HAUTE ÉCOLE
D'INGÉNIERIE ET DE GESTION
DU CANTON DE VAUD
www.heig-vd.ch

Haute Ecole Spécialisée de Suisse occidentale
Fachhochschule Westschweiz
University of Applied Sciences and Arts Western Switzerland

Cours TAL – Labo 8 : modèles de langage et *Cloze test*

Distribué le mardi 4 juin 2019

Objectif et plan

L'objectif de ce labo est d'entraîner des modèles de langues (EN : *language models*) sous NLTK en utilisant le package **nltk.lm**. Les modèles seront entraînés sur des romans de Balzac en français (fournis par le projet Gutenberg), et leurs performances seront mesurées par leur perplexité sur de nouveaux textes. Les modèles serviront également à deviner des mots cachés dans un texte, et ici leurs performances seront comparées à celles des humains sur cette même tâche.

1. Obtenir les données

- Télécharger les dix romans de la *Comédie humaine* de Balzac en français, enlever les notices en anglais au début et à la fin, puis les assembler en un corpus.
- Utiliser une procédure à laquelle on donnera la liste de noms de fichiers ou des URLs.
- Sauvegarder le texte résultant : quelle taille fait-il en Ko ou Mo?
- Extraire un fragment (environ 2000 mots) qui servira de donnée de test (pas le bon format, utiliser C.H. 10).
- Segmenter et tokenizer chacun des deux corpus (train/test) en une *liste de listes* de mots, y compris les ponctuations (une liste = une phrase).
- Sauvegarder aussi chaque liste dans un fichier *pickle*.

2. Entraîner un premier modèle de langage de NLTK

- En utilisant les instructions disponibles pour le module NLTK LM, entraîner un modèle de langage sur les données d'entraînement. Attention, le package lm n'est disponible qu'à partir de NLTK version 3.4.
- Commencer avec un modèle MLE à l'ordre 2, comme montré dans le tutoriel.
- Puis, par exemple, aller jusqu'à l'ordre 3 ou 4, avec l'algorithme de Laplace (ou de Kneser-Ney).
- Calculer la perplexité du modèle sur l'ensemble de test.
- Générer quelques phrases dans le style de Balzac selon les explications de NLTK.

3. Entraîner un second modèle de langage de NLTK

- En utilisant par exemple le modèle de Laplace.
- Le tester aussi sur le corpus de test et comparer les scores.

4. Cloze test

- Supprimer un mot sur 7 dans le corpus de test.
- En utilisant la méthode **lm.generate** avec du contexte, demander au modèle de langage de prédire ces mots (en utilisant donc 2-4 mots précédents). Quelle est la performance du système ? Comparer les modèles générés en (2) et en (3).
- Générez un texte avec les mots en question (1 mot sur 7) remplacés par des blancs, et passez-le à un autre binôme : quelle est leur performance pour deviner les mots ?

Merci d'envoyer votre notebook Jupyter par email au professeur avant le **vendredi 14 juin à 23h59**.