Expérience 3: Compression sans perte de textes

Question

Comment des textes peuvent-ils être stockés sous une forme compressée?

Matériel nécessaire

- Navigateur ou copies de tableaux
- Tableau ASCII
- Annexes électroniques

Description de l'exercice

Il faut stocker un proverbe éthiopien de la manière la plus compacte possible.
Je kleiner die Eidechse, umso grösser ist ihre Hoffnung, ein Krokodil zu werden . (Plus un lézard est petit, plus son espoir est grand de devenir un crocodile.)

A l'aide d'un tableau ASCII, chaque signe de ce texte peut être codé avec huit un ou zéro. Le proverbe étant constitué de 80 signes, il peut donc être écrit en code ASCII de la manière suivante.

```
01001010 01100101 00100000 01101011 01101100 01100101 01101001 01101110
01100101 01110010 00100000 01100100 01101001 01100101 00100000 01000101
01101001 01100100 01100101 01100011 01101000 01110011 01100101 00101100
00100000 01110101 01101101 01110011 01101111 00100000 01100111 01110010
11110110 01110011 01110011 01100101 01110010 00100000 01101001 01110011
01110100 00100000 01101001 01101000 01110010 01100101 00100000 01001000
01101111 01100110 01100110 01101110 01110101 01101110 01100111 00101100
00100000 01100101 01101001 01101110 00100000 01001011 01110010 01101111
01101011 01101111 01100100 01101001 01101100 00100000 01111010 01110101
00100000 01110111 01100101 01110010 01100100 01100101 01101110 00101110
```

Le code est constitué de 80 x 8 un et zéro, c'est-à-dire de 640 bits. En 1952, David A. Huffman a créé une méthode raffinée pour stocker des textes de manière beaucoup plus compacte. A la place d'un tableau ASCII, un tableau de codage spécial est généré pour chaque texte. Les signes qui se répètent fréquemment sont définis par un code court.

L'annexe électronique permet de générer un tableau – respectivement un arbre – de Huffman pour chaque texte choisi et de produire des textes à l'aide de la méthode de Huffman.

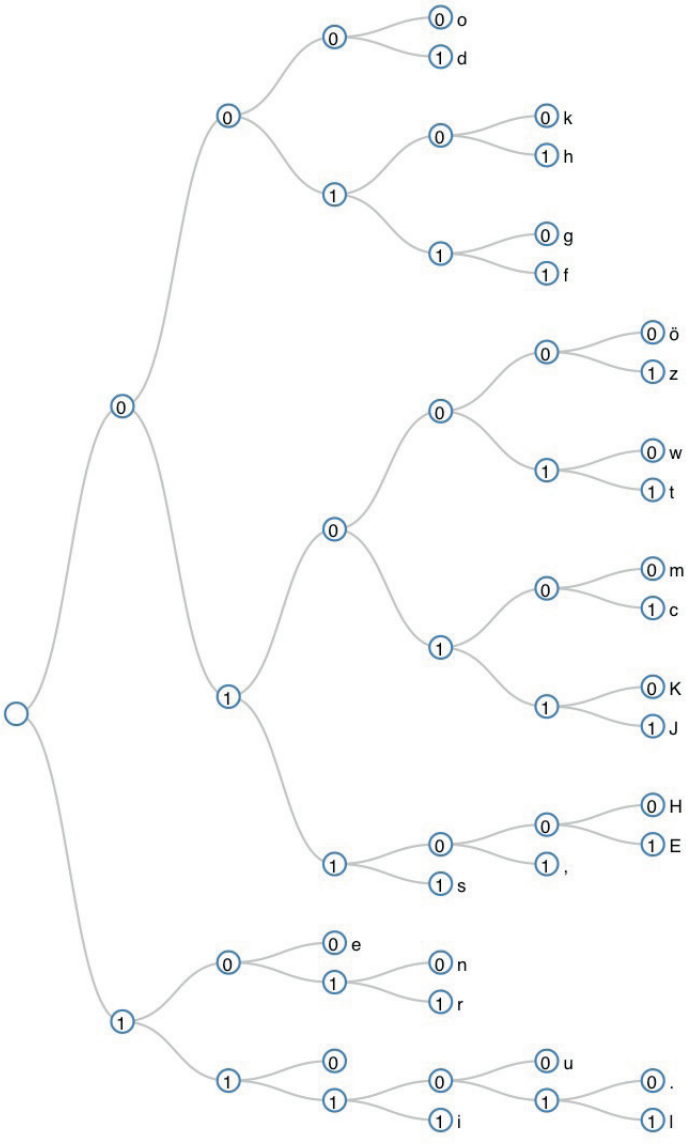
Le proverbe éthiopien peut être stocké de manière plus compacte par le code suivant:

```
0101111001100010011101110011111010100101111000011111100110011001111100011000
0101010010101111000110111011100010100011100001100011010110100000111011110010-
0111101111011101001111011110010110111001100110000000001110011110101110010100
0011001101110100111110101100101101011000000100000000011111111011110010001111-
1001100100101001011000110010101110100
```

Le proverbe est stocké avec 336 un et zéro au lieu des 640 du code ASCII. On économise ainsi presque 50% de place de stockage.

Décodage du proverbe

Pour le décodage, on utilise l'arbre de Huffman suivant.



Lors du décodage, on commence par la racine, à gauche. Si le chiffre du code est un 1, on suit la branche du bas et si c'est un 0, la branche du haut. On procède de même jusqu'à la fin de la branche où se trouve un signe ou un caractère.

La première partie du code 0100101001100011101010 est codée de la manière suivante:

010111	= J
100	= e
110	= (espace)
00100	= k
111011	= I

Exercice 1: Décoder à l'aide de l'arbre de Huffman

Il faut trouver les signes correspondant aux codes suivants. Pour le faire, utilise l'arbre de Huffman de la page précédente.

011000
1111
111011
00111
100

Exercice 2: Coder son propre texte

Utilise l'annexe électronique pour coder ton propre texte.
Calcule combien il faut de bits (1 et 0) avec le codage ASCII et combien il en faut avec le codage de Huffman.
Selon quel facteur le texte peut-il être compressé?

L’illustration suivante montre un codage de Huffman à l’aide de l’annexe électronique.