UV Data Mining
Compte-Rendu de Travaux Pratiques : Probabilités et Suites de caractères

Maxime Chambreuil
maxime.chambreuil@insa-rouen.fr

5 février 2003

Chapitre 1

Explication

1.1 Apparition d'un caractère dans un texte

On a d'abord calculer les probabilités d'apparition d'un caractère dans un texte. On obtient ainsi les lois de probabilités marginales, qui vont nous servir par la suite.

1.2 Apparition d'un caractère en début de ligne

On a ensuite réexécuter le même algorithme, non pas sur tout le texte mais sur les caractères en début de ligne.

1.3 Apparition d'un couple de caractères

On modifie notre algorithme pour ne considérer non plus un caractère mais une fenêtre de 2 caractères. Pour tous les couples, on obtient ainsi une matrice des probabilités. A l'intersection i et j, on obtient la probabilité jointe des caractères i et j.

1.4 Génération d'un texte

Pour générer notre texte, on va d'abord génerer notre premier caractère. Pour cela, on tire un nombre entre 0 et 1 et on regarde quel est le caractère qui a cette probabilité d'apparaitre en début de ligne.

Maintenant qu'on connait notre premier caractère, on va trouver le suivant en fonction des probabilités conditionnelles. On va la aussi se servir d'un nombre aléatoire pour faire le choix de la probabilité, qui va nous donner ensuite le caractère. On continue le processus itérativement en fonction du nombre de caractère de notre nouveau texte à définir au préalable.

Chapitre 2

Interprétation

On remarque qu'avec un couple de caractères, on est capable de retrouver des syllabes qui ont du sens.
On peut très bien imaginer retrouver des mots compréhensibles en augmentant notre fenêtre d'étude : 3, 4 , etc... caractères.

Chapitre 3

Conclusion

Ce TP m'a permis de manipuler le calcul de probabilités sous Matlab, notamment à l'aide des fonctions cumsum et find, la manipulation des fichiers (fgetl, double, fopen, fclose, feof, etc...), la manipulation des codes ASCII (char, num2str, etc...) et l'affichage (imagesc).

J'ai aussi pu bien comprendre le calcul des probabilités conditionnelles à partir des probabilités jointes et marginales, et appliquer ainsi le théorème de Bayes.