Université de Paris

# Mathématiques discrètes
# Projet : Codage, entropie et mots typiques

**Consignes** Le but du projet est de présenter une application dans laquelle les mathématiques discrètes jouent un rôle fondamental.

Le rendu final du projet consistera en un article destiné au grand public au format pdf de 800-1000 mots plus une annexe numérique, qui pourra contenir par exemple une démonstration interactive, une vidéo explicative et/ou des graphiques générés par du code écrit par vous-même ; cette annexe sera rendue sous la forme d'un lien vers un dépôt en ligne. La forme exacte et la technologie utilisée pour l'annexe peut varier et est donc laissé au libre choix des étudiants. L'article et son annexe seront jugés non seulement sur le contenu mais aussi sur la clarté de la présentation, la qualité de rédaction, et la créativité.

**Contenu** Le sujet détaille quelques points à développer mais ceux-ci sont seulement proposés comme point de départ de votre travail. Vous êtes encouragés à développer d'autres pistes en lien avec les mathématiques discrètes. De même, la bibliographie conseillée est un point de départ. Vous pouvez vous appuyer sur d'autres sources sur lesquelles vous porterez un œil critique et que vous prendrez soin de citer correctement.

**Charte de bonne conduite** Lisez attentivement la charte de bonne conduite. Portez une attention particulière à citer toutes vos sources, y compris les exemples et les images que vous utiliserez. L'utilisation d'outils d'IA tels que ChatGPT est formellement interdite. L'équipe pédagogique sera très attentive à tous ces aspects lors de la correction.

**Calendrier** Consultez la page Moodle du cours pour les dates des principales étapes du projet.

## Bref descriptif du sujet

L'entropie en informatique mesure la quantité d'information portée par une source de symboles aléatoires. Un exemple de source de symboles est l'ensemble des lettres dans un texte en français pour lesquels on connaît la fréquence avec laquelle apparaît chaque lettre. Un autre exemple serait les mots du dictionnaire avec la fréquence à laquelle ils apparaissent dans un texte. Plus l'entropie est grande, plus il faudra de bits par symbole, en moyenne, pour encoder la source. Une source qui émet presque toujours le même symbole ne porte pas beaucoup d'information : son entropie est proche de 0, alors que si un symbole est choisi aléatoiement parmi 16 mots, son entropie peut aller jusqu'à 4 bits d'information.
L'entropie est une notion centrale en compression car elle quantifie précisément le nombre moyen de bits nécessaires pour transmettre un mot provenant d'une source. L'algorithme de Huffman est un exemple d'algorithme dont le taux de compression est exactement l'entropie de la source dont il code les mots.

# Bibliographie conseillée

- `https://fr.wikipedia.org/wiki/Entropie_de_Shannon`
- `http://images.math.cnrs.fr/Claude-Shannon-et-la-compression-des-donnees.html`

# Pistes de développement

1. La définition de l'entropie pour une source aléatoire.
2. Écrire un programme qui calcule l'entropie des lettres à partir d'un texte en français et comparer avec l'entropie des lettres dans une autre langue de votre choix.
3. Tracer la courbe de l'entropie des sources binaires (qui émettent 2 symboles, 0 ou 1) en fonction de $p$ la probabilité d'émettre le symbole 0 (la source émet 1 avec probabilité $1 - p$).
4. Compter le nombre de mots de longueur $N$ comportant $pN$ symboles 1 et $N - pN$ symboles 0.
5. La notion de « mot typique ». Quels sont les mots typiques de longueur $N$ pour une source binaire qui émet 1 avec probabilité $p$ ?
6. Estimation de $\binom{n}{k}$ avec l'entropie de la source qui émet 0 avec probabilité $k/n$, et 1 avec probabilité $1 - k/n$.