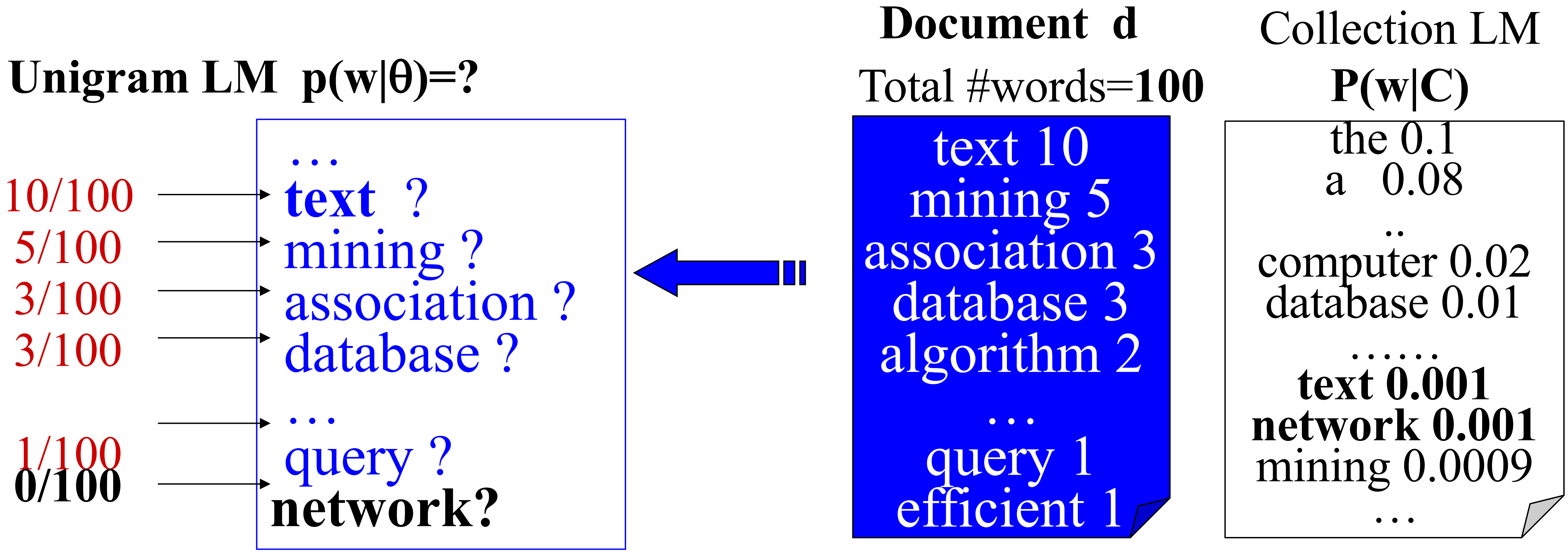
Linear Interpolation (Jelinek-Mercer) Smoothing



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
…

$$p(w\,|\,d) = (1-\lambda)\frac{c(w,d)}{|d|} + \lambda\, p(w\,|\,C)$$

$$\lambda \in [0,1]$$

$$p(\text{"text"}\,|\,d) = (1-\lambda)\frac{10}{100} + \lambda * 0.001$$

$$p(w\,|\,d) = (1-\lambda)\frac{c(w,d)}{|d|} + \lambda\, p(w\,|\,C)$$