Générateur de noms par chaîne de Markov

Prérequis

- Python 3
- Aucune dépendance

Objectif

Concevoir un algorithme générateur de noms reposant sur le procédé des chaînes de Markov.

Explication

Une chaîne de Markov est un processus stochastique selon lequel la prédiction d'un événement futur ne dépend que de l'état présent du processus et ne dépend pas des états précédents : le système est dit *sans mémoire*, l'historique des prédictions précédentes n'a aucun impact sur les prédictions futures. Un exemple d'application parmi d'autres est la génération de mots.

Cependant, ce modèle *sans mémoire* (dit d'ordre 0) est assez limité et génère des mots qui ne représentent aucune réalité alphabétique. Des modèles de Markov d'ordre élevé permettent de prendre en compte une partie de l'historique dans la génération de la prédiction. Dans le contexte de la génération de mots, l'ordre d'une chaîne de Markov est le nombre de lettres déjà prédites prises en compte pour la prédiction de la lettre suivante.

Exemple: Un mot commence par la lettre 'q'. L'état actuel est la lettre 'u' (le mot en cours de génération est donc 'qu'). Un modèle d'ordre 0 pourrait prédire la lettre 't' après un 'u' car il existe une probabilité non négligeable dans la langue française que la lettre 't' suive la lettre 'u'. Le mot généré serait alors 'qut' qui aboutirait après quelques prédictions supplémentaires à un mot décousu sans aucun sens. Un modèle d'ordre 1 utiliserait en plus l'état précédent (la lettre 'q') dans la génération et prendrait ainsi en compte la paire 'qu' pour prédire la lettre suivante. Il est alors fort probable que des lettres comme 'a', 'e' ou 'i' soient prédite et quasi-improbable qu'une consonne le soit. Un ordre élevé (mais pas trop) augmente la cohérence du résultat tout en assurant une part d'aléatoire suffisante.

Déroulement

Pour créer un générateur de noms, il faut donc connaître l'état actuel (une lettre) et disposer de la fréquence d'occurrence des lettres dans la langue française pour prédire la lettre suivante. Dans ce travail, on cherche à spécialiser un modèle de Markov pour la génération de nom de ville française. L'INSEE tient à jour sur son site internet la liste des noms de toutes les communes françaises en Métropole et Outre-Mer. Grâce à plus de 36 000 noms différents, ce document constitue une base de données intéressante pour constituer un générateur de nom de commune.

- Télécharger cette archive, y extraire le fichier *communes-01042019.csv* et le parser.
- Construire le tableau d'occurrence des lettres pour ce jeu de données.
- Programmer un algorithme de génération de nom de commune française en utilisant le procédé de chaînes de Markov.
- Proposer une liste amusante de quelques noms générés de communes très ... "rurales".
  ex : Saint Juliets-en-son, Villedieu-d'Espisseleu, Courteuil-aux-choux

Aide

- Les modules *csv*, *random* et *string* seront d'un grand secours... Les dictionnaires aussi...
- Il faut probablement donner une longueur limite à la chaîne de caractère générée.
- Il y a un juste milieu à trouver dans le choix de l'ordre de la chaîne de Markov. Un ordre trop petit donnera un nom absurde, un ordre trop grand ne fera que retrouver un nom de commune déjà existant.
- Veillez à exclure de la génération les noms de communes issues du jeu de données.
- Libre à vous de spécialiser votre algorithme sur d'autres jeux de données (génération de prénoms, de noms de planètes, d'aliments...) sous réserve d'avoir une base de données suffisante.