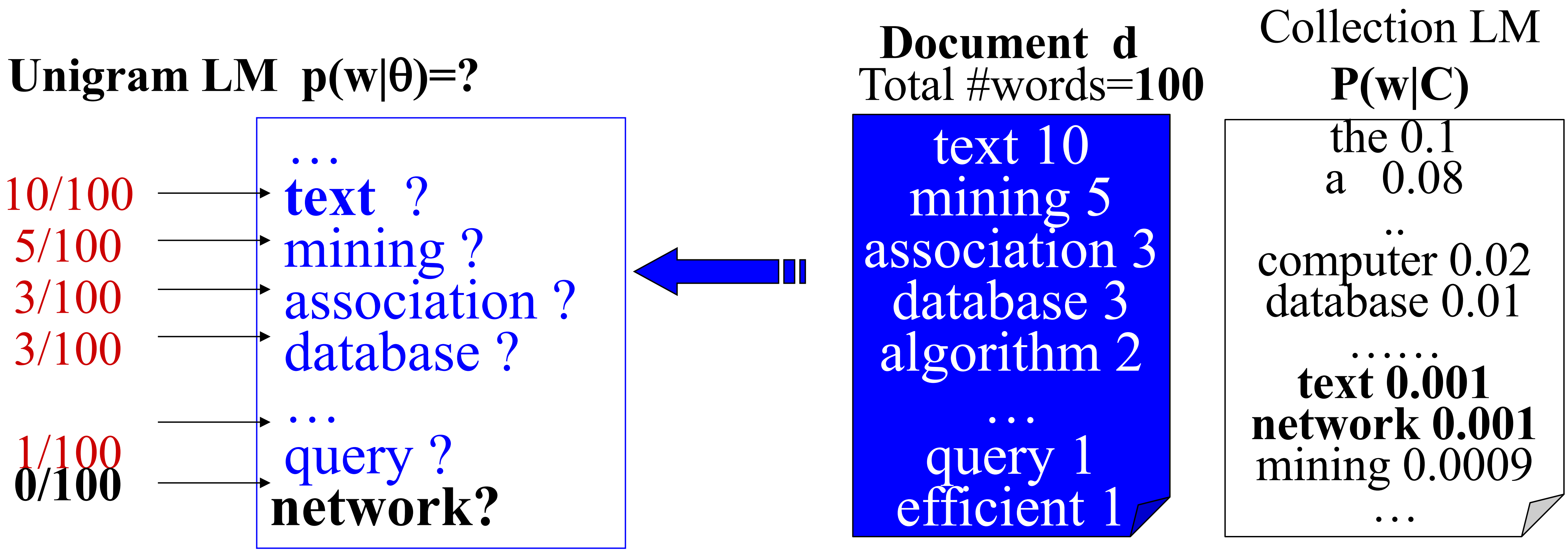
Dirichlet Prior (Bayesian) Smoothing



Unigram LM p(w|θ)=?

10/100 → **text** ?

5/100 → mining ?

3/100 → association ?

3/100 → database ?

→ …

1/100 → query ?

0/100 → **network?**

Document d

Total #words=**100**

text 10
mining 5
association 3
database 3
algorithm 2
…
query 1
efficient 1

Collection LM

P(w|C)

the 0.1
a 0.08
..
computer 0.02
database 0.01
……
text 0.001
network 0.001
mining 0.0009
…

$$p\,(w\,|\,d) = \frac{c(w;d) + \mu\, p(w\,|\,C)}{d\,|+\mu} = \frac{|\,d\,|}{|\,d\,|+\mu}\,\frac{c(w,d)}{|\,d\,|} + \frac{\mu}{|\,d\,|+\mu}\,p(w\,|\,C)$$

$$\mu \in [0,+\infty)$$

$$p(\text{"text"}\,|\,d) = \frac{10 + \mu * 0.001}{100 + \mu} \qquad p(\text{"network"}\,|\,d) = \frac{\mu}{100+\mu} * 0.001$$