Assignment #3

B05902120 / Yu-Ting, TSENG

Apr 26, 2019

Smoothing

Given a document collection with a vocabulary of $w_1, ..., w_6$, and a query $q = w_1 w_3$, we are planning to rank two documents d_1 and d_2 by using language model (LM). The following table shows the statistical information about the word counts for d_1, d_2 and REF, which is a reference corpus for smoothing.

Word Count	d_1	d_2	REF
w_1	2	7	8000
w_2	0	1	100
w_3	3	1	1000
w_4	1	1	400
w_5	1	0	200
w_6	3	0	300
sum	10	10	10000

If we do not smooth the LM, which document would be ranked higher? Show your calculation.

$P(q|d_1) = P(w_1 w_3|d_1) = P(w_1|d_1) * P(w_3|d_1) = \frac{2}{10} * \frac{3}{10} = 0.06$
$P(q|d_2) = P(w_1 w_3|d_2) = P(w_1|d_2) * P(w_3|d_2) = \frac{7}{10} * \frac{1}{10} = 0.07$
Document d_2 would be ranked higher if we do not smooth the LM.

If we smooth the LM with the Dirichlet prior smoothing and set $\mu = 10$, which document would be ranked higher? Show your calculation.

$P'(w_1|d_1) = \frac{|d_1|}{|d_1|+\mu} P(w_1|d_1) + \frac{\mu}{|d_1|+\mu} P(w_1|REF) = \frac{10}{10+10}\frac{2}{10} + \frac{10}{10+10}\frac{8000}{10000} = 0.10+0.40 = 0.50$
$P'(w_3|d_1) = \frac{|d_1|}{|d_1|+\mu} P(w_3|d_1) + \frac{\mu}{|d_1|+\mu} P(w_3|REF) = \frac{10}{10+10}\frac{3}{10} + \frac{10}{10+10}\frac{1000}{10000} = 0.15+0.05 = 0.20$
$P(q|d_1) = P(w_1 w_3|d_1) = P(w_1|d_1) * P(w_3|d_1) = 0.50 * 0.20 = 0.10$

$P'(w_1|d_2) = \frac{|d_2|}{|d_2|+\mu} P(w_1|d_2) + \frac{\mu}{|d_2|+\mu} P(w_1|REF) = \frac{10}{10+10}\frac{7}{10} + \frac{10}{10+10}\frac{8000}{10000} = 0.35+0.40 = 0.75$
$P'(w_3|d_2) = \frac{|d_2|}{|d_2|+\mu} P(w_3|d_2) + \frac{\mu}{|d_2|+\mu} P(w_3|REF) = \frac{10}{10+10}\frac{1}{10} + \frac{10}{10+10}\frac{1000}{10000} = 0.05+0.05 = 0.10$

$P(q|d_2) = P(w_1w_3|d_2) = P(w_1|d_2) * P(w_3|d_2) = 0.75 * 0.10 = 0.075$
Document d_1 would be ranked higher if we smooth the LM with the Dirichlet prior smoothing and set $\mu = 10$.

Compare the answers of (a) and (b). Which ranking is more reasonable? Explain your answer.

The answer of (b) is more reasonable. In the sake of the sparsity of the data, there might be some unseen events appearing. In other words, the word out of the training data does not mean that the word does not exist; therefore, we couldn't set the probability of the unseen events to be 0, instead, we apply smoothing on the data.