Projet numérique MAP 432 : langage et chaînes de Markov

L'objectif de ce projet est d'utiliser les corrélations dans le langage afin d'analyser des textes par des méthodes probabilistes.

Étape 1. On va construire un fichier de référence qui va servir à l'apprentissage. Pour cela, extraire plusieurs livres (du même auteur) de la base de données :

http://www.fullbooks.com/

et les concaténer dans un unique fichier txt. Pour réduire la complexité, il faut traiter ces données brutes en supprimant les majuscules et la ponctuation (mais en conservant les espaces) afin d'obtenir une très longue chaîne de caractères avec 27 valeurs possibles.

Étape 2. A partir du fichier précédent, construire les matrices de transition pour des chaînes de Markov avec mémoire $0, 1, 2, 3, 4$. Comparer ensuite les textes qu'elles génèrent.

Dans un livre, les mots sont séparés par des espaces pour permettre la lecture. Ce n'est pas le cas dans le langage parlé et pourtant nous sommes capables de distinguer les mots. Nous allons maintenant utiliser les corrélations entre les lettres pour deviner la position des espaces.

Étape 3. Reprendre le fichier de la première étape et supprimer tous les espaces pour obtenir une suite de caractères notée $\{a_i\}_{i\leq N}$. À partir de $\{a_i\}_{i\leq N}$, calculer les statistiques de tous les motifs de 2 et de 4 caractères. On notera $\mathbb{P}_2$ et $\mathbb{P}_4$ les probabilités associées. La fonction

$$F : n \mapsto \log \frac{P_4\big(a_n, a_{n+1}, a_{n+2}, a_{n+3}\big)}{P_2\big(a_n, a_{n+1}\big) P_2\big(a_{n+2}, a_{n+3}\big)}$$

permet de tester la dépendance entre les caractères. En effet, si F a un minimum local en n, alors il est probable qu'un espace devrait figurer entre a_{n+1} et a_{n+2}. En utilisant cette heuristique, réintroduire des espaces dans le texte et tester ensuite la pertinence de cette procédure. On pourra ajouter un seuil pour ne sélectionner que les minima locaux suffisamment bas. Pour affiner la méthode, on peut aussi décider d'ajouter un espace en fonction de critères supplémentaires, par exemple en examinant aussi les valeurs

$$\frac{P_2\big(a_n, a_{n+1}\big) P_3\big(a_{n+2}, a_{n+3}, a_{n+4}\big)}{P_3\big(a_n, a_{n+1}, a_{n+2}\big) P_2\big(a_{n+3}, a_{n+4}\big)}.$$

Pour en savoir plus sur l'analyse statistique des signaux (langage, musique, images), on pourra consulter le livre :

D Mumford, A Desolneux, *Pattern theory: the stochastic analysis of real-world signals*, A. K Peters (2010)