TD – Stockage – correction – 2021-2022

Correction exercice 3 – Indexation vectorielle

Considérons les textes suivants :
Document 1 : « le professeur parle de la recherche d'information textuelle »
Document 2 : « la recherche d'information est un domaine de recherche qui s'intéresse à des nombreux problèmes »
Document 3 : « le modèle vectoriel de recherche d'information est un modèle simple à comprendre »

1. En considérant un anti-dictionnaire composé des termes :
{à, au, d', de, du, des, elle, elles, est, je, il, ils, le, la, les, lui, qui, son, s', sa, ses, tu, un, une}
représenter l'ensemble des termes d'indexation de chacun des documents ci-dessus. Attention aux majuscules !
2. Quel vocabulaire V est associé à cette collection dans le cas de l'utilisation de la troncature à base de 3 règles :
    1. s → /
    2. e → /
    3. ll → l
3. Calculer les pft à base de tf de chacun de ces termes pour chaque document.
4. Calculer les pdf à base d'idf de chacun des termes présents dans les documents
5. Donner les représentations vectorielles des documents sans normalisation.
6. En déduire le tableau de l'index inversé pour ce corpus.
7. Calculer les normes de chaque vecteur document.

Question 2 et Question 3

Indication du ptf (à base the tf) et du pdf (avec ln) pour chaque terme.

Document 1 : professeur (1/1.10), parl (1/1.10), recherch (1/0), information(1/0), textuel (1/1.10).
Document 2 : recherch (2/0), information (1/0), domain (1/1.10), intér (1/1.10), nombreux (1/1.10), problèm (1/1.10).
Document 3 : modèl (2/1.10), vectoriel (1/1.10), recherch (1/0), information (1/0), simpl (1/1.10), comprendr (1/1.10)

Note : Les valeurs de pdf possibles, comme N=3 (3 documents dans le corpus), sont ln(3/df) : ln(3/1) = 1.10 , ln(3/2) = 0.41 , ln(3/3) = 0

Question 4

On remarque que les termes recherch et information ont un pdf=0 donc ils auront des poids égaux à 0 pour tous les doucments, on peut les éliminer du vocabulaire, pour obtenir :

t1 = comprendr, t2 = domain, t3 = intér, t4 = modèl, t5 = nombreux, t6 = parl, t7 = problèm, t8 = professeur, t9 = simpl, t10 = textuel, t11 = vectoriel

Ce qui donne l'index inversé (sous forme de tableau):

	d1	d2	d3
t1	0	0	1.10
t2	0	1.10	0
t3	0	1.10	0
t4	0	0	2.20
t5	0	1.10	0
t6	1.10	0	0
t7	0	1.10	0
t8	1.10	0	0
t9	0	0	1.10
t10	1.10	0	0
t11	0	0	1.10

Avec de vocabulaire qui contient 11 termes, on a les vecteurs :
d1 = (0 0 0 0 0 1.10 0 1.10 0 1.10 0)
d2 = (0 1.10 1.10 0 1.10 0 1.10 0 0 0 0)
d3 = (1.10 0 0 2.20 0 0 0 0 1.10 0 1.10)

Calculons les normes des vecteurs documents :
$||d1|| = (1.10^2+1.10^2+1.10^2)^{1/2} = 1.90$
$||d2|| = (1.10^2+1.10^2+1.10^2+1.10^2)^{1/2} = 2.20$
$||d3|| = (1.10^2+2.20^2+1.10^2+1.10^2)^{1/2} = 2.91$