Language models

EXERCICE 1

Consider the description of two documents using language models :

Model D1 (θ_{D1})		*Model D2 (θ_{D2})*	
t	***$P(t\|\theta_{D1})$***	***t***	***$P(t\|\theta_{D2})$***
opacity	*0.2*	*opacity*	*0.15*
on	*0.1*	*on*	*0.12*
the	*0.01*	*the*	*0.0002*
posterior	*0.01*	*posterior*	*0.0001*
part	*0.04*	*part*	*0.03*
of	*0.03*	*of*	*0.04*
right	*0.005*	*right*	*0.04*
lob	*0.001*	*lob*	*0.01*
…	…	…	…

Suppose that we process the following query Q:
opacity on the right lob

Question : Compute the similarity between Q et the documents D1 et D2 by using the multinomial language model of IR :

P(opacity on the right lob | θ_{D1}) = ?

P(opacity on the right lob | θ_{D2}) =?

We use the last formula of slide 36:

$$P(q|\theta_d) \propto \prod\nolimits_{t\in q} p(t|\theta_d)^{w_t^q}$$

P(opacity on the right lob | θ_{D1}) $\propto$ 0.2 * 0.1 * 0.001 * 0.005 * 0.001 = 0,000 000 000 1

P(opacity on the right lob | θ_{D2}) $\propto$ 0.15 * 0.12 * 0.002 * 0.04 * 0.01 = 0,000 000 014 4

So, D2 answers better than D1 to the query Q.

Using personal calculator? Hard as we manipulate very small numbers!
Same with a computer!

Weh using logs (in base 10, for instance) pour avoir :
log (P(opacity on the right lob | θ_{D1})) = log(0.2) + log(0.1) + log(0.001) + log(0.005) + log(0.001) = -10

log P(opacity on the right lob | θ_{D2}) = log(0.15) + log(0.12) + log(0.002) + log(0.04) + log(0.01) = -7.841637508

We rank then values according to decreasing order:
log P(opacity on the right lob | θ_{D2}) > log (P(opacity on the right lob | θ_{D1}))
So, still D2 is a better answer that D1 for the query.

EXERCICE 2

This exercise focuses on the smoothing of probabilities, using Jelinek-Mercer.

Consider a corpus composed of 4 documents:

D1: opacity on the posterior part of the right lob of the lung

D2: tumor on the anterior part of the temporal lob of the brain

D3: tumor in the left lung

D4: tumor in the brain

1. Define the vocabulary T, with the stop-list {on, of, the}.
Similar to other models (like vector space).
T= {anterior, brain, left, lob, lung, opacity, part, posterior, right, temporal, tumor}
(We can call them t_1 to t_{11})

2. Compute the probabilities, by maximum likelihood, of the 11 terms for the 4 documents.

Term ti	D1	D2	D3	D4
anterior	0,00	0,17	0,00	0,00
brain	0,00	0,17	0,00	0,50
left	0,00	0,00	0,33	0,00
lob	0,17	0,17	0,00	0,00
lung	0,17	0,00	0,33	0,00
opacity	0,17	0,00	0,00	0,00
part	0,17	0,17	0,00	0,00
posterior	0,17	0,00	0,00	0,00
right	0,17	0,00	0,00	0,00
temporal	0,00	0,17	0,00	0,00
tumor	0,00	0,17	0,33	0,50

3. Compute the matching, using only the probabilities of question 2., with the query Q "tumor on the right lob and the left lob of the brain".
Same as exercice 1: we can go fast as no document contain all the query terms => all the documents have a rsv of 0 !

matching Q

D1	D2	D3	D4	D5
0	0	0	0	0

4. Compute the smoothed probabilities with Jelinek-Mercer smoothing of the terms for the 4 documents. We use λ*= 0.1*

We need first to compute the ML probabilities for each term on the whole corpus p(t|C):

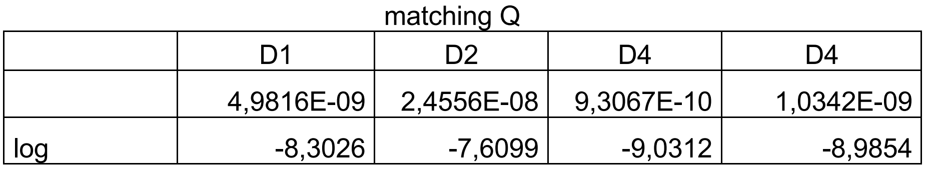
Term ti	P(t|C)
anterior	0,06
brain	0,12
left	0,06
lob	0,12
lung	0,12
opacity	0,06
part	0,12
posterior	0,06
right	0,06
temporal	0,06
tumor	0,18

Then, we use the Jelinek-Mercer formula for the smoothed probabilities (cf. slides), by giving 0.9 for the ML probability from the document, et 0.1 for the probabilities from the corpus. For instance, for "anterior" for the document D1, we get : 0.9*0+0.1*0.059=0.0059.We get:

Jelinek Mercer smooting with Lambda=0.1

Term ti	D1	D2	D3	D4
anterior	0,0059	0,1559	0,0059	0,0059
brain	0,0118	0,1618	0,0118	0,4618
left	0,0059	0,0059	0,3059	0,0059
lob	0,1618	0,1618	0,0118	0,0118
lung	0,1618	0,0118	0,3118	0,0118
opacity	0,1559	0,0059	0,0059	0,0059
part	0,1618	0,1618	0,0118	0,0118
posterior	0,1559	0,0059	0,0059	0,0059
right	0,1559	0,0059	0,0059	0,0059
temporal	0,0059	0,1559	0,0059	0,0059
tumor	0,0176	0,1676	0,3176	0,4676

5. *Redo question 3 with smoothed probabilities. Here again we use logs (in base 10 for instance).*



matching Q

	D1	D2	D4	D4
	4,9816E-09	2,4556E-08	9,3067E-10	1,0342E-09
log	-8,3026	-7,6099	-9,0312	-8,9854

With or without logs, the ranking of result is :
D2, D1, D4, D3.