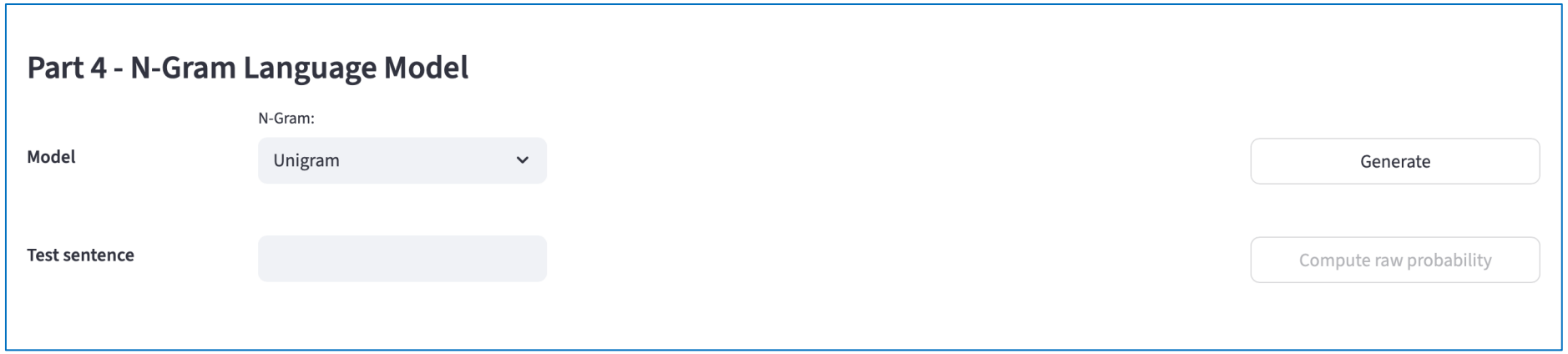
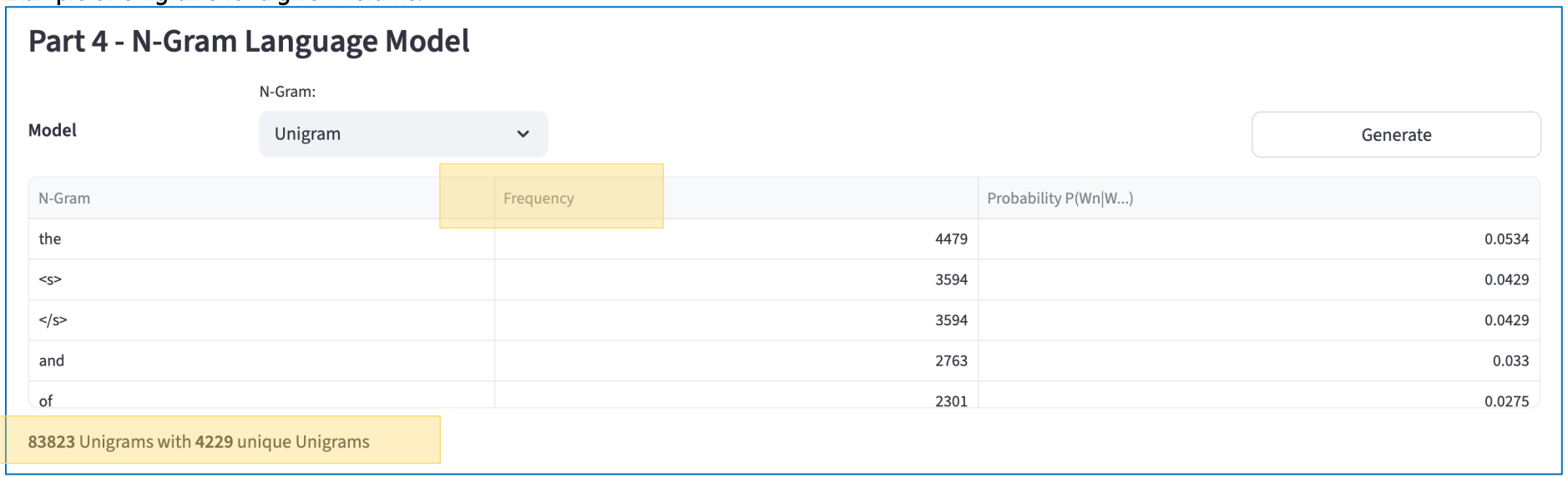
I. Generate Unigram, Bigram and Trigram:

For each Article (or a given Volume, or all the Corpus of articles), generate the Unigrams, Bigrams and Trigrams with or without normalization (Porter, Lancaster and Snowball).



Example of Unigrams for a given Volume:



</s> is either a ".", "?" or "!" (see the example provided with TP n°1)

Example of Trigrams for a given Volume:

The probabilities are estimated based on Markov Hypothesis and using Chain Rule, as mentioned below :

$$P(w_{1:n}) = \prod_{k=1}^{n} P(w_k|w_{k-1})$$

For Unigram :

$$P(w_n) = \frac{C(w_n)}{N}$$

For Bigram :

$$P(w_n|w_{n-1}) = \frac{C(w_{n-1}\, w_n)}{\sum_w C(w_{n-1}\, w)} = \frac{C(w_{n-1}\, w_n)}{C(w_{n-1})}$$

For N-Gram :

$$P(w_n|w_{n-N+1\,:\,n-1}) = \frac{C(w_{n-N+1\,:\,n-1}\, w_n)}{C(w_{n-N+1\,:\,n-1})}$$

Where:

w_n : is a token.

$C(w_n)$: is the number of occurrences of the unigram w_n.

N: indicates the number of tokens.

$C(w_{n-1}\, w_n)$: is number of occurrences of the bigram $w_{n-1}\, w_n$.

$\sum_w C(w_{n-1}\, w)$: is number of occurrences of bigrams starting with w_{n-1}.