Université de Rouen - Normandie	Année 2022–2023
UFR Sc. & Tech.	Master info 1re année
U.E. d'Algorithmique du texte	Fiche de TP N°3

Vous rendrez au plus tard le jour de l'examen une archive compressée (`.gz` ou `.zip` pas de `.rar`) contenant :

— un rapport au format **PDF** comprenant une description des structures de données et des fonctions utilisées et un commentaire détaillé de vos résultats ;

— les sources de vos programmes contenant IMPÉRATIVEMENT un `makefile` ;

— un fichier `README.txt` contenant une aide pour utiliser vos programmes ;

et ne contenant ABSOLUMENT pas :

— les exécutables ;

— les fichiers générés.

Travail demandé

1. Écrire un générateur pseudo-aléatoire de textes. Votre exécutable doit IMPÉRATIVEMENT être nommé `genere-texte` et prendre UNIQUEMENT deux paramètres d'abord la longueur du texte à générer puis la taille de l'alphabet. Il doit UNIQUEMENT écrire ses résultats sur la sortie standard.
2. Écrire un générateur pseudo-aléatoire de mots. Votre exécutable doit IMPÉRATIVEMENT être nommé `genere-mots` et prendre UNIQUEMENT quatre paramètres d'abord le nombre de mots à générer puis la longueur minimale et la longueur maximale des mots et enfin la taille de l'alphabet. Il doit UNIQUEMENT écrire ses résultats sur la sortie standard (un mot par ligne).
3. Implanter l'algorithme d'Aho-Corasick (avec les complexités annoncées dans le cours) pour compter le nombre d'occurrences exactes d'un ensemble de k mots dans un texte. En particulier lors du calcul des suppléants de chaque nœud du *trie* le parcours en largeur du *trie* s'effectue avec une file. Les opérations Enfiler et Défiler doivent s'exécuter en temps constant.

   Utiliser deux méthodes pour représenter l'arbre :

   — une matrice de transitions ;

   — une table de hachage ;

   Vos exécutables doivent IMPÉRATIVEMENT être nommés `ac-matrice` et `ac-hachage` respectivement et prendre UNIQUEMENT deux paramètres, d'abord le nom du fichier qui contient les mots à rechercher (un par ligne) puis le nom du fichier qui contient le texte. Ainsi les exécutions de :

   — `ac-matrice mots.txt texte.txt`

   — `ac-hachage mots.txt texte.txt`

   devront afficher UNIQUEMENT 80 avec les fichiers `mots.txt` et `texte.txt` fournis.

4. Utiliser `genere-texte` pour générer pseudo-aléatoirement des textes de longueur 5 000 000 sur des alphabets de taille 2, 4, 20 et 70.
5. Pour chacun des alphabets utiliser `genere-mots` pour générer pseudo-aléatoirement 3 ensembles de 100 mots de longueur entre 5 et 15, entre 15 et 30 et entre 30 et 60 respectivement.
6. Pour chacun des textes, utiliser `ac-matrice` et `ac-hachage` pour effectuer la recherche des 3 ensembles dans les 4 textes avec les 2 implantations de l'algorithme d'Aho-Corasick.
7. Relever les temps d'exécution de chacune de ces recherches, faire des courbes pertinentes et commenter.

Aide : vos programmes seront exécutés avec le script `script.sh` fourni.