ECAM
BRUSSELS ENGINEERING SCHOOL

Winter 2020
Sébastien Combéfis

AI611μ Word Prediction with N-Grams Model using Python

Mission 2: Bigram model training with `nltk`

This assessment evaluates the following competencies:

- *AI201 – Train an N-Grams model from a given text corpus* *(+1)*
- *AI501 Write an application that solves the word prediction problem with N-Grams models* *(+1)*
- *AI102 Formally describe N-Grams models thanks to probabilities* *(+1)*
- *AI103 Preprocess a corpus and compute basic statistics on it* *(+1)*

You may also be assessed on the following competencies:

- *AI502 – Evaluate the quality of a given N-Grams model* *(+2)*

In this mission, you have to use the `nltk` Python module to train a bigram model for a given corpus. You have to directly use the `MLE` object defined in the `nltk.lm` module to train a bigram model. To succeed the mission, you have to:

1. Write a program that train a bigram model for a given sentence.
2. Print one probability of your model and compare with results obtained by hand.
3. Present to the teacher your program and how it works and make a demonstration.

For example, for the `text = 'i love chinese food.  chinese people love food.'` sentence, if you print the probability `model.score('chinese', ['love'])` obtained thanks to the `model` variable which is a `MLE` object, you should obtain `0.5`.

Optionally, you are asked to compute a trigram model, and why not a 4-grams model, and compare the quality of the obtained models. Think about using a test sentence not in the training set, or compute the perplexity for one test sentence.