APPLICATIONS DU TRAITEMENT AUTOMATIQUE DES LANGUES

TD3 : Analyse statistique – Lissage des modèles N-gramme

Exercice 1 :

Considérer le corpus d'apprentissage suivant :

Training data:
I am Sam. Sam I am. Sam I like. Sam I do like. Do I like Sam.

Nous utilisons un modèle bi-gramme avec un lissage de Laplace.

1. Donner les probabilités bi-grammes suivantes estimées par ce modèle:
   - P(do/<s>)
   - P(do/Sam)
   - P(Sam/<s>)
   - P(Sam/do)
   - P(I/Sam)
   - P(I/do)
   - P(like/I)

2. Calculez les probabilités des séquences suivantes selon le modèle. Laquelle des deux séquences est la plus probable selon notre estimateur?
   - <s> do Sam I like
   - <s> Sam do I like

Exercice 2 :

Soit le corpus d'apprentissage suivant :

« he is he is he is going abroad is going to study in the field »

Calculer P(is going abroad to study) avec :

1. Le modèle bi-grams sans lissage ?
2. Le modèle bi-grams avec lissage de Laplace?
3. Le modèle bi-grams avec lissage de Good Turing?