TP HMM et Viterbi

Objectif : créer un détecteur d'entités nommées en utilisant un HMM (modèle de Markov caché). Les états représenteront les étiquettes des entités. Les observations seront les mots.

Données : 2 corpus contenant des transcriptions automatiques d'émission de radio. Ces transcriptions sont annotées en entités nommées. Elles sont disponibles à cette adresse : https://cloud-ic2.univ-lemans.fr/s/5Di4RYHwD559Br7

- Un corpus d'apprentissage : train.ester1.cut.bio
- Un corpus de développement : dev.ester1.cut.bio

À rendre : un notebook python contenant vos travaux commentés.

Les données sont organisées en colonne. La première colonne contient le mot transcrit, la seconde l'étiquette décrivant l'entité nommée.

Question 1 : statistiques sur le corpus d'apprentissage
- Combien de mots différents existe-t-il ?
- Quels sont les 10 mots les plus fréquents et les 10 mots les moins fréquents ?
- Combien d'étiquettes BIO différentes existe-t-il ? Donnez la fréquence de chaque étiquette.
- Combien d'étiquettes « begin » BIO différentes existe-t-il ?
- Combien d'étiquettes « inner » BIO différentes existe-t-il ?

Question 2 : Apprentissage
- Décrivez les paramètres de l'HMM
- Calculer les log probabilités initiales, les log probabilités de transition, les log probabilités d'émission. Les transitions et les mots rares dont la fréquence est inférieure à 5 (à vérifier) devront être lissés en utilisant la méthode présentée en cours d'initiation à la recherche.

Question 3 : Avec les log probabilistes, écrire l'algorithme de Viterbi et l'appliquer au corpus de développement. Attention, penser à retirer la colonne des étiquettes ! Celle-ci servira à évaluer la performance du système.
Évaluer la performance du système en adaptant les méthodes du script suivant : https://cloud-ic2.univ-lemans.fr/s/8KLCb8oFY2xGJEj

Question 4 : Développer un second système qui répond pour chaque mot l'étiquette la plus probable. Évaluer et comparer les résultats avec les systèmes Viterbi.