

Mini projet: Analyse textuelle

Quel est le classement des lettres de l'alphabet selon leur fréquence dans un texte français ? Est-ce que ce ranking est différent dans d'autres langues ? Quel est le quota de ponctuations par rapport aux lettres ? Est-ce que « dieu » est le mot le plus utilisé dans la bible ? Combien de fois le mot "amour" apparaît-il dans « Roméo et Juliette » ? Nous allons développer un logiciel qui permet de répondre à toutes ces questions.

L'objectif consiste à créer un programme d'analyse textuelle, c'est-à-dire un outil informatique capable de parcourir un texte puis d'effectuer un certain nombre de statistiques. Le programme sera basé sur le principe de listes chaînées. Les données à traiter seront lues dans un fichier. Afin de ne pas compliquer la tâche, nous nous limitons dans le cadre de ce mini-projet aux textes codés en format ASCII simple. Vous disposerez de quelques exemples de fichier texte, ainsi que d'un programme corrigé (sans code source bien entendu) qui permettra de vérifier la pertinence des résultats de votre application.

Cahier des charges

Pour réussir ce projet, l'outil doit au moins comporter les caractéristiques suivantes:

- L'application doit être judicieusement décomposée en modules (*.c et *.h).
- La compilation se fait à l'aide d'un makefile
- Le programme est lancé à partir du terminal
- Il ouvre et lit un fichier texte ASCII spécifié sous forme de paramètre
- Si aucun paramètre n'est fourni, le programme affiche une petite aide puis s'arrête.
- Si ce fichier indiqué n'existe pas, le programme affiche un message d'erreur puis s'arrête.
- Si le fichier indiqué existe, le programme est capable de calculer les statistiques suivantes:
 - a) compter le nombre de caractères,
 - b) déterminer les caractères les plus fréquents, affichés par ordre décroissant, en donnant le nombre absolu et le pourcentage relatif aux autres caractères,
 - c) compter le nombre de mots,
 - d) déterminer les mots les plus fréquents, affichés par ordre décroissant, en donnant le nombre absolu et le pourcentage relatif aux autres caractères,
 - f) calculer la moyenne des caractères par mot.

Bonus

Pour une meilleure note, de nombreuses améliorations peuvent être effectuées à partir du programme de base, par exemple des options supplémentaires. Comme d'habitude, nous faisons appel à votre créativité. Voici quelques suggestions:

- une option qui indique que le programme doit confondre majuscules et minuscules
- une option qui spécifie que les statistiques sont écrites dans un fichier de sortie
- une option qui permet compter le nombre de mots *uniques*, c'est-à-dire les mots différents les uns des autres
- etc.