M2 ATAL
Algorithmique et Alignement de Chaînes

Université de Nantes
2020/2021

Distanciel
Recherche de Motif avec pré-traitement du texte

Automne 2020

— *ce projet est à réaliser seul.e ou en binôme*
— *le langage de programmation est celui de votre choix, mais votre code doit pouvoir fonctionner sous Linux (Ubuntu)*
— *le tout est à déposer sur Madoc au plus tard le mercredi 2 décembre 2020 à 20h20, sous la forme d'***un ficher compressé .zip** *contenant :*
  — *le code source correctement organisé et commenté (sa provenance, s'il n'est pas de vous, doit apparaître à la fois dans les fichiers source et dans le rapport)*
  — *le rapport* **au format PDF**
  — *le/s fichier/s texte utilisé/s, ou un lien (par ex. sur UNCLOUD) vers ces fichiers s'ils sont trop volumineux (le .zip ne peut dépasser 20Mo sur Madoc)*
  — **un ficher texte README.txt contenant les instructions de compilation et d'exécution**

Travail demandé Le but est d'implémenter et de tester trois méthodes vues en cours pour la recherche exacte de motifs avec pré-traitement du texte T, à savoir :

1. arbre des suffixes
2. tableau des suffixes
3. transformée de Burrows-Wheeler et FM-index

En particulier, on s'attachera à évaluer et commenter les trois aspects suivants :
— la taille (en mémoire) de la structure,
— le temps de construction de la structure, et
— le temps de la recherche de motifs.

Pour cela, vous ferez des tests sur :
— un nombre $k \geq 20$ de textes de longueurs différentes (en nombre de caractères), sachant que (a) les k textes doivent être exprimés sur le même alphabet, (b) la longueur n_{max} du plus grand texte doit être au moins 200 fois supérieure à la longueur n_{min} du plus petit et (c) les tailles des k textes doivent être réparties à intervalles réguliers entre n_{min} et n_{max}.

  Quelques exemples de textes, que vous pouvez utiliser, vous sont fournis dans la partie Distanciel sur Madoc, mais il n'est pas obligatoire de les utiliser (attention : certains textes sont à pré-traiter – notamment ADN et protéines – pour retirer les passages à la ligne inutiles).

— pour chaque texte, vous procéderez à au moins 1000 recherches de motifs différentes, pour des tailles de motifs allant de 4 à 13 (donc 100 motifs par taille), et plus précisément :
  — au moins 800 motifs (= 80 motifs $\times$ 10 tailles) *présents* dans le texte
  — au moins 200 motifs (= 20 motifs $\times$ 10 tailles) générés au hasard (pouvant donc être présents ou non dans le texte)

Vous présenterez et discuterez des résultats de la manière qui vous convient le mieux, mais on s'attend au moins à ce qu'il y ait :
— trois courbes "*taille mémoire*" présentant, pour chacune des trois structures étudiées, l'évolution de la taille mémoire de cette structure en fonction de la taille du texte ;
— trois courbes "*temps de création de l'index*", réalisée sur le même principe que la précédente ;

— trois fois trois courbes "*temps de recherche de motifs*", réalisées sur le même principe que la précédente, mais en détaillant pour chaque méthode de recherche de motif les trois cas suivants : (a) motifs présents dans T, (b) motifs générés au hasard et (c) l'ensemble des deux.

Rapport On s'attend à ce que le rapport soit concis (maximum 12 pages), et à ce qu'il aborde les points ci-desosus (liste non exhaustive) :

— choix du langage ;
— description (rapide) de chacun des algorithmes testés ainsi que de ce qu'ils fournissent en sortie (ex : recherche de toutes les positions du motif dans le texte vs recherche de la simple existence du motif dans le texte) ;
— provenance de chacun des algorithmes testés (s'ils n'ont pas été écrits par vous-mêmes) ;
— complexité en temps de la construction de l'index pour chacun des algorithmes testés (attention à bien vérifier cette complexité, notamment si les algorithmes ne sont pas écrits par vous-mêmes) ;
— complexité en temps de la recherche de motifs ;
— complexité en espace requise pour faire fonctionner les algorithme testés ;
— choix des textes et des motifs, et description des méthodes mises en oeuvre pour les générer ;
— figures et tableaux de résultats (dont les courbes évoquées ci-avant, mais il n'est pas interdit d'en proposer d'autres) ;
— commentaires et analyses sur les résultats ;
— conclusion, incluant les difficultés rencontrées et les améliorations/extensions possibles.

Remarques diverses

Pour toute question, utiliser le forum dédié au distanciel sur Madoc, appelé "Échanges et questions autour du distanciel Recherche de Motifs". Ce forum vous permet d'échanger entre vous, et aussi de poser vos questions à l'enseignant.

Il n'est pas interdit de réutiliser du code existant, mais dans ce cas il est **impératif** de l'indiquer clairement, et de citer vos sources. Par exemple, vous pourrez trouver des ressources en Python aux URL suivantes (liste non exhaustive) :

```
http://www.langmead-lab.org/teaching-materials/
https://clementj01.users.greyc.fr/alea2019/exos/
```

Si vous expérimentez des limitations (en mémoire, en temps), il faut l'indiquer dans le rapport (dans les difficultés rencontrées). Il faut également indiquer comment vous avez contourné le problème (ex : textes plus courts, moins de recherches de motifs par texte que demandé), en étant précis dans votre description.

A l'inverse, il n'est pas interdit d'étendre vos tests au-delà de ce qui vous est demandé (ex : nombre de textes testés, nombre de recherche de motifs par texte). Dans ce cas également, il convient de l'indiquer précisément dans le rapport.