parce que les nombres premiers sont en nombre infini). Ainsi $u \in L$ et $|u| > N$.



Tout découpage $u = xyz$ avec $y \neq \varepsilon$ et $|xy| \leqslant N$ est de la forme :
- $x = a^i$ pour un certain i ;

– $y = a^j$ pour un certain $j \neq 0$;
– $z = a^{P-i-j}$.

Voici un exemple où le choix de l'entier k est un peu plus compliqué. Il faut le choisir de sorte qu'en itérant k fois le facteur y, on obtienne en mot de longueur un nombre non premier. Dans ce cas le choix de k dépend de P (et donc de N).

Si on choisit $k = P + 1$ on a $xy^{P+1}z = a^i a^{Pj} a^j a^{P-i-j} = a^{Pj+P} = a^{P(j+1)}$.

Pour justifier que le mot obtenu est en dehors du langage L, il faut montrer que $P(j + 1)$ n'est pas un nombre premier.

Puisque $j > 0$, on a $j+1 \geqslant 2$, donc $P(j+1)$ est bien le produit de deux nombres supérieurs à 1 et donc il n'est pas premier.

Donc $a^{P(j+1)} \notin L$ et le langage n'est pas rationnel.