Projet n°1 (synthèse de textes)

Le but de ce projet est de mettre en œuvre différentes techniques permettant de produire artificiellement des suites de caractères "imitant" un texte en français.

1. **Modèle par lettres d'ordre 0**. À partir du fichier goriot.txt (voir le site web du cours pour télécharger ce fichier et récupérer les commandes Scilab qui permettent de le lire), où de tout autre texte en français assez long (voir par exemple le site du projet Gutenberg, http://www.gutenberg.org), concevoir un programme qui calcule la distribution empirique des symboles courants rencontrés dans le texte. On se limitera aux minuscules non accentuées (caractère "a" à "z"), et à un seul symbole séparateur, l'espace (on pourra convertir les majuscules, minuscules accentués, etc. en symboles admissibles). À partir de cette distribution empirique, concevoir un programme qui synthétise un texte de 10 lignes (environ 800 symboles) comme réalisation successives de variables aléatoires i.i.d. suivant cette distribution. Calculer l'entropie moyenne par symbole de ce modèle.

2. **Modèle par lettres d'ordre 1**. On cherche maintenant à affiner le modèle précédent en tenant compte, lorsque l'on synthétise une nouvelle lettre, de la lettre précédente. On modélise donc un texte synthétique comme une suite de variables aléatoires $(X_1, X_2, \dots X_n)$ dont la distribution jointe de probabilité est donnée par

$$p(x_1, x_2, \dots x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\dots p(x_n|x_{n-1}), \tag{1}$$

où $p(x)$ est la distribution de chaque symbole (calculée précédemment) et $p(x_k|x_{k-1})$ est la distribution conditionnelle de X_k sachant X_{k-1} (cette distibution étant supposée indépendante de k), qui peut être estimée empiriquement à partir du texte d'apprentissage $(t_1, t_2, \dots t_n)$ par

$$\tilde{p}(u|v) = \frac{\tilde{p}(u,v)}{\tilde{p}(v)},$$

où

$$\tilde{p}(u,v) = \frac{1}{n-1}\,\#\{k \in \{2,\dots n\},\ t_k = u \text{ et } t_{k-1} = v\}$$

et

$$\tilde{p}(v) = \frac{1}{n-1}\,\#\{k \in \{2,\dots n\},\ t_{k-1} = v\}.$$

Concevoir un programme qui calcule successivement les tableaux $\tilde{p}(v)$, $\tilde{p}(u,v)$ et $\tilde{p}(u|v)$, puis un programme qui synthétise un texte selon la distribution (1), en tirant la première lettre selon $\tilde{p}(v)$ puis les lettres suivantes à l'aide de la distribution conditionnelle $\tilde{p}(u,v)$. Calculer l'entropie moyenne par symbole de ce modèle, à savoir

$$H(X_k|X_{k-1})$$

(on rappellera la définition de cette quantité avant d'en évaluer la valeur numérique). Commenter la différence obtenue avec la question précédente.

Comme pour la question précédente, synthétiser un texte de 10 lignes par cette méthode.

3. **Modèle par lettres d'ordre 2**. Pour obtenir un modèle encore plus précis, on peut tenir compte des dépendance par rapport aux deux lettres précédentes, ce qui revient à modéliser un texte synthétique comme une suite de variables aléatoires $(X_1, X_2, \ldots X_n)$ dont la distribution jointe est

$$p(x_1, x_2, \ldots x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_2, x_3)\ldots p(x_n|x_{n-2}, x_{n-1}). \quad (2)$$

Donner une formule permettant d'estimer empiriquement $p(u|v,w)$ sur le texte d'apprentissage, puis concevoir comme à la question précédente un programme qui réalise cette estimation, puis un autre qui synthétise des textes selon ce nouveau modèle. Donner une formule pour l'entropie moyenne par symbole de ce modèle, en calculer la valeur numérique et la commenter.

Comme pour les questions précédentes, synthétiser un texte de 10 lignes par cette méthode.

4. **Modèle par mots d'ordre 0**. On reprend la question 1, mais les variables aléatoires ne sont plus les lettres mais les mots. Ceci suppose que le texte d'apprentissage soit suffisamment riche, disons au minimum quelques centaines de milliers de lettres (quelques centaines de ko en texte brut, ie au format ASCII). Étant donné le nombre important de mots possibles à manipuler, plutôt que de travailler sur des tableaux comme pour les lettres, on va travailler directement sur le texte d'apprentissage lui-même. En effet, pour tirer un mot au hasard selon la distribution empirique du texte d'apprentissage, il suffit de calculer la suite $i_1, i_2, \ldots i_p$ de tous les indices de la première lettre de chaque mot dans le texte initial, puis de tirer au hasard un entier j entre 1 et p (qui correspondra donc au mot commençant à l'incice i_j dans le texte d'apprentissage). Concevoir un programme qui synthétise un texte selon ce modèle. Calculer ensuite l'entropie moyenne par symbole, définie par

$$H = \sum_{m \in M} -\frac{p(m)}{l(m)} \log_2 p(m),$$

où M est l'ensembme des mots, $p(m)$ la probabilité du mot m et $l(m)$ sa longueur (nombre de lettres).

(et non par mot !) de ce modèle, et la calculer (attention, cette question mérite réflexion, et sera évaluée en conséquence).

Comme pour les questions précédentes, synthétiser un texte de 10 lignes par cette méthode.

5. **Modèle par mots d'ordre 1**. On reprend le modèle d'ordre 1 pour les lettres (question 2), mais avec les mots. Pour effectuer la synthèse du mot x_k à partir du mot x_{k-1}, on repère toutes les occurences du mot x_{k-1} dans le texte d'apprentissage, on choisit au hasard (uniformément) l'une de ces occurences et l'on prend pour x_k le mot suivant cette occurence. Concevoir un programme qui synthétise un texte selon ce modèle. Définir enfin l'entropie moyenne par symbole de ce modèle, la calculer et commenter le résultat obtenu.

   Comme pour les questions précédentes, synthétiser un texte de 10 lignes par cette méthode.

6. **Influence des lettres accentuées**. Reprendre la question précédente, mais en autorisant les lettres accentuées (les mots "côté", "côte" et "cote", confondus à la question précédente, sont donc maintenant considérés comme distincts). Comparer les résultats obtenus avec ceux de la question précédente, du point de vue de la synthèse et du point de vue de l'entropie moyenne par symbole. Conclure.