USTOMB -2021- Faculté MI Département Informatique

Recherche d'Informations L3 (ISIL)

TD2

Exercice 1 :

Pondération des termes selon TF/IDF
Soit le tableau ci-dessous représentant le fichier inverse.

Terme	D1	D2	D3	D4	D5
Algo		1			1
Informa		1			1
Programm	3	2	2		1
lang	1		1	1	
fonct			1	1	1
const	1				

Calculer les poids $W(t_i, d_j)$ du terme t_i dans le document d_j selon *TF *IDF* donnée selon les formules suivantes :

$$TF = \frac{freq(t_i, d_j)}{\sum_{\forall t' \in d_j} freq(t', d_j)}$$

$$IDF = \log(N/Nt)$$

Soient :
t_1= Algo, t_2= Inform, t_3= Programm , t_4= lang , t_5= fonc , t_6= const
freq(t_1, D1)=0, freq(t_1, D2)=1, freq(t_1, D3)=0, freq(t_1, D4)=0 , freq(t_1, D5)=1
freq(t_2, D1)=0, freq(t_2, D2)=1, freq(t_2, D3)=0, freq(t_2, D4)=0 , freq(t_2, D5)=1
freq(t_3, D1)=3, freq(t_3, D2)=2, freq(t_3, D3)=2, freq(t_3, D4)=0 , freq(t_3, D5)=1
freq(t_4, D1)=1, freq(t_4, D2)=0, freq(t_4, D3)=1, freq(t_4, D4)=1, freq(t_4, D5)=0
freq(t_5, D1)=0, freq(t_5, D2)=0, freq(t_5, D3)=1, freq(t_5, D4)=1 , freq(t_4, D5)=1
freq(t_6, D1)=1, freq(t_6, D2)=0, freq(t_6, D3)=0, freq(t_6, D4)=0 , freq(t_6, D5)=0

$tf_{1,1}$= 0/5 ; $tf_{1,2}$= 1/4 ; $tf_{1,3}$= 0/4 ; $tf_{1,4}$= 0/2 ; $tf_{1,5}$= 1/4
$tf_{2,1}$= 0/5 ; $tf_{2,2}$= 1/4 ; $tf_{2,3}$= 0/4 ; $tf_{2,4}$= 0/2 ; $tf_{2,5}$= 1/4
$tf_{3,1}$= 3/5 ; $tf_{3,2}$= 2/4 ; $tf_{3,3}$= 2/4 ; $tf_{3,4}$= 0/2 ; $tf_{3,5}$= 1/4
$tf_{4,1}$= 1/5 ; $tf_{4,2}$= 0/4 ; $tf_{4,3}$= 1/4 ; $tf_{4,4}$= 1/2 ; $tf_{4,5}$= 0/4
$tf_{5,1}$= 0/5 ; $tf_{5,2}$= 0/4 ; $tf_{5,3}$= 1/4 ; $tf_{5,4}$= 1/2 ; $tf_{5,5}$= 1/4

$tf_{6,1}$= 1/5 ; $tf_{6,2}$= 0/4 ; $tf_{6,3}$= 0/4 ; $tf_{6,4}$= 0/2 ; $tf_{6,5}$= 0/4

idf_1= log(5/2)= 0.4
idf_2= log(5/2)=0.4
idf_3= log(5/4)=0.097
idf_4= log(5/3)=0.221
idf_5= log(5/3)=0.221
idf_6= log(5/1)=0.67

$w_{1,1}$= $tf_{1,1}$* idf1 =0*0.4=0
$w_{1,2}$= $tf_{1,2}$* idf1 =0.25*0.4=0.10
$w_{1,3}$= $tf_{1,3}$* idf1 =0*0.4=0
$w_{1,4}$= $tf_{1,4}$* idf1 =0*0.4=0
$w_{1,5}$= $tf_{1,5}$* idf1 =0.25*0.4=0.10

$w_{2,1}$= $tf_{2,1}$* idf2 =0.2*0.4=0
$w_{2,2}$= $tf_{2,2}$* idf2 =0*0.4=0.10
$w_{2,3}$= $tf_{2,3}$* idf2 =0*0.4=0
$w_{2,4}$= $tf_{2,4}$* idf2 =0*0.4=0
$w_{2,5}$= $tf_{2,5}$* idf2 =0.25*0.4=0.10

$w_{3,1}$= $tf_{3,1}$* idf3 =0.6*0.097=0.0582
$w_{3,2}$= $tf_{3,2}$* idf3 =0.5*0.097=0.0485
$w_{3,3}$= $tf_{3,3}$* idf3 =0.5*0.097=0.0485
$w_{3,4}$= $tf_{3,4}$* idf3 =0*0.097=0
$w_{3,5}$= $tf_{3,5}$* idf3 =0.25*0.097=0.02425

$w_{4,1}$= $tf_{4,1}$* idf4 =0.2*0.221=0.0442
$w_{4,2}$= $tf_{4,2}$* idf4 =0*0.221=0
$w_{4,3}$= $tf_{4,3}$* idf4 =0.25*0.221=0.055
$w_{4,4}$= $tf_{4,4}$* idf4 =0.5*0.221=0.11
$w_{4,5}$= $tf_{4,5}$* idf4 =0*0.221=0

$w_{5,1}$= $tf_{5,1}$* idf5 =0*0.221=0
$w_{5,2}$= $tf_{5,2}$* idf5 =0*0.221=0
$w_{5,3}$= $tf_{5,3}$* idf5 =0.25*0.221=0.055
$w_{5,4}$= $tf_{5,4}$* idf5 =0.5*0.221=0.11
$w_{5,5}$= $tf_{5,5}$* idf5 =0.25*0.221=0.055

$w_{6,1}$= $tf_{6,1}$* idf6 =0.2*0.67=0.134

$w_{6,2}$= $tf_{6,2}$* idf6 =0*0.67=0
$w_{6,3}$= $tf_{6,3}$* idf6 =0*0.67=0
$w_{6,4}$= $tf_{6,4}$* idf6 =0*0.67=0
$w_{6,5}$= $tf_{6,5}$* idf6 =0*0.67=0

Soit la requête Q =«fonctionlangage de programmation »
Calculer le degré de correspondance **R(D_i, Q) =**$\sum_{i \in Q} w(t_q, D_i)$ représentant la somme des poids des termes de la requête t_q dans le document D_i suivant les cas de calcul de *TF *IDF* précédents.

Q= t3 et t4 et t5
R(D_1, Q) = $w_{3,1}$+ $w_{4,1}$ + $w_{5,1}$ = 0.0582 + 0.0442 + 0 = 0.1024
R(D_2, Q) = $w_{3,2}$+ $w_{4,2}$+ $w_{5,2}$ = 0.0485 + 0 + 0 = 0.0485
R(D_3, Q) = $w_{3,3}$+ $w_{4,3}$+ $w_{5,3}$ =0.0485 + 0.055 + 0.055 = 0.1585
R(D_4, Q) = $w_{3,4}$+ $w_{4,4}$+ $w_{5,4}$ = 0 + 0.11 + 0.11 = 0.22
R(D_5, Q) = $w_{3,5}$+ $w_{4,5}$+ $w_{5,5}$ = 0.0242 + 0 + 0.055 = 0.0792

Quel est le document qui sera classé en haut lors de la recherche ?
Le document D4 est celui qui sera classé en premier

Exerccie 2:

Modèle booléen classique

Considérons deux documents D1 et D2, représentés par l'ensemble des termes d'indexation $T = \{t_1, t_2, ..., t_{10}\}$.

Les poids des termes dans D1

ti	*t*1	*t*2	*t*3	*t*4	*t*5	*t*6	*t*7	*t*8	*t*9	*t*10
*W (ti; D*1)	0.5	0	0.8	0	1	0	0.6	0.8	0	0.9

Les poids des termes dans D2

ti	*t*1	*t*2	*t*3	*t*4	*t*5	*t*6	*t*7	*t*8	*t*9	*t*10
*W (ti; D*2)	1	0.7	0	0	1	0	0	0	0	0.9

1-Donnez les formules de conjonction de D1 et D2.

Dans le modèle classique les poids des termes sont considérés soit 1 soit 0.
Nous prenons la considération que puisque un terme existe dans le document son poids sera égal à 1.
Les poids seront transformés comme suit :

t_i	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
W (t_i ;D1)	0	0	0	0	1	0	0	0	0	0

t_i	t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
W (t_i ;D2)	1	0	0	0	1	0	0	0	0	0

2-Traiter les requêtes suivantes en utilisant le modèle booléen classique

$$Q1 : t1 \wedge t5$$

$$Q2 : (t1 \wedge t5) \vee (t8 \wedge t10)$$

Q1 :
D1 : t1 et t5 = 1 et 0 =0
D2 : t1 et t5 = 1 et 1 =1

Q2 :
D1 : (t1 et t5) ou (t8 et t10) = (1 et 0) ou (0 et 0)=0
D2 : (t1 et t5) ou (t8 et t10)= (1 et 1) ou (0 et 0)= 1

Exercice 3

Considérons deux documents D1 et D2, représentés sur un vocabulaire T={t_1, …, t_{10}}.
La formule logique de D1, est :
- W_{D1} est défini par

t	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
$W_{D1}(t)$	0.5	0	0.8	0	1	0	0.6	0.8	0	0.9

La formule logique de D2 est :
- W_{D2} est défini par :

t	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	t_9	t_{10}
$W_{D2}(t)$	1	0.7	0	0	1	0	0	0	0	0.9

Traiter les deux requêtes suivantes :
Q1 : t_1 et t_5
Q2 : (t_1 et t_5) ou (t_8 et t_{10})

En utilisant une similarité basée sur la logique floue

Soit l'application de la formulation de Zadeh :

Q1 :
R(D1, Q1)= min(R(D1,q1) ,R(D1,q2))= min (0.5, 1) = 0.5
R(D2, Q1)= min(R(D2,q1) ,R(D2,q2))= min (1,1) = 1