Compression & Algo du Texte

– M1 –
2014-2015

Feuille 1 : codage & algorithme de Huffman

► **Exercice 1** ◄ On considère l'alphabet $A = \{a, b\}$.

a. Listez tous les mots non-vides de longueur au plus 2 de A et de B.

b. En déduire que si un des mots de taille 2 est compressé par ϕ, alors il y a au moins un mot dont la taille augmente quand on applique ϕ.

c. Généralisez à des mots plus grands : si ϕ compresse au moins un mot, alors ϕ augmente la taille pour au moins un mot.

► **Exercice 2** ◄ On souhaite compresser un texte qui représente de l'ADN, codé comme un mot sur l'alphabet $\{A, G, C, T\}$ et enregistré dans un fichier où chaque caractère est encodé en ASCII sur un octet.

a. Comment gagner facilement en espace en utilisant le nombre réduit de caractères différents ? On supposera dans un premier temps que le mot u représentant l'ADN est de longueur multiple de 4.

b. Comment gérer les autres longueurs ?

c. Est-ce que la décompression est difficile à effectuer ?

► **Exercice 3** ◄ On a un texte u sur l'alphabet avec 23 lettres $L = \{a, \ldots, w\}$. On note $|u|_\alpha$ le nombre d'occurences de la lettre $\alpha \in L$ dans le mot u. On suppose que $|u|_a = 1000$, $|u|_b = 400$, $|u|_c = 200$, et $|u|_\alpha = 20$ pour les autres.

a. Quelle est la taille optimale ℓ des mots pour un codage de longueur fixe de u ?

b. Combien de bits faut-il pour coder u avec un tel codage ϕ ?

On propose de créer un nouveau codage ψ à partir de ϕ de la façon suivante : $\psi(a) = 0$ et $\psi(\alpha) = 1\phi(\alpha)$ (on rajoute un 1 devant) pour tout $\alpha \neq a$.

c. Est-ce que ψ est un codage de longueur fixe ? un codage préfixe ?

d. Combien de bits faut-il pour encoder u avec ψ ?

e. Comment décoder un message encodé par ψ ?

f. Généralisez la construction en utilisant 2 bits pour distinguer certains caractères des autres. Est-ce qu'on y gagne en compression pour u ?

► **Exercice 4** ◄ On considère le message suivant :

$$m = tacuitracriatuactaura$$

a. Quelle sera la taille du codage du message en code ASCII ?

b. Quelle sera la taille minimale du codage du message en code de longueur fixe ?

c. Appliquez le procédé de Huffman et proposez un code pour représenter le message. Ecrivez le codage associé au message. Quelle est sa taille ?

d. Comment encoder l'arbre de Huffman ? Quelle est la taille totale de l'encodage de l'arbre et du message ?

e. Le message à compresser est maintenant consituté de 100 occurrences de m mises bout à bout. Quelle est la taille totale de l'encodage ?

► **Exercice 5** ◄ On considère le codage de Huffman suivant :

$$a \mapsto 1 \qquad b \mapsto 011 \qquad c \mapsto 000$$
$$r \mapsto 010 \qquad t \mapsto 001$$

Décodez le message $m = 011101000110111000$. On suppose qu'une erreur a été commise en transmettant m et que le 4eme bit a été transformé en 0. Quel devient le nouveau décodage du message ?

► **Exercice 6** ◄ On considère une source qui émet continument des 0 avec une probabilité $p_0 = \frac{1}{4}$ et des 1 avec une probabilité $p_1 = \frac{3}{4}$. On désire utiliser la méthode de huffman pour compresser l'information reçue. Pour cela on regroupe par blocs de 2 bits 00, 01, 10 et 11. On calcul la probabilité que chaque bloc apparaisse et on applique la méthode de Huffman. Quel est le taux moyen de compression (taille moyenne du message compressé/bit de la source) ?

On regroupe maintenant les bits 3 par 3. Quel est le taux moyen de compression ?