PROJ631 – Projet algorithmique

Titre : Compression de données par codage de Huffman

Descriptif général

Le codage de Huffman, du nom de son concepteur, est une méthode statistique de compression de données. Son principe est de remplacer un caractère (ou symbole) par une suite de bits de longueur variable. L'idée sous-jacente est de coder ce qui est fréquent sur peu de bits et au contraire ce qui est rare sur des séquences de bits plus longues. Il permet une compression sans perte, c'est-à-dire qu'une suite de bits strictement identique à l'originale est restituée par décompression. Il nécessite cependant que soit connues (ou estimées) les fréquences d'apparition des différents symboles à coder. Il existe ainsi plusieurs variantes de l'algorithme de Huffman (statique, semi-adaptatif ou adaptatif) aujourd'hui utilisées dans des algorithmes de compression de fichiers tels que gzip.

Ce sujet concerne les versions statique et semi-adaptative de l'algorithme. Dans le premier cas l'ensemble des symboles codables (l'alphabet) et leur fréquence sont fixées indépendamment du texte à coder. Dans le second cas, le texte à coder est tout d'abord lu intégralement de façon à construire l'alphabet et les fréquences d'apparition des éléments de l'alphabet.

Descriptif détaillé

Votre programme devra réaliser la phase d'encodage d'un texte selon les trois étapes suivantes :

1. Détermination de l'alphabet et des fréquences des différents caractères codables
2. Construction de l'arbre de codage
3. Codage du texte

Etape 1 : Détermination de l'alphabet et des fréquences

Dans la version statique, lecture d'un fichier texte (format similaire à celui du fichier freq.dat fourni).

Dans la version semi-adaptative, lecture du texte à coder et détermination de l'alphabet et des fréquences.

Etape 2 : Construction de l'arbre

L'algorithme est décrit dans l'article de son créateur publié en 1952. Il repose sur une structure d'arbre binaire où tous les nœuds internes ont exactement deux successeurs. Les feuilles sont étiquetées avec les caractères originaux, les branches par 0 (fils gauche) et 1 (fils droit). Les chemins depuis la racine jusqu'aux feuilles constituent les codes des caractères. La construction de l'arbre est réalisée de la manière suivante :

Créer un arbre (feuille) pour chaque caractère codable avec la fréquence associée
Répéter
    Extraire les 2 arbres t_1 et t_2 de fréquence minimale
    Créer un nouvel arbre t avec t_1 et t_2 comme sous-arbres, la fréquence associée à t étant la somme des fréquences de t1 et t2
Jusqu'à ce qu'il ne reste plus qu'un seul arbre

Etape 3 : Codage du texte

Le code de chaque caractère est obtenu par un parcours en profondeur de l'arbre.

Le texte codé sera affiché à l'écran et sauvegardé dans un fichier de texte de sortie sous forme d'une succession d'octets.

Références

D.A. Huffman, A method for the construction of minimum-redundancy codes, Proceedings of the I.R.E., septembre 1952, pp. 1098-1102.

http://www.data-compression.info/Comparisons/index.html

Ressources

- freq.dat : Fichier texte de fréquences (format à respecter)
- montexte.txt, extrait_alice.txt, alice29.txt : Textes à coder, difficulté croissante

Livrables

Evaluation

Référent

Sylvie Galichet

Sur le plagiat

Le plagiat est une forme de fraude définie dans la charte **anti-plagiat** adoptée par l'Université Savoie Mont Blanc - https://dsi.univ-smb.fr/profil/pers/charte-anti-plagiat-2014.pdf - pouvant mener à des sanctions disciplinaires. Pour lutter contre ce phénomène, l'établissement s'est doté d'un outil de détection du plagiat permettant d'évaluer le degré d'authenticité d'un document.

En particulier, dans ce module il n'est pas admissible

- de présenter un code trouvé sur internet et/ou copié d'un autre projet sans le mentionner explicitement
- de présenter un code non compris