Université des Sciences et de Technologie Mohamed Boudiaf USTOMB

Faculté Des Mathématiques et Informatique MI

Examen L3 ISIL **Recherche d'informations RI** **14/05/2023**

1h30

Questions de cours (8 points)

1- Quels sont les éléments essentiels d'un SRI ? **1-les documents, 2-le contenu des documents, 3-les besoins de l'utilisateur et 4-la satisfaction. 1 pt**

2- Quelle était la méthode naïve utilisée dans l'indexation des documents avant l'émergence des SRI ? expliquez brièvement son fondement. **La méthode de zipf ; son principe est de ranger les mots selon la fréquence observée et par la suite éliminer ceux dont la fréquence est en dessous d'un certain seuil et ceux dont la fréquence est en dessus d'un autre seuil. 2 pts**

3- Quelle est l'utilité d'indexer les documents dans un corpus ? peut-on considérer de la même façon l'importance d'indexer les requêtes ?argumentez. **Pour ne considérer que les termes importants et aussi pour un gain d'espace de stockage concernant les documents. Concernant les requêtes c'est différent car on ne peut pas parler de corpus, l'indexation de la requête permet au SRI de répondre à une forme de langage défini. 2 pts**

4- Comment définissez vous l'expression « Terme discriminent » par rapport à un corpus ou une requête**? Un terme discriminent distingue un document par rapport à un autre c'est-à-dire le différencie par rapport à un autre, même chose pour une requête. 1 pt**

5- Selon vous, dans quel cas peut-on affirmer que l'appariement exact pourrait donner une satisfaction à l'utilisateur qui connait bien le langage d'interrogation? argumentez. **Dans le cas où le corpus est asssez volumineux car autrement, il est fort probable que les résultats de recherche aboutissent à un nombre de documents avoisinant 0. 1 pt**

6- Comment est évaluée la performance des résultats au niveau du SRI ? **Par les valeurs précision et rappel. 0.5 pt**

7- A quoi sert la normalisation (annulation d'un terme) utilisée dans la méthode d'indexation dite « basée sur les valeurs discriminantes » ? **à distinguer l'importance du terme dans le corpus en mesurant l'impact de son élimination dans ce corpus**.

8- Citez 2 modèles basés sur l'appariement rapproché utilisés dans les SRI. **Vectoriel, booléen étendu, …. 0.5 pt**

Exercice 1 (6 points)

Soient les requêtes suivantes :

Q_1= t_1 ou (t_2 et t_3)

Q_2=t_2 ou t_3

Un SRI basé sur l'appariement exact a été appliqué à un corpus de 10 documents et a donné les résultats suivants :

Q1 : Résultats obtenus

doc	Pertinence
1	P
2	P
3	P
4	
5	P
6	P
7	P
8	P
9	P

Q2 : Résultats obtenus

doc	Pertinence
1	
2	
3	P
4	
5	
6	P
7	
8	
9	

1- Représentez l'organigramme général d'un SRI. **1 pts**
2- Comment jugez-vous la méthode d'indexation de documents appliquée à ce corpus ? est-elle efficace ? argumentez. **En analysant Q1 et Q2 avec les résultats obtenus, Il y a une faille dans l'indexation car on constate que presque tous les documents ont été rapportés par rapport à Q1, ceci est du principalement à la présence et à la considération du terme t_1 dans le corpus et dans les requêtes. 2 pts**
3- Que peut-on dire du terme t_1 ? **il n'est pas discriminent 1pt**
4- Comment pouvons-nous améliorer ce SRI ? et à quel niveau de l'organigramme ? **Au niveau de l'indexation dans l'organigramme en reconsidérant la liste des « mots vides en y incluant le terme t_1 » 2 pts**

Exercice 2 (Modèle vectoriel) (6 points)

Soit un corpus contenant les 2 documents d1 et d2 représentés comme suit:

d1=(1,0,1,0,0,0)

d2=(1,2,3,0,1,0)

Soient les 2 requêtes représentées comme suit :

q1=(2,0,2,0,0,0)

q2=(0,0,0,2,0,2)

1- Donnez la matrice Termes-Documents

	t_1	t_2	t_3	t_4	t_5	t_6
d1	1	0	1	0	0	0
d2	1	2	3	0	1	0

0.25 pts

2- Donnez la matrice Termes-Requêtes

	t_1	t_2	t_3	t_4	t_5	t_6
Q1	2	0	2	0	0	0
Q2	0	0	0	2	0	2

0.25 pts

3- Choisir 2 mesures de similarités en donnant leurs formulations exactes.

Le produit scalaire :	$RSV(d_i,q) = \sum_{j=1}^{n} w_{ij} \times w_{qj}$
Distance euclidienne :	$RSV(d_i,q) = \sqrt{\sum_{j=1}^{n}(w_{ij} - w_{qj})^2}$
La mesure cosinus :	$RSV(d_i,q) = \frac{\vec{d_1}}{\lvert\vec{d_1}\rvert} \cdot \frac{\vec{q}}{\lvert\vec{q}\rvert} = \frac{\sum_{j=1}^{n} w_{ij} \times w_{qj}}{\sqrt{\sum_{j=1}^{n} w_{ij}^2} \times \sqrt{\sum_{j=1}^{n} w_{qj}^2}}$
La mesure de Dice :	$RSV(d_i,q) = \frac{2 \times \sum_{j=1}^{n} w_{ij} \times w_{qj}}{\sum_{j=1}^{n} w_{ij}^2 + \sum_{j=1}^{n} w_{qj}^2}$
La mesure de Jacard :	$RSV(d_i,q) = \frac{\sum_{j=1}^{n} w_{ij} \times w_{qj}}{\sum_{j=1}^{n} w_{ij}^2 + \sum_{j=1}^{n} w_{qj}^2 - \sum_{j=1}^{n} w_{ij} \times w_{qj}}$
	$\sum_{j=1}^{n} w_{ij} \times w_{qj}$

0.5 pt / formule = 1 pts

4- Considérons uniquement la requête q1 : appliquer les 2 mesures choisies dans (3), donnez le résultat de classement des documents obtenus dans chaque cas.

Pour chaque mesure : 0.5 X 2 pour le calcul de RSV(d,Q) = 1 pt

0.25 pour la conclusion c à d classement.

1.25 x 2 =2.5 pour la question 4

(la plupart vont choisir cosinus et distance euclidienne)