Below are some formulas that we can use to assign a document into one of two categories (authors, topics, languages, etc) $A$ or $B$.

**Word rates** The rate of occurrence of a word $w_i$ in all known texts of type $A$ is

$$\mu_{Ai} = \frac{\text{number of occurrences of } w_i \text{ in all known texts of type } A}{\text{total number of words in all known texts of type } A}$$

Similarly, we get $\mu_{Bi}$ for all known texts of category $B$.

**Selecting words with high discriminatory power** To decide if word $w_i$ should be used for discrimination, we can use the measure

$$z_i = \frac{(\mu_{Ai} - \mu_{Bi})^2}{\mu_{Ai} + \mu_{Bi}}$$

High values indicate a good discriminator.

**Deciding between the two categories** If the evidence in the document that we are assigning is $\mathcal{E}$ and we want to select between the hypotheses $A$ (the document belongs to category $A$) and $B$ (the document belongs to category $B$), the Bayesian approach specifies that

$$\frac{P(A|\mathcal{E})}{P(B|\mathcal{E})} = \frac{P(A)}{P(B)} \cdot \frac{P(\mathcal{E}|A)}{P(\mathcal{E}|B)}$$

$P(\mathcal{E}|A)$ are the final odds, $P(A)$ are the prior odds (independent of the evidence $\mathcal{E}$, and $P(\mathcal{E}|A)$ is the likelihood ratio.

Using logarithms (ln stands for the natural logarithm with base $e = 2.718 \ldots$):

$$\ln \frac{P(A|\mathcal{E})}{P(B|\mathcal{E})} = \ln \frac{P(A)}{P(B)} + \ln \frac{P(\mathcal{E}|A)}{P(\mathcal{E}|B)}$$

Values of the ratios above 1 or of their logarithms above 0 favor category $A$. Values below 1 (or 0 for the logarithms) favor category $B$. Values exactly equal to 1 (or 0) indicate no preference for any of the two categories.

**The Poisson distribution** The probability of a word $w_i$ appearing $x_i$ times in a document with total length $s$ words, under the hypothesis that the mean rate of occurrence of $w_i$ in the document is $\mu$ is

$$P(x_i) = \frac{(\mu s)^{x_i} \cdot e^{-\mu s}}{x_i!}$$
The likelihood ratio under the assumption of Poisson distribution  The likelihood ratio from evidence on one word $w_i$ appearing $x_i$ times in the new document of total length $s$ words is

$$K_i = \left( \frac{\mu_{A_i}}{\mu_{B_i}} \right)^{x_i} e^{-s(\mu_{A_i} - \mu_{B_i})}$$

The log-likelihood from word $w_i$ is

$$l_i(x_i) = x_i \ln \left( \frac{\mu_{A_i}}{\mu_{B_i}} \right) - s(\mu_{A_i} - \mu_{B_i})$$

and from many words $w_1, w_2, \ldots, w_n$

$$\sum_{i=1}^{n} l_i(x_i) = \sum_{i=1}^{n} [x_i \ln \left( \frac{\mu_{A_i}}{\mu_{B_i}} \right) - s(\mu_{A_i} - \mu_{B_i})]$$