calcul d'entropie (tableaux, niveau 2)

On se propose ici de calculer l'entropie

- d'une chaîne de caractères telle que définie dans le cours d'ICC;
- plus généralement, d'une distribution de probabilité.

Distribution de probabilité

Une distribution de probabilités est simplement un tableau de nombres réels compris entre 0 et 1 et dont la somme est 1.0

Définissez un type `Distribution` pour représenter une distribution de probabilités.

Entropie d'une distribution de probabilité

L'entropie d'une distribution de probabilités est l'opposé de la somme de ses valeurs multipliées par leur logarithme. Si l'on veut l'exprimer en bits, on prendra le logarithme en base 2 :

$$-\sum_i p_i \log_2(p_i)$$

Définissez une fonction `entropie` qui calcule l'entropie d'une distribution de probabilités.

Attention à traiter correctement le cas *p*=0, pour lequel *p log(p)* vaut 0.

Entropie d'une chaîne de caractères (définition ICC)

Dans le cours ICC, nous avons défini l'entropie d'une chaîne de caractères comme l'entropie de la distribution de probabilité empirique, i.e. la distribution de probabilité résultant du décompte du nombre de chacune des lettres dans cette même chaîne.

Il nous faut donc commencer par calculer cette distribution. Définissez pour cela une fonction `calcules_probas` qui reçoit une chaîne de caractères et retourne la distribution de probabilité de ses lettres : pour chacune des lettres, on compte le nombre de fois qu'elle apparaît dans la chaîne puis on divise par le nombre total de lettre.

Pour rendre cette fonction un tout petit peu plus générale :

- on ajoutera un paramètre booléen supplémentaire qui, lorsqu'il est `true`, prendra en compte également les espaces (i.e. 27 lettres) alors que sinon le calcul est fait comme dans le cours ICC en ignorant les espaces ;
- on ignorera la casse (i.e. la différence majuscule, minuscule) : utiliser pour cela la fonction `toupper()` de la bibliothèque `cctype` qui retourne la version majuscule (ou lui-même) d'une lettre (alphabétique) reçue en paramètre ; on utilisera par ailleurs au préalable la fonction `isalpha` (de la même bibliothèque `cctype`) pour déterminer si un caractère donnée est une lettre de l'alphabet ou non.

Pour calculer cette distribution de probabilité empirique, je vous conseille de tout d'abord créer un tableau de taille fixe de 27 nombre pour compter le nombre d'occurrence de chacune des vingt-sept lettre. Utilisez ensuite ce tableau pour construire une distribution de probabilité avec uniquement les lettres qui apparaissent, i.e. les probabilités non nulles (cf exemple ci-dessous).

Terminez en définissant une fonction `entropie` qui reçoit en paramètre une chaîne de caractères et retourne son entropie (i.e. l'entropie de sa distribution empirique des lettres).

Tests

Voici quelques entropies qui vous permettrons de tester votre code :

distribution	entropie (en bit)
{ 0.0, 0.1 }	0
{ 0.3, 0.7 }	0.881291
{ 0.5, 0.5 }	1
{ 0.1, 0.2, 0.3, 0.4 }	1.84644
{ 0.25, 0.25, 0.125, 0.0625, 0.0625, 0.0625, 0.0625, 0.0625, 0.0625 }	2.875

Et pour rappel (cf cours ICC), l'entropie de la chaîne « IL FAIT BEAU A IBIZA », en ignorant les espaces, est de 2.875

bit, sa distribution de probabilités étant la dernière donnée ci-dessus.