Language Models G&F Sec. 2.3

a term-document matrix

	D_1	D_2	D_3
a	1	1	1
arrived	0	1	1
damaged	1	0	0
delivery	0	1	0
fire	1	0	0
gold	1	0	1
in	1	1	1
of	1	1	1
shipped	1	0	1
silver	0	2	0
truck	0	1	1
length	7	8	7

22 tokens

Given a query q we want to calculate for each document d_i the pobability of d_i generating q

$$sc(q, d_i) = \prod_{t_j \in q} P(t_j | M_{D_i}) \prod_{t_j \notin q} (1 - P(t_j | M_{D_i}))$$

$P(t_j | M_{D_i})$: prob. of M_{D_i} generating a term in the query

$(1 - P(t_j | M_{D_i}))$: prob. of M_{D_i} NOT generating a term NOT in the query

How to estimate $P(t_j | M_{D_i})$ for a given term and document i?

$$P(t_j | M_{D_i}) = P_{M_i}(t_j | M_{D_i}) = \frac{tf(t_j, D_i)}{length(D_i)}$$

which is easy enough to compute, except that if a term t_j doesn't happen to occur in document d_i, $P(t_j | M_{d_i}) = 0$ and that makes the whole product zero :-(

So what do we do when this happens? Choices include:

1. $P(t_j | M_{d_i}) = \frac{\sum_{\text{all } d_i} tf_j}{\sum_{\text{all } d_i} \sum_{\text{all } t_j} tf_j} = \frac{\text{occurrences of } t_j}{\text{occurrences of all terms}}$

G&F call this $\frac{cf}{cs}$

OR

2. $P(t_j|M_{di})$ ≠ risk t_j in d_i based on as shown on p. 4 3/

For each term t_i in document d_j, calculate

$$\frac{tf\ i}{\text{length}(d_j)}$$

G&F Table 3.12 p. 49

$P_{ML}(t|M_d)$	D_1	D_2	D_3	Pavg(t) Average Prob.
a	1/7	1/8	1/7	$(2/7+1/8)/3$
arrived	0	1/8	1/7	$(\frac{1}{8}+\frac{1}{7})/2$
damaged	1/7	0	0	1/7
delivery	0	1/8	0	1/8
fire	1/7	0	0	1/7
gold	1/7	0	1/7	1/7
in	1/7	1/8	1/7	$(2/7+1/8)/3$
of	1/7	1/8	1/7	$(2/7+1/8)/3$
shipment	1/7	0	1/7	1/7
silver	0	1/4	0	1/4
truck	0	1/8	1/7	$(1/8+1/7)/2$

To calculate average probability,

$$\sum_{d:\,t \in d} P_{ML}(t|M_d)$$

$$= \frac{\text{sum of each row}}{\#\text{ of non-zero entries}}$$

↑ G&F Table 2.13 p. 50

but we're not done!

Given these average probabilities, and the length of each document, we can compute

$$\overline{f_{t_i}} = P_{avg}(t_i) \times length(d_j)$$

	$\overline{f_t}$	D_1	D_2	D_3	$P_{avg}(t_i)$
a					
arrived			$8 \cdot \left(\frac{1}{8}+\frac{1}{7}\right) / 2$		$\left(\frac{1}{8}+\frac{1}{7}\right)/2$
⋮					
shipment					1/7
silver					1/4
⋮					
length		$\frac{1}{7}$	$\frac{2}{8}$	$\frac{}{7}$	

Given $\overline{f_t}$ for term t_i and document d_j,

$$R_{t,d} = \left(\frac{1}{1+\overline{f_t}}\right)\left(\frac{\overline{f_t}}{1+\overline{f_t}}\right)^{tf_{t,d}}$$

$$R_{\substack{arrived,\\ D_2}} = \left(\frac{1}{1+1.071}\right)\left(\frac{1.071}{1+1.071}\right)^{1}$$

$$= \frac{1}{2.071} \cdot \frac{1.071}{2.071} = \frac{1.071}{4.289} = .24$$

$$R_{\substack{silver,\\ D_2}} = \left(\frac{1}{1+2}\right)\left(\frac{2}{1+2}\right)^{2} = \frac{1}{3}\cdot\left(\frac{2}{3}\right)^{2} = \frac{4}{27} = .148$$

Now we can (Finally!) calculate

$$P(t|M_d) = P_{ml}(t|M_d)^{(1-R_{t,d})} \times P_{avg}(t)^{R_{t,d}}$$

($P_{ml}(t|M_d)$ from p. 3; $(1-R_{t,d})$ and $R_{t,d}$ from p. 4)

and this gives G&F Table 2.16

For q = "gold silver truck" and document d_1, for example,

$$P(q|M_{d_1}) = \prod_{t \in q} P(t|M_{d_1}) \times \prod_{\text{other terms}} (1-P(\cdot)) \cdot \prod_{\text{all other terms}} (1-P(t,M_{d_1})) =$$

.143 × .091 × .091

gold

silver

truck

$$(.00118) \times (1-.141)(1-.09)(1-.143)(1-.045)(1-.143)(1-.141)(1-.141)(1-.143)$$

(a, arrived, damaged, delivery; size, in, of, shipment)

$$= (.00118)(.86)(.91)(.86)(.95)(.86)(.86)(.86)(.86)(.91)(.91) =$$

$$.0003$$

which is close to Table 2.17