Exemple Examen Recherche d'information

Documents autorisés

Exercice 1

Soit $q = q_1, \ldots, q_m$ une requête, d un document et $P(q_i|d)$ la probabilité du mot q_i dans le modèle de langue de d. On suppose que nous disposons d'une collection de documents comportant au total 8 mots *w1, ..., w8*.

La Table ci-dessous liste pour chaque mot sa probabilité dans le modèle de langue de référence, P(w|REF), estimé sur la collection (2ème colonne), la fréquence du terme c(w; d) dans un document (3ème colonne). Les colonnes 4 et 5 représentent les probabilités du terme dans les modèles langue du document *d*, estimés respectivement selon le maximum de vraisemblance et Dirichlet avec le paramètre μ.

Mots	P(w|REF)	c(w; d)	$P_{ml}(w\|d)$	$P_{\mu}(w\| d)$
w1	0.3	2		
w2	0.15	1		
w3	0.1	2		0.125
w4	0.1	4		
w5	0.05	1		
w6	0.1	0		
w7	0.1	0		
w8	0.1	0		

1- Remplir la colonne 4 ($p_{ml}(w|d)$),
2- La colonne 5 représente la probabilité du terme calculée après un lissage de Dirichlet effectuée sur la collection. Seule la probabilité de *w3* est donnée dans le tableau, déduire la valeur de μ ?
3- Sans effectuer les calculs de probabilités de la colonnes 5, indiquer pour chacun des mots de cette colonne si sa probabilité lissée ($P_{\mu}(w|d)$) est {>;=;<} à celle non lissée, calculée selon $p_{ml}(w|d)$, c'est-à-dire celle de la colonne 4.
4- Quelle condition doit satisfaire c(w; d) pour que la probabilité lissée du mot w soit toujours la même que la valeur non lissée quelque soit le paramètre μ.