Théorie de l'Information

Mini-projet # 2
16 octobre 2006

SIC-SICOM

Maria-João Rendas

Le fichier CandideVoltaire.txt contient deux Chapitres de *Candide* de Voltaire. Vous allez utiliser ce texte (cette séquence de symboles émis par une très respectable source, Voltaire!) pour illustrer la notion de ensemble δ-informatif S_δ, de sa taille H_δ et en général de codage avec pertes.

1. Considérez dans un premier temps que les lettres de ce texte (les symboles ASCII dans le fichier) sont des échantillons statistiquement indépendants et identiquement distribués, et utilisez-le pour estimer la loi de probabilité de la source.

2. Tracez la valeur de H_δ en fonction de δ. Déterminez combient de bits N il faut pour coder le texte avec un code de longueur fixe, avec une probabilité d'erreur inférieure à 0.4. Déterminez l'ensemble S_δ de symboles qui possèdent un code.

3. Implémentez un codeur C_δ qui associe un mot binaire de longueur N pour tout symbole dans S_δ et qui, pour tous les autres, utilise un mot de code chosi au hasard. Implémentez également le décodeur D_δ correspondant. Le code C_δ n'est pas *complet* (il existent des mots de code qui ne sont pas utilisées). Construisez le codeur complet $\overline{C}_\delta$ (et le décodeur correspondant $\overline{D}_\delta$)qui a la même longueur de mots de code et la probabilité d'erreur la plus petite possible.

4. Utilisez C_δ et D_δ pour coder et décoder le fichier *CandideVoltaire.txt*. Etes-vous capable de le lire? Interprétez.

5. Tout en maintenant un code de longueur fixe, avec le même nombre N de bits par lettre du texte original, comment pouvez-vous faire pour diminuer δ, la probabilité pour qu'il n'existe pas de code pour les symboles de la source?

6. Considérez maintenant l'ensemble de toutes les séquences de **deux symboles**. Répétez l'analyse dans la première question, d'abord en considérant que les lettres adjacentes sont statistiquement indépendantes, et dans un deuxième temps, estimez directement leur loi conjointe. Comparez H_δ dans les deux cas. Interprétez leur différence en termes des mesures introduites en cours.

Suggestion: vous pouvez partir du script CountCandide.m qui est fourni.