Linear Interpolation (Jelinek-Mercer) Smoothing

Unigram LM p(w|θ)=?
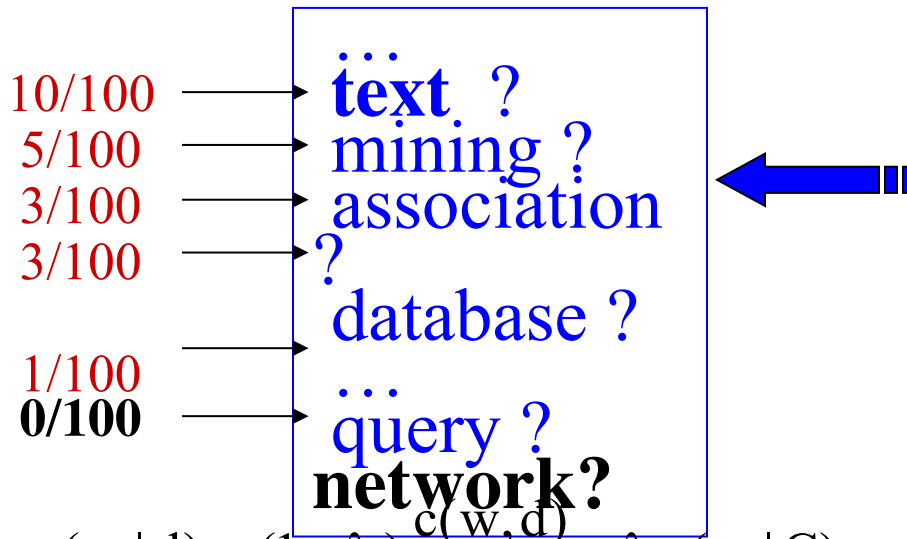
Document d

Total #words=**100**

Collection LM
P(w|C)

10/100
5/100
3/100
3/100

1/100
0/100

...
text ?
mining ?
association ?
database ?
...
query ?
network?

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

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

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

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

$\lambda \in [0,1]$

$$p(\text{"network"} \mid d) = \lambda * 0.001$$

4