# Recherche Textuelle - ALGORITHME DE BOYER-MOORE-HORSPOOL

*L'algorithme de Boyer-Moore-Horspool ou Horspool est un algorithme de recherche de sous-chaîne publié en 1980 par Nigel Horspool, un professeur à l'université de Victoria au Canada. Il consiste en une simplification de l'algorithme de Boyer-Moore qui ne garde que la première table de saut . Notion que nous allons ci-après expliciter.*

Exemple : motif = ''chasse''

| U | n | | c | h | a | s | s | e | u | r | | s | a | c | h | a | n | t | | c | h | a | s | s | e | r | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| c | h | a | s | s | e | | s | a | n | s | | s | o | n | | c | h | i | e | n | | | | | | | |
| | | | | | | | | | | | | | | | | | | | | | c | h | a | s | s | e | |

Dans le cas de l'algorithme naif version 2 : combien fait-on de comparaisons : ……..

L'algorithme de Boyer-Moore-Horspool repose sur 2 idées :

— La première idée consiste à comparer le motif avec la portion du texte qui apparaît dans la fenêtre de droite à gauche, et non pas de gauche à droite. Ainsi, on fait décroître j à partir de M-1 jusqu'à trouver que le caractère qui lui fait face dans le texte, c'est-à-dire x = texte[i + j], est différent du caractère y = motif[j] du motif.

*Par exemple : s'il commence la recherche du motif CTGCGA au début d'un texte, il vérifie d'abord la 6ème position en regardant si elle contient un A. Ensuite, s'il a trouvé un A, il vérifie la 5ème position pour regarder si elle contient le dernier G du motif, et ainsi de suite jusqu'à ce qu'il ait vérifié la 1ère position du texte pour y trouver un A.*

— La deuxième idée consiste à opérer un décalage de la fenêtre qui varie en fonction de la paire de caractères qui ont révélé **la non-correspondance**, c'est-à-dire en fonction de (x, y). On va construire ce que l'on nomme une table des sauts, pour chaque caractère du motif.

Celle table traduit en fait l'écart minimal entre une lettre du motif et la fin du motif. La dernière lettre du mot est traitée à part, elle renvoie un écart maximal si elle n'est pas présente ailleurs dans le mot.

pour le motif de longueur 6 : CTGCGA, la table des sauts sera :

| Motif | CTGCGA | | | |
|---|---|---|---|---|
| Lettres | A et autres lettres | G | C | T |
| Distance à la fin du motif | 6 | 1 | 2 | 4 |

pour le motif de longueur 6 : ATGCGA, la table des sauts sera :

| Motif | ATGCGA | | | | |
|---|---|---|---|---|---|
| Lettres | autres lettres | A | G | C | T |
| Distance à la fin du motif | 6 | 5 | 1 | 2 | 4 |

TD - implémentation sous Python ouvrir le notebook

1. Création de la table de sauts
   a) Écrire une fonction qui donne la table des sauts d'un motif. Utiliser un dictionnaire dont les clés seront les lettres du motif et les valeur le saut associé.
   b) Effectuer des tests appropriés
   Par exemple : table_sauts('bonjour') renvoie {'b': 6, 'o': 2, 'n': 4, 'j': 3, 'u': 1}
   table_sauts('ACTGACTGACTG') renvoie {'A': 3, 'C': 2, 'T': 1, 'G': 4}

2. Compléter le code l'algorithme de Boyer More Horspool donné et faire des tests