Another Reason for Smoothing

Content words

Query = "the	algorithms	for	data	mining"
$p_{DML}(w\|d1)$:	*0.04*	*0.001*	*0.02*	*0.002*	*0.003*
$p_{DML}(w\|d2)$:	*0.02*	*0.001*	*0.01*	*0.003*	*0.004*

p("algorithms"|d1) = p("algorithm"|d2)
p("data"|d1) < p("data"|d2)
p("mining"|d1) < p("mining"|d2)

→ Intuitively, d2 should have a higher score, but p(q|d1)>p(q|d2)…

So we should make p("the") and p("for") less different for all docs, and smoothing helps achieve this goal…

After smoothing with $p(w|d) = 0.1 p_{DML}(w|d) + 0.9 p(w|REF)$, $p(q|d1) < p(q|d2)$!

Query	= "the	algorithms	for	data	mining"
P(w|REF)	0.2	0.00001	0.2	0.00001	0.00001
Smoothed p(w|d1):	0.184	0.000109	0.182	0.000209	0.000309
Smoothed p(w|d2):	0.182	0.000109	0.181	0.000309	0.000409