Corrigé type Contrôle 1, Module : RI, 3eme année Licence ISIL, Date : 2023

Questions de compréhension (06pts)

Q1 Les éléments d'un corpus de test significatif : (**1,5)**

- Un nombre de documents élevé
- Un ensemble de requêtes ;
- Une liste de documents pertinents pour chaque requête.

Q2 : Dans le processus d'indexation, le fichier inverse est une structure très utilisée. Quel est l'inconvénient majeur de cette structure ? **Structure Lourde (0,5)**

Q3 : La notion de pertinence peut être appréhendée à deux niveaux, expliquer brièvement ces deux niveaux ? (**01pt)**

- **Niveau utilisateur :** la pertinence correspond à la **satisfaction** de l'utilisateur par apport à l'ensemble des documents restitués par le SRI. (**Pertinence subjective, cognitive)**
- **Niveau système :** le système **mesure** un degré de pertinence, une valeur de similitude entre un document et une requête. (**Pertinence algorithmique, objective)**

Q4 : Quels est le but de tout système de recherche d'information ? **(01 pt**)

Le but de tout SRI est de rapprocher la **pertinence système de la pertinence utilisateur**.

Q5 : Quels est le principe de l'algorithme de ***Porter*** **? Tronquer les suffixes des termes selon les règles spécifiées par l'algorithme (01pt)**

Q6 : Quels la relation entre un SRI et une Base de données ? (**01pt)**

Les BD ne permettent de réaliser qu'une partie de fonctionnalités de la RI.

Exercice 1 : (07pts)

Question 1 : 04pts:

Etape 1 et 2 =

D1 = "computing, programs, written ,software, development, programming ",
D2 = "programming, language, softwares" ;
D3 = "computer, software, program" ;
D4 = "information, retrieval"

Etape 3 (application des règles) :

D1 = "comput, program, written, software develop, program",
D2 = "program, language, software" ;
D3 = "comput, software, program" ;
D4 = "information, retrieval"

Etape 4 (Fréquences des termes) :

D1 = "comput, 1; program, 2; written, 1:software,1; develop1",
D2 = "program, 1, language, 1; software,1" ;
D3 = "comput,1; software,1; program, 1" ;

D4 = "information,1 ; retrieval, 1 "

N=4

Terme	n_t	(tf)				IDF	TF*IDF				
		D1	D2	D3	D4		D1	D2	D3	D4	W
T1 comput	**2**	**1**	**0**	**1**	**0**	**Log(4/2)=0,3**	**0,3**	**0**	**0,3**	**0**	**0,6**
T2 Develop	**1**	**1**	**0**	**0**	**0**	**Log(4/1)=0,6**	**0,6**	**0**	**0**	**0**	**0,6**
T3 information	**1**	**0**	**0**	**0**	**1**	**Log(4/1)=0,6**	**0**	**0**	**0**	**0,6**	**0,6**
T4 language	**1**	**0**	**1**	**0**	**0**	**Log(4/1)=0,6**	**0**	**0,6**	**0**	**0**	**0,6**
T5 program	**3**	**2**	**1**	**1**	**0**	**Log(4/3)=0,12**	**0,24**	**0,12**	**0,12**	**0**	**0,48**
T6 retrieval	**1**	**0**	**0**	**0**	**1**	**Log(4/1)= 0,6**	**0**	**0**	**0**	**0,6**	**0,6**
T7 software	**3**	**1**	**1**	**0**	**0**	**Log(4/2)= 0,3**	**0,3**	**0,3**	**0**	**0**	**0,36**
T8 written	**1**	**1**	**0**	**0**	**0**	**Log(4/1)= 0,6**	**0,6**	**0**	**0**	**0**	**0,6**

Question 2 : 01pt

T= {comput,; develop; information ; language; program. retrieval; software; written}

Question 3:

Matrice documents –termes (1pt) : 0,25 pour chaque vecteur correcte

	T1	T2	T3	T4	T5	T6	T7	T8
D1	0,6	0,6	0	0	0,48	0	0,36	0,6
D2	0	0	0	0,6	0,48	0	0,36	0
D3	0,6	0	0	0	0,48	0	0,36	0
D4	0	0	0,6	0	0	0,6	0	0

Question 4 : 01pt

Vecteur Requête –termes

	T1	T2	T3	T4	T5	T6	T7	T8
Q1	0,6	0	0	0	0,48	0	0	0

$$R(d,q) = \frac{\sum_{i=1}^{n} di \times qi}{}$$

R(d1,q1)= **0,59 (0,25pt)**

R(d2,q1)= **0,23 (0,25pt)**

R(d3,q1)= **0,59 (0,25pt)**

R(d4,q1)= **0 (0,25pt)**

Exercice 3 : (05 pts)

Matrice documents –termes (1pt) : 0,25 pour chaque vecteur correcte

	T1	T2	T3	T4	T5	T6
D1	2	0	2	0	0	0
D2	3	0	2	1	0	0
D3	1	2	3	0	1	0

Vecteur Requête –termes (0,5pt)

	T1	T2	T3	T4	T5	T6
Q1	3	0	3	0	1	0

1) Mesure de Cosinus (0 ,5pt)

$$R(d,q) = \frac{\sum_{i=1}^{n} di \times qi}{\sqrt{\sum_{i=1}^{n} di^2} \times \sqrt{\sum_{i=1}^{n} qi^2}}$$

R(d1,q)=2*3+2*3/$\sqrt{2^2+2^2} * \sqrt{3^2+3^2+1}$=12/ 2,82*4,35= **0,97** **(0,5pt)**

R(d2,q)= 3*3+2*3/$\sqrt{3^2+2^2+1} * \sqrt{3^2+3^2+1}$=15/ 3,74*4,35= **0,92** **(0,5pt)**

R(d3,q)=3+9+1 /$\sqrt{1^2+2^2+3^2}+1 * \sqrt{3^2+3^2+1}$= 13 / 4,69 = **0,77** **(0,5pt)**

Le document d1 est plus pertinent que les documents d1 et d3 (0 ,5pt)

La liste ordonnée des documents est : d1, d2, d3

2) La distance euclidienne. (0 ,5pt)

$$\boldsymbol{R(d,q) = \sqrt{\sum_{i=1}^{n} (di - qi)^2}}$$

R(d1, q)=$\sqrt{1+1+1}$ = 1,73 (0 ,5pt)

R(d2, q)= $\sqrt{1+1+1}$ = 1,73 (0 ,5pt)

R(d3, q)=$\sqrt{(2)^2+(2)^2} = 2,82$ (0,5pt)

Les documents **d1 et d2 sont** les plus pertinents que **d3 (0 ,5pt)**

La liste ordonnée des documents est d1, d2, d3 OU d2, d1, d3

Exercice 3 (2pts) :

Q : (t2 ET t5) OU ⅂ (t3 ET t5)

t	t1	t2	t3	t4	t5	T2 ET t5	T3 ET t5	l (t3 ET t5)	Q
W*D1*(t)	0.5	0	0.8	0	0,7	0	0,7	0,3	**0,3 (0 ,5 pt)**
WD2(t)	1	0.7	0	0	0,2	0,2	0	1	**1 (0,5 pt)**
WD3(t)	0	0.6	0.3	0.2	0.9	0,6	0,3	0,7	**0,7 (0,5 pt)**
$W_{D4}(t)$	0	0.8	0,1	0	0.4	0,4	0,1	0,9	**0,9 (0,5 pt)**