TD – Recherche d'information - correction

Exercice 5 – indexation dans le modèle vectoriel

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.
2. Calculer les tf de chacun de ces termes pour chaque document.
3. Calculer l'idf de chacun des termes présents dans les documents
4. En déduire le tableau du fichier inverse pour ce corpus.
5. Calculer les normes des vecteurs documents.

Question 1 (cf Question 2)

Questions 2
Indication du tf et du df et de l'idf (avec log10) pour chaque terme.
Document 1 : professeur (1/1/0.48), parle(1/1/0.48), recherche(1/3/0), information(1/3/0), textuelle(1/1/0.48).
Document 2 : recherche(2/3/0), information(1/3/0), domaine(1/1/0.48), intéresse(1/1/0.48), nombreux(1/1/0.48), problèmes(1/1/0.48).
Document 3 : modèle(2/1/0.48), vectoriel(1/1/0.48), recherche(1/3/0), information(1/3/0), simple(1/1/0.48), comprendre(1/1/0.48)

Pour les idf : valeurs avec $\log_{10}(3/df)$. Pour log10(3/1) = 0.48 , log10(3/2) = 0.17 , log10(3/3) = 0.
Question 3

Ce qui donne le vocabulaire :

t1 = comprendre, t2 = domaine, t3 = information, t4 = intéresse, t5 = modèle, t6 = nombreux, t7 = parle, t8 = problèmes, t9 = professeur, t10 = recherche, t11 = simple, t12 = textuelle, t13 = vectoriel.

Et le fichier inverse :

	d1	d2	d3
t1	0	0	0.48
t2	0	0.48	0
t3	0	0	0
t4	0	0.48	0
t5	0	0	0.95
t6	0	0.48	0
t7	0.48	0	0
t8	0	0.48	0
t9	0.48	0	0
t10	0	0	0
t11	0	0	0.48
t12	0.48	0	0
t13	0	0	0.48

En fait, on peut et on doit éliminer du vocabulaire les termes qui n'indexent aucun document, d'où :
t'1 = comprendre, t'2 = domaine, t'3 = intéresse, t'4 = modèle, t'5 = nombreux, t'6 = parle, t'7 = problèmes, t'8 = professeur, t'9 = simple, t'10 = textuelle, t'11 = vectoriel.

	d1	d2	d3
t'1	0	0	0.48
t'2	0	0.48	0
t'3	0	0.48	0
t'4	0	0	0.95
t'5	0	0.48	0
t'6	0.48	0	0
t'7	0	0.48	0
t'8	0.48	0	0
t'9	0	0	0.48
t'10	0.48	0	0
t'11	0	0	0.48

Calculons les normes des documents :
$||d1|| = (0.48^2+0.48^2+0.48^2)^{1/2} = 0.83$
$||d2|| = (0.48^2+0.48^2+0.48^2+0.48^2)^{1/2} = 0.96$
$||d3|| = (0.48^2+0.95^2+0.48^2+0.48^2)^{1/2} = 1.26$

Exercice 6 – pondération dans le modèle vectoriel

Fournir les résultats des requêtes suivantes pour le corpus de l'exercice 5 :
Q0 : pomme de terre
Q1 : recherche d'information
Q2 : recherche d'information textuelle
Q3 : domaine du modèle vectoriel
Commencer par analyser les requêtes comme les documents (anti-dictionnaire), et utiliser une pondération des requêtes par le tf uniquement.

--- Avec utilisation de fichier inverse :

Pour Q0 :
pomme ~~de~~ terre
// //
pas de terme du vocabulaire donc requête vide donc pas de document qui répond !

Pour Q1 :
recherche ~~d~~ information
// //
même résultat que pour Q0, avec la différence qu'ici les termes ont été enlevés du vocabulaire car ils ne sont pas utilisés pour la RI.

Pour Q2 :
recherche ~~d~~ information textuelle
// // (t10') tf=1

||Q2|| = rac(1) = 1
donc le vecteur requête non-normalisé est Q = (0 0 0 0 0 0 0 0 0 1 0)

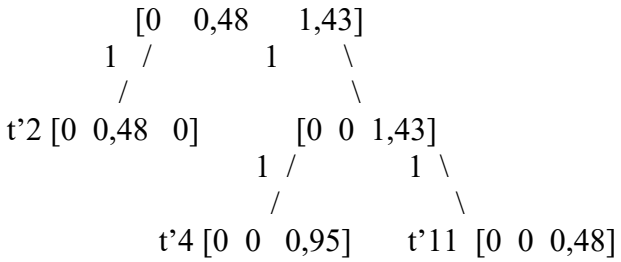
[0.48 0 0]
1 |
t'10 [0.48 0 0]

Résultat après normalisation :

Sim(D1,Q2)=0.48/(0.83*1)=0.578
Sim(D2,Q2)= Sim(D3,Q2)=0
Donc il n'y a que D1 qui répond.

Pour Q3 :
domaine ~~du~~ modèle vectoriel
t'2 (tf=1) t'4 (tf=1) t'11 (tf=1),
donc le vecteur requête non normalisé est Q3 = (0 1 0 1 0 0 0 0 0 0 1)

||Q3|| = rac(3) = 1.73



Normalisation :
Sim(d1,Q3) = 0
Sim(d2,Q3) = 0.48/(0.96*1.73) = 0.289
Sim(d3,Q3) = 1.43/ (1.26*1.73) = 0.656
Résultat :
d3, d2. (d1 ne répond pas).

Refaire les calculs avec cosinus sans fichier inverse et vérifier que l'on obtient exactement les même résultats :
Exemple pour Q3 :

Sim(d1, Q3) = (0*0 + 1*0 + 0*0 + 1*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*0) / (0.83 * 1.73) = 0
Sim(d2,Q3) = (0*0 + 1*0.48 + 0*0 + 1*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*0)/(0.96*1.73) = 0.289
Sim(d3,Q3) = (0*0 + 1*0 + 0*0 + 1*0.95 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 1*0.48)/ (1.26*1.73) = 0.656

On obtiend donc exactement les mêmes résultats.