ECAM
BRUSSELS ENGINEERING SCHOOL

Winter 2020
Sébastien Combéfis

AI611μ Word Prediction with N-Grams Model using Python

Coding 1: Corpus statistics

This assessment evaluates the following competencies:

- *AI101 – Understand the N-Grams model* *(+1)*
- *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)*
- *AI103 – Preprocess a corpus and compute basic statistics on it* *(+2)*

In this coding assessment, you have to complete an existing Python program that analyses a text file, whose path is passed as a command-line argument, and produce a list of unique words present in the text with their corresponding number of occurrence [1]. To succeed the assessment, you have to:

1. Complete the program to make it produce the intended result.
2. Explain to the teacher how you designed your code and make a demonstration.

You can assume that the text file will only contain english words and you can ignore punctuation signs, just considering that they are word delimiters. You have to choose the format of the result produced by your program (JSON document, simple flat file, SQL queries, etc.) and it should be either printed on the standard output or saved in a text file, depending on the optional `out` argument.

[1] The code can be found here: `https://github.com/ukonline/uCourse/blob/master/AI611%C2%B5/code/corpustats.py`