Science de l'information : Série 4

Exercice 4.1 :

1. a. Pour calculer l'entropie du mot de Homer, il nous suffit de calculer l'entropie d'un caractère, puis de le multiplier par la longueur du mot voulu (ici 8). Cela nous est permis, car les caractères sont indépendants entre eux, car avec remise. L'entropie du premier caractère est donc de :

   $2*(2/8*\log_2(4))+6*(1/8*\log_2(8)) = 3.25$ bits.

   L'entropie étant la même pour chaque lettre suivante, l'entropie totale est de 8*(l'entropie d'un caractère) = 8*3.25 = **26 bits**

   b. Le meilleur code binaire se trouve avec un code de Huffman, qui nous donne une longueur de 2 pour E et une longueur de 3 pour les autres lettres. Le code serait par exemple **: E{00}, N{010}, T{011}, R{100}, O{101}, P{110}, I{111}.** Pour la longueur moyenne, comme les caractères sont à nouveau indépendants, on peut calculer la longueur d'un caractère (en moyenne) et la multiplier par le nombre de caractères. La longueur du premier caractère est de

   2/8*2 + 6*1/8*3 = 2.75 bits.

   La longueur moyenne totale est donc de 2.75*8 = **22 bits.** Notons que cela revient à calculer l'entropie du bloc, en trouvant le code du bloc par encodage des caractères.

2. Cette question demande un peu de logique. En effet, Bart va tirer 8 lettres sans remise d'un sac contenant 8 lettres. Donc tirer toutes les lettres. Mais cela revient a réassembler les lettres du mot ENTROPIE. E étant de longueur 2, et les autres de longueur 3, cela revient à additionner la longueur de chaque lettre (en comptant 2x E, vu qu'il est 2x dans le sac), donc

   2x2 + 6*3 = **22 bits**

3. Ici, au lieu d'encoder par lettre, nous avons meilleur intérêt à encoder les mots par mot entier. En tirant les lettres à la manière de Bart, il y aura 8! mots différents (le premier caractère a 8 possibilités, le second 7, le troisième 6, etc.), soit 40'320 mots possibles. Maintenant nous allons ''décider'' que tous les mots sont équiprobables. En prenant le $\log_2$ du nombre de mots possibles (donc appliquer Shannon-Fano), nous trouvons **15.299… bits**. Ainsi, avec une longueur moyenne de 15.299, il est possible d'encoder tous les mots. Ce chiffre indique que certains mots auront une longueur moyenne de 16, d'autres moins. **Lisa a donc raison**.

En revanche, si nous décidons d'encoder sur un nombre entier de bits (pour des raison informatiques par exemple), alors Lisa a tort car 2^{15} bits ne sont pas suffisants, et 2^{16} suffit, donc 16 bits.

Exercice 4.2

1. En tirant 2 D6, les résultats totaux possibles sont 2, 3, 4, 5,…12 avec les probabilités respectives suivantes : 1/36, 2/36, 3/36, 4/36, 5/36, 6/36, 5/36, 4/36, 3/36, 2/36, 1/36, ce qui vaut respectivement 1/36, 1/18, 1/12, 1/9, 5/36, 1/6, 5/56, 1/9, 1/12, 1/18, 1/36.

   L'entropie $(B_2|B_1=2)$ (soit d'arriver sur une case, en partant de la case 2) vaut $1/36*\log_2(36) + 1/18*\log_2(18) + 1/12*\log_2(12) + 1/9*\log_2(9) + 5/36*\log_2(36/5) + 1/6*\log_2(6) + 5/36*\log_2(36/5) + 1/9*\log_2(9) + 1/12*\log_2(12) + 1/18*\log_2(18) + 1/36*\log_2(36) = 3.274$ bits. En calculant la même chose pour $(B_2|B_1=3)$ (soit d'arriver sur une case en partant de la case 3) $(B_2|B_1=4)$ (etc.), $(B_2|B_1=5)$,…etc. on s'aperçoit qu'ils valent tous la même chose, soit 3.274 bits. L'entropie de B2 sachant B1 étant la moyenne de ces entropies, on trouve l'entropie totale valant également 3.274 bits.

   De manière récursive, on trouve que $(B_3|B_1=2, B_2=2)$, $(B_3|B_1=3, B_2=2)$, $(B_3|B_1=4, B_2=2)$, $(B_3|B_1=5, B_2=5)$,…, $(B_3|B_1=12, B_2=12)$ valent également tous 3.274 bits. Toujours récursivement, on trouve que $(B_n|B_1, B_2,\ldots B_{n-1})$ vaut exactement **3.274 bits**

   L'entropie par symbole est ici l'entropie conditionnelle, car elle répond à la définition (vu que n tend vers ''l'infini'', soit le nombre total de cases). C'est pourquoi les deux sont identiques, car leur définition est similaire.

2. L'entropie conditionnelle de la source de Lisa ne diffère pas de celle de Bart. En effet, aucune information supplémentaire n'est fournie chez Lisa en fonction de la case sur laquelle elle atterrit. Le fait de faire revenir sur la case 10 ne change rien du tout. En effet, avant de lancer le dé, Lisa a toujours autant de chances de tomber sur la case 30 (donc 10) que sur les autres, ce qui laisse l'entropie inchangée.

3. c.f. moodle

4. idem

5. Entre Homer et Lisa, Lisa se rapproche plus de la vérité, car sa séquence est plus ''aléatoire'' que celle d'Homer. En effet, les jets de dés d'Homer arriveront à chaque tir a une zone fixe (même si sur autant de tirs l'aléatoire devient présent). Donc même si pour avoir un résultat exact il faudrait une infinité de tirs, la méthode de Lisa est plus correcte.

   Ces entropies sont supérieurs à l'entropie par symbole des points 1 et 2 car le fait de lancer le dé, et de calculer ''manuellement'' apporte plus d'information aux cases que par calcul théorique.