Travaux Dirigés d'informatique linguistique n°5

Cours d'Introduction à l'informatique linguistique

—Licence Informatique 3ème année—

---

n-grammes

Dans ce TP, nous allons manipuler la notion de n-grammes.

---

1 Ressources

Pour cette séance, nous vous fournissons un certain nombre de ressources se trouvant dans le répertoire suivant :
`http://igm.univ-mlv.fr/~mconstan/enseignement/l3/infolingu/data/`

- `ngrams.py`, un module python qui permet de manipuler les n-grammes.
- `tagging.py`, un module python qui permet de charger et manipuler des textes étiquetés au format `Treetagger`,
- `suzanne.zip`, une archive de textes étiquetés au format `Treetagger`.

2 n-grammes sur lettres

Dans cet exercice, l'unité minimale de traitement est le caractère.

1. Écrire une fonction qui prend comme paramètre un nom de fichier texte. Cette fonction doit stocker et compter les unigrammes de lettres de ce fichier. Vous vous aiderez notamment de la classe `ngrams`.
2. Écrire une autre fonction qui stocke et compte les bigrammes de lettres de ce fichier.
3. Écrire encore une autre fonction qui stocke et compte les trigrammes de lettres de ce fichier.
4. Écrire une fonction qui prend comme paramètres une chaîne de caractères `prefix` et trois éléments de type `ngrams` représentant les unigrammes, bigrammes et trigrammes calculés dans les trois questions précédentes. Le paramètre `prefix` correspond à une séquence de lettres déjà tapées par un utilisateur. La fonction doit renvoyer la lettre suivante la plus probable étant donné `prefix`.
5. Étant donne un mot avec une lettre manquante marquée par le symbole `*`, écrire une fonction qui permet de prédire cette lettre.

3 n-grammes sur les étiquettes grammaticales

Dans cet exercice, l'unité minimale de traitement est l'étiquette grammaticale d'un mot.

1. Stocker et compter les unigrammes, bigrammes et trigrammes d'étiquettes grammaticales à partir de quelques textes étiquetés de l'archive `suzanne.zip`.
2. Étant donné une séquence de deux étiquettes, prédire l'étiquette suivante la plus probable.

4 Étiquetage morphosyntaxique simple

1. Pour chaque mot de l'archive de textes étiquetés, calculer son étiquette la plus fréquente
2. Écrire une fonction qui prend un texte en entrée et assigne à chacun de ses mots son étiquette la plus fréquente dans l'archive de textes étiquetés.
3. Évaluer automatiquement votre étiquetage à l'aide de la méthode écrite à cet effet dans le TP précédent.