Programmation fonctionnelle
Projet 3
Autocomplétion de texte

Le but de ce projet est de programmer une fonction qui, pour un début de phrase donné, suggère le mot suivant le plus probable. Ce type de fonctionnalité est utilisé pour aider la saisie de texte (sur un téléphone, dans une barre de recherche) et est aussi, dans une version plus élaborée, à la base du fonctionnement des grands modèles de langage.

Apprentissage Vous devrez définir une fonction qui prend un texte en entrée (de préférence long, et semblable au type de texte que l'on souhaite produire ensuite) et génère la distribution de probabilité dans ce texte : pour chaque mot qui apparaît dans le texte, on doit calculer la probabilité qu'un autre mot apparaisse à la suite. Par exemple dans ce paragraphe, la probabilité que le mot "la" soit suivi du mot "probabilité" est de 40%, 20% pour qu'il soit suivi de "distribution", "suite" ou "soit", et 0% pour les autres mots.

On obtiendra donc un dictionnaire, qui aura pour clés des mots et pour valeurs des distributions de probabilité. Pour améliorer la cohérence des suggestions, on peut calculer la probabilité d'un mot suivant non pas pour chaque mot mais pour chaque paire ou triplet de mots (un *n-gramme*), ce qui permettra de donner plus de contexte aux suggestions.

Stockage des suggestions Il faudra stocker les probabilités dans une structure de données adaptée. Pour cela vous devrez implémenter un *trie* (aussi appelé *arbre préfixe*), un arbre où chaque arc est décoré par un caractère, les caractères permettant d'accéder rapidement aux nœuds où sont stockés les valeurs proprement dîtes.

Fonctionnalités à implémenter

- fonctions pour construire un *trie*, y insérer des associations mot/valeur, accéder aux valeurs pour un mot donné
- fonction qui prend en entrée un texte et apprend sa distribution de probabilité pour chaque mot ou n-gramme.

— fonction qui prend une phrase incomplète, une distribution de probabilité, et retourne le mot suivant le plus probable (ou bien les 3 mots les plus probables).

Pour aller plus loin Cette partie est optionnelle : vous pouvez utiliser la fonction précédente pour coder un générateur automatique de texte (les textes produits n'auront aucun sens, mais ils auront l'apparence d'un texte à peu près correct grammaticalement).