APPLICATIONS DU TRAITEMENT AUTOMATIQUE DES LANGUES

TD2 : Analyse statistique – Modèle 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.

1. Normaliser le corpus
2. En utilisant un modèle 2-gramme, quel est le mot suivant le plus probable prédit par le modèle pour les séquences de mots suivantes :
   - <s> Sam . . .
   - <s> Sam I do . . .
   - <s> Sam I am Sam . . .
   - <s> do I like . . .

3. Laquelle des phrases suivantes est la meilleure, c'est-à-dire, aura une probabilité plus élevée avec le modèle

   - <s> Sam I do I like </s>
   - <s> Sam I am </s>
   - <s> I do like Sam I am </s>

Exercice 2 :

Considérons à nouveau les mêmes données d'apprentissage et le même modèle bi-gramme de l'exercice 1.

Calculer la perplexité de :

<s> I do like Sam