M2 ISN – Chaînes de Markov et modélisation

Responsable : Adrien Hardy, email : adrien.hardy@univ-lille.fr

Projet

Ce projet est à faire par groupes de deux. Envoyer un compte rendu détaillé au format pdf ainsi que le script au format ipynb par email au plus tard le 9 février à 23h59.

Le chiffrement par décalage, dont l'utilisation a été popularisée par Jules César, consiste à coder un texte en opérant un décalage sur les 26 lettres de l'alphabet[1]. Ainsi, avec un décalage +3, A devient D, E devient H et Y devient B dans l'alphabet codé. Ce n'est pas une méthode de chiffrement robuste vu qu'il suffit de tester les 25 possibilités de décalage, ce qu'un ordinateur moderne peut faire très rapidement.

Exercice 1. Écrire une fonction qui, à un texte donné, renvoie les 30 premiers caractères des 25 textes obtenus par tous les décalages possibles de l'alphabet. À l'aide de ce script, décoder le texte suivant :

```
LD L CFWP, XPY HZCCJ XZCP LMZFE HSLE ESPJ NLYYZE DPP ESLY LMZFE
HSLE ESPJ NLY.  UFWTFD NLPDLC
```

Une méthode de chiffrement plus solide consiste à appliquer une permutation sur les lettres de l'alphabet. Avoir la clef du code c'est donc connaître la permutation qui a été utilisée. Comme il y a $26! \simeq 4.10^{26}$ clefs, il n'est pas envisageable de les tester une à une. Par exemple, on cherche à décoder le texte suivant :

```
XDP XESVR XTT,  NXMBY NAMY HTAMMRKADH RQRZ MYR NYSTR NSKTP XKSWDP QSW
ERBXWZR MYR HKRXMRZM ZRBKRMZ XKR XTNXQZ YAPPRD AD MYR ISZM WDTAFRTQ
JTXBRZ  MYSZR NYS PSD M ERTARVR AD IXHAB NATT DRVRK OADP AM.  KSXTP PXYT
```

On va utiliser une méthode approchée basée sur les chaînes de Markov pour casser ce type de chiffrement. Pour cela, on a à notre disposition de nombreux textes longs écrits en anglais ; par exemple le Project Gutenberg recense de nombreux livres libres d'accès que l'on peut trouver au format plain text (txt). En partant d'un texte (le plus long possible, voir la concaténation de plusieurs textes), on peut calculer la fréquence $A(x, y)$ d'apparition du symbole[2] y après le symbole x ; on a donc $\sum_y A(x, y) = 1$ pour tout x.

[1] Dans la suite, on fait abstraction des accents, cédilles, etc. Pour faire simple, on ne s'intéressera qu'à des textes écrits en anglais.

[2] Ici, par symbole on veut dire une lettre de l'alphabet mais aussi une ponctuation ou une espace. Pour simplifier, on pourra regrouper tous les symboles non-alphabet en le symbole "espace".

Exercice 2. A partir d'un texte assez long (par exemple Guerre et Paix de Léon Tolstoï, en version anglaise), créer la matrice A des fréquences et représenter-la graphiquement ; on pourra associer à un nombre entre 0 et 1 un niveau de gris. Repérer quelques suites de deux symboles les moins probables et identifier des mots qui les contiennent. En utilisant la matrice A comme matrice de transition d'une chaîne de Markov, générer un mot aléatoire de 6 lettres qui commence par a et qui n'existe pas dans la langue anglaise.

Etant donné une suite de symboles $w = x_1 x_2 \cdots x_n$, la quantité

$$L(w) = \prod_{i=1}^{n-1} A(x_i, x_{i+1})$$

représente donc la vraisemblance de voir cette suite apparaître, au regard du texte qu'on a choisi pour construire la matrice A. Si σ est une permutation sur l'ensemble des symboles, on note $w_\sigma := \sigma(x_1)\sigma(x_2)\cdots\sigma(x_n)$ la suite obtenue après permutation des symboles initiaux. À partir de maintenant, on prend pour suite w les trois lignes de texte codé au-dessus. Une stratégie pour décoder w, c'est-à-dire trouver une permutation σ telle que w_σ est le message initial, consiste à simuler des permutations aléatoires de loi

$$\mu(\sigma) := \frac{1}{Z} L(w_\sigma) \qquad \text{où} \quad Z := \sum_\sigma L(w_\sigma).$$

Exercice 3. Expliquer pourquoi cette stratégie a des chances de marcher, puis simuler de telles permutations aléatoires à l'aide de l'algorithme de Metropolis-Hastings. Proposer une version décryptée de w.

On pourrait rajouter un paramètre $\beta > 0$ et simuler des permutations de loi

$$\mu_\beta(\sigma) := \frac{1}{Z_\beta} L(w_\sigma)^\beta \qquad \text{où} \quad Z_\beta := \sum_\sigma L(w_\sigma)^\beta.$$

Exercice 4. Implémenter cette variante et expliquer comment le paramètre β influence le résultat. Mettre en application sur un texte anglais plus long (une demi-page ou plus) de votre choix que vous aurez chiffré à l'aide d'une permutation de l'alphabet.

Bonus : Proposer d'autres stratégies pour décoder de tels chiffrements et les illustrer.

Remarque : On pourrait chiffrer un texte à l'aide de symboles plus généraux qu'une permutation de l'alphabet (par exemple $A \to \star, B \to \boxplus, C \to \therefore$, ...) mais décoder un message revient alors à trouver une bijection σ entre ces symboles abstraits et les symboles de l'alphabet ; on peut donc procéder de la même façon qu'au dessus pour casser ces codes.