Université de Bouira
Département d'Informatique
Module : Traitement Automatique des Langues
Niveau M1-ISIL & IA

Série TP 3

Analyse lexicale – Exercices.

Exercice 1

Soit le corpus C annoté (POS taggé) en anglais. Chaque mot est associé à une étiquette morphosyntaxique (POS tag) :

Book a car	**V DT N**	Book/V a/DT car/N
Park the car	**V DT N**	Park/V the/DT car/N
The book is in the car	**DT N V P DT N**	The/DT book/N is/V in/P the/DT car/N
The car is in a park	**DT N V P DT N**	The/DT car/N is/V in/P a/DT park/N
Book bag	**N N**	Book/V a/DT car/N

Un modèle de Markov caché (HMM) pour le POS Tagging repose sur deux types de probabilités :

- Probabilités de transition : La probabilité de passer d'un POS tag à un autre.
- Probabilités d'émission : La probabilité qu'un mot soit généré par un POS tag donné.

À partir du corpus C, l'entrainer afin de :

1. Calculer les probabilités (table) de transition entre les POS tags.
2. Calculer les probabilités (table) d'émission pour chaque mot du corpus.
3. Calculer les probabilités initiales.

Vous devez maintenant étiqueter une nouvelle séquence de mots en utilisant le modèle HMM que vous avez appris : « **Book the park** »

4. Utilisez l'algorithme de Viterbi pour déterminer la séquence de POS tags la plus probable pour cette phrase.

Exercice 2

Dérouler l'algorithme Soundex pour les mots suivants : Robert, Algeria, Processing. Avec :

- 1 = B, F, P, V
- 2 = C, G, J, K, Q, S, X, Z
- 3 = D, T
- 4 = L
- 5 = M, N
- 6 = R