

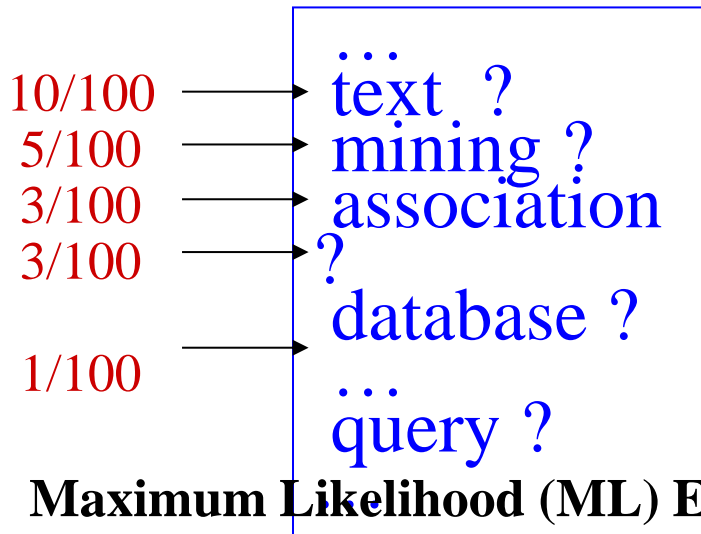
Estimation of Unigram LM

Unigram LM $p(w|\theta)=?$

Estimation

Text Mining Paper d

Total #words=**100**



Maximum Likelihood (ML) Estimator:

$$p(w | \theta) = p(w | d) = \frac{c(w, d)}{|d|}$$



A blue box containing the word counts from a document. The words and their counts are listed in white: 'text 10', 'mining 5', 'association 3', 'database 3', 'algorithm 2', 'query 1', and 'efficient 1'. The box has a folded corner at the bottom right.

Is this the best estimate?