

Basic symbols

V : Vocabulary (set of types) or its size

N : size of the training text (number of tokens)

$C(w)$: count (frequency) of $w \in V$

K : size of test text

Sanity checks:

$$\sum_{w \in V} C(w) = N$$

$$\sum_{w \in V} P(w) = 1$$

Perplexity:

$$\begin{aligned}(\prod_{i=1}^K P(w_i | h))^{\frac{-1}{K}} &= \sqrt[K]{\frac{1}{\prod_{i=1}^K P(w_i | h)}} \\ &= 2^{-\frac{1}{K} \sum_{i=1}^K \lg P(w_i | h)} \\ &= B^{-\frac{1}{K} \sum_{i=1}^K \log_B P(w_i | h)}\end{aligned}$$

Model families:

Maximum likelihood
estimation

$$P_{RF}(w) = \frac{C(w)}{N}$$

Laplace (add-one)
smoothing

$$P_L(w) = \frac{C(w) + 1}{N + V} = \frac{C_L^*(w)}{N}$$

$$C_L^* = (C(w) + 1) \cdot \frac{N}{N + V}$$

Constant

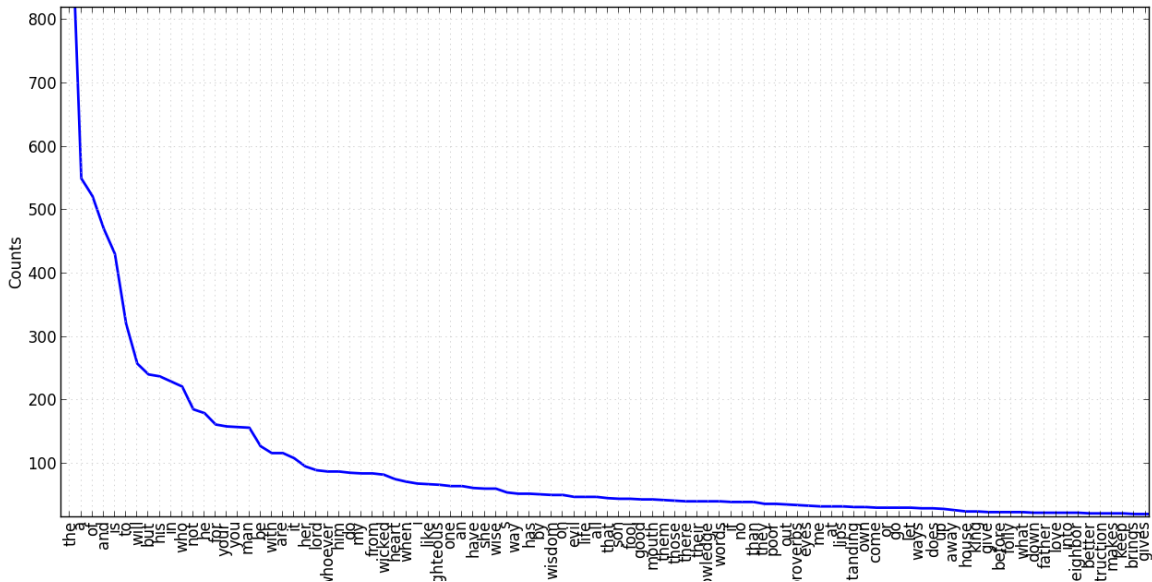
$$P_{\text{const}}(w) = \frac{1}{V} = \frac{C_{\text{const}}^*}{N}$$

$$C_{\text{const}}^*(w) = \frac{N}{V}$$

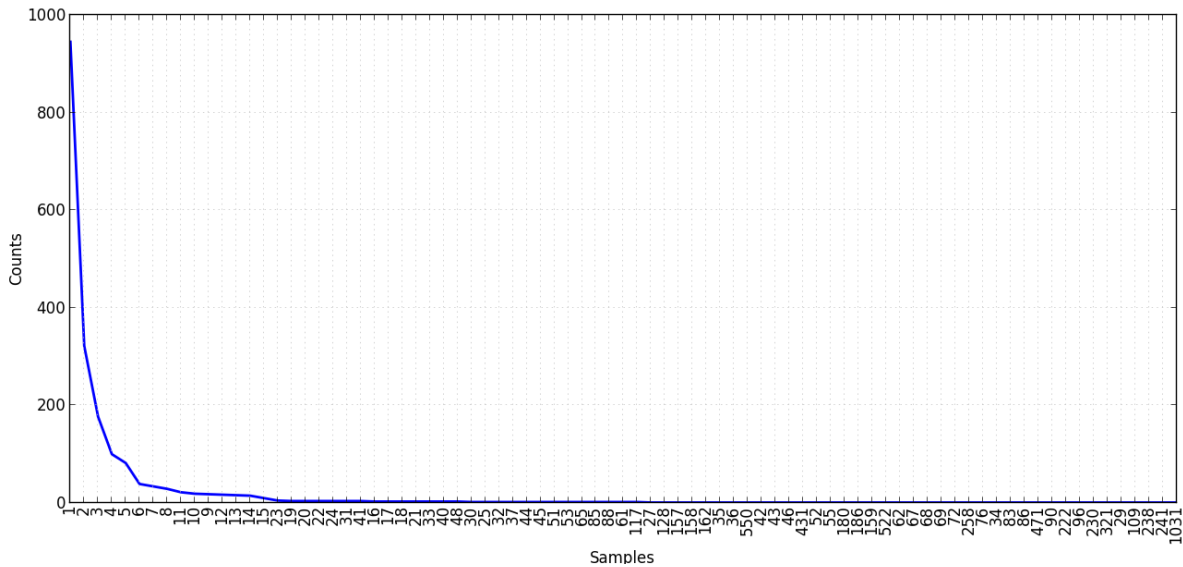
Frequencies:



Frequencies, zoomed:



Frequencies of frequencies:



Frequencies of frequencies, zoomed:

