ECAM
BRUSSELS ENGINEERING SCHOOL

Winter 2020
Sébastien Combéfis

AI611μ Word Prediction with N-Grams Model using Python

Quizz 3: N-Grams model training and evaluation

This assessment evaluates the following competencies:

- *AI101 – Understand the N-Grams model* *(+1)*
- *AI102 – Formally describe N-Grams models thanks to probabilities* *(+1)*
- *AI502 – Evaluate the quality of a given N-Grams model* *(+1)*

Three affirmations are given for each assessed competency. For each of them, you have to decide whether it is true or false. To get a star for the competency, you must have the correct answer for the three affirmations.

AI101	**True**	**False**
With an N-grams model, it is not possible to estimate the probability of a sentence whose some words are not in the vocabulary of the training set.	☐	☐
An N-grams model can capture syntactic facts of the language such as the fact that a determinant is often followed by a noun.	☐	☐
It is possible to design an N-grams model without having to analyse a corpus of texts.	☐	☐

AI102	**True**	**False**
If we use a bigram model, the probability $P(w_2\|w_0w_1)$ is equal to $P(w_2\|w_0)$.	☐	☐
The bigram Laplace smoothing consists in adding 1 to all the zero counts before normalising by unigram counts to get the bigram probabilities.	☐	☐
The probability of the sentence $w_0w_1w_2$ following a trigram model is obtained with the following formula: $P(w_0w_1w_2) = P(w_0)P(w_1\|w_0)P(w_2\|w_0w_1)$.	☐	☐

AI502	**True**	**False**
A lower perplexity measured on a test set indicates a better N-grams model, according to the test set.	☐	☐
An intrinsic evaluation is completely independent of the language model and measures if it is well designed for the task at hand of the user application.	☐	☐
An N-grams model generally better models the training corpus as N decreases.	☐	☐